



Multivariate Data Analysis and Modelling Basic Course

Contents



Multivariate Data Analysis and Modelling

Basic MVDA

1. Introduction
2. Principal Component Analysis (PCA)
3. Applications of PCA
4. Partial Least Squares Projections to Latent Structures (PLS)
5. Multivariate Characterisation
6. Multivariate Calibration
7. Process Data Analysis
8. Conclusions

Multivariate Data Analysis and Modelling

Additional Topics

11. Additional Topics I – Pre-processing Methods
12. Additional Topics II – Method Extensions and Special Cases
13. Additional Topics III – Process Applications
14. Exercises

Table of Contents

Basic MVDA	5
1. Introduction	5
2. Principal Component Analysis, PCA	21
3. Applications of PCA	43
4. Partial Least Squares Projections to Latent Structures, PLS	61
5. Multivariate Characterisation	83
6. Multivariate Calibration	95
7. Process Data Analysis	111
8. Conclusions	125
Additional Topics	129
11. Additional Topics I – Pre-processing methods	129
• Scaling and Centering	130
• Transformation and Expansion of data	144
• Signal Correction and Compression	154
12. Additional Topics II – Method extensions and special cases	169
• PLS and one response variable	170
• PLS Discriminant Analysis (PLS-DA)	175
13. Additional Topics III – Process Applications	181
• Multivariate Statistical Process Control, MSPC	182
• Batch Statistical Process Control, BSPC	199
Exercises	219
Getting Started	
FOODS	221
IRIS	227
Easy PCA	
ARCHAEOLOGY	237
METABONOMICS	243
Easy PLS	
LOWARP	253
USDVOLVO	259

Quality Control

THICKNESS	269
CUPRUM	279

Multivariate Characterisation

SURFACTANT	287
PULP	301

Multivariate Calibration

SUGAR	305
NIR_CHIP	315
CELLULOSE	323

Process Applications

SOVRING	331
PROC1A	345
Baker's Yeast	353

<i>Flow-chart for MVDA in SIMCA</i>	367
--	------------

<i>References</i>	369
--------------------------	------------

Multivariate Data Analysis and Modelling Basic Course

Chapter 1 Introduction



Introduction

- Three Types of Problems
- Difficulties with Complex R&D Problems
- Data Tables and Notation
- Example
- Philosophy of Modelling
- Principle of Projections

Three types of Problems

Multivariate Data Analysis provides tools for:

- Overview & Summary
 - Structure
 - Similarity/Dissimilarity
 - Outliers
- Classification
 - Recognise groups
- Relationship between Y and X
 - Not one x/y at a time, but all x's/y's simultaneously
 - Finding how the x-variables affect the responses
 - Finding how the x- and y-variables correlate to each other
 - How to set X to get the best profile of Y

Problem I --- Overview & Summary

Example: FOODS

Problem:

To investigate the food consumption pattern in Europe; the relative amount of twenty common provisions were collected for 16 countries.

We would like to examine the similarity/dissimilarity between the countries based on these data.

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
	Prima	ONAM	Gr_Co	Inst_C	Tea	Sweet	Biscui	Pa_So	Ti_Sol	In_Pot	Fro_Fi	Fro_Ve	Apples	Orangi	Ti_Fruj	Jam	Garlic	Butter	Margar	Olive_C	Yoghu	Crisp
1	Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	30	26	
3	Italy	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24	94	5	18	
4	France	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47	36	57	3	
5	Holland	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97	13	53	15	
6	Belgium	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80	83	20	5	
7	Luxembo	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94	84	31	24	
8	England	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94	57	11	28	
9	Portugal	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78	92	6	9	
10	Austria	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72	28	13	11	
11	Switzerl	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48	61	48	30	
12	Sweden	97	13	93	31	43	43	39	54	45	56	78	53	75	9	68	32	48	2	93		
13	Denmark	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91	30	11	34	
14	Norway	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63	94	28	2	62	
15	Finland	98	12	84	20	64	27	10	8	18	12	50	57	22	37	15	96	94	17	64		
16	Spain	70	40	40	62	43	2	14	23	7	59	77	30	38	86	44	51	91	16	13		
17	Ireland	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	25	31	3	9	

Problem II --- Classification

Example: IRIS (A classical data set in statistics)

- **Data:**

- The data table contains petal and sepal lengths and widths of 50 specimens each of *Iris setosa*, *Iris versicolor* and *Iris virginica*. This data set was introduced by the great statistician Fisher as early as 1936. It is commonly known as "The Fisher Iris Data".

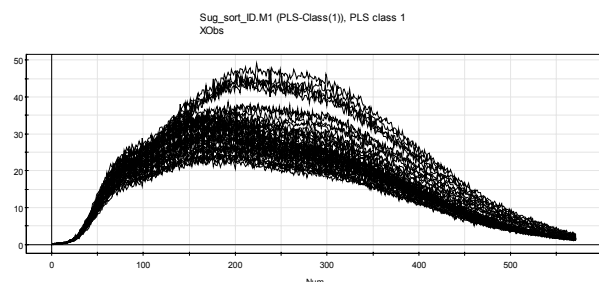
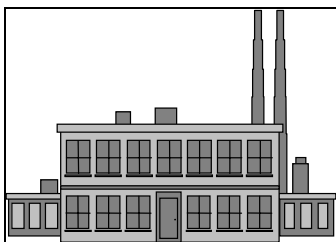
- **Goal:**

- To find a multivariate model that classifies a new Iris specimen in the correct class according to its petal and sepal lengths and widths

Problem III --- Quantification & Prediction

Example: SUGAR

Multivariate calibration at sugar production plant



Problem: In sugar production two important product quality properties are *impurity* and *colour* of the sugar, but measuring these quality measures are laborious and time-consuming. It was desired to try to replace conventional wet-chemistry measurements with rapid on-line fluorescence measurements.

Data: 106 time points in the process, 571 X-variables, 2 Y-variables

Multivariate data analysis

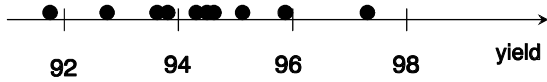
- Extracting information from data with many variables using them all simultaneously.
 - Deals mainly with:
 - **HOW** to get information out of existing multivariate data
 - Deals less with:
 - **How** to structure the problem
 - **Which** variables to measure
 - **How** to collect data (Design of experiments, DOE)
- } Needs process knowledge

Data

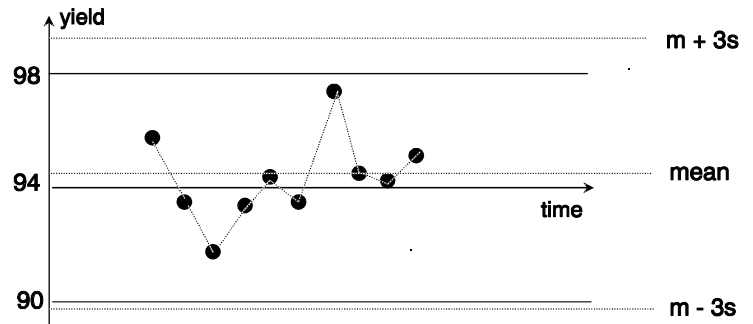
- Measurements of interest on process or system
- Data are not information, they are collected to **extract information**
- Important features of data
 - Variability
 - Types of data

Variability

Ten measurements of yield, under identical conditions



Ten measurements of yield as a function of time



Any measurement and experiment is influenced by noise

Under stable conditions, any process or system varies around its mean, and stays within "control limits"; ± 3 standard deviations in 99.4% of the observations

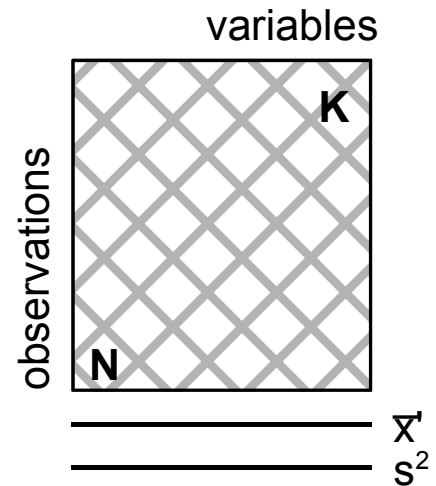
Types of data

What types of data for **Modelling** and **Analysis** are there?

- | | |
|---|--|
| <ul style="list-style-type: none"> • Univariate data $K = 1$ • Bivariate data $K = 2$ | <ul style="list-style-type: none"> • Quantitative • Qualitative |
| <ul style="list-style-type: none"> • Few-variate data $K \leq 5$ • Multivariate data $K \geq 6$ | <ul style="list-style-type: none"> • Processes (Continuous/Batch) • Time Series (Stationary/Dynamic) |
| | <ul style="list-style-type: none"> • Controlled/Uncontrolled |

Data with many variables (multivariate)

- Interested in **many variables** $K \geq 6$
- Controlled variables (**input or x-variables**) and/or characteristics (**output or y-variables**)
- Most systems and processes are characterised by a multitude of variables \Rightarrow large data matrices



K variable averages
K variable variances

$K*(K-1)/2$ covariances
Very many
Dominate when $K \geq 6$

Difficulties with complex R&D problems

- Dimensionality
 - K (number of X-variables) large
 - M (number of Y-variables) large
 - N large, medium, or small

Data table short and fat, or square and very large



- Collinearity
 - X's often not independent
 - Y's often not independent

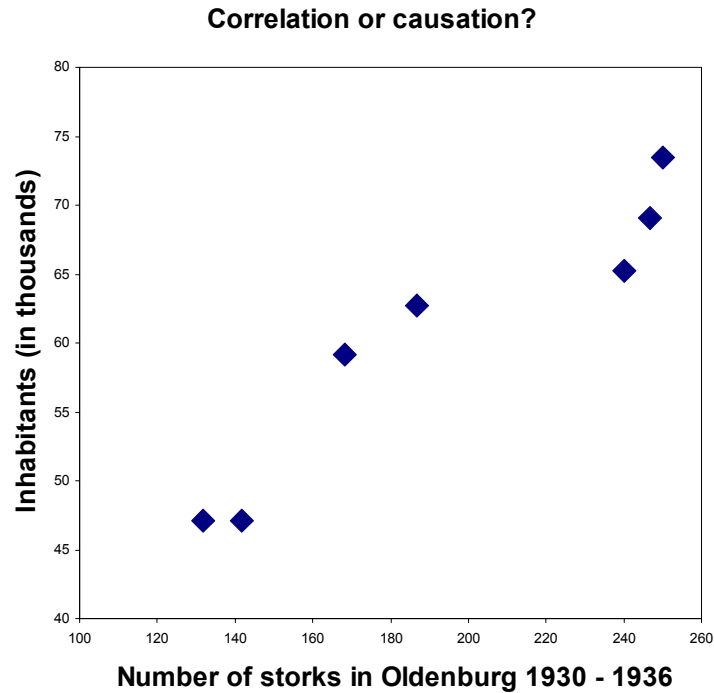
Only a few variables affect the system! Which?

Correlation vs causality!

- Noise
 - Individual measurements are noisy with large variability
- Missing data

Correlation and Causality

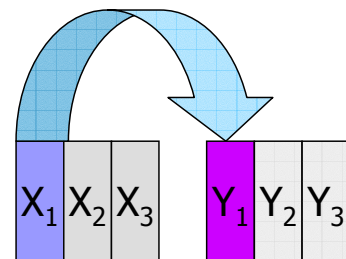
Although the two variables are correlated, this does not imply that one causes the other!



Methods of Analysis

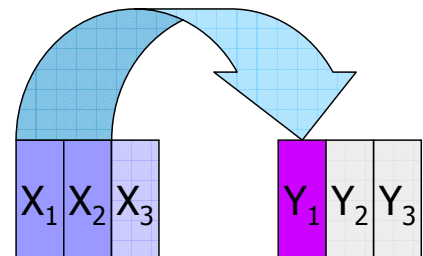
- **COST Approach**

- Plot and evaluate one variable or a pair of variables at time
- OK 50 years ago (few variables)



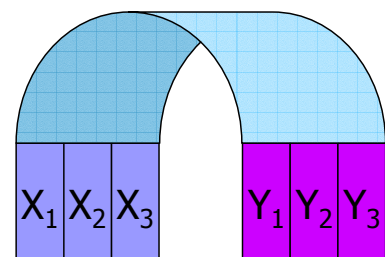
- **Classical Statistics**

- Find a relationship between a few of the X's and one Y at a time
- OK 50 years ago (few and essentially uncorrelated variables)



- **Multivariate Analysis**

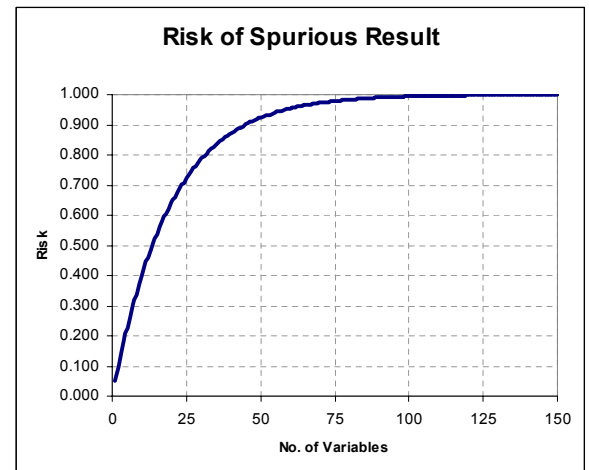
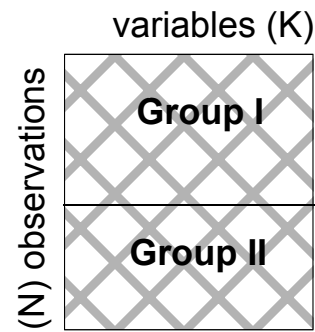
- Model all the variables together to find relationships between **all** the X's and all the Y's



Risks with traditional statistical methods

- Comparing one group against another
- Typically 95% confidence level used
 - Type I errors (False positives – spurious results)
 - Type II errors (False negatives – missed opportunities, risk of **not** seeing the information)
- Risk = $1 - 0.95^k$

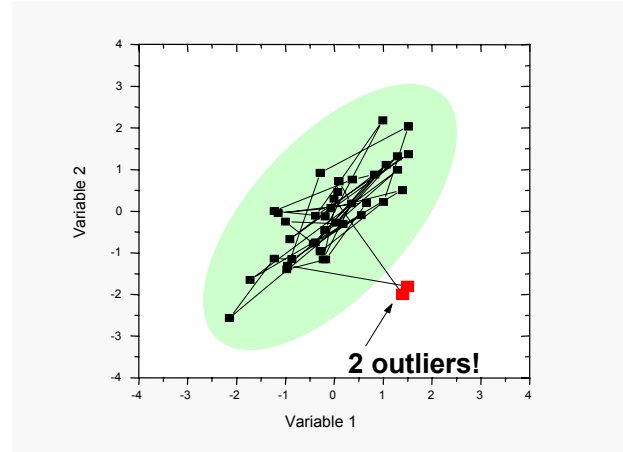
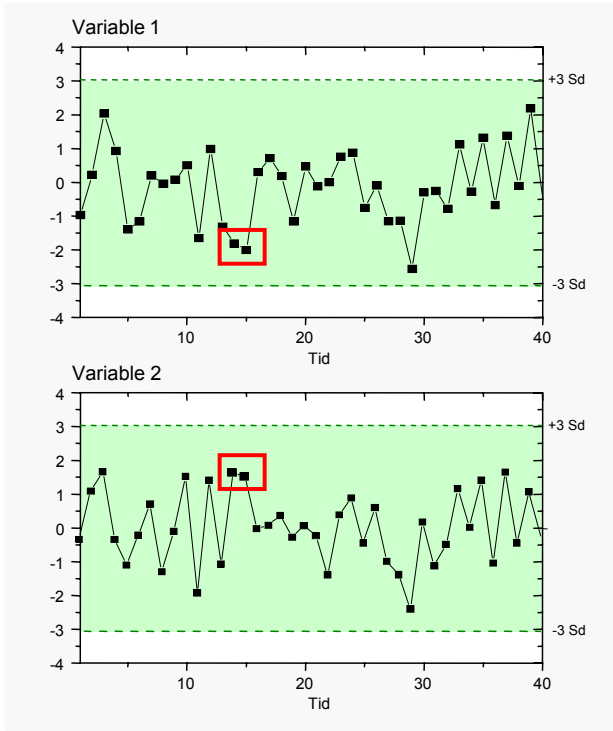
K	5	10	30	100
Risk	23%	40%	79%	99.4%



All data are needed

- In Shewhart's days (1930) process engineers were lucky to have one measurement of product quality
- Today we may get 10 or more quality measurements on each sample
- Most outliers remain undetected with the use of classical SPC techniques!
 - No covariance information

Fast and Correct Decision Making



- The outliers are not detected until you look at the combination of the variables
- The information is found in the correlation pattern - not in the individual variables!

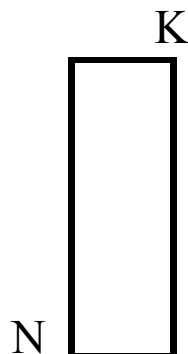
Classical methods of statistics ~1930

- Multiple regression
- Canonical Correlation
- Linear Discriminant Analysis
- Analysis of variance
- & all Maximum Likelihood

Assumptions and limitations:

- Independent X's (Rank $X = K$) => Many Cases, Few variables
- Precise X's ("errors" only in Y)
- Regression analysis one Y at a time (independent Y's)
- No missing data

Tables are long and lean



Chemistry , Biology, Engineering ... ~2000

- Experimental Costs (+ ethics, regulations)
- Instrumental & electronics revolution

- Projection methods: PCA, PLS, PLS-DA, PCR

- Each "classical" method has a projection correspondence

Few cases (observations), Many variables

Chemometrics
Short & fat data tables



Example: FOODS

Problem I --- Overview

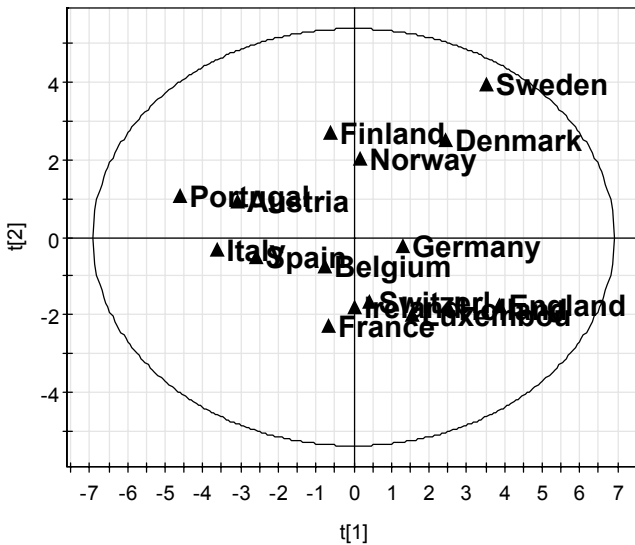
Problem: To investigate the food consumption pattern in Europe; the relative amount of twenty common products are given for 16 countries.

Perform a multivariate analysis (PCA) to overview data!!!

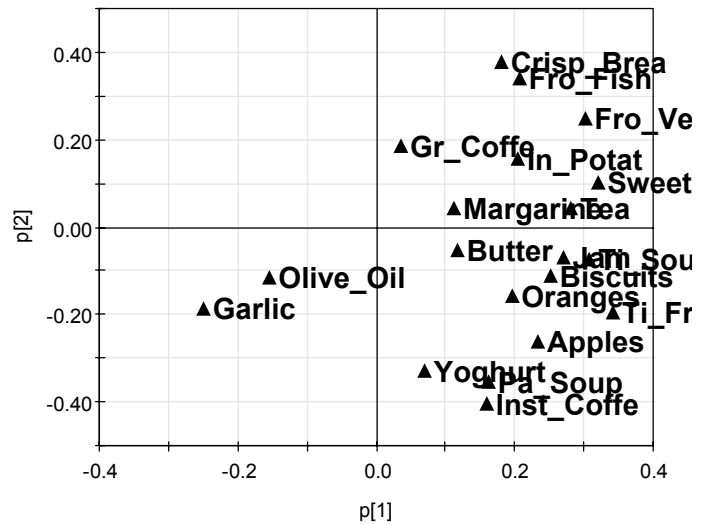
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	Prima	ONAM	Gr_Co	Inst_C	Tea	Sweet	Biscu	Pa_So	Ti_So	In_Pot	Fro_Fj	Fro_Ve	Apples	Orangi	Ti_Fru	Jam	Garlic	Butter	Margar	Olive_C	Yoght	Crisp
2	1	Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	30	26
3	2	Italy	82	10	60	2	55	41	3	2	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	Luxembo	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	Switzerl	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

FOODS: PC model, component 1 and 2

FOODS.M1 (PCA-X), PCA for overview
t[Comp. 1]/t[Comp. 2]



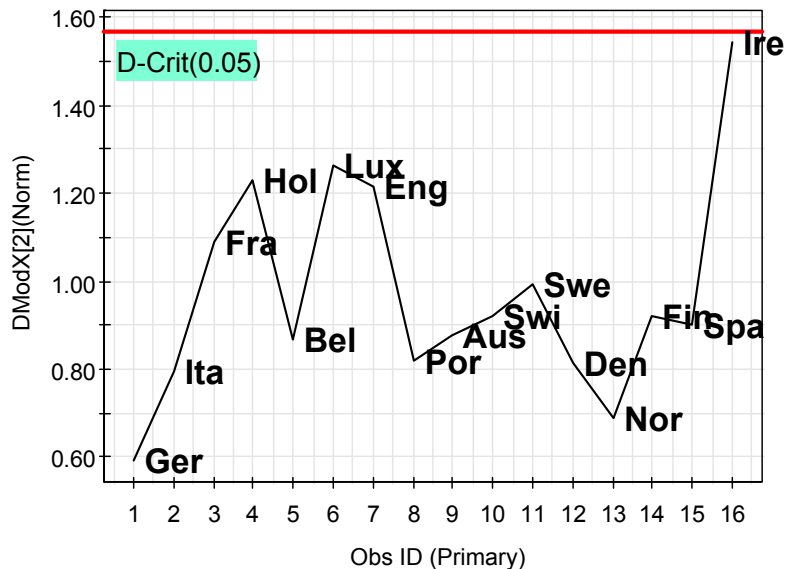
FOODS.M1 (PCA-X), PCA for overview
p[Comp. 1]/p[Comp. 2]



FOODS: Distance to model (DModX) after 2 components

- DModX represents the unexplained variation ("residuals")
- Ireland is farthest away from the model plane \Leftrightarrow displays largest portion of unexplained variation

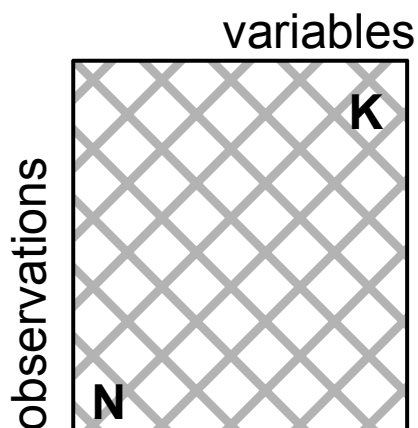
FOODS.M1 (PCA-X), PCA for overview
DModX[Comp. 2]



M1-D-Crit [2] = 1.566

Projection Methods

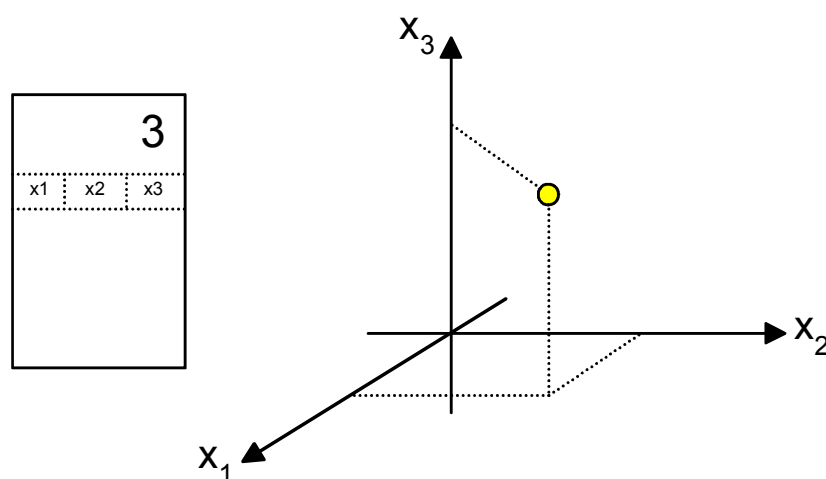
We have a large data table of dimensions $N \times K$



- **Observations** might be:
 - Analytical samples
 - Compounds
 - Experimental runs (trials)
 - Reactions
 - Process time points
 - Individuals
 - ...

- **Variables** might be:
 - From spectra: NMR, IR, NIR, UV, MS, X-ray, ...
 - From separation: HPLC, GC, TLC, Electrophoresis
 - Process: T, pH, P, flows, ...
 - Other: Curve forms, structure descriptors, thermodynamics, quantum mechanics, elemental compositions,...

The Principles of Projections



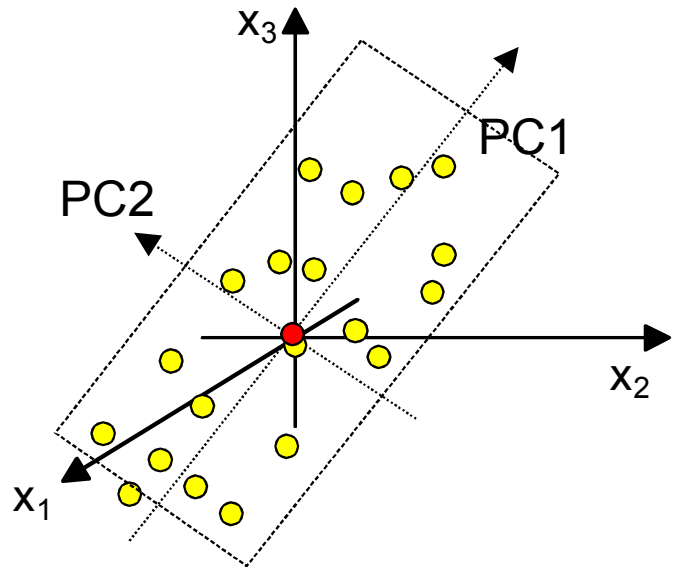
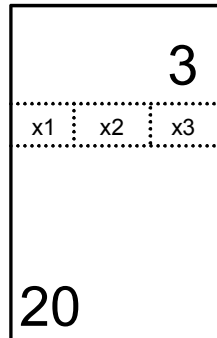
- Each row in a data table corresponds to a point in K -space, here 3-dimensional space

The Principles of Projections

- We are looking for windows in K-space, which can be used to overview the data and interpret their meaning

- **Convert data tables to plots**

- Projection onto line in K-space
- Projection onto plane
- Projection onto hyperplane



Philosophy of modelling

- **Models**

WHAT ? Approximation of reality

Analogy: Toy train

Map

Math function

WHY ? Simplifies study of reality

faster

simpler

cheaper

HOW ? Mathematical models

empirical $y = a + bx + e$

semi empirical $y = a + b \log x + e$

fundamental $H\psi = E\psi + e$
 $pV = nRT + e$

Models founded on analogy and projection principles!

All models are approximations of reality – always founded on some assumptions !

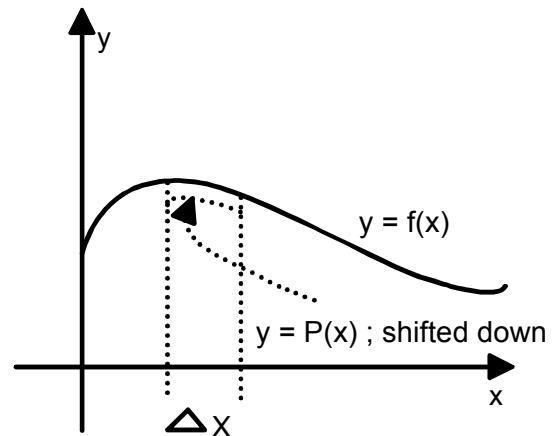
Semi-Empirical modelling: Taylor series.

- In a limited interval, Δx , any continuous and differentiable function, $f(x)$, can be arbitrarily well approximated by a polynomial, a Taylor series:

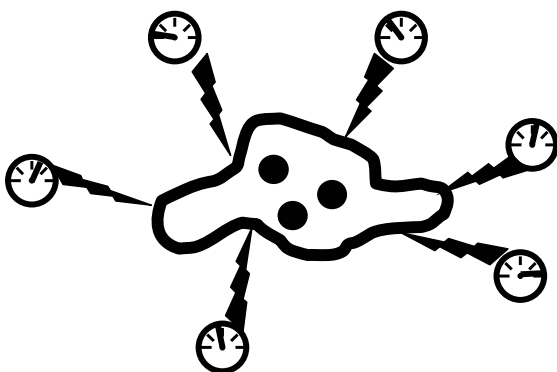
$$y = P(x) = b_0 + b_1x + b_2x^2 + \dots + b_px^p + \dots + E(p + 1)$$

- For a given degree, p , the approximation is better, the smaller the interval Δx
- For a given interval Δx , the approximation is better, the higher the degree, p
- The above can be generalised to functions of many variables:

$$y = P(x_1, x_2, \dots) + E = b_0 + b_{11}x_1 + b_{22}x_2 + \dots + b_{11}x_1^2 + b_{22}x_2^2 + \dots + b_{12}x_1x_2 + b_{13}x_1x_3 + \dots + b_{11\dots 1}x_1^p + \dots + E(p+1)$$



Basic Conceptual Projection Model



Any set of variables (X or Y) measured on a system or process with limited variation (observations are similar) can be approximated by a bilinear model

- Fictitious process
 - 3 latent variables ●
 - 6 measured variables 📊

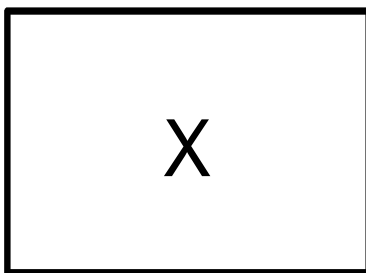
$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}$$

Summary - Multivariate Analysis by Projections (PCA & PLS)

- Deals with the dimensionality problem
- Handles many variables and few observations
 - Short and wide data tables
- Handles few variables and many observations
 - Long and lean data tables
- 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

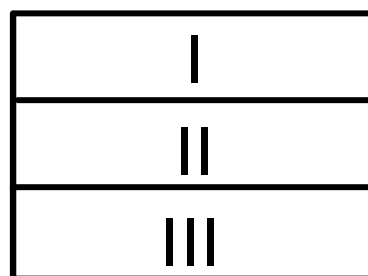
Summary - Three types of problems

Any question to a data table has a projection method solution



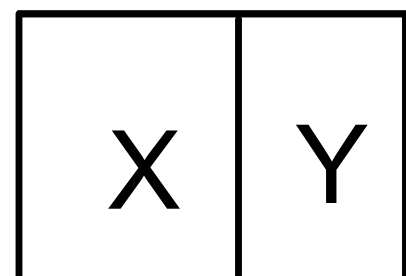
Overview and Summary

- PCA, Principal Components Analysis



Classification

- SIMCA, one PC model per class
- PLS-DA, PLS discriminant analysis
- MSPC (one class + "outliers")



Relation between blocks of variables, X & Y

- PLS analysis
- PLS-DA
- Multiple regression

Multivariate Data Analysis and Modelling Basic Course

Chapter 2 Principal Component Analysis (PCA) – Overview of data tables

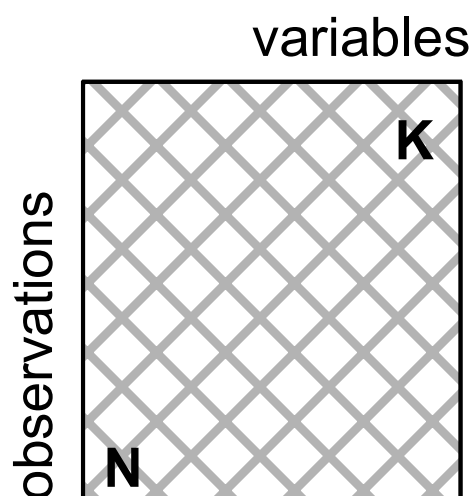


Contents

Principal Components Analysis (PCA)

- Notation
- Scaling of Variables
- Geometric Interpretation
- Algebraic Solution
- Example
- Diagnostics
 - Outliers
 - Residuals
 - Model complexity & Predictive power (Cross-validation)
- Conclusions

Notation



Observations

- Analytical samples
- Chemical compounds
- Trials (experimental runs)
- Process time points
- Chemical reactions
- Biological individuals
- Etc...

Variables

- From spectra
 - NMR, IR, UV, MS, ESCA, X-ray, ...
- From separation
 - HPLC, GC, TLC, Electrophoresis, Trace elements, ...
- Process
 - T, P, pH, flows, ...
- Others
 - Curves, theory, ...

Notation

N = number of observations

K = number of variables

A = number of principal components

index $i = 1, 2, 3, \dots, N$ is used for observations

index $k = 1, 2, 3, \dots, K$ is used for variables

index $a = 1, 2, 3, \dots, A$ is used for principal components

w_s = scaling weights

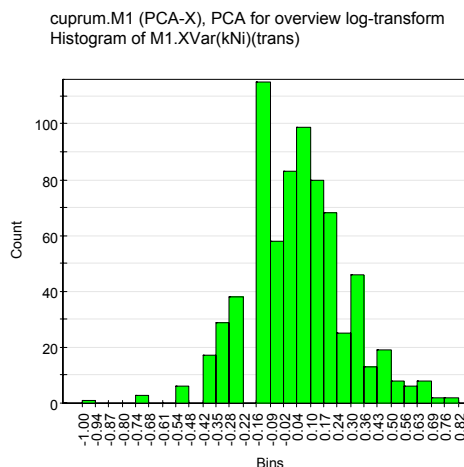
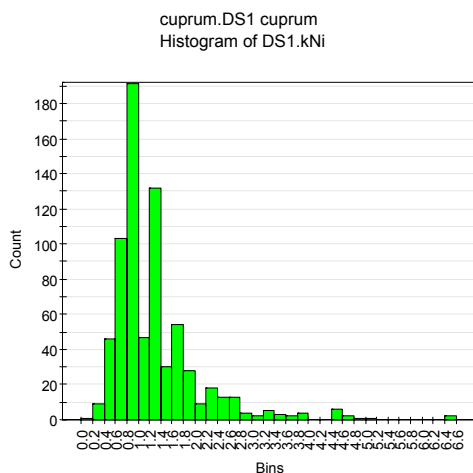
T = scores matrix, t_1, \dots, t_A ; score vectors (column vectors)

P' = loading matrix, p_1', \dots, p_A' ; loading vectors (row vectors)

Pre-treatment of data - Transformations

If the data are not approximately normally distributed, transformation may be needed to get a good model

- Data without transformation
 - skew distribution
- Data after log-transformation
 - closer to normal distribution



Pre-treatment of data - Scaling

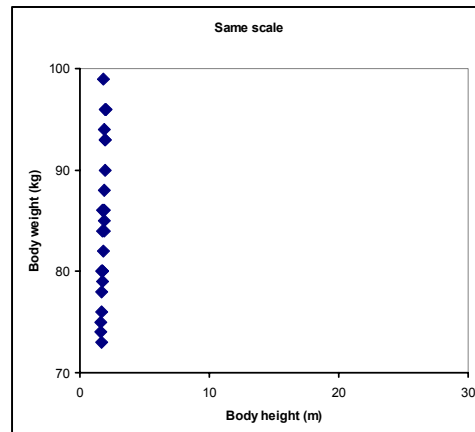
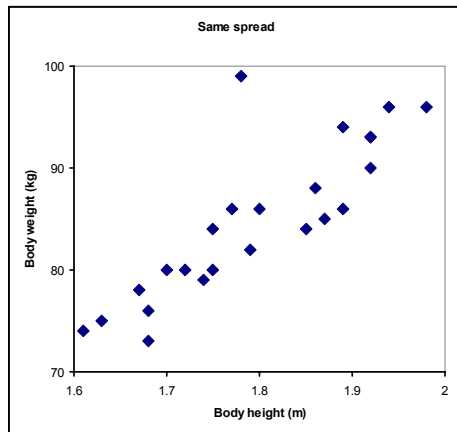
- **Problem:** Variables can have substantially different ranges
- **Example:** LOWARP, polymers characterised with regards to strength and warp
- Response wrp3 varies between 0.2 and 1.0
- Response st4 ranges from roughly 17000 to 30000

Response																	
5	6	7	8	9	10	11	12	13	14	15	16	17	18				
wrp1	wrp2	wrp3	wrp4	wrp5	wrp6	st1	st2	wrp7	st3	st4	wrp8	st5	st6				
0.9	5.0	0.2	1.0	0.3	4.2	232	15120	1.2	2190	26390	1.3	2400	0.7				
3.7	7.3	0.7	1.8	2.5	5.4	150	12230	1.8	905	20270	2.1	1020	0.6				
3.6	6.9	0.9	2.1	4.8	9.4	243	15550	1.2	1740	21180	1.4	1640					
0.6	3.1	0.3	0.4	0.4	1.1	188	11080	1.0	1700	17630	1.0	1860	0.5				
0.3	2.1	0.3	0.3	0.8	1.1	172	11960	1.2	1810	21070	1.3	1970	0.5				
1.2	5.0					245	15600	1.1	2590	25310	1.3	2490	0.6				
2.3	3.9	0.3	0.4	0.7	1.4	242	13900	1.5	1890	21370	1.6	1780					
2.6	5.9	0.4	0.2	0.7	1.2	243	17290	1.6	2130	30530	1.6	2320	0.7				
2.2	5.3	0.2	0.7	0.6	2.0	204	11170	1.0	1670	19070	1.1	1890	0.6				
5.8	7.0	0.9	1.0	5.6	11.8	262	20160	1.6	1930	29830	1.8	1890					
0.8	2.9	0.5	0.6	1.1	2.0	225	14140	1.3	2140	22850	1.3	2110	0.7				
2.8	5.1	1.0	1.2	2.7	6.1	184	15170	1.9	1230	23400	2.1	1250	0.6				
1.1	4.7	0.6	0.9	1.3	3.5	198	13420	1.4	1750	23790	1.4	1930	0.7				
1.9	4.7	1.0	1.0	2.8	5.4	234	16970	1.5	1920	25010	1.6	1790	0.7				
2.9	5.9	0.5	0.6	1.0	6.6	239	15480	1.5	1800	23140	1.6	1730					
5.5	7.9	0.8	2.4	5.5	9.3	256	18870	1.5	1880	28440	1.8	1790					
3.2	6.0	0.3	0.5	1.5	5.2	249	16310	1.5	1860	24710	1.7	1780					

Scaling, Example Body height/Body weight

Data for 23 individuals (22 players + referee; football game)

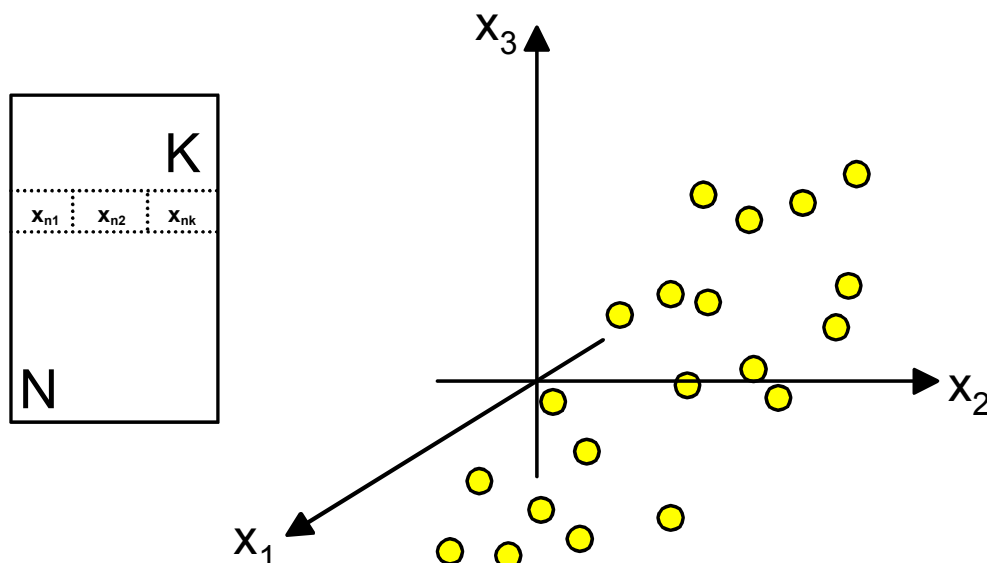
Height (m)	1.8	1.61	1.68	1.75	1.74	1.67	1.72	1.98	1.92	1.7	1.77	1.92
Weight (kg)	86	74	73	84	79	78	80	96	90	80	86	93
Height (m)	1.6	1.85	1.87	1.94	1.89	1.89	1.86	1.78	1.75	1.8	1.68	
Weight (kg)	75	84	85	96	94	86	88	99	80	82	76	



**Right:
The
outlier
not so
easy to
spot!**

Scaling of variables

- Defining/Selecting the length of variable axes
- **Recommended:** To set variability along each axis equal to one (unit variance)



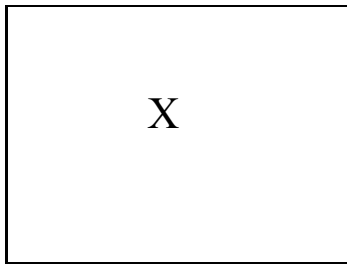
Unit variance scaling (UV-scaling)

- PCA is scale dependent

Variance of a variable \Leftrightarrow "importance"

Drawback ?

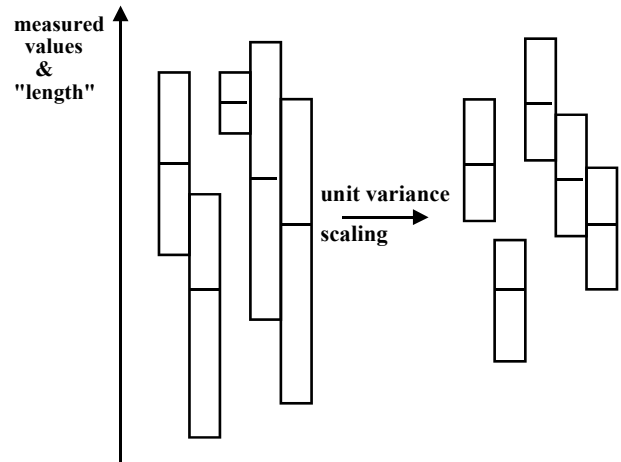
Asset?



WS

$$x_{ik} = x_{ik} * ws_k; \quad ws_k \text{ is scaling weight of var. } k$$

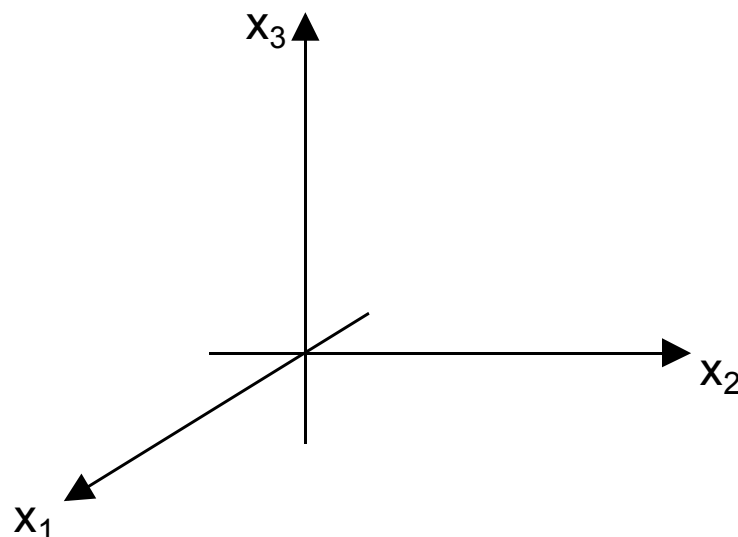
$$ws \Rightarrow 1/s_k$$



All variables $\Rightarrow s^2(x_k) = 1$ (variance = 1)

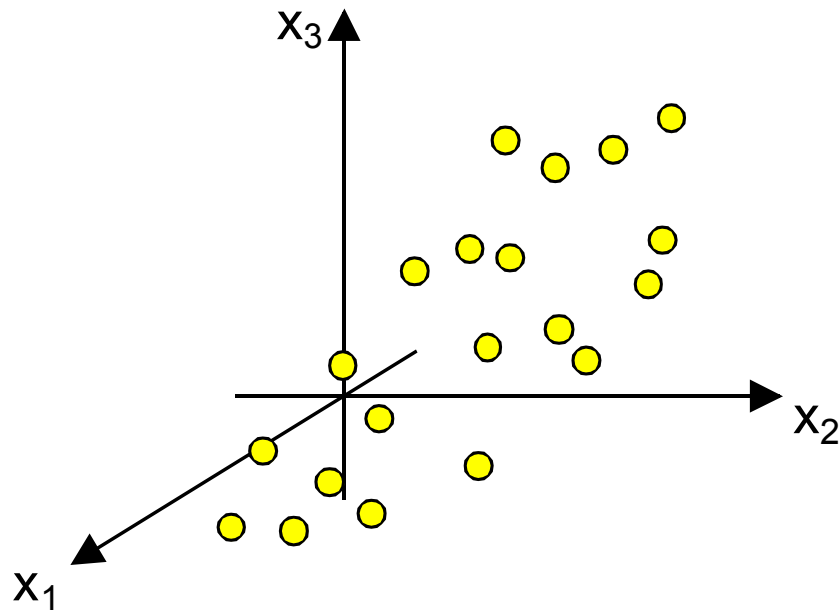
- The scaling can be modified (advanced topic)

PCA -- Geometric Interpretation, 1



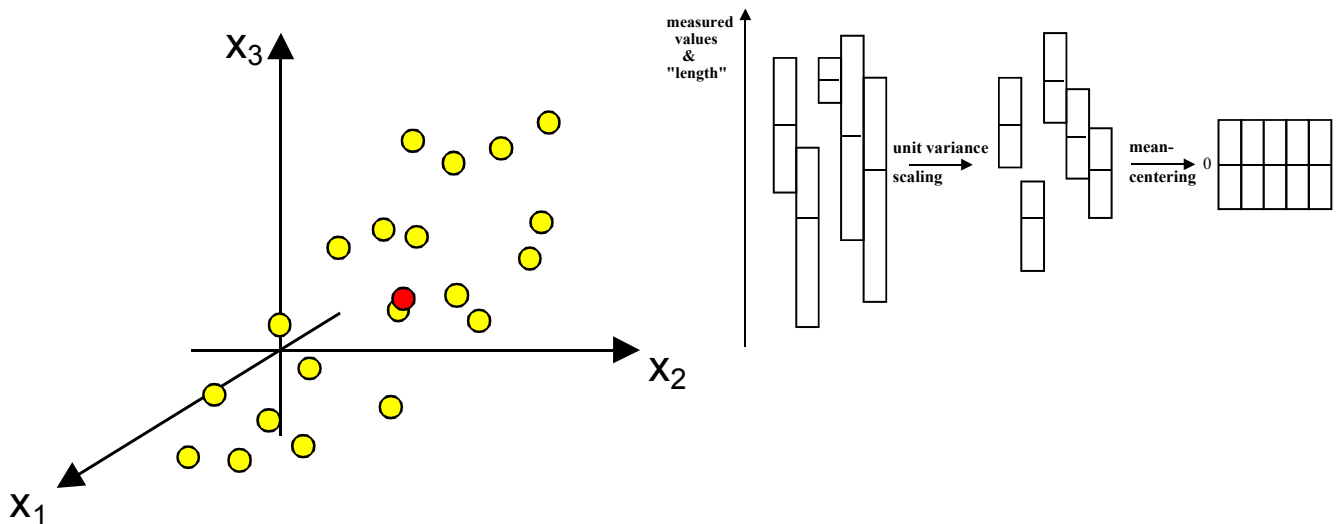
- For the matrix X we construct a space with K dimensions (we see, however, only three of these)
- Each variable has one co-ordinate axis with the length determined by scaling, usually unit variance

PCA -- Geometric Interpretation, 2



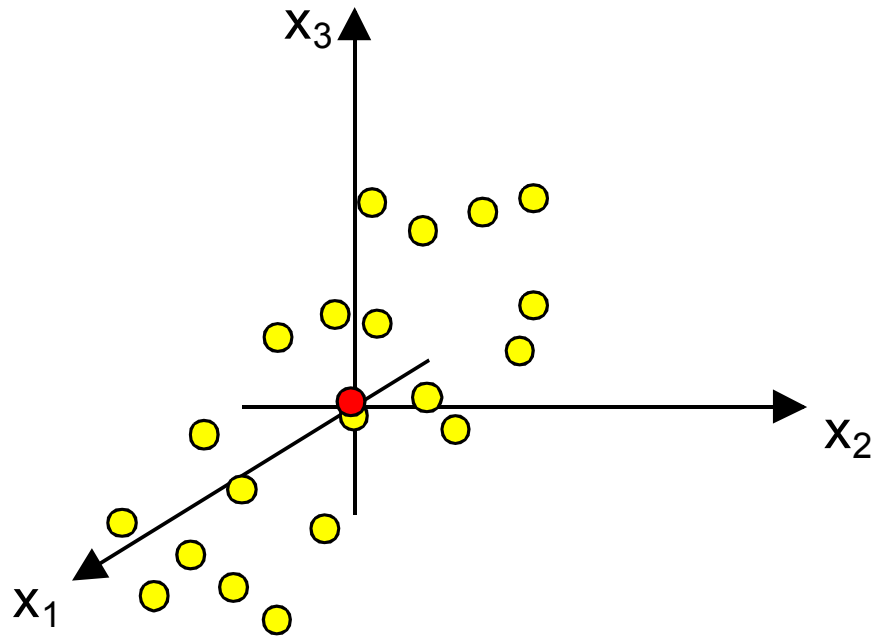
- Each observation is represented by one point in K-space
- Hence, the data matrix X is a swarm of points in this space

PCA -- Geometric Interpretation, 3



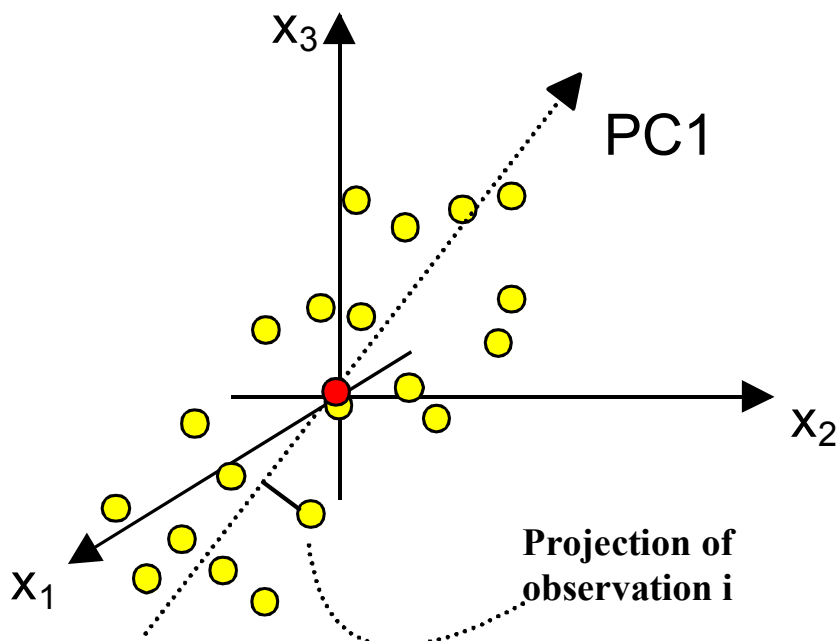
- First we calculate the average of each variable. The vector of averages is a point in K-space. The average is subtracted.

PCA -- Geometric Interpretation, 4



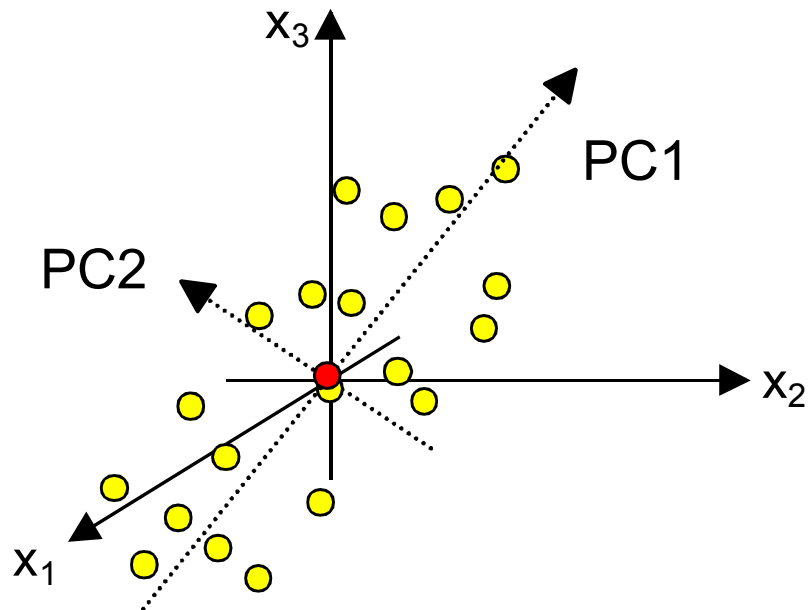
- The mean-centering procedure corresponds to moving the co-ordinate system

PCA -- Geometric Interpretation, 5



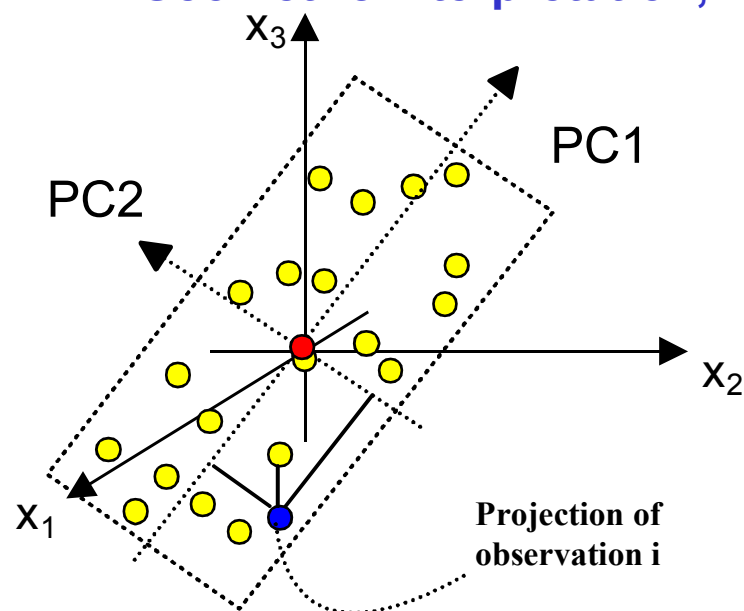
- The first principal component is the line in X-space that best approximates the data (least squares). The line goes through the average point.

PCA -- Geometric Interpretation, 6



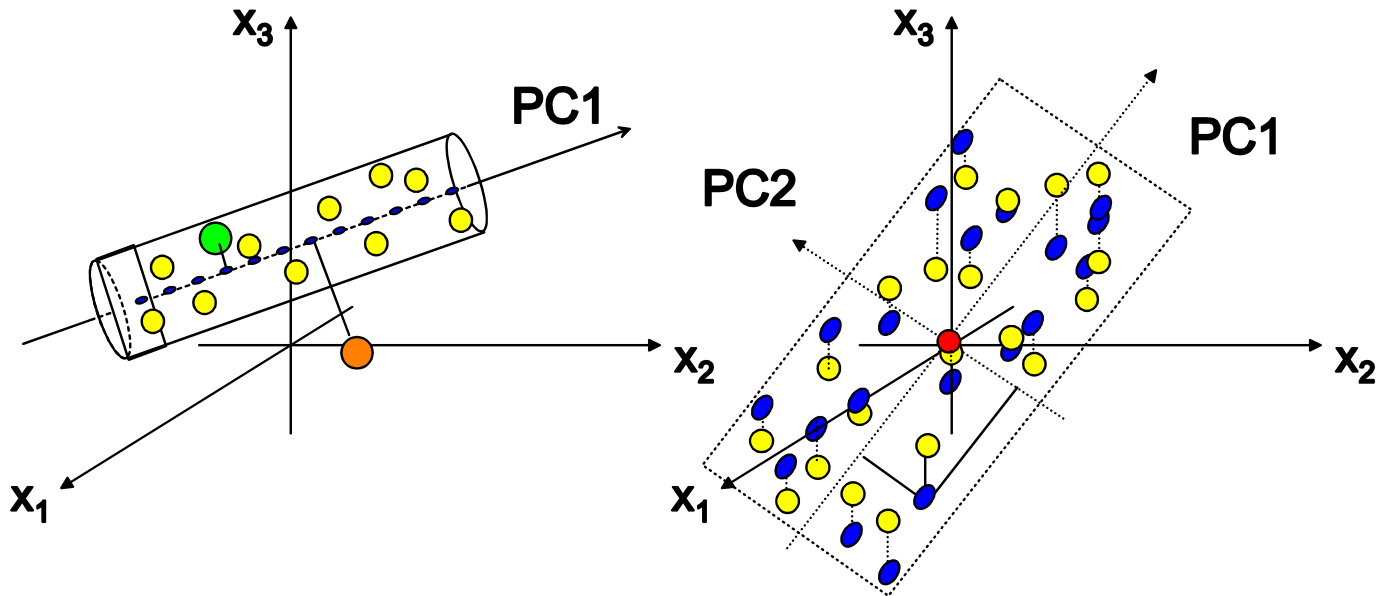
- The second PC is represented by a line in X-space orthogonal to the first PC, also passing through the average point. The second PC improves the approximation of X as much as possible.

PCA -- Geometric Interpretation, 7



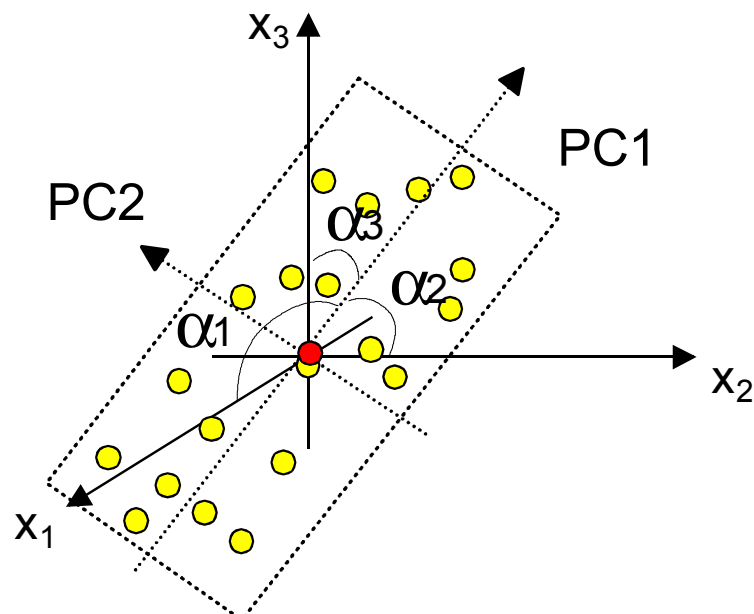
- The two principal components form a plane in the X-space. This plane is a window into the multidimensional space, which can be visualised graphically.

PCA -- Geometric Interpretation, 8



- Yellow points are the observed values. Blue points are their approximations. Projected locations on the model (line, plane, or hyperplane) are given by the *scores* (t).

PCA -- Geometric Interpretation, 9



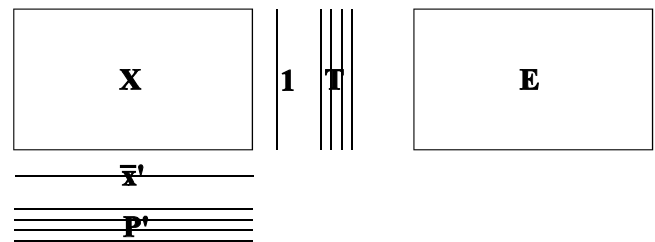
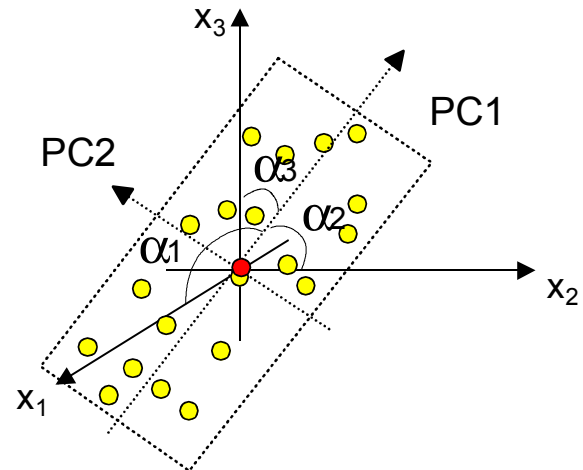
- The direction of, for example, PC1 (p_1) is given by the cosine of the angles α_1 , α_2 and α_3 . These values indicate how the variables x_1 , x_2 and x_3 "load" into PC1. Hence they are called loadings.

PCA, overview of a data table (data set)

- X is modelled as

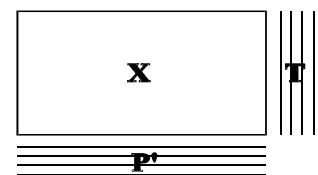
$$X = \mathbf{1} * \bar{X} + T * P' + E$$

- Each PC (score vector) is associated with a loading vector
- **Scores, (t)** are co-ordinates in the (hyper)-plane (columns in T)
- **Loadings, (p)** define the orientation of the (hyper)-plane (rows in P')
- **DModX**, is the distance between the observations and the model plane (residual row SD)



Scores & Loadings

- The scores, t_{ia} , are new variables that summarise the old ones
- The scores are sorted in descending importance, t_1, t_2, t_3, \dots
- Typically 2-5 principal components are sufficient to summarise a data table well
- The loadings, p_{ak} , express how the old variables are linearly combined to form the scores; scores are combinations of the initial variables
- The loadings are used to interpret the scores. They unravel the **magnitude** (large/small correlation) and the **manner** (positive/negative correlation) in which the variables contribute to the scores (principal components)



Example - Overview (FOODS)

Problem I --- Overview

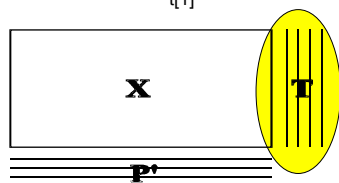
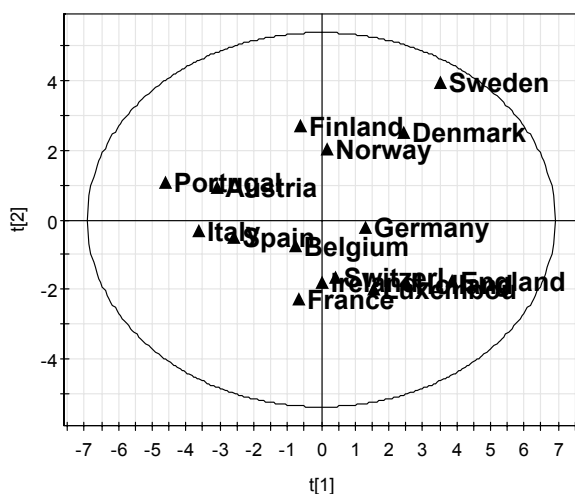
Problem: To investigate the food consumption pattern in Europe; the relative amount of twenty common products are given for 16 countries.

Perform a multivariate analysis (PCA) to overview data!!!

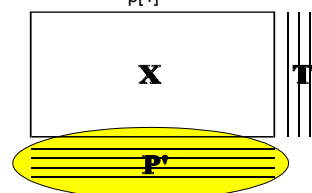
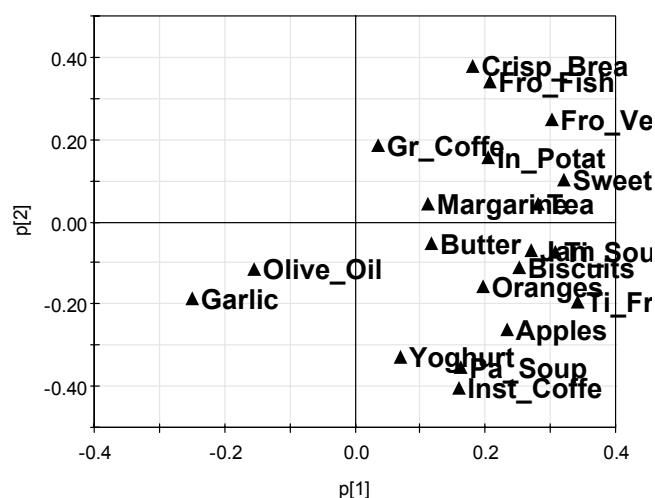
Dataset - FOODS																						
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	Prima	ONAM	Gr_Co	Inst_C	Tea	Sweet	Biscu	Pa_So	Ti_Sou	In_Pot	Fro_F	Fro_Ve	Apples	Orang	Ti_Fru	Jam	Garlic	Butter	Margar	Olive_C	Yogh	Crisp
2	1	Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	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	Luxembo	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	Switzerl	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

Example - Overview (FOODS)

FOODS.M1 (PCA-X), PCA for overview
t[Comp. 1]/t[Comp. 2]

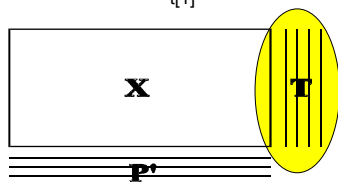
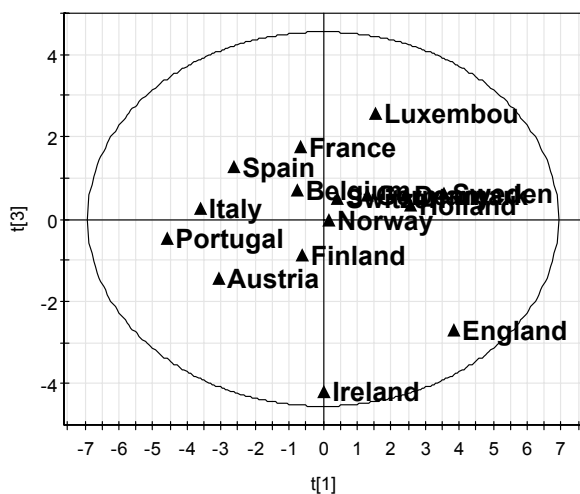


FOODS.M1 (PCA-X), PCA for overview
p[Comp. 1]/p[Comp. 2]

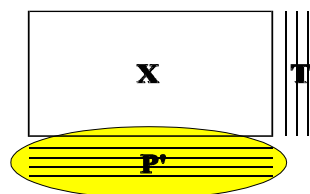
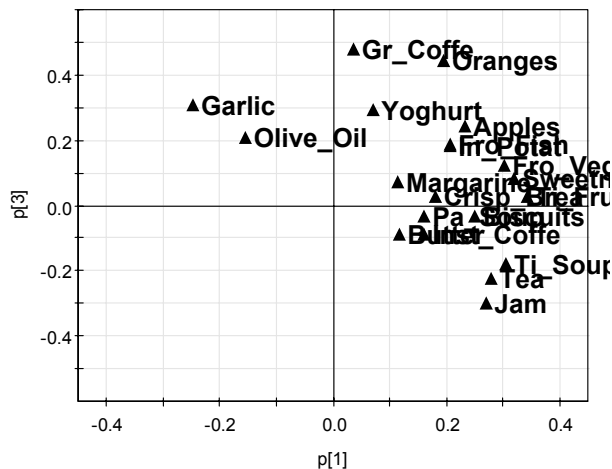


Example - Overview (FOODS)

FOODS.M1 (PCA-X), PCA for overview
t[Comp. 1]/t[Comp. 3]



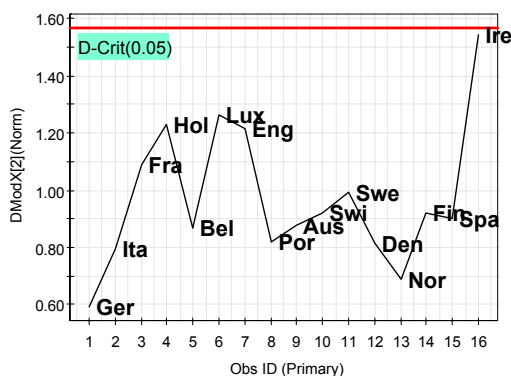
FOODS.M1 (PCA-X), PCA for overview
p[Comp. 1]/p[Comp. 3]



Example - Overview (FOODS)

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

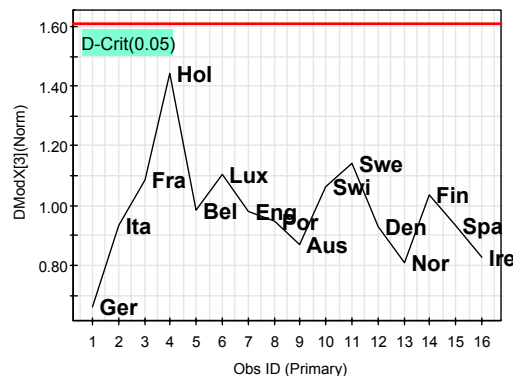
FOODS.M1 (PCA-X), PCA for overview
DModX[Comp. 2]



M1-D-Crit [2] = 1.566

A = 2

FOODS.M1 (PCA-X), PCA for overview
DModX[Comp. 3]



M1-D-Crit [3] = 1.608

A = 3

How to use PCA

Example: LOWARP

- **Problem:** To develop a new polymer with a given profile of warp (low warp) and strength (high strength)

- The polymers consist of 4 constituents (ingredients)

- glas
- crtp
- mica
- amtp

- 17 polymers were made according to a mixture design and 14 responses were determined for each polymer

- Responses: wrp1-wrp8, st1-st6

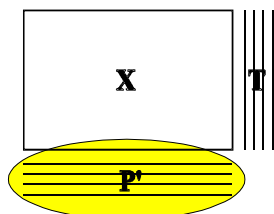
Application: PCA on LOWARP responses

- 17 polymers, 14 responses; warp and strength measured

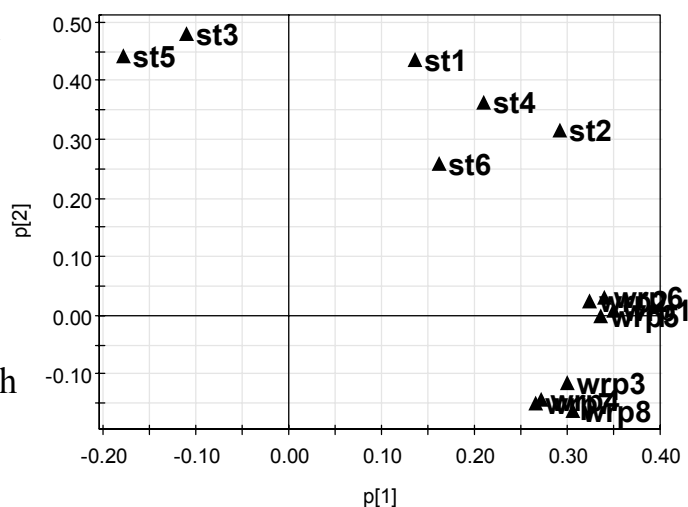
- 14 responses summarised by two principal components (projected onto a plane)

- 1st PC 49% SS
- 2nd PC 27% SS

- Conclusion:
 - PC1 interpreted as mainly warp
 - PC2 interpreted as mainly strength



lowarp.M1 (PCA-Y), PCA of 14 responses
p[Comp. 1]/p[Comp. 2]



PCA - Diagnostics

- **Observation** diagnostics
 - strong and moderate outliers
 - groups
 - trends
- **Variable** diagnostics
 - correlation
 - contribution
 - which variables are well explained
- **Model** diagnostics
 - fit (R^2)
 - prediction (Q^2), Cross-validation

Observation diagnostics

PCA can be used to unravel

- Strong outliers (groupings, trends)
- Moderate outliers (groupings, trends)

Strong outliers:

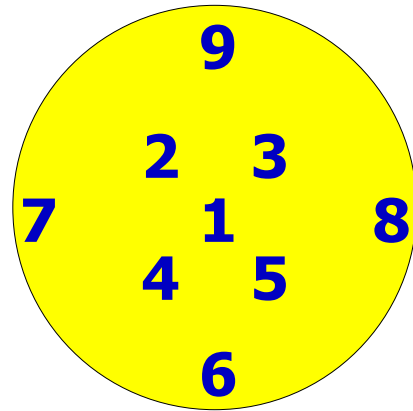
- Found in scores
- Detection tool: Hotelling's T^2
 - a method to establish the “normal” area in the score plot

Moderate outliers:

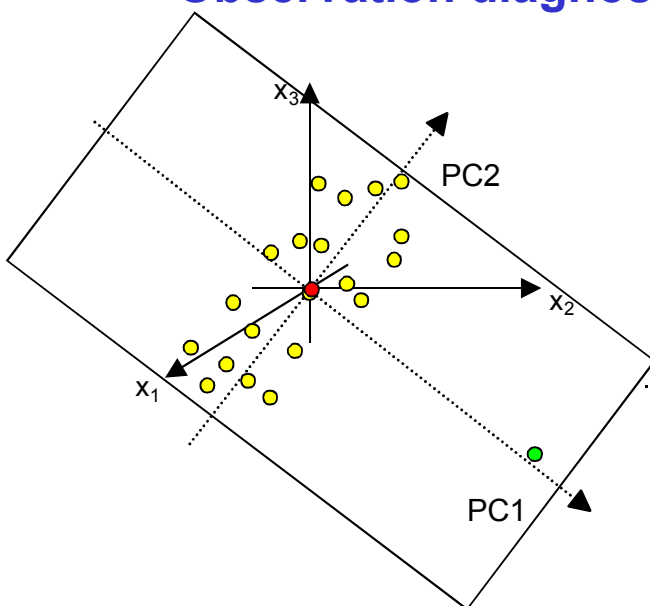
- Found in observation residuals
- Detection tool: DModX

Observation diagnostics; Example

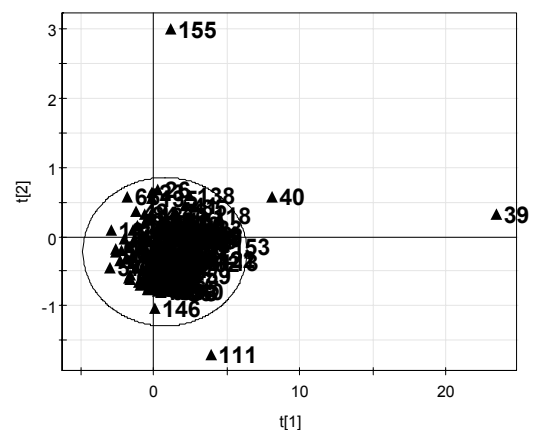
- Thickness: Quality control scheme for polymer disk manufacturing
- The objective was to produce disks with uniform thickness within given specification
- Problems stemmed from small but expensive increases in the number of disks discarded
- Nine thickness measurements were taken on the disks produced



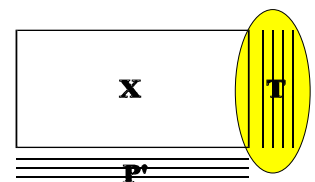
Observation diagnostics - Strong Outliers



thicknes.M1 (PCA-X), PCA for overview
t[Comp. 1]/t[Comp. 2]



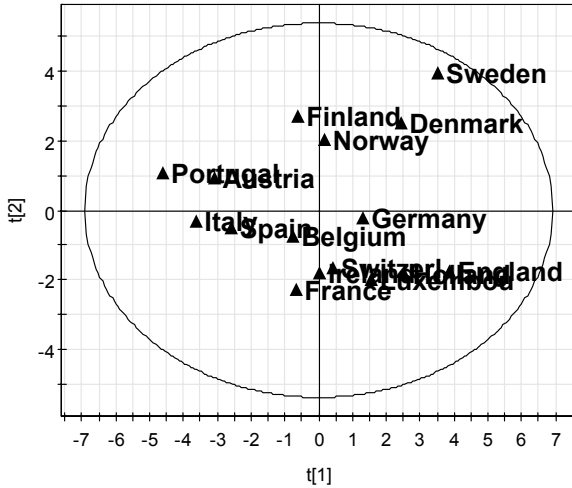
- **Outliers** are serious and interesting, and easy to find
- **Strong outliers** are found in score plots



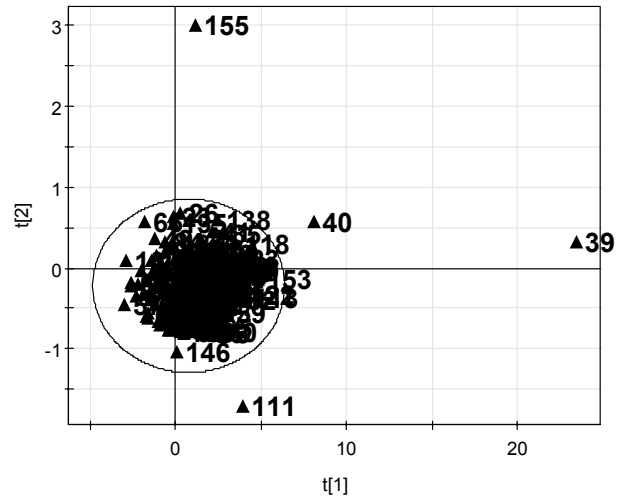
Strong outliers - detection tool Hotelling's T²

- Hotelling's T² is a multivariate generalisation of Student's t-test
- It provides a tolerance region for the data in a two-dimensional score plot, e.g., t₁/t₂

FOODS.M1 (PCA-X), PCA for overview
t[Comp. 1]/t[Comp. 2]



thicknes.M1 (PCA-X), PCA for overview
t[Comp. 1]/t[Comp. 2]



Observation diagnostics - Moderate Outliers

- Moderate outliers can be detected by inspecting the residual for each observation (DModX)

- Residual observation variation (SSOX)

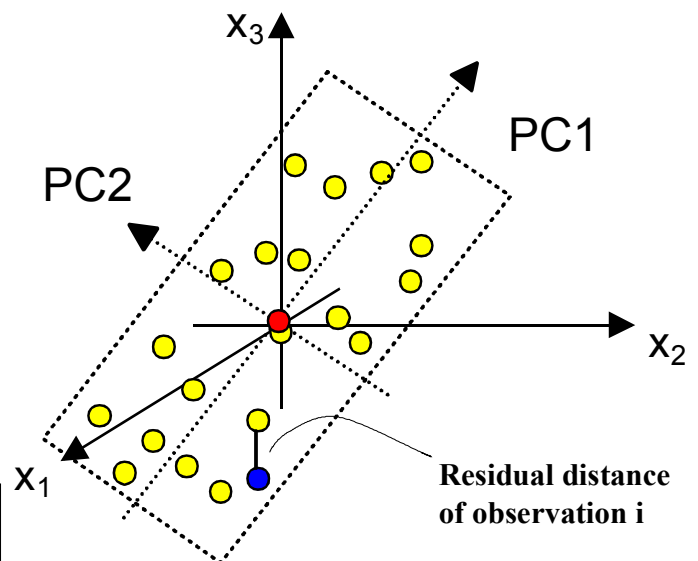
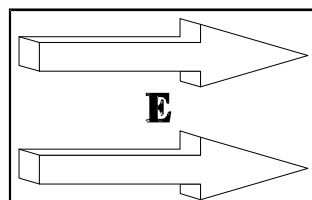
$$\sum_k e_{ik}^2$$

- Residual observation variance (S2OX)

$$\sum_k e_{ik}^2 / DF$$

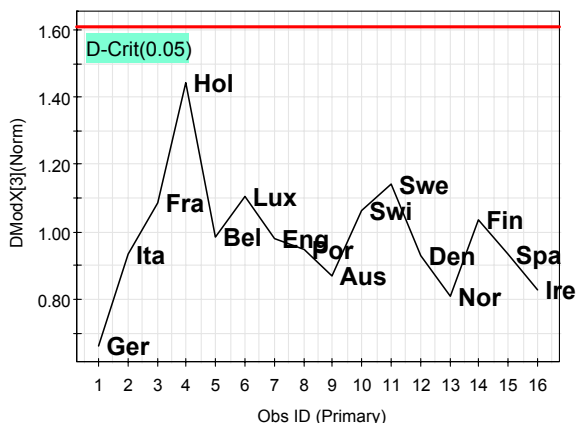
- DModX, normalised distance
[S2OX/variance (E)]^{1/2}

- DModX, absolute distance
[S2OX]^{1/2}



Moderate Outliers - Detection tool DModX

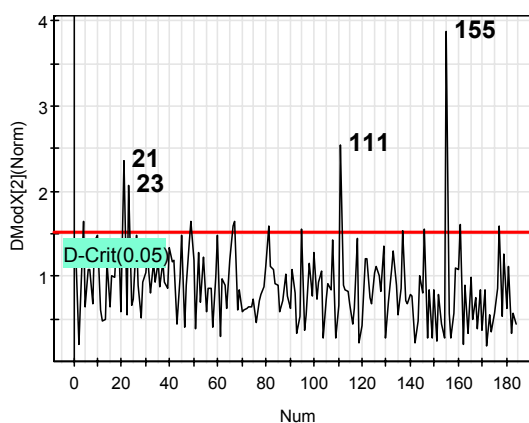
FOODS.M1 (PCA-X), PCA for overview
DModX[Comp. 3]



M1-D-Crit [3] = 1.608

No moderate outliers

thicknes.M1 (PCA-X), PCA for overview
DModX[Comp. 2]



M1-D-Crit [2] = 1.512

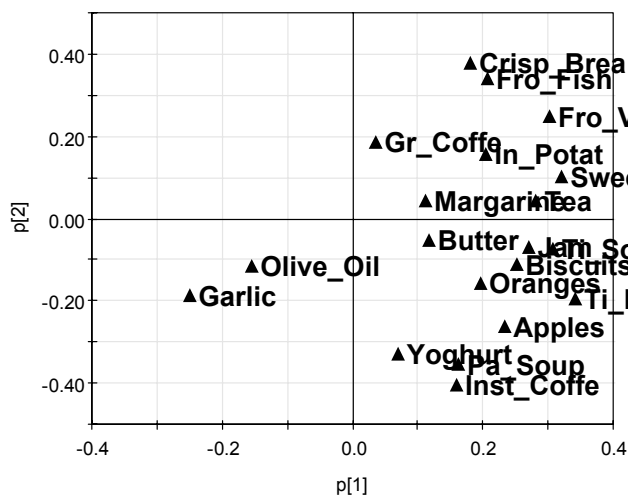
One to four moderate outliers

Critical distance (DCritX) is derived from the approximate F-distribution of DModX²

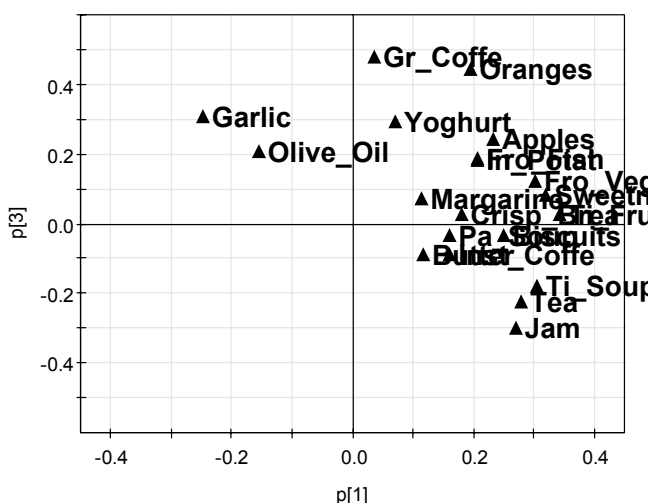
Interpretation of variables

- Variable correlations and model contribution can be seen in loading plots

FOODS.M1 (PCA-X), PCA for overview
p[Comp. 1]/p[Comp. 2]



FOODS.M1 (PCA-X), PCA for overview
p[Comp. 1]/p[Comp. 3]



Residuals of variables

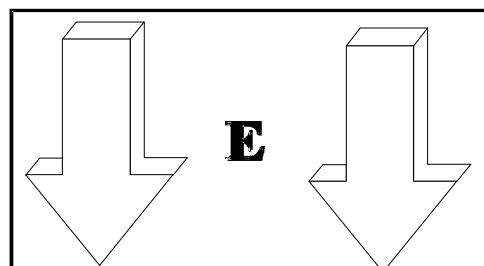
- The residuals tell us the extent to which each variable is modelled (ranges from 0 to 1)
 - Residuals of matrix E pooled column-wise
- SSVX, residual variable variation

$$\sum_i e_{ik}^2$$
- S2VX, residual variable variance

$$\sum_i e_{ik}^2/DF$$
- R2VX (cum), explained variation

$$1 - SSVX[A]/SSVX[0]$$
- R2VXadj(cum), explained variance

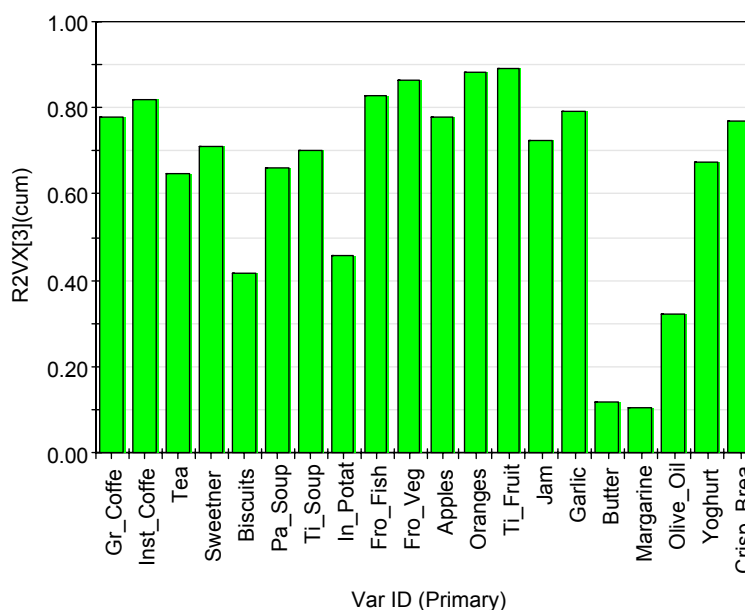
$$1 - S2VX[A]/S2VX[0]$$



Variable diagnostics - Well explained variables

- R2VX and R2VX_{adj} can be used to assess which variables are well explained and which are not
 - R² increases with increased number of principal components

FOODS.M1 (PCA-X), PCA for overview
R2VXcum[Comp. 3]

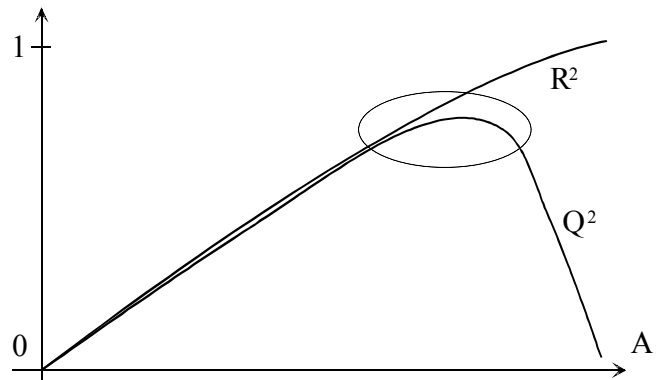


Model diagnostics - Validity vs model complexity

- Trade-off between fit and prediction ability

- **Question:** How can we determine the appropriate number of principal components to include in a model?

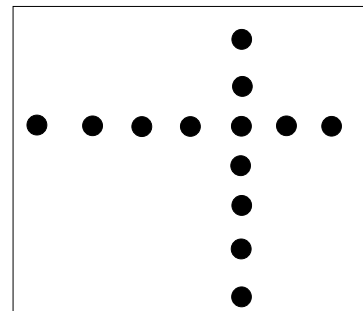
- **Method:** Cross-validation (CV); CV simulates the predictive power of a PC-model.



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

Cross-validation, PCA

- Data are divided into G groups (SIMCA default is 7) and a model is estimated for the data devoid of one group
- The deleted group is predicted by the model \Rightarrow partial PRESS (Predictive Residual Sum of Squares or Prediction Error SS)
- This is repeated G times; then all partial PRESS's are summed to PRESS
- If a new PC_a enhances the predictive power compared with PC_{a-1} (i.e., $PRESS < SS$), the new PC_a is kept in the model



- Cross validation is done in two phases and several deletion rounds:
 - First removal of observations (rows)
 - Then removal of variables (columns)

Model diagnostics - Evaluation of R² and Q²

• **PRESS** is the sum of squared differences between predicted and observed x-elements.

$$\text{PRESS} = \sum (x_{ik} - \hat{x}_{ik})^2$$

• PRESS can be transferred into a dimensionless quantity, Q², which resembles R²

$$Q^2 = 1 - \text{PRESS} / \text{SSX}_{\text{total}}$$

$$R^2 = 1 - \text{SSX}_{\text{resid}} / \text{SSX}_{\text{total}}$$

Q² > .5 Good (Depending on application)

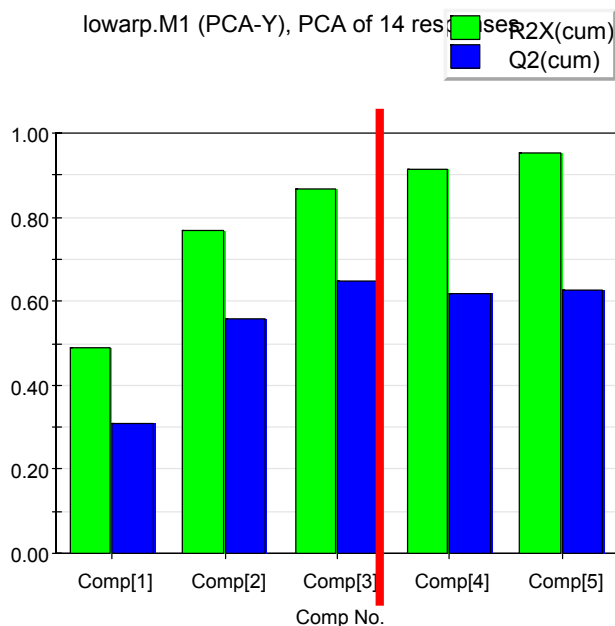
Q² > .9 **Excellent** (Depending on application)

Important:

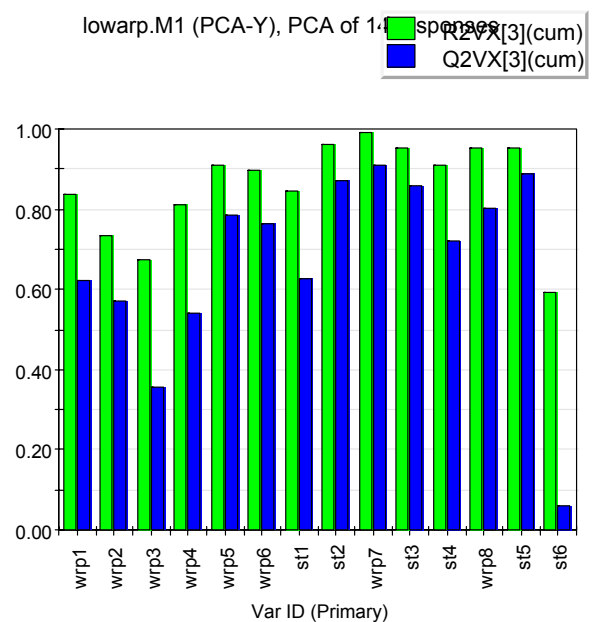
1. R² is always larger than Q²
2. High R² and high Q² is good
3. The difference between R² and Q² should not be too large

Model Diagnostics - Example Lowarp

Model overview



Variable Overview



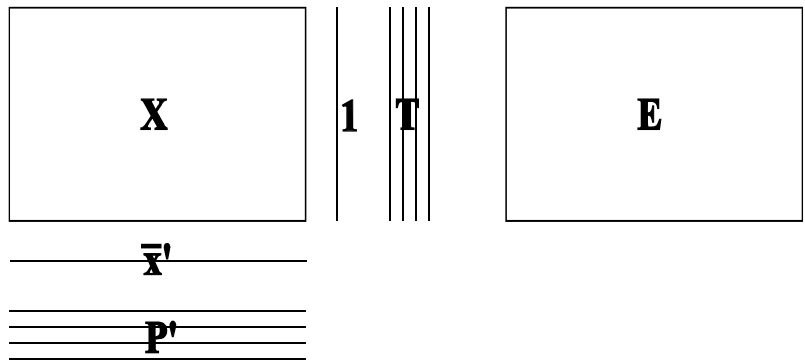
PCA Summary

Principal Component Analysis

- **Modelling:**

Data table X is approximated by a least squares (hyper)-plane + residuals (E)

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



- **Calculations:**

One PC at a time -- NIPALS

All PC.s together -- SVD (can often not handle missing data)

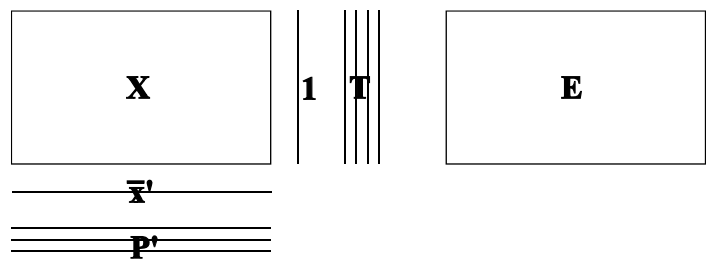
Conclusions

Principal Component Analysis

models the correlation structure of a data set and is used for:

Overview of a data set (data table):

- dominating variables
 - trends
 - outliers, groups, clusters
 - similarities / dissimilarities
- | | |
|--------------|-----------------|
| observations | scores |
| variables | loadings |



Summary of a data set (data table):

- scores ↔ latent variables
principal properties
- loadings ↔ influence of variables

- **Classification:** A new observation is similar to the training observations if it is found within the tolerance volume of the model
- In processes, PCA is used for overview of data and for monitoring

Multivariate Data Analysis and Modelling Basic Course

Chapter 3 Applications of PCA

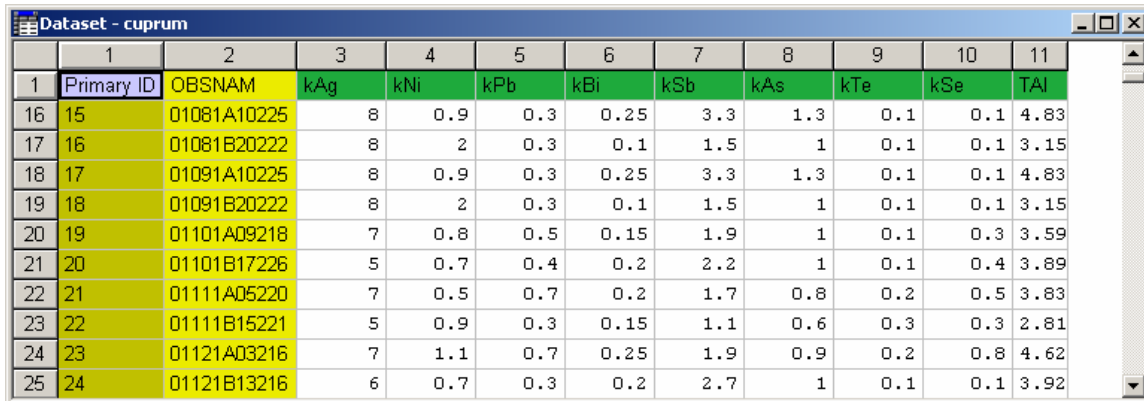


Contents

- **Overview**
 - Quality control
 - Assessment of drug exposure
- **Classification**
 - 3 different types of IRIS
 - 4 classes of wood chips from particleboard industry

Example – Quality control (CUPRUM)

- Electrolytic production of Copper
 - Boliden AB produces approximately 300 tonnes of Copper every day
 - extremely pure (99.998 %) Copper
 - impurity testing twice a day to ensure quality (TAI, Total Analysis Index)
 - TAI is a weighted sum of 8 different impurities (PPM-level)



	1	2	3	4	5	6	7	8	9	10	11
	Primary ID	OBSNAM	kAg	kNi	kPb	kBi	kSb	kAs	kTe	kSe	TAI
16	15	01081A10225	8	0.9	0.3	0.25	3.3	1.3	0.1	0.1	4.83
17	16	01081B20222	8	2	0.3	0.1	1.5	1	0.1	0.1	3.15
18	17	01091A10225	8	0.9	0.3	0.25	3.3	1.3	0.1	0.1	4.83
19	18	01091B20222	8	2	0.3	0.1	1.5	1	0.1	0.1	3.15
20	19	01101A09218	7	0.8	0.5	0.15	1.9	1	0.1	0.3	3.59
21	20	01101B17226	5	0.7	0.4	0.2	2.2	1	0.1	0.4	3.89
22	21	01111A05220	7	0.5	0.7	0.2	1.7	0.8	0.2	0.5	3.83
23	22	01111B15221	5	0.9	0.3	0.15	1.1	0.6	0.3	0.3	2.81
24	23	01121A03216	7	1.1	0.7	0.25	1.9	0.9	0.2	0.8	4.62
25	24	01121B13216	6	0.7	0.3	0.2	2.7	1	0.1	0.1	3.92

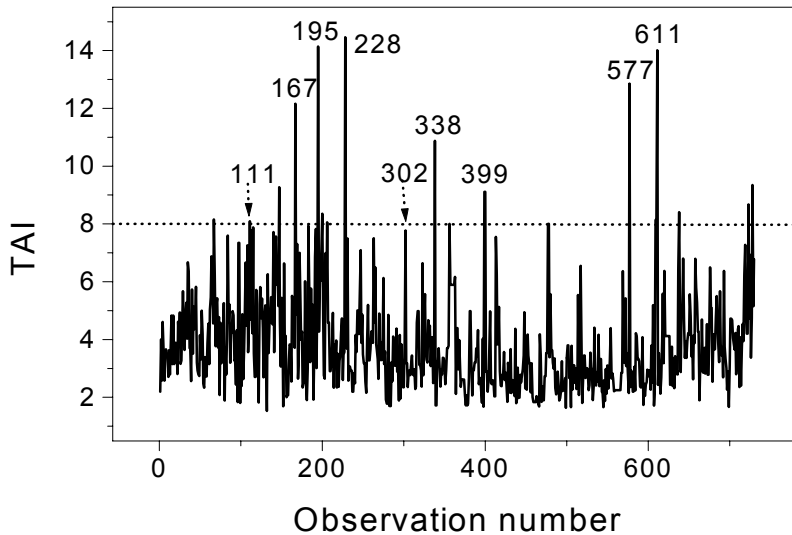
Example – CUPRUM

- The data - 9 variables, 730 observations
 - 8 measured variables (Ag-Se)
 - 1 calculated variable (TAI)
 - data sampled twice a day over one year giving 730 observations
 - all variables were log-transformed
- The Copper industry uses only the TAI value to determine the quality and thereby the price. Copper products with TAI over 8.0 are discarded.
- Question:
 - Can we do better with projection methods?

CUPRUM – Time series plot of TAI

- Quality control limit corresponding to $TAI = 8$
- Samples 111 ($TAI = 8.1$) and 302 ($TAI = 7.8$) have approximately the same TAI value

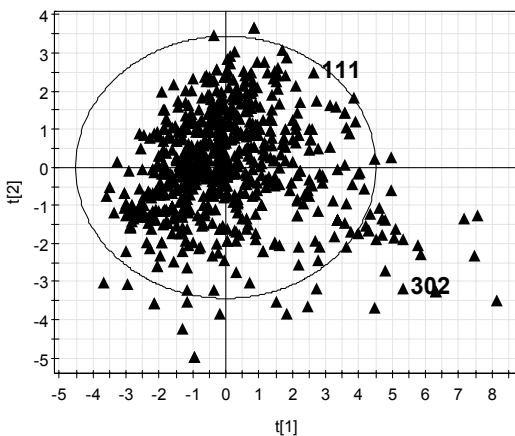
Time series plot of TAI



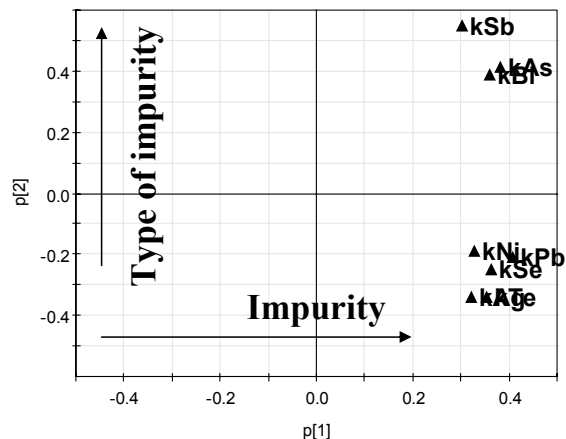
CUPRUM – Scores and loadings of PC-model

- A PC-projection of the table was made. The 8-dimensional table (the TAI variable excluded) was thus projected onto a two-dimensional plane, showing 67% of the variability in the data
- Samples 111 and 302 are situated far apart!
- The corresponding loading plot revealed two types of impurities

cuprum.M1 (PCA-X), pca for overview all vars log-transformed
t[Comp. 1]/t[Comp. 2]



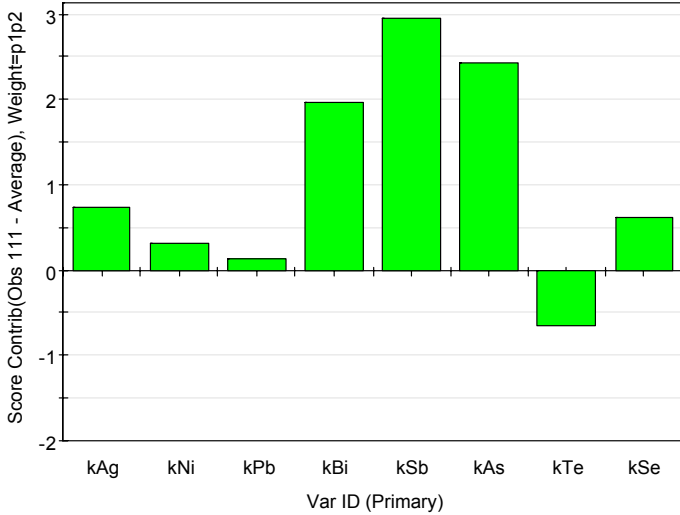
cuprum.M1 (PCA-X), pca for overview all vars log-transformed
p[Comp. 1]/p[Comp. 2]



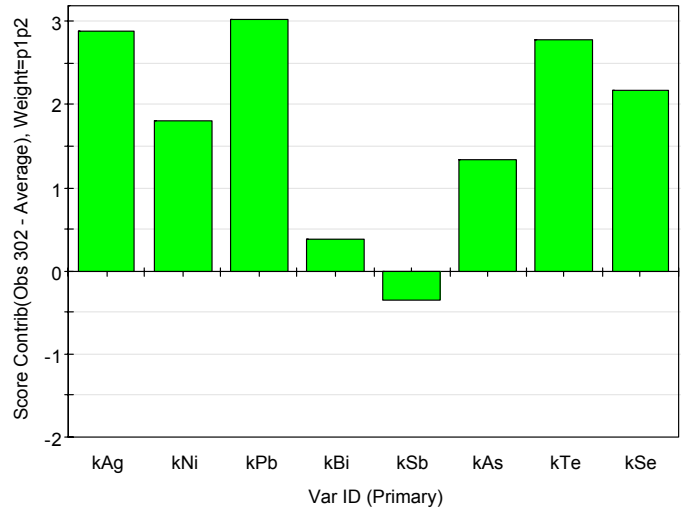
CUPRUM – Contribution plots

- Contribution plots “zoom in” on a single sample. Here, the variable profiles of samples 111 and 302 are shown.

cuprum.M1 (PCA-X), PCA for overview log-transform
Score Contrib(Obs 111 - Average), Weight=p1p2

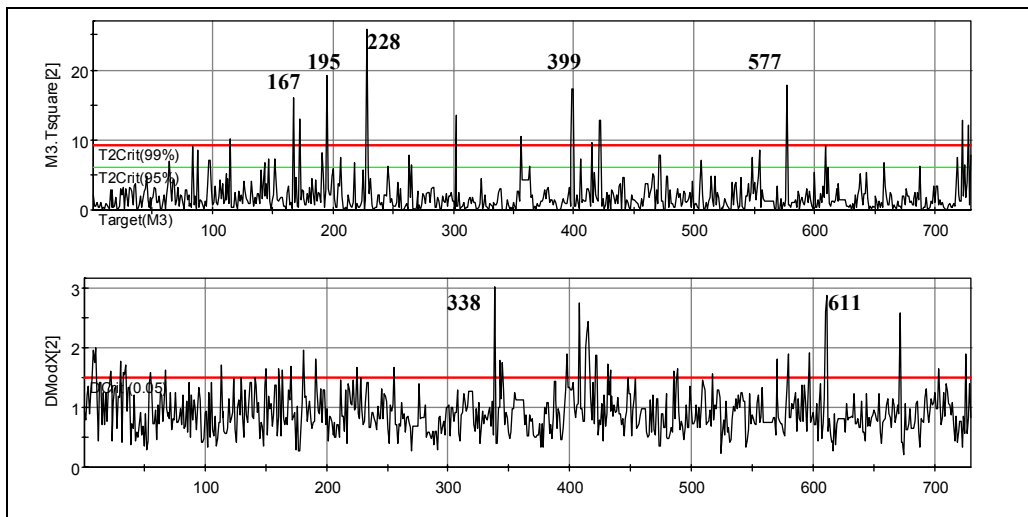


cuprum.M1 (PCA-X), PCA for overview log-transform
Score Contrib(Obs 302 - Average), Weight=p1p2



Statistical process control (SPC) charts

- SPC uses quality and/or process data to monitor the process. Walther Shewhart (father of SPC) introduced the concept for defining the "normal" region of variation for one variable. Today: Multivariate SPC, MSPC.
- Illustrated by Hotelling's T^2 and DModX.



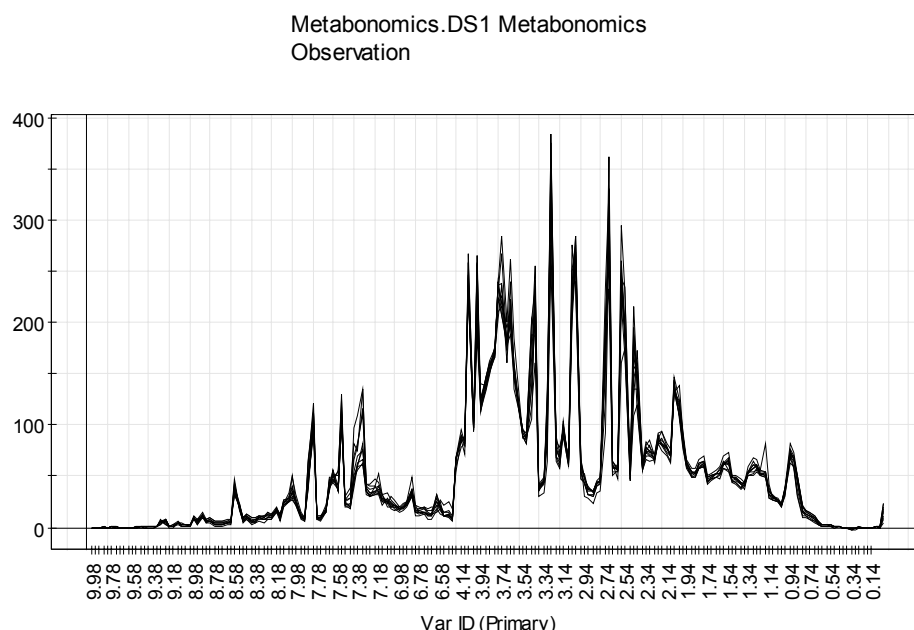
CUPRUM - Summary

The overview of the Copper data

- The PC-analysis
 - The score plot showed that two samples (111 and 302) are **not** similar even though their TAI-values indicate this
 - Using the loading plot, the two samples could be ascribed to two different types of impurities
- Summary
 - The overview of the data provided by the PC-analysis is much more powerful in describing the impurities than the TAI-variable. The TAI-scale does not distinguish between different types of impurity combinations, and therefore leads to loss of information.

Example – Assessment of drug exposure (Metabonomics)

- Metabonomics: monitoring of complex time-related metabolite profiles that are present in biofluids, e.g., urine, plasma, saliva, etc.



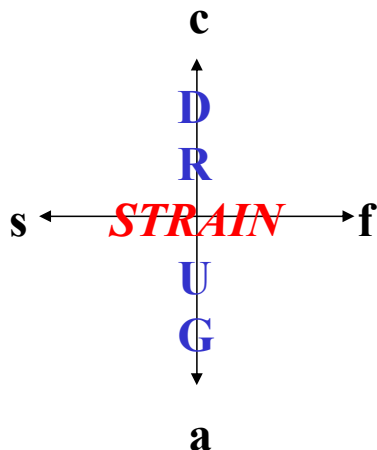
- Proton-NMR spectra of urinary profiles of drug-exposed rats

Metabonomics – The Data

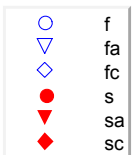
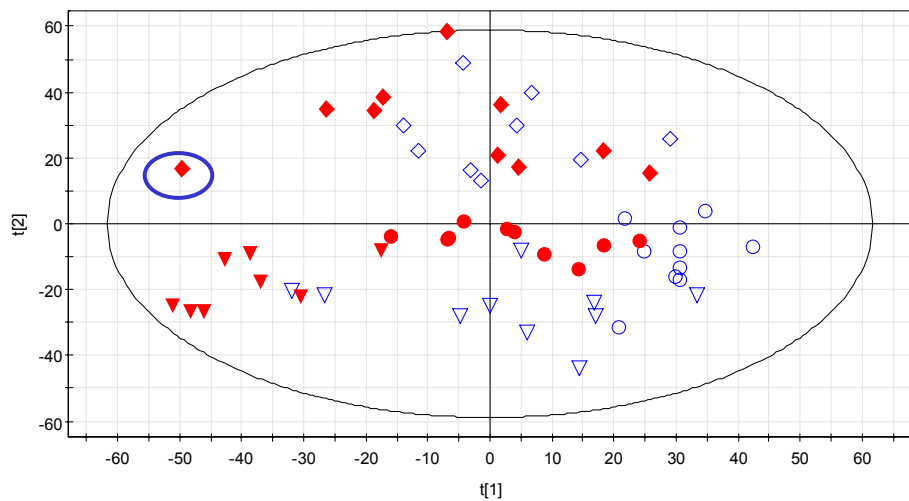
- Rats exposed to chloroquine (an antimalarial) or amiodarone (an antiarrhythmic)
- Observations: $N = 57$ rats
Variables: $K = 194$ variables ($^1\text{H-NMR}$ shift regions)
- Six groups (“classes”):
 - Control Sprague-Dawley, 10 rats, “s”
 - Sprague-Dawley treated with amiodarone, 8 rats, “sa”
 - Sprague-Dawley treated with chloroquine, 10 rats, “sc”
 - Control Fisher, 10 rats, “f”
 - Fisher treated with amiodarone, 10 rats, “fa”
 - Fisher treated with chloroquine, 9 rats, “fc”

Metabonomics – PCA to overview

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



Metabonomics.M1 (PCA-X), Overview with Pareto scaling
t[Comp. 1]/t[Comp. 2]

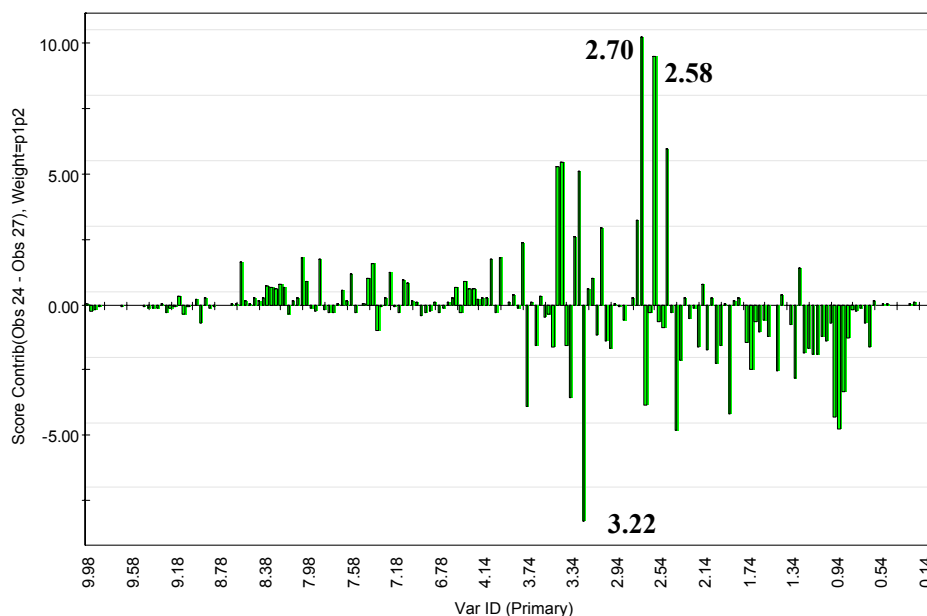


- One outlier, rat 27, encircled
 - Measurement error ?
 - Handling/environmental differences ?
 - Slow responder ?

Metabonomics – Contribution plot to reveal differences

Metabonomics.M1 (PCA-X), Overview with Pareto scaling
Score Contrib(Obs 24 - Obs 27), Weight=p[1]p[2]

- How is rat 27 different from a “normal” sc-rat?
- Chemical shift regions 2.58, 2.70 and 3.22



Metabonomics – Conclusions

- Multivariate analysis of NMR-data creates one or several maps (i.e., score plots, loading plots) that show trajectories of biochemical changes in biofluids induced by toxin exposure or disease
- Through this technology it is possible
 - (i) to detect target organs or pathways of dysfunction
 - (ii) to uncover likely chemical mechanisms of toxicity, and
 - (iii) to identify useful biomarkers indicative of onset, development, and decay of abnormal animal health conditions.

Leading reference: Nicholson, J.K., Connelly, J., Lindon, J.C., and Holmes, E., Metabonomics: A Platform for Studying Drug Toxicity and Gene Function, Nature Reviews, 2002; 1:153-161.

Classification (IRIS)

- **IRIS** *A classical data set in statistics*

- **Data:**

The data table contains petal (sw: kronblad) and sepal (sw: foderblad) lengths and widths of 50 specimens each of *Iris setosa*, *Iris versicolor* and *Iris virginica*. This data set was introduced by the great statistician Fisher as early as 1936. It is commonly known as "The Fisher Iris Data"

- **Objective:**

A multivariate model that classifies a new Iris specimen in the correct group according to its petal and sepal lengths and widths

Training data

Prediction data

	K = 4	K = 4
N = 75	25 Iris Se. (1 - 25)	25 Iris Se. (76 - 100)
	25 Iris Ve. (26 - 50)	25 Iris Ve. (101 - 125)
	25 Iris Vi. (51 - 75)	25 Iris Vi. (126 - 150)

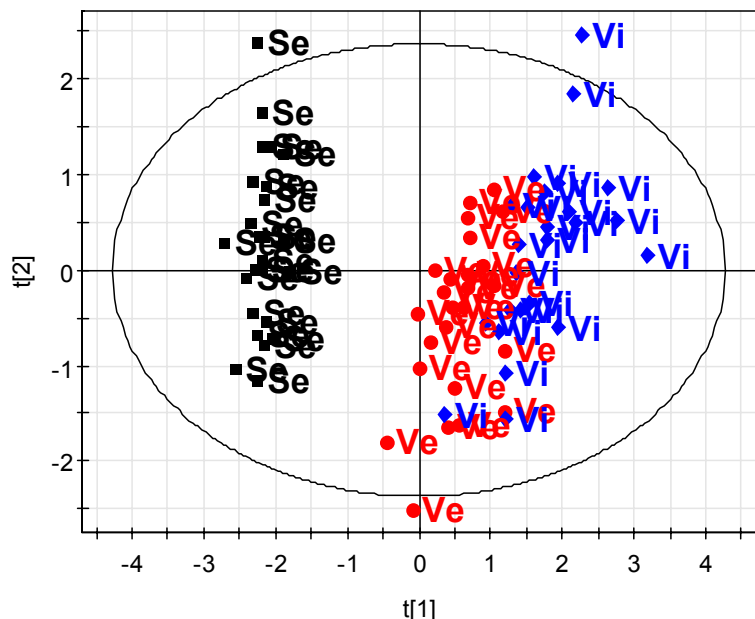
IRIS: Step 1, Overview of training data (PCA)

IRIS training.M1 (PCA-X), PCA entire training set
t[Comp. 1]/t[Comp. 2]

- The PCA score plot shows Setosa well separated from Versicolor and Virginica

- The latter two classes are partly separated

- $R^2 = 0.96$ (A = 2)
- $Q^2 = 0.75$ (A = 2)

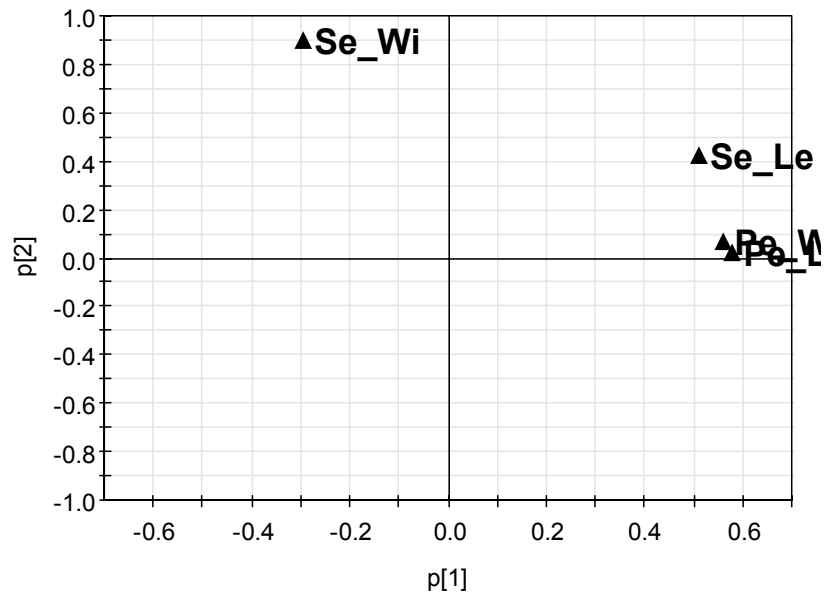


IRIS: Step 1, Overview of data

IRIS training.M1 (PCA-X), PCA entire training set
p[Comp. 1]/p[Comp. 2]

- The loading plot shows that Setosa specimens are smaller (shorter and slimmer) than Virginica and Versicolor samples.

- The variable Se_Wi is mainly responsible for the within class separation of samples.



IRIS: Step 1, A look at the raw data

Sepal Length Sepal Width Petal Length Petal Width

/ Setosa

	Sepal Length	Sepal Width	Petal Length	Petal Width
/ Min	4.30	2.30	1.00	0.10
/ Max	5.80	4.40	1.90	0.60

/ Versicolor

	Sepal Length	Sepal Width	Petal Length	Petal Width
/ Min	4.90	2.00	3.00	1.00
/ Max.	7.00	3.40	5.10	1.80

/ Virginica

	Sepal Length	Sepal Width	Petal Length	Petal Width
/ Min	4.90	2.20	4.50	1.40
/ Max.	7.90	3.80	6.90	2.50

- Conclusion:** Setosa is easy to separate from Virginica and Versicolor

IRIS: Step 2, PC modelling of Ve/Vi

- **Question**

How do we separate Virginica and Versicolor ?

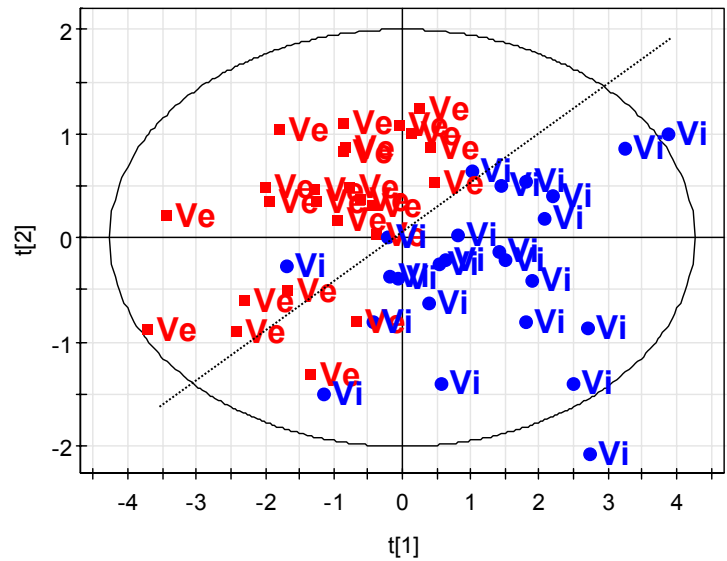
- **First attempt**

A PC model with only Virginica and Versicolor ?

- **Conclusion**

Some but not complete separation

IRIS training.M2 (PCA-X), Setosa excluded
t[Comp. 1]/t[Comp. 2]



IRIS: Step 3, SIMCA

- A separate PC model is made for each class (Se/Ve/Vi)
- Then all prediction data (75 obs) are subjected to each model

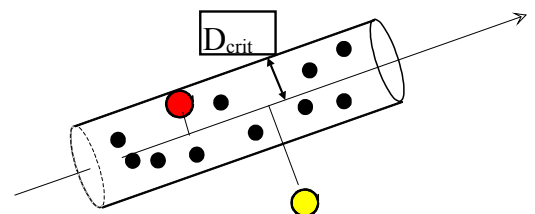
IRIS training

Observations (N) = 75, Variables (K) = 4

Models:

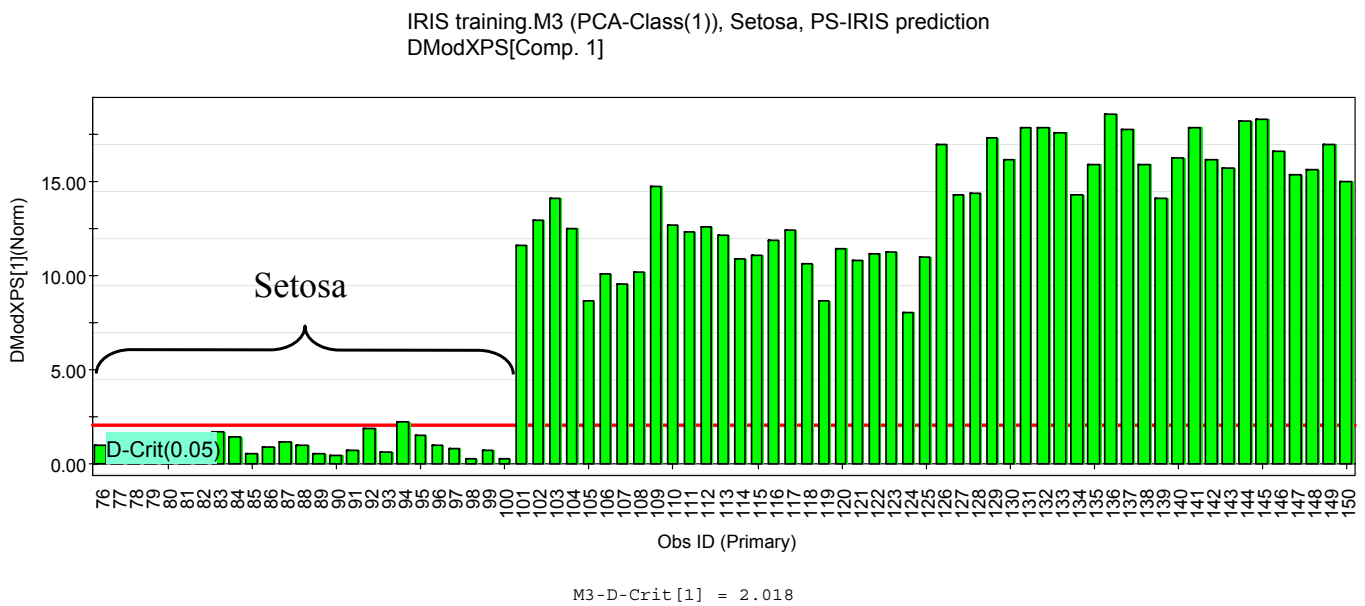
No.	Model	Type	A	R2X	R2Y	Q2(cum)	Date	Title	Hiera...
3	M3	PCA-Class(1)	1	0.589		0.214	2002-10-14	Setosa	
4	M4	PCA-Class(2)	1	0.671		0.387	2002-10-14	Versicolor	
5	M5	PCA-Class(3)	2	0.896		0.452	2002-10-14	Virginica	

- Assignment of class membership is based on a comparison of DModX and Dcrit



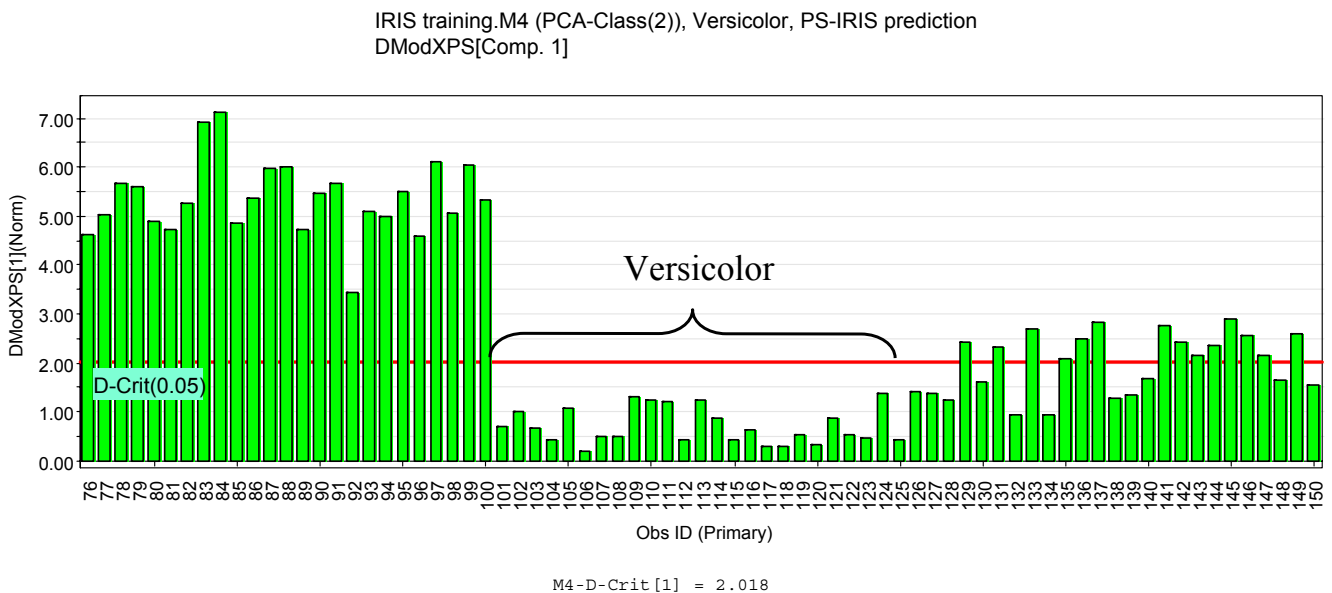
IRIS: Prediction (classification) by Setosa model

- 24 of 25 Setosa's are correctly classified; V_e/V_i fundamentally different



IRIS: Prediction (classification) by Versicolor model

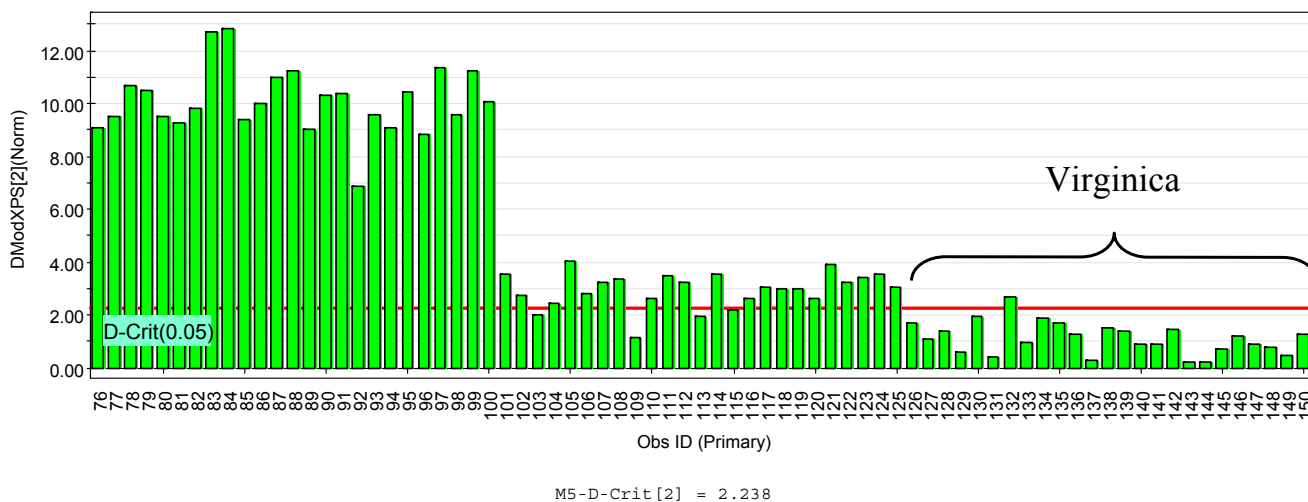
- V_e prediction set observations correctly classified; however, some V_i samples are false positives



IRIS: Prediction (classification) by Virginia model

- Vi prediction set observations correctly classified; however, most Ve samples are also inside Dcrit of Vi-model

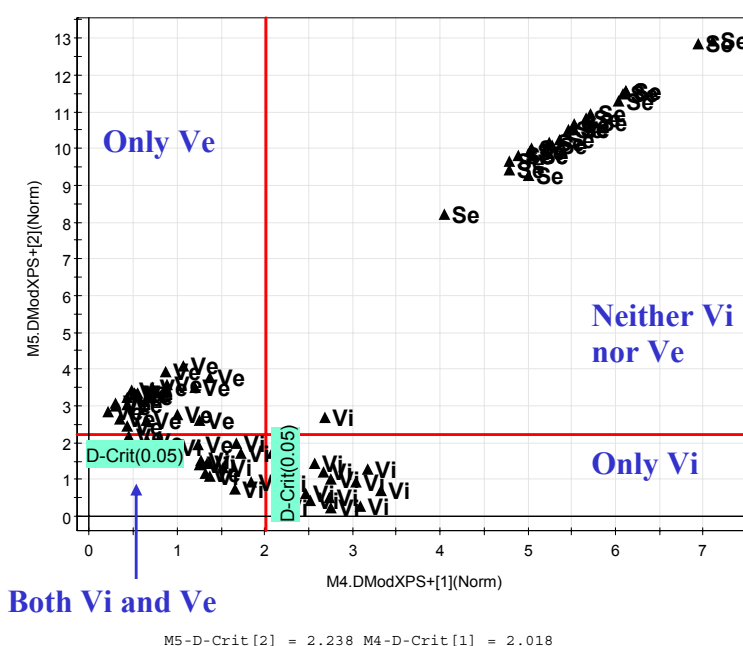
IRIS training.M5 (PCA-Class(3)), Virginia, PS-IRIS prediction
DModXPS[Comp. 2]



IRIS: Coomans plot derived from Vi/Ve models

- Named after the Belgian chemometrician Danny Coomans (~ 1980)
- DModX of two models are plotted in a scatter plot
- Four zones of diagnostic interest are created (see plot)

IRIS training
M4.DModXPS+[Comp. 1]/M5.DModXPS+[Comp. 2]



IRIS - Summary

- The conclusions that may be drawn are thus the following:
 - (i) *Setosa* specimens are quite different from *Versicolor* and *Virginica* observations.
 - (ii) There is an overlap between the *Versicolor* and *Virginica* classes, and they cannot be completely separated. However, it is possible to predict if an unknown sample is
 - (a) definitely *Versicolor*;
 - (b) definitely *Virginica*;
 - (c) definitely neither;or
 - (d) *Virginica* or *Versicolor*
- using the SIMCA methodology.

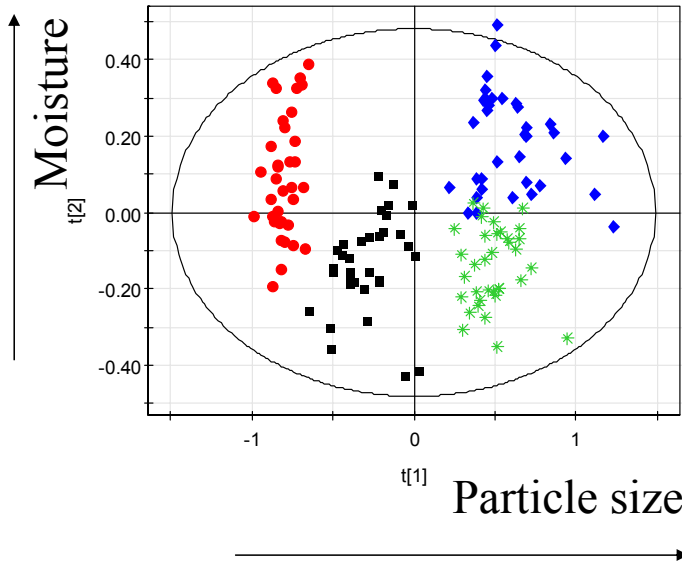
Industrial example: NIR_Chip

- NIR measurements from particleboard industry
- NIR data in the range 400-2498 nm (1050 spectral variables) were measured on four types of wood chips – differing in particle size and moisture content
- 140 (4*35) training set observations and 78 prediction set samples
- *Objective*: To study whether the wood chips could be distinguished from each other using NIR and SIMCA

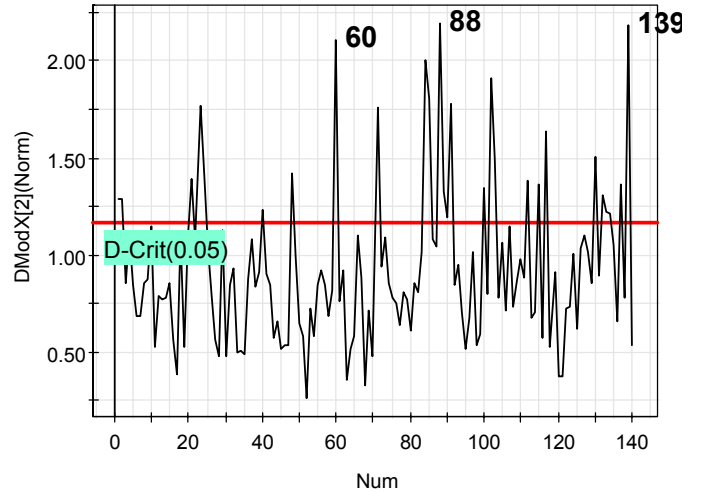
NIR_Chip: Step 1 – PCA for overview

- The two first components explain 94% of the spectral variation; Some separation among the four classes is seen

NIRChipT.M1 (PCA-X), PCA for overview
t[Comp. 1]/t[Comp. 2]



NIRChipT.M1 (PCA-X), PCA for overview
DModX[2] (Norm)



M1-D-Crit [2] = 1.163

NIR_Chip: Step 2 – SIMCA

- Four class-specific PCA models were computed
- Complexity always $A = 3$
- R^2X in the range 0.86 to 0.94

NIRChipT M2								
Type: PCA-Class(1) Observations (N)=35, Variables (K)=1050 (X=1050, Y=0)								
Components:								
A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Signific...	Iterations
0	Cent.							
1	0.504	0.504	17.6	0.389	0.0295	0.389	R1	46
2	0.392	0.896	13.7	0.774	0.0303	0.862	R1	6
3	0.0348	0.931	1.22	0.28	0.0312	0.901	R1	17

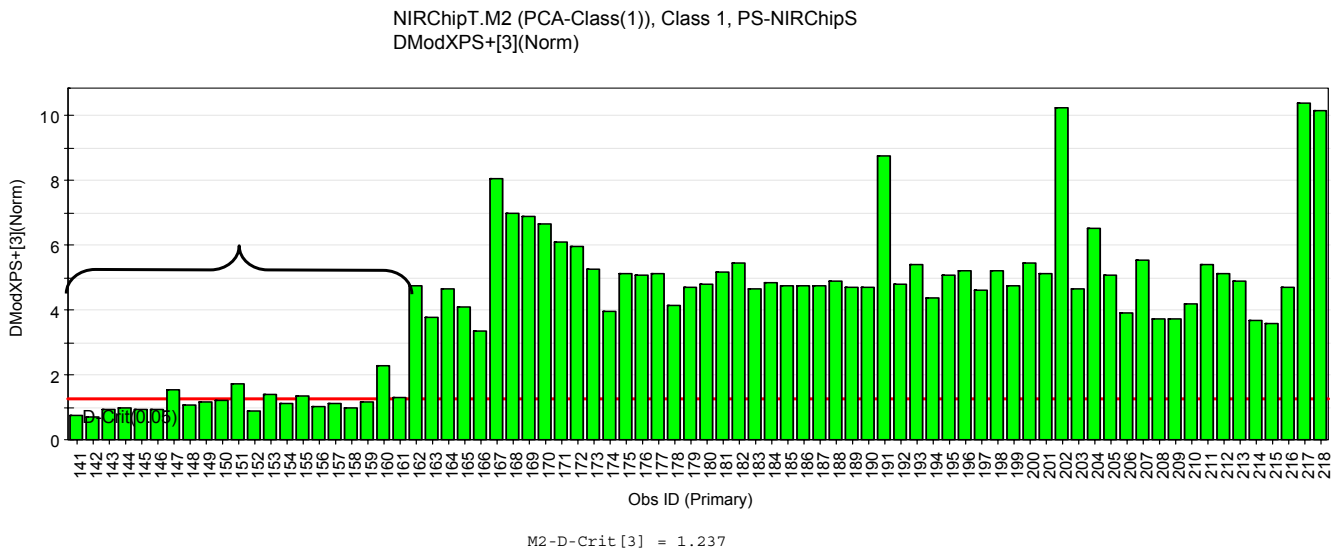
NIRChipT M3								
Type: PCA-Class(2) Observations (N)=35, Variables (K)=1050 (X=1050, Y=0)								
Components:								
A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Signific...	Iterations
0	Cent.							
1	0.686	0.686	24	0.661	0.0295	0.661	R1	9
2	0.156	0.841	5.45	0.272	0.0303	0.753	R1	23
3	0.0947	0.936	3.32	0.568	0.0312	0.893	R1	7

NIRChipT M4								
Type: PCA-Class(3) Observations (N)=35, Variables (K)=1050 (X=1050, Y=0)								
Components:								
A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Signific...	Iterations
0	Cent.							
1	0.608	0.608	21.3	0.566	0.0295	0.566	R1	12
2	0.189	0.797	6.61	0.429	0.0303	0.752	R1	23
3	0.117	0.915	4.11	0.543	0.0312	0.887	R1	8

NIRChipT M5								
Type: PCA-Class(4) Observations (N)=35, Variables (K)=1050 (X=1050, Y=0)								
Components:								
A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Signific...	Iterations
0	Cent.							
1	0.381	0.381	13.3	0.0805	0.0295	0.0805	R1	94
2	0.339	0.72	11.9	0.52	0.0303	0.559	R1	14
3	0.142	0.862	4.97	0.475	0.0312	0.769	R1	11

NIR_Chip: Classification of prediction set samples

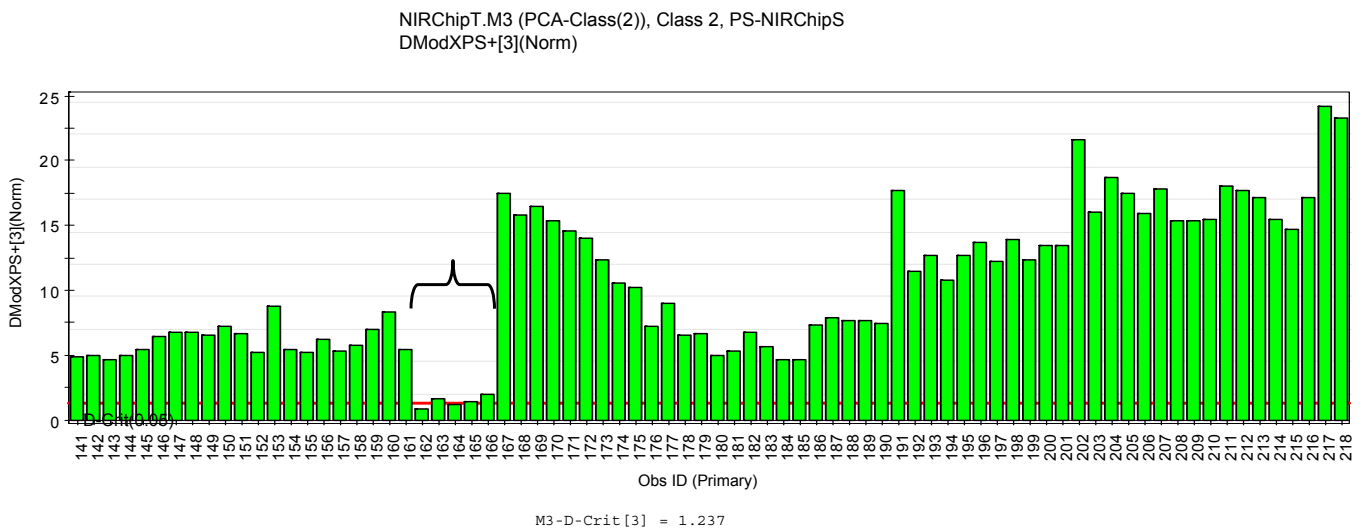
- Class 1 model



- Observations 141-161 are classified as close to Class 1

NIR_Chip: Classification of prediction set samples

- Class 2 model

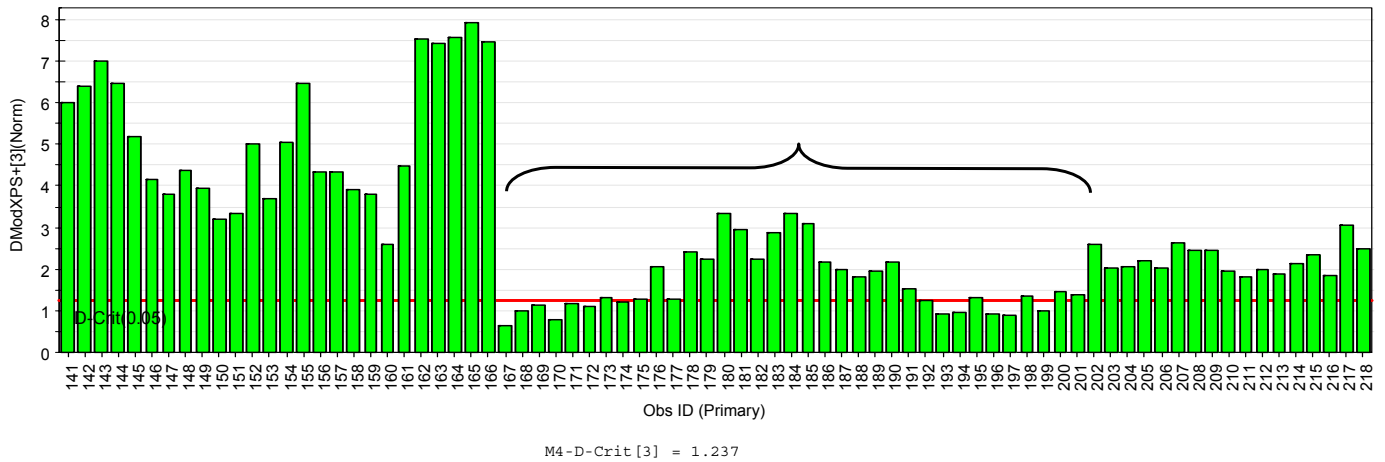


- Observations 162-166 are classified as close to Class 2

NIR_Chip: Classification of prediction set samples

- Class 3 model

NIRChipT.M4 (PCA-Class(3)), Class 3, PS-NIRChipS
DModXPS+[3](Norm)

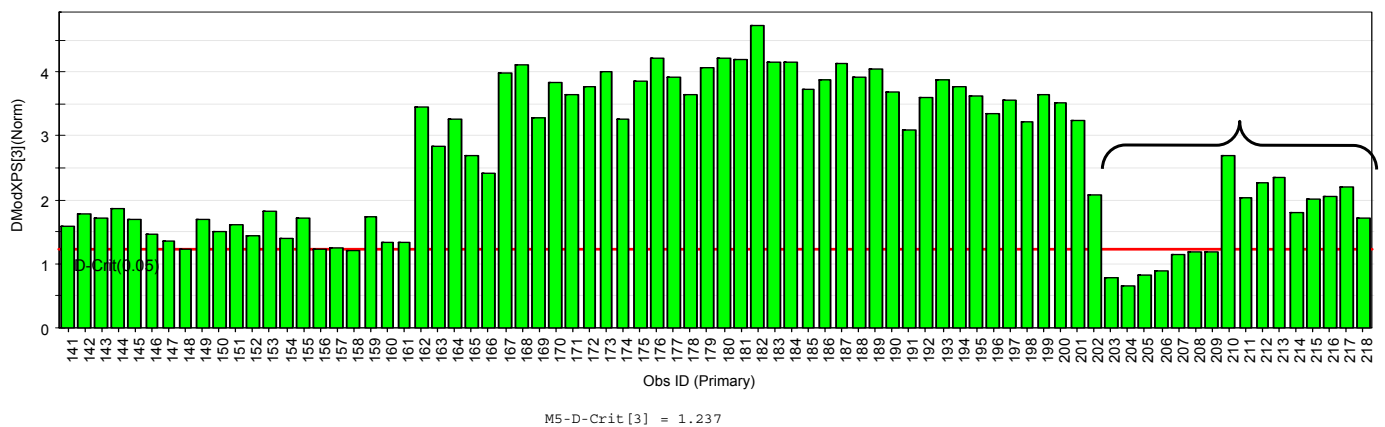


- Observations 167-175, 177, and 192-201 are classified as close to Class 3. Samples 176, 178-191, and 202-218 are also classified as rather similar to class 3.

NIR_Chip: Classification of prediction set samples

- Class 4 model

NIRChipT.M5 (PCA-Class(4)), Class 4, PS-NIRChipS
DModXPS[3](Norm)



- Observations 203-209 are classified as close to class 4. Samples 202 and 210-218 are classified as rather close to class 4

Conclusions of NIR_Chip investigation

- NIR characterisation coupled with multivariate data analysis is useful for on-line discrimination of four types of starting material in the particleboard industry.
- The correct class memberships are as follows:
 - Observations 141-161 \Leftrightarrow class 1
 - Observations 162-166 \Leftrightarrow class 2
 - Observations 167-201 \Leftrightarrow class 3
 - Observations 202-218 \Leftrightarrow class 4.
- Classification results ranged from good to excellent for the prediction data. The worst classification was for class 3 samples.
- This study hints at how multivariate characterisation for classification of raw materials can be carried out when the same starting material is delivered in different batches, or supplied by different manufacturers. This approach is common practice in the pharmaceutical and particleboard industries.

Multivariate Data Analysis and Modelling Basic Course

Chapter 4

Partial Least Squares Projections to Latent Structures (PLS) – Relating X to Y



Contents

Partial Least Squares Projection to Latent Structures (PLS)

- Notation
- Scaling of Variables
- Geometric Interpretation
- Algebraic Solution
- Example
- Diagnostics
 - Outliers
 - Residuals
 - Cross-validation
- Predictions
- Conclusions

Quantitative Modelling, PLS

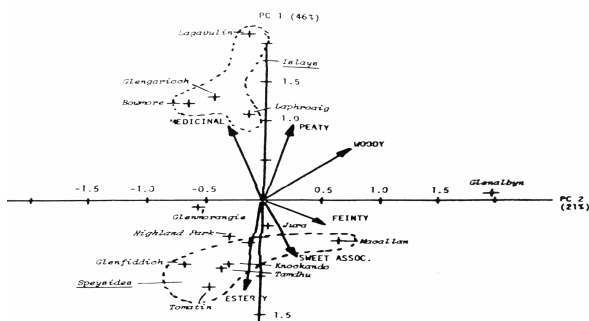
- Find relationships between sets of multivariate data X and Y
- Predict one set from other for new observations x_i

Applications:

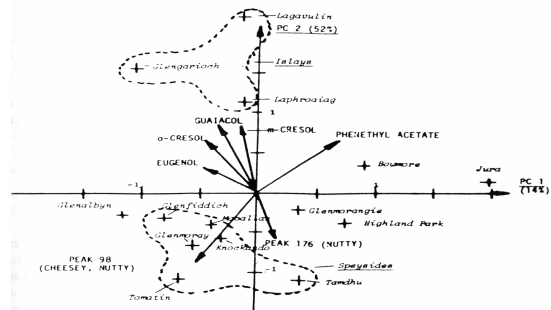
- Process modelling and optimisation
- Chemical composition \Leftrightarrow Quality
Physical measurements \Leftrightarrow Biol. Activity
- Chemical structure \Leftrightarrow Reactivity
Properties
Biol. Activity
- Multivariate calibration
Signals (spectra) \Leftrightarrow Concentrations
Energy content
Age; Taste.....

Example; Whisky

Correlation of Sensory & Analytical Data in Flavour Studies into Scotch Malt Whisky;
Swan & Howie, 1984



Principal Component Plot of Sensory data of Fourteen Scotch Malt Whiskies



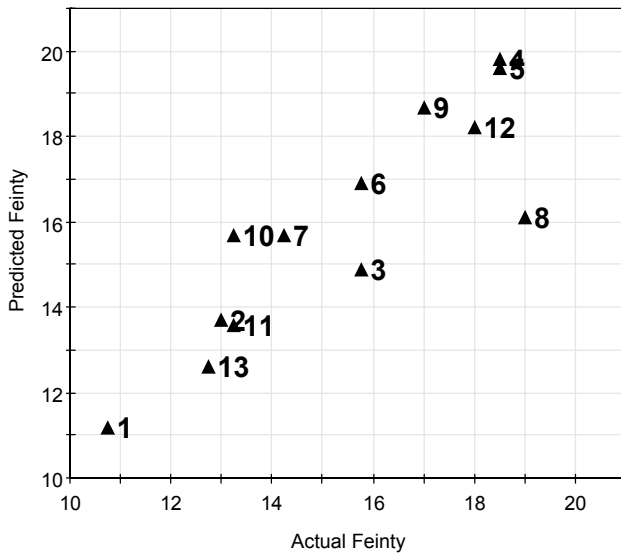
Principal Component Plot of Data from Thirteen Selected g.c. Peaks of Fourteen Scotch Malt Whiskies

- PCA of sensory data of 14 Scotch malt whiskies

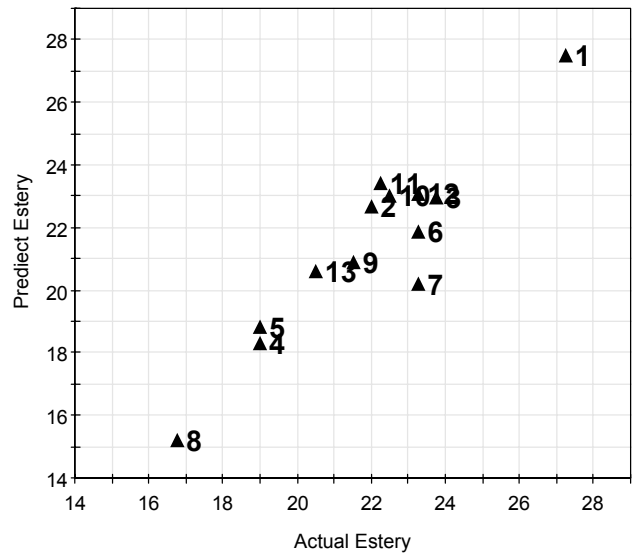
- PCA of GC data of 14 Scotch malt whiskies

Example; Whisky

Whisky - PLS predictions for test set

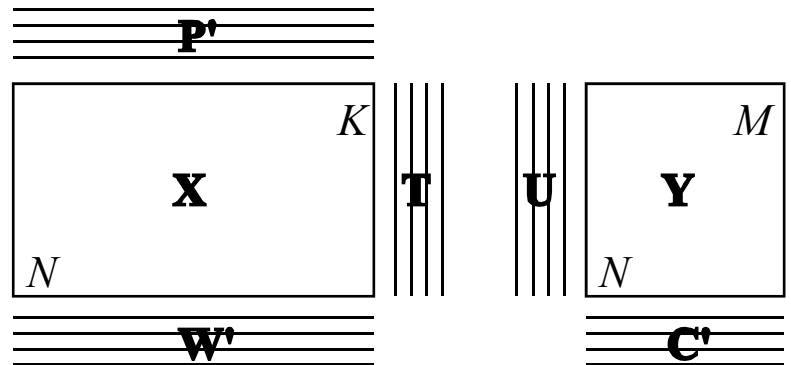


Whisky - PLS predictions for test set



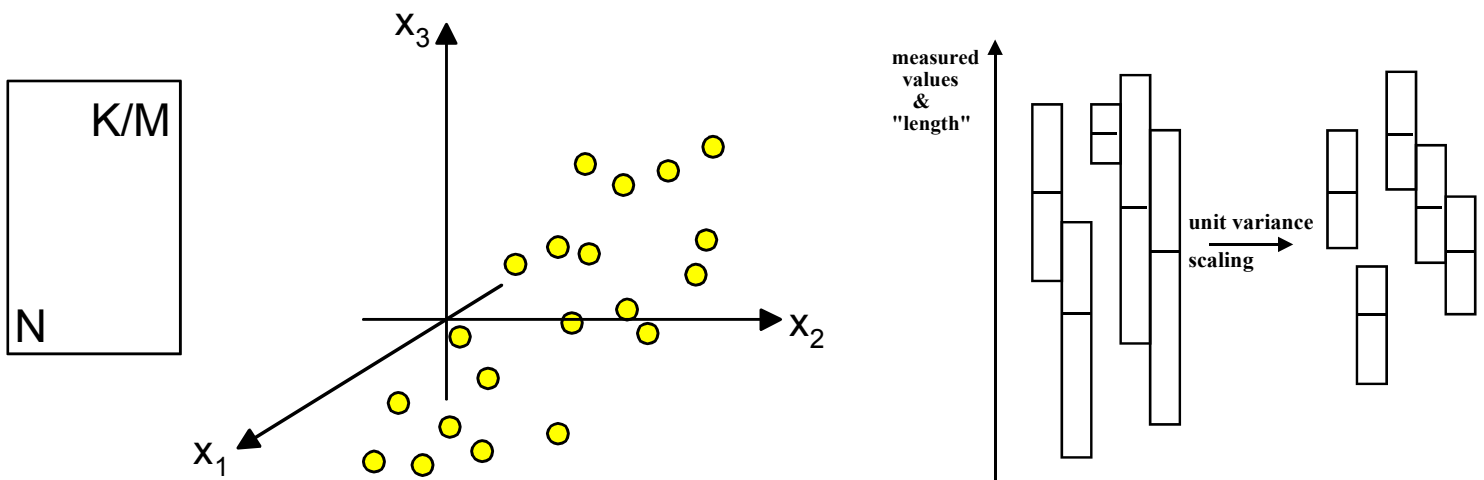
Notation

- K = number of X variables
- M = number of Y variables
- N = number of observations
- A = number of PLS components



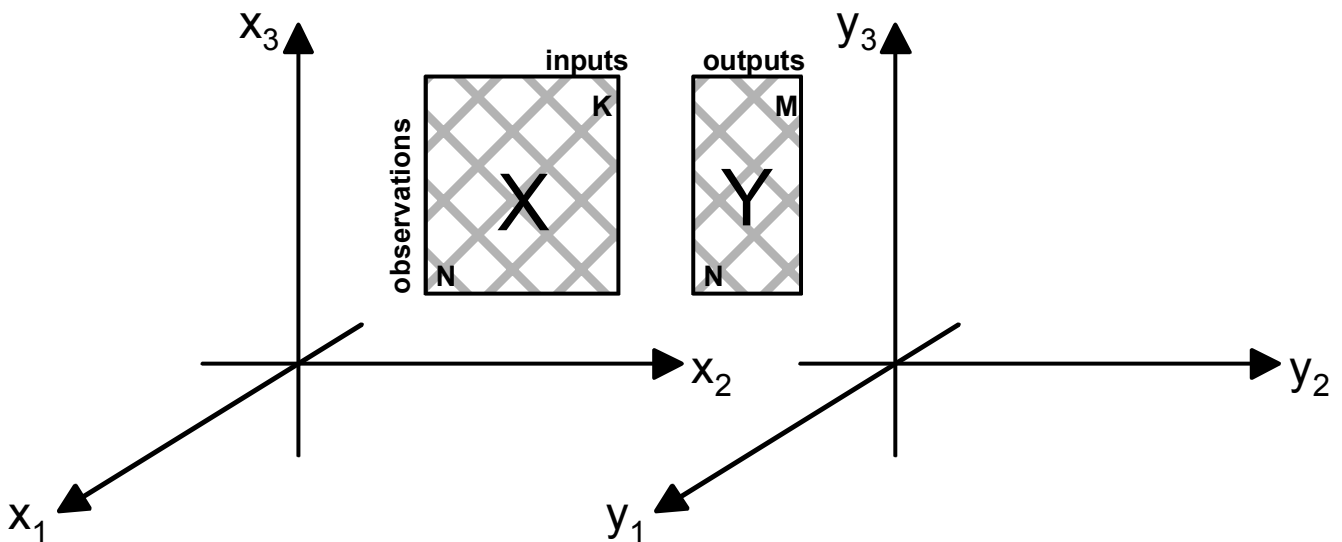
- T = matrix of X-scores with columns t_1, \dots, t_A (vectors)
- P = matrix of X-loadings with columns p_1, \dots, p_A (vectors)
- W = matrix of X-weights with columns w_1, \dots, w_A (vectors)
- U = matrix of Y-scores with columns u_1, \dots, u_A (vectors)
- C = matrix of Y-weights with columns c_1, \dots, c_A (vectors)

Scaling of variables



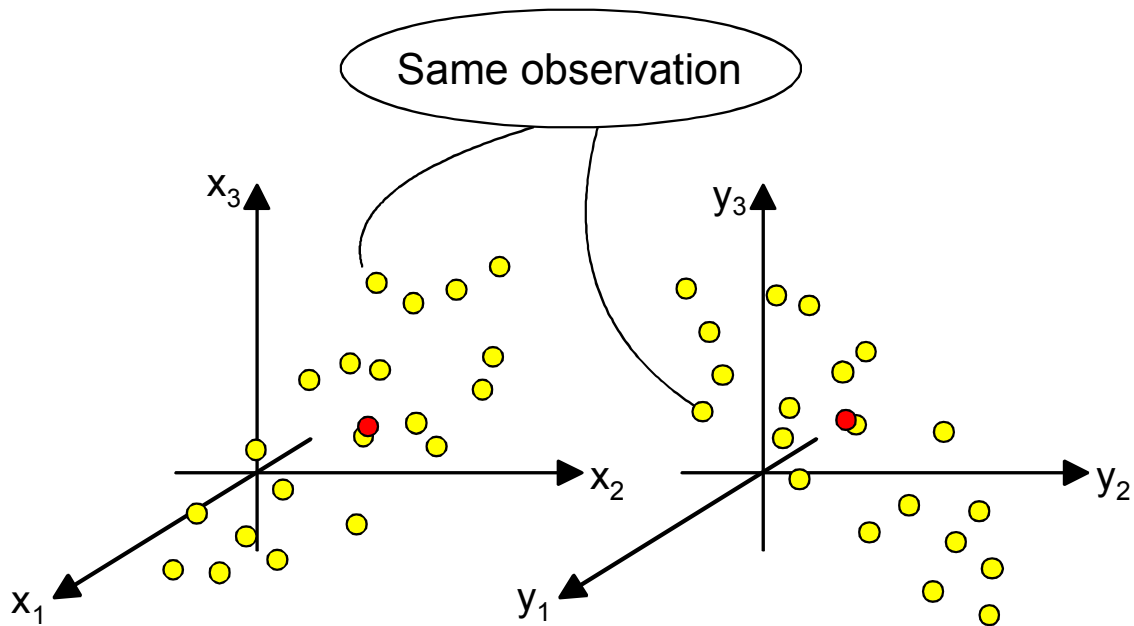
- Defining the length of variable axes (X- and Y-spaces)
- Usually, unit variance scaling is initially used to set each axis length (length one)

PLS -- Geometric Interpretation, 1



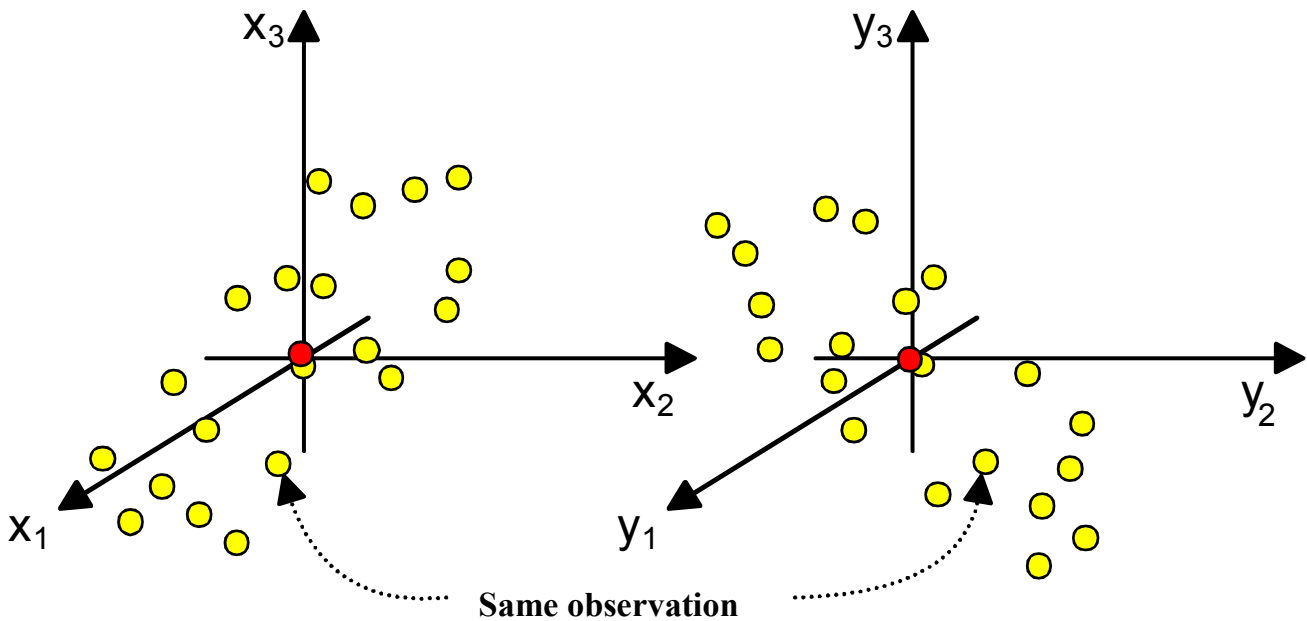
- For each matrix, X and Y, we construct a space with K and M dimensions, respectively
- Each X- and Y-variable has one coordinate axis with the length defined by its scaling, typically unit variance

PLS -- Geometric Interpretation, 2



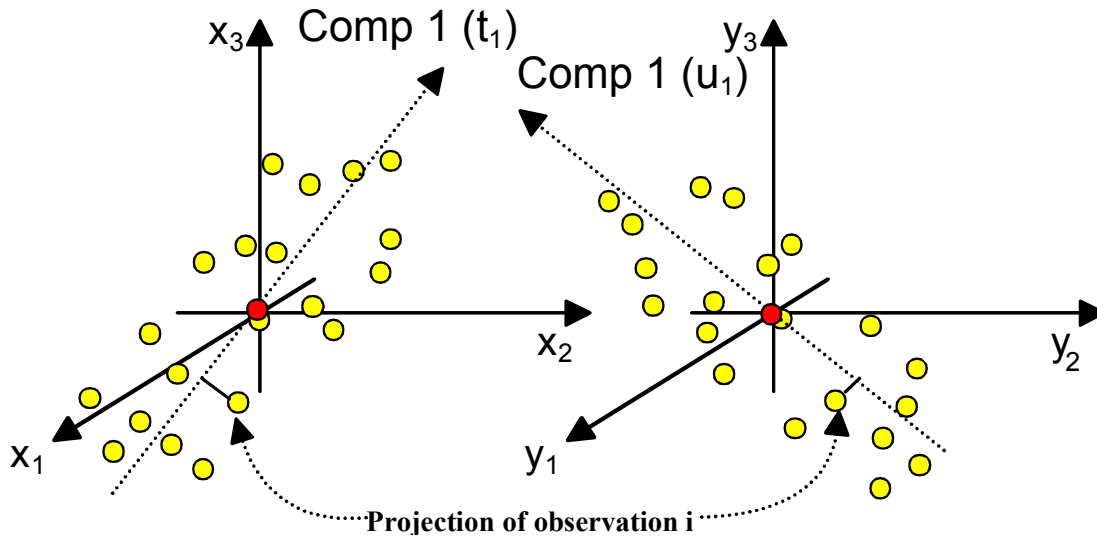
- Each observation is represented by one point in the X-space and one in the Y-space
- As in PCA, the initial step is to calculate and subtract the averages; this corresponds to moving the coordinate systems

PLS -- Geometric Interpretation, 3



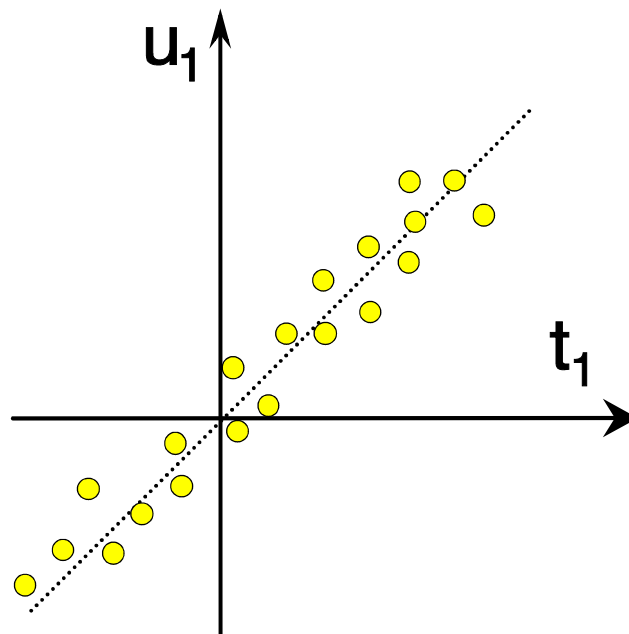
- The mean-centering procedure implies that the origin of each coordinate system is re-positioned

PLS -- Geometric Interpretation, 4



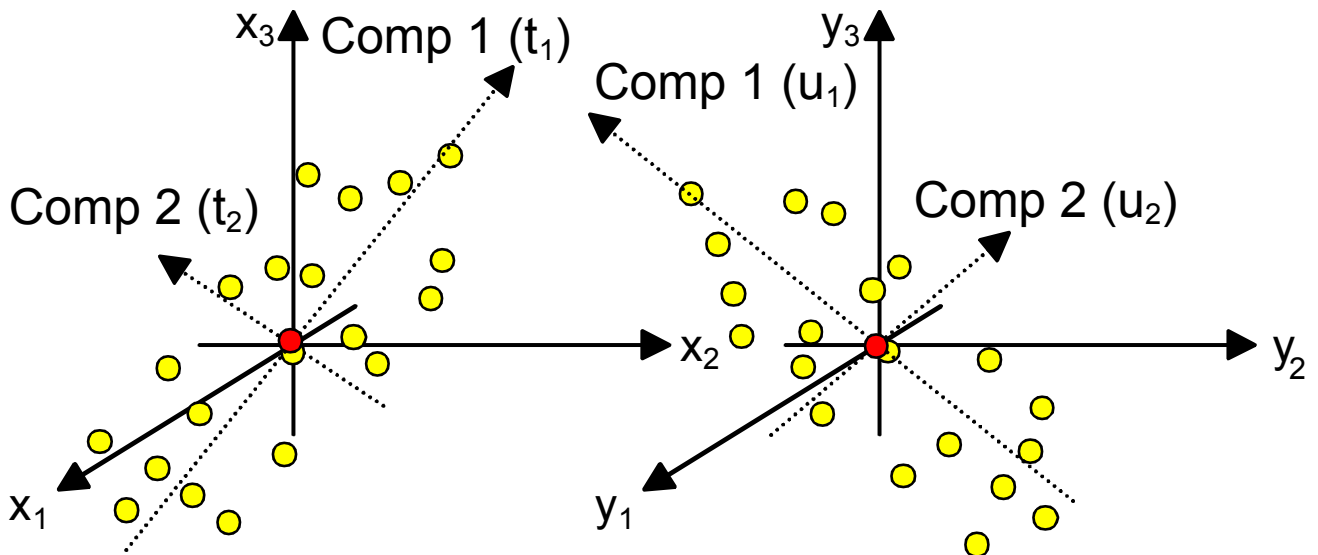
- The first PLS-component is a line in the X-space and a line in the Y-space, calculated to
 - a) approximate the point-swarms well in X and Y and also
 - b) provide a good correlation between the projections (t_1 and u_1)
- Directions are w_1 and c_1 and co-ordinates along these vectors are t_1 and u_1 , respectively.

PLS -- Geometric Interpretation, 5



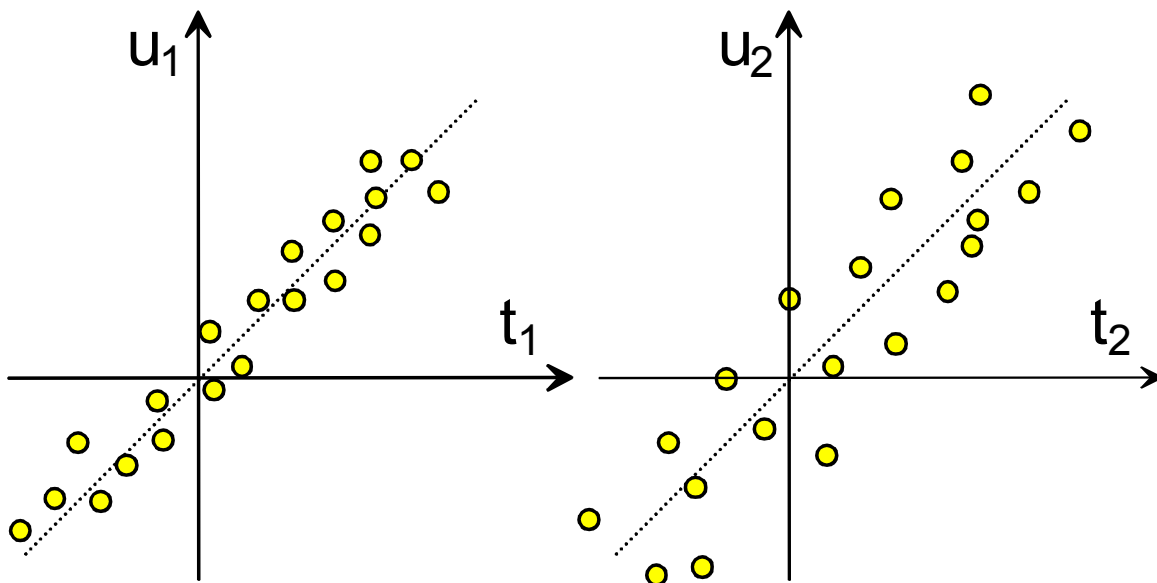
- The projection coordinates, t_1 and u_1 , in the two spaces, X and Y, are connected and correlated through the **inner relation** $u_{i1} = t_{i1} + h_i$ (h_i is a residual)
- The slope of the dotted line is 1.0

PLS -- Geometric Interpretation, 6



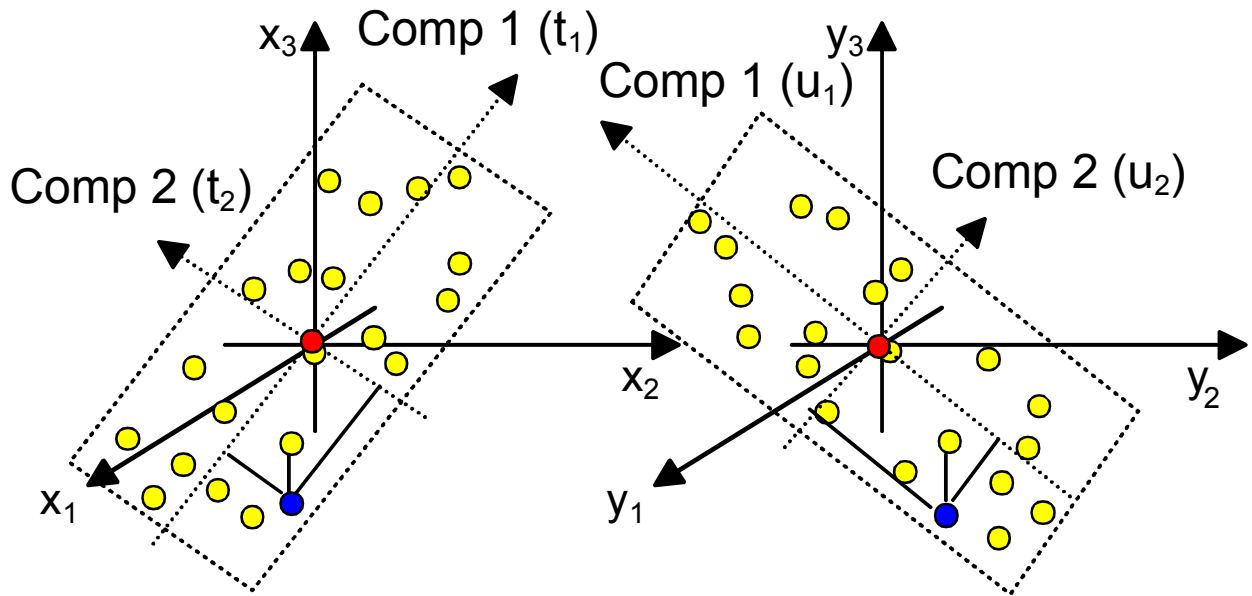
- The second PLS component is represented by lines in the X- and Y-spaces. X-lines are orthogonal. Y-lines may or may not be orthogonal.
- These lines, with directions w_2 and c_2 and projection co-ordinates t_2 and u_2 , improve the approximation and correlation as much as possible.

PLS -- Geometric Interpretation, 7



- The second projection coordinates (t_2 and u_2) correlate, but usually less well than the first pair of latent variables
- When the correlation is better between t_2 and u_2 than between t_1 and u_1 this indicates a strong structure in X that is not present in (related to) Y

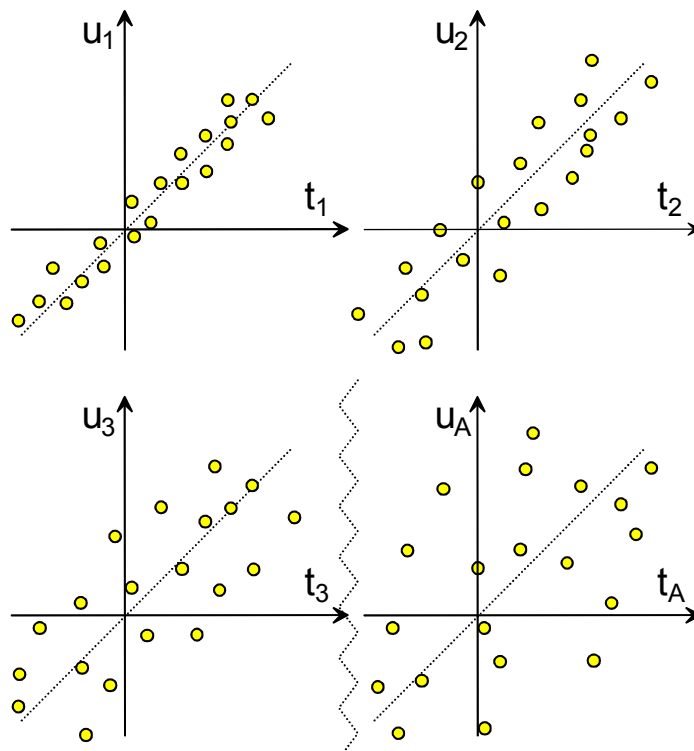
PLS -- Geometric Interpretation, 8



- The PLS components form planes in the X- and Y-spaces
- The variability around the X-plane is used to calculate a **tolerance interval** within which new observations similar to the training set will be located. This is of interest in classification and prediction.

PLS -- Geometric Interpretation, 9

- Repeated plotting of successive pairs of latent variables will give a good appreciation of the correlation structure



PLS, Overview

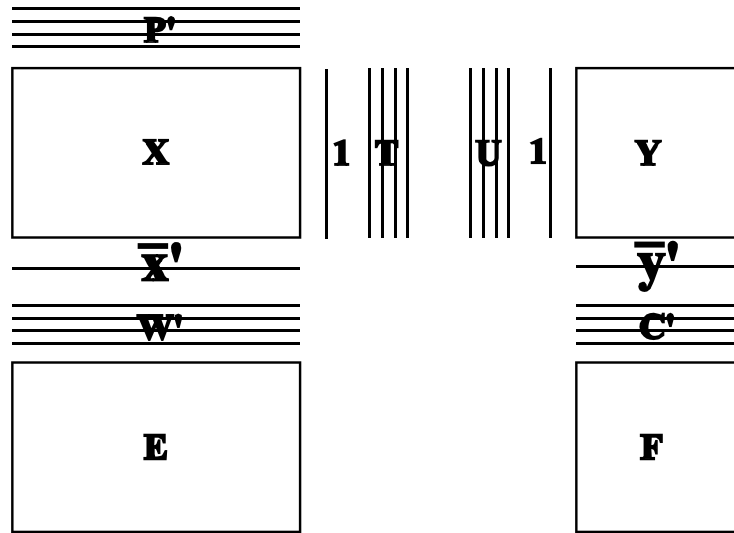
$$X = 1 * \bar{x}' + T * P' + E$$

$$Y = 1 * \bar{y}' + U * C' + F$$

$$= 1 * \bar{y}' + T * C' + G$$

(because $U = T + H$)

(inner relation)



PLS

differences to

PCA

Projection of X that
both
approximates X well,
and correlates well with Y

Projection of X that
is an **optimal**
approximation of X
(least squares fit)

PLS, Parameter properties

• For each component:

- 1) **t** are linear combinations of **X** with weights **w**
 - **t** is a **summary** of the **X**-variables that are **correlated with Y**
- 2) **u** are linear combinations of **Y** with weight **c**
 - **u** is a **summary** of the **Y**-variables

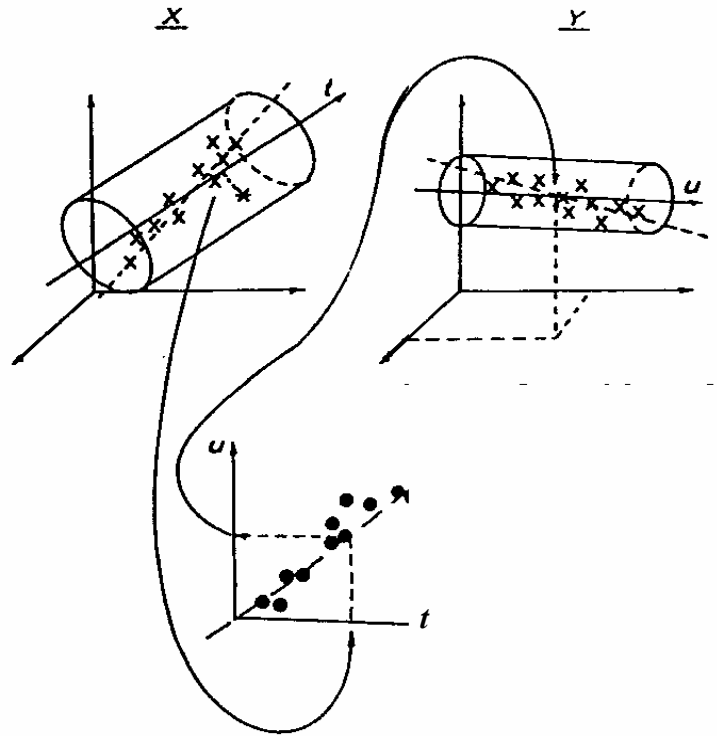
- 3) **w** are the covariances between the **x**'s and **u**
 - Columns of **X** highly correlated with **Y** have high weights

4) At Convergence for the Orthogonality:

- **p** is computed so that **t*p'** is the "best approximation of **X**"
- **t*p'** is removed from **X** for the next component

PLS predictions

- A new observation is similar to the training set if it is inside the tolerance cylinder in X-space
- Then its projection on the X-model (t) can be entered into the T-U-relation giving a u -value for each model dimension
- These values define a point on the Y-space model, which, in turn, corresponds to a predicted value for each y -variable



Example - Understanding relationships (LOWARP)

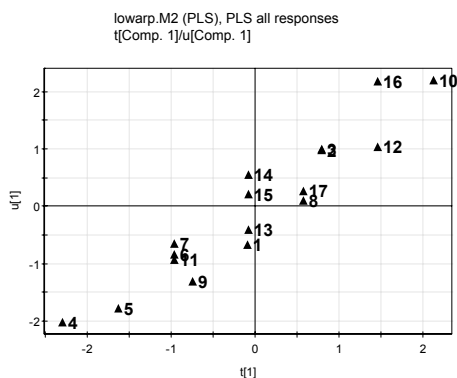
- Experimental production of new polymer
 - R&D environment
 - Many responses to consider
- The development of a polymer with a certain profile of properties was desired: low warp and shrinkage and high strength. To obtain this a number of polymer formulations were made with four constituents
 - Glass 20 to 40 %
 - Crtp 0 to 20 %
 - Mica 0 to 20 %
 - Amtp 40 to 60 %

Example - Understanding relationships (LOWARP)

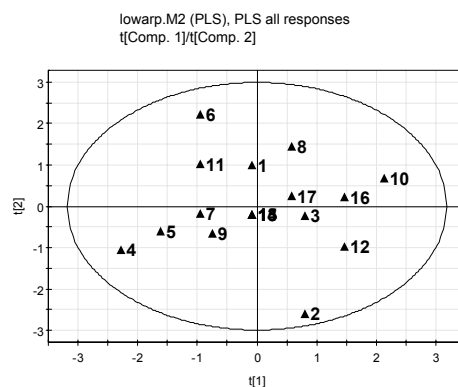
- The Data - 18 variables, 17 observations
 - 4 manipulated variables (X)
 - 14 quality variables (Y)
- Statistical experimental design was used
 - the design was a mixture, extreme vertices design
 - 14 runs + 3 centre points
- The use of design of experiments (DOE) enables us to use the PLS-model as a causal model
 - establish cause and effect relationships

PLS score plots: relationships among observations

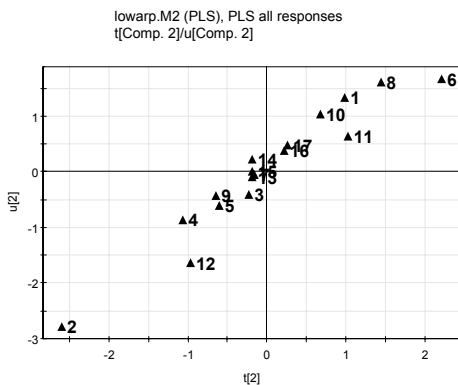
t_1/u_1 shows relationships among observations between X (t_1) and Y (u_1)



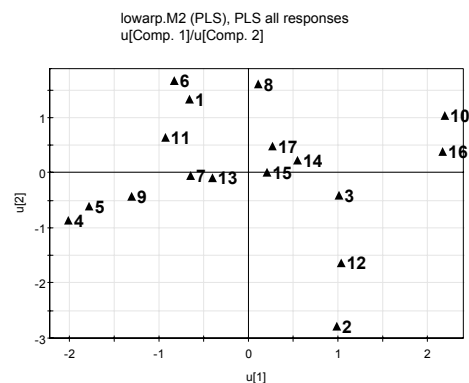
t_1/t_2 shows relationships among observations in the X-space ($r = 0.0$)



t_2/u_2 shows relationships between X and Y in the second dimension



u_1/u_2 shows relationships among observations in the Y-space ($r = 0.0$)



PLS - Interpretation of variable influence

a) Loadings ($w \cdot c$)

- Provides an overview of the relationships among all X-variables and Y-variables at the same time. Often plotted as scatter plots. Line plot representation used with spectral data (see Chapter 6).

b) Regression coefficients

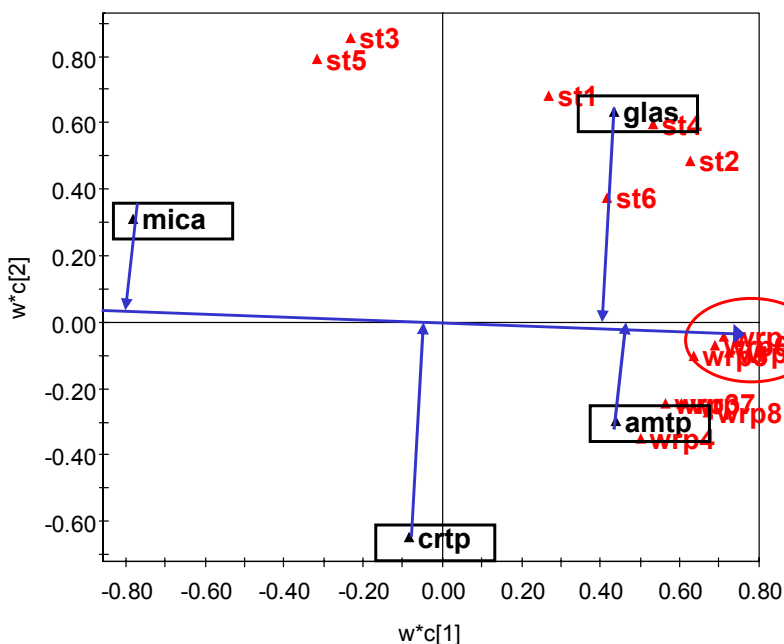
- The size and sign of the scaled and centred regression coefficient indicates relation of term to y.

c) VIP, variable influence on projection.

- Cumulative measure of the influence of term k on the model. Terms with VIP larger than around 0.8 are the most meaningful.

Interpretation of PLS-model – loadings

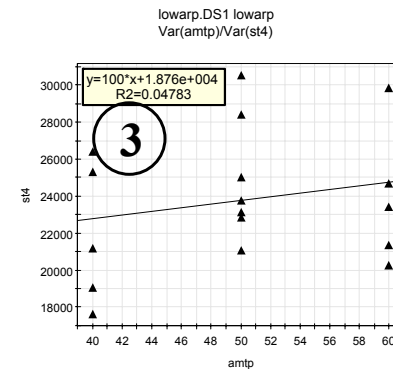
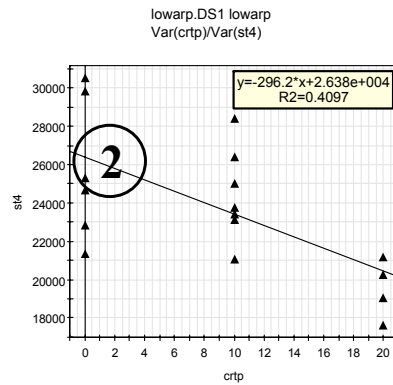
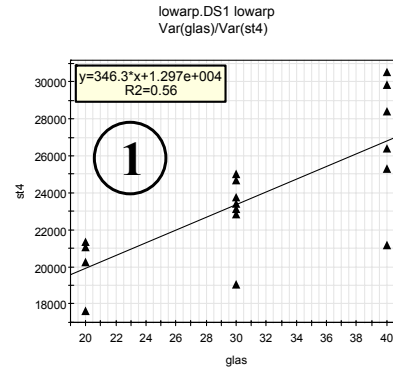
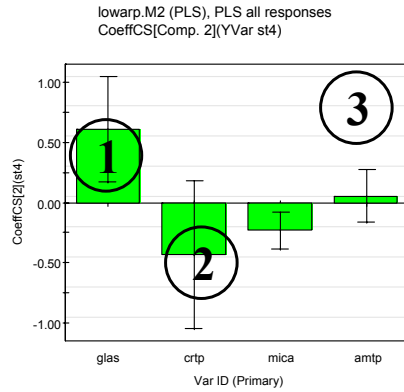
lowarp.M2 (PLS), pls no expansion, Work set
Loadings: $w \cdot c[1]/w \cdot c[2]$



- Find the point (0,0), marked with lines.
- Find an important y-variable (e.g. wrp2)
- Imagine a line from this y through (0,0)
- Project all x-variables down on this line
- The x-variables far out from (0,0) are important
- x.s on same side of (0,0) to y have positive influence
- x.s on other side of (0,0) to y have negative influence
- ANALOGY: A See-saw: close to the fulcrum (the pivotal point) no influence, but far out large impact

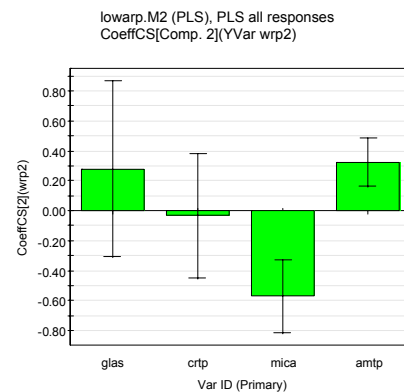
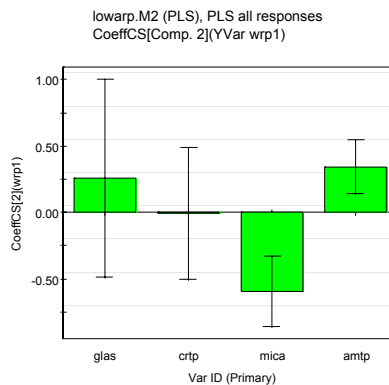
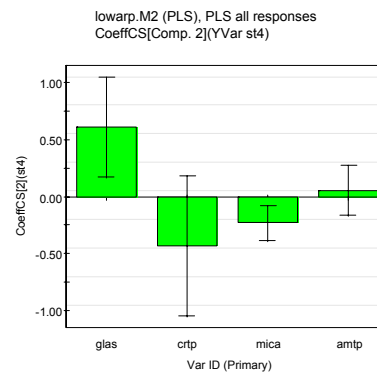
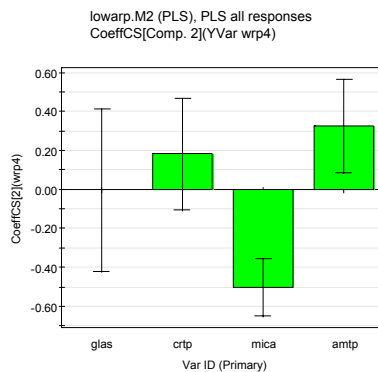
Interpretation of PLS-model – regression coefficients

- Positive (1), negative (2), and near zero (3) coefficient



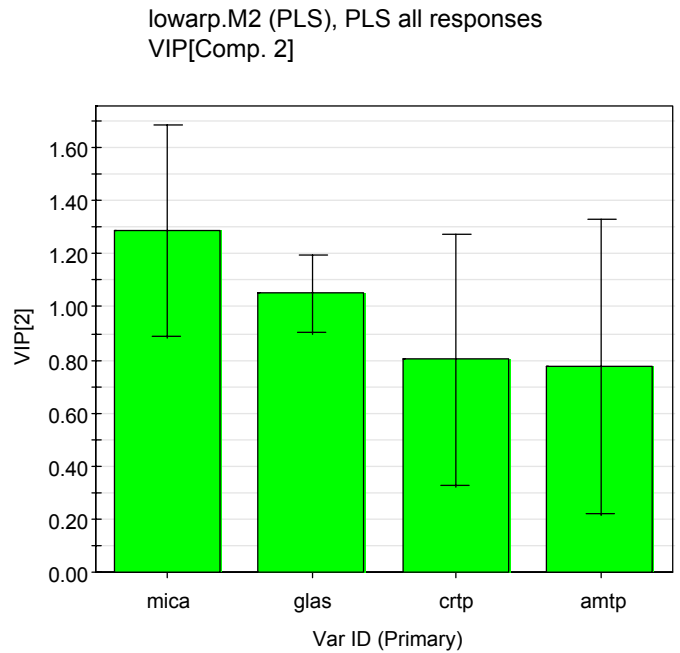
Interpretation of PLS-model – regression coefficients

- Regression coefficients - uncorrelated responses
- Regression coefficients - correlated responses



Interpretation of PLS-model – VIP

- VIP is a condensed summary of a PLS model showing the influence of each X-variable on the model
- Only one VIP expression regardless of number of responses and Y-variables



Variable related parameters – summary

a) Loadings

$$w * c$$

b) regression coefficients

$$Y = X * B_{PLS} + F$$

$$B = W * (P' * W)^{-1} * C'$$

The size and sign of the scaled and centred regression coefficient indicates relation of term to y.

Note: In PLS the regression coefficients are generally not mathematically independent

c) **VIP**, variable influence on projection. Cumulative measure of the influence of term k on the model. Terms with VIP larger than around 0.8 are the most meaningful.

$$[VIP(k)]^2 = \frac{[\sum_a [(W_{ak}^2) * SSY\%(a)]]}{SSY\%(cum)} * K$$

SSY%(a) = % SS of Y explained by the a.th PLS component

SSY%(cum) = Total SS of Y explained by the model

K = Number of terms in the model

$$\sum VIP(k)^2 = K$$

PLS - Diagnostics

- Observation diagnostics - strong and moderate outliers
- Variable diagnostics - which variables are well explained?
- Model diagnostics - cross-validation & response permutation test

PLS - Observation Diagnostics

- **Strong outliers, groups, inhomogeneity,...**

PLS plots:

- 1) X space (t_1, t_2, \dots)
- 2) Y space (u_1, u_2, \dots)
- 3) X, Y space (t_1, u_1, \dots)

- **Moderate Outliers, trends, in X and Y**

Plot DModX vs Num of observation (not for designed data):

Observation X \rightarrow RSD: Distance to Model (DModX)

Check that no observation has large DModX

Plot DModY vs Num of observation

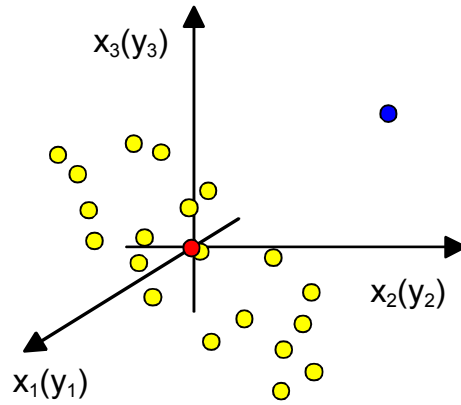
Check that no observation has large DModY

- **Observation Risk**

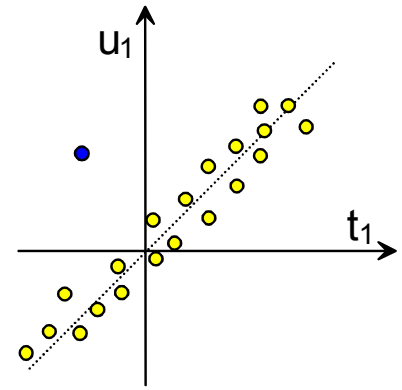
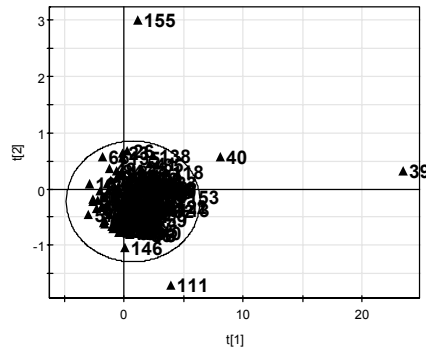
Sensitivity measure

PLS - Strong outliers

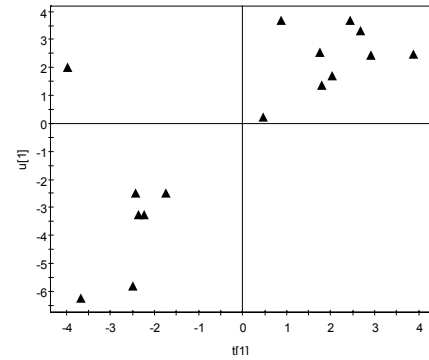
- An observation may be an outlier in X, in Y, and in the X/Y relation (a few examples are shown)



thicknes.M1 (PCA-X), PCA for overview
t[Comp. 1]/t[Comp. 2]



Scores: t[1]/u[1]



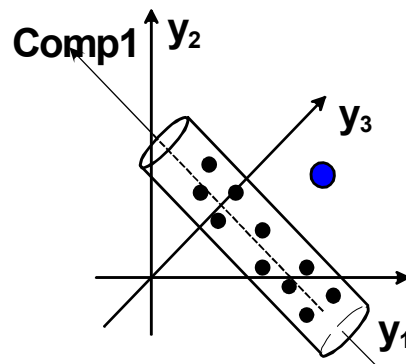
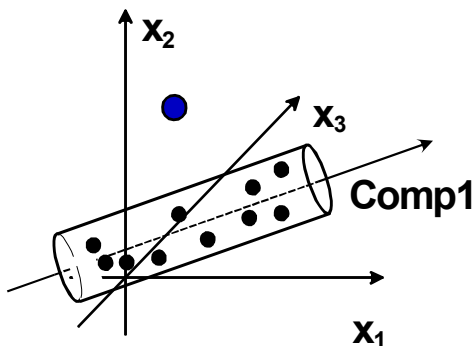
PLS - Moderate outliers

- Moderate outliers can be found when examining PLS residuals, E and F:

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

$$Y = \mathbf{1} * \bar{y}' + U * C' + F$$

$$= \mathbf{1} * \bar{y}' + T * C' + G$$



- The E and F residual matrices are related to the diameters of the "beer cans" surrounding the data points

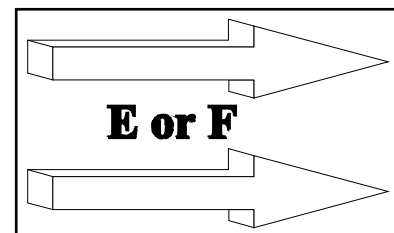
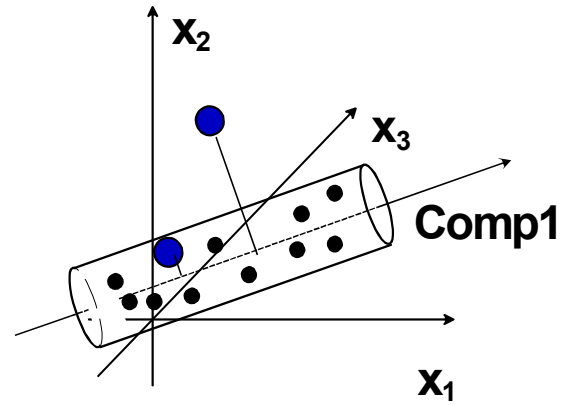
PLS - Moderate outliers

- Moderate outliers can be detected by inspecting the residual SD for each observation (DModX)
- Residual observation variance (S2OX)

$$\sum_k e_{ik}^2 / DF$$
- DModX, normalised distance

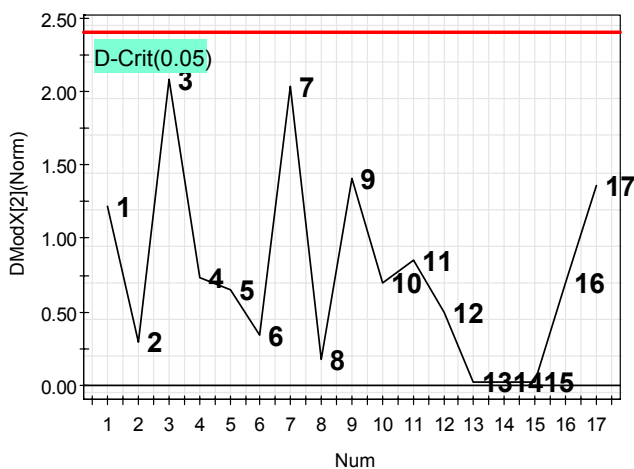
$$[S2OX / \text{variance}(E)]^{1/2}$$
- DModX, absolute distance

$$[S2OX]^{1/2}$$
- Formulas above are analogous for Y-space, only e_{ik} should be replaced with f_{im}



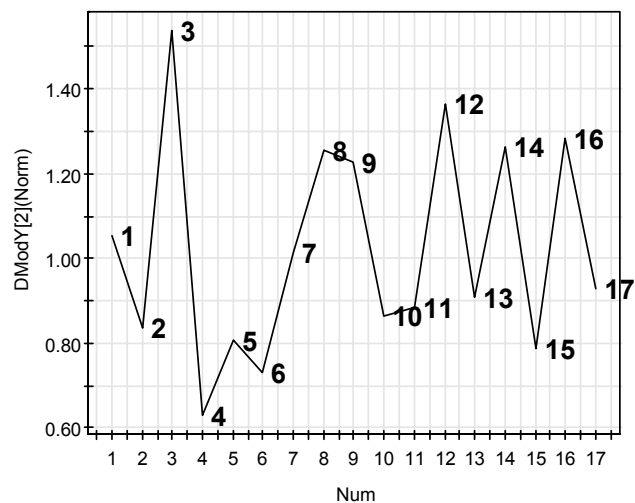
PLS - Moderate outliers

lowarp.M2 (PLS), PLS all responses
DModX[Comp. 2]



M2-D-Crit [2] = 2.403

lowarp.M2 (PLS), PLS all responses
DModY[Comp. 2]

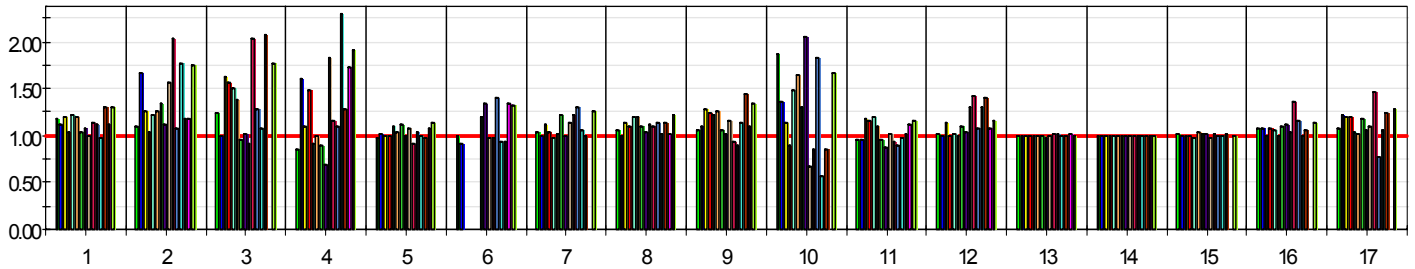


- There are no moderate outliers in the LOWARP example

PLS – Observation risk

- A measure of how sensitive a model is to the inclusion of an observation
- LOWARP – PLS model most sensitive to observations 2, 4, and 10

lowarp.M2 (PLS), PLS all responses
Observation Risk for all the Y's and Pooled Y's

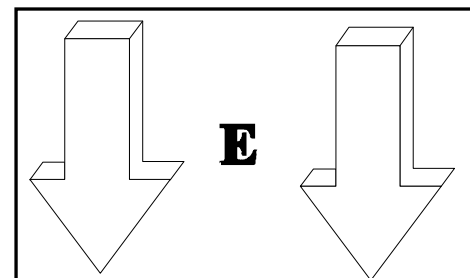


PLS - Variable diagnostics

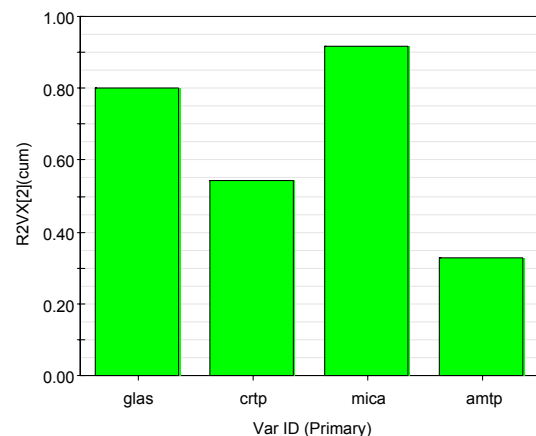
- **Information from size of residuals of X-variables:**

Shows how much of the variation of a variable that the model explains

- SSVX, residual variable variation
 $\sum_i e_{ik}^2$
- S2VX, residual variable variance
SSVX/DF
- R2VX (cum), explained variation
 $1 - \text{SSVX}[A]/\text{SSVX}[0]$
- R2VX_{adj}(cum), explained variance
 $1 - \text{S2VX}[A]/\text{S2VX}[0]$



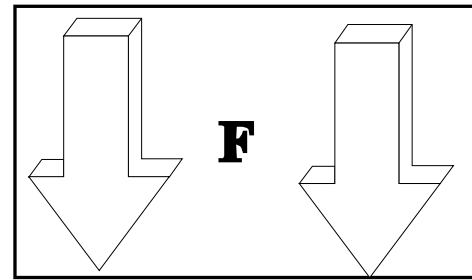
lowarp.M2 (PLS), PLS all responses
R2VXcum[Comp. 2]



PLS - Variable diagnostics

- **Information from size of residuals of Y-variables:**

Shows how well a Y-variable is modelled



- SSVY, residual variable variation

$$\sum_i f_{im}^2$$

- S2VY, residual variable variance

$$SSVY/DF$$

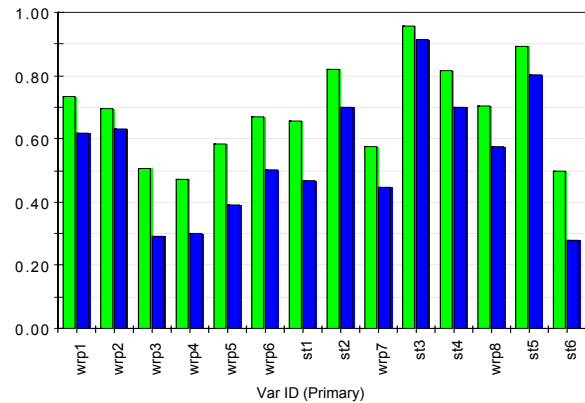
- R2VY (cum), explained variation

$$1 - SSVY[A]/SSVY[0]$$

- R2VY_{adj}(cum), explained variance

$$1 - S2VY[A]/S2VY[0]$$

lowarp.M2 (PLS), PLS all responses: ■ R2VY[2](cum) ■ Q2VY[2](cum)



PLS - Model diagnostics

- SIMCA supports two internal model validation strategies

1. Cross validation

To estimate the optimal model complexity

2. Response permutation test (Validate-option)

To check the degree of overfit (Discussed in Chapter 6)

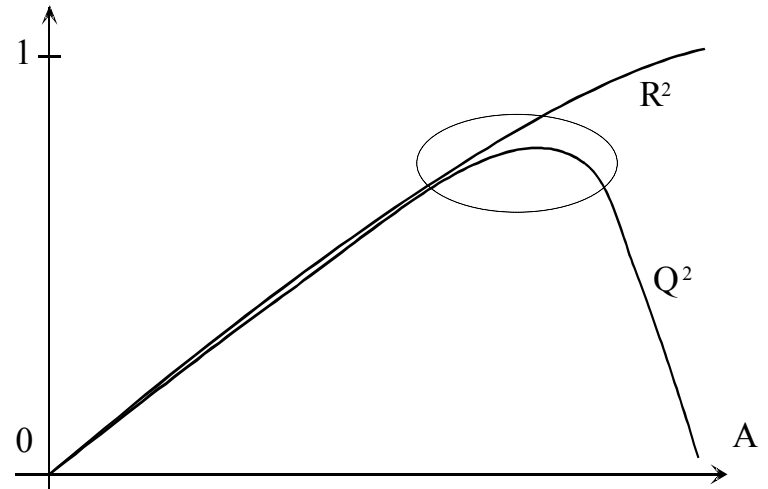
Model validity vs model complexity

- Trade-off between Fit and Prediction ability

- A model must not be overfitted, i.e. modelling noise

- **Question:** How can we determine the appropriate number of PLS components to include in a model?

- **Method:** Cross-validation (CV); CV simulates the predictive power of a PLS-model.



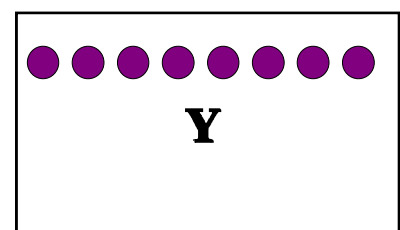
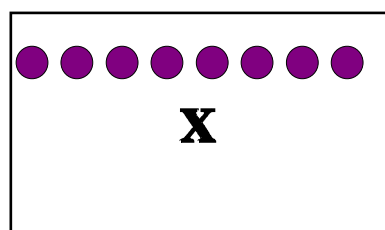
R^2 estimate of goodness of fit

Q^2 estimate of goodness of prediction

Cross-validation, PLS

- Data are divided into G groups (SIMCA-P default = 7)
- A model is estimated for the data devoid of one group
- The deleted group is predicted by the model \Rightarrow partial PRESS
- This is repeated G times; then all partial PRESS's are summed to PRESS
- If a new PLS component enhances the predictive power compared with the preceding PLS component, i.e. $PRESS < SS$, the new PLS component is kept in the model
- **NOTE:** In PCA cross-validation estimates Q^2X , in PLS Q^2Y .

Data are deleted row-wise



Evaluation of R² and Q²

• **PRESS** is the sum of squared differences between predicted and observed y-elements.

$$P R E S S = \sum (y_{im} - \hat{y}_{im})^2$$

• PRESS can be transferred into a dimensionless quantity, Q², which resembles R²

$$Q^2 = 1 - PRESS/SSY_{total}$$

$$R^2 = 1 - SSY_{resid}/SSY_{total}$$

Q² > .5 Good (Depending on application)

Q² > .9 **Excellent** (Depending on application)

Important:

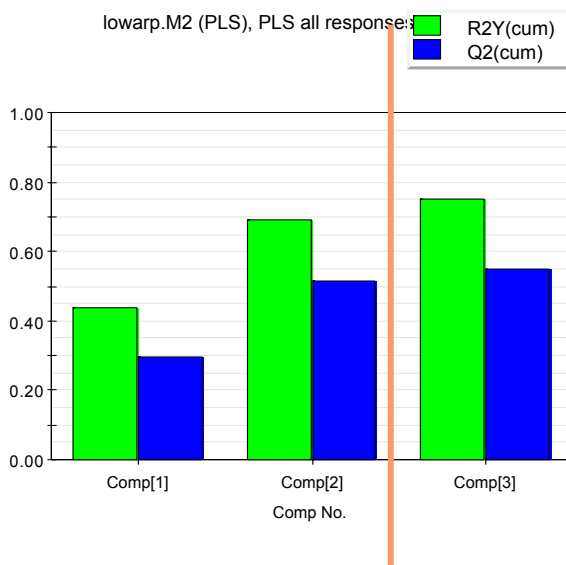
1. R² is always larger than Q²

2. High R² and high Q² is good

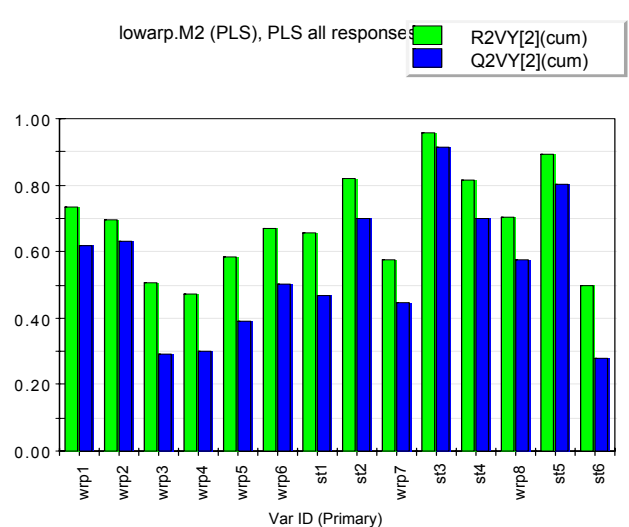
3. The difference between R² and Q² should not be too large

Model Complexity: Lowarp Example

For each PLS component



For each Y-variable



PLS, Summary

• **Modelling:** Data tables X and Y are approximated by (hyper)-planes + residuals (E, F), and an inner relation between U and T

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

$$Y = \mathbf{1} * \bar{y}' + U * C' + F$$

Calculations: One PLS-dim at a time -- NIPALS

$$U = T + H$$

The number of model dim.s (A)

Cross-validation (predictive significance)

(H is a residual matrix)

Residuals: Std. dev. of X- or Y-residuals of one observation = distance of obs. to (hyper)plane in X- or Y-space

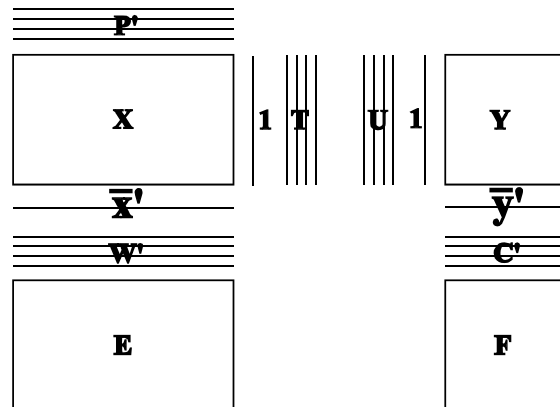
Predictions: New obs. projected onto (hyper)-plane in X-space \Rightarrow X-scores (t) \Rightarrow inner relation \Rightarrow Y-scores (u) \Rightarrow C \Rightarrow predicted Y (and confidence interval)

PLS, Summary

PLS

models the relation between two data tables X \rightarrow Y and is used for:

- dominating factors in X and Y
 - selection of relevant variables (X and/or Y)
- outliers
 - in X-space
 - in Y-space
 - in inner relation (t \rightarrow u)
- groups, clusters in X and Y



- similarities / dissimilarities
 - observations **scores, t, and u**
 - variables **loadings, p, and weights, w, and c**
- predictions
 - x \rightarrow y (what y is obtained by x)
 - y \rightarrow x (how to set x to get y)



Multivariate Data Analysis and Modelling Basic Course

Chapter 5 Multivariate Characterisation



Contents

- Introduction to multivariate characterisation
 - Each observation is characterised by many variables
 - The multivariate data are summarised by PCA (or PLS)
- Example: Solvents
- Example: Surfactants

Introduction to multivariate characterisation

- Multivariate characterisation is a method of *quantifying qualitative -- discrete -- changes*
- \Rightarrow A qualitative change can be described by means of quantitative latent variables (principal properties)
- Multivariate characterisation is useful when
 - Variability is introduced by uncontrolled qualitative factors, e.g., batches of raw material
 - Qualitative factors have many levels, e.g., choices among solvents
catalysts
additives
substituents
compounds for biological testing
stationary phases in HPLC, TLC, GC

Procedure

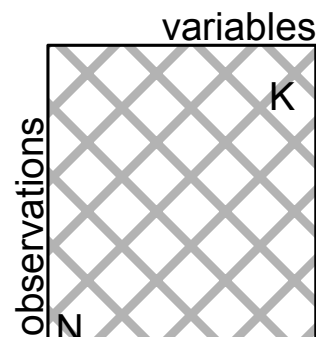
- For each varying constituent/ingredient/batch in the system:

1) Measure & calculate a battery of properties using relevant model systems (chemical, physical, biological, ...)

This will give you a multivariate data table \rightarrow

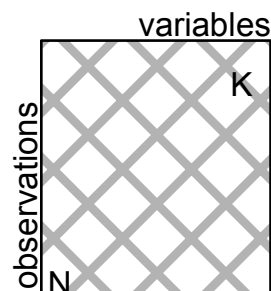
2) Use PCA to derive the “principal properties” for each constituent/ingredient/batch

3) Use the resulting PPs as factors to model the changes among the observations, and for DOE



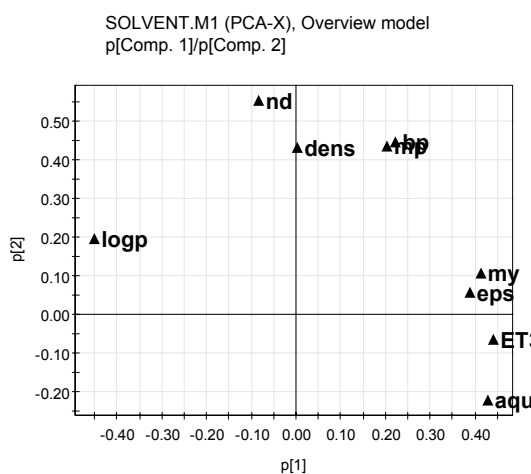
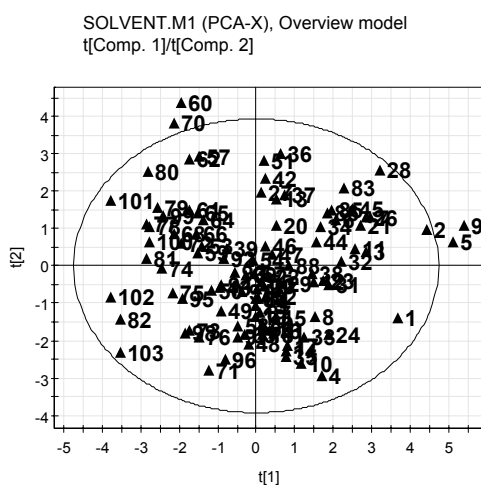
Example: Solvents

- How do we select an appropriate solvent for chemical synthesis ?
- Assumption: solvent “effects” in the “real” system can, at least partly, be seen also in well chosen properties
- Example: For 103 solvents, 9 property values were compiled:
 - 1) melting point 2) boiling point
 - 3) dielectr. const. 4) dipole moment
 - 5) refract. index 6) ET30 (λ_{\max} of UV of organic dyes)
 - 7) density 8) log P
 - 9) water solubility



Solvents - Some results

- PCA gave 2 significant components modelling 70% of the variance
- These two factors, “principal properties”, can be used as quantitative variables for the selection of *representative* solvents

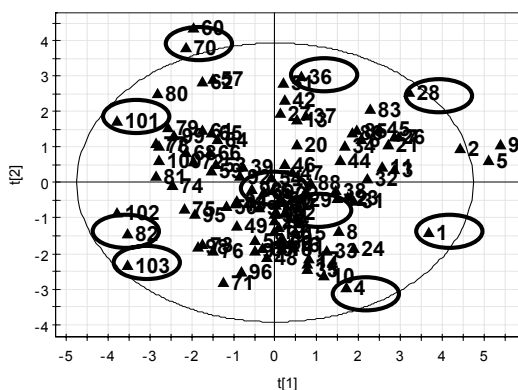


- 1st component reflects hydrophilicity/water solubility; 2nd component reflects polarizability

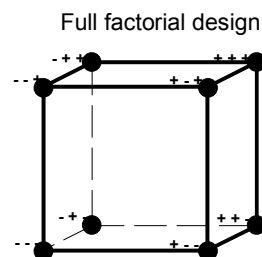
How to select representative solvents

- 1) Create an experimental design, e.g., a fractional factorial design expanded with some center-points
- 2) Select solvents with PPs matching the design table as closely as possible (or use D-optimal design)

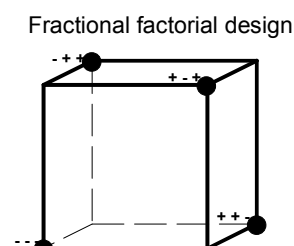
SOLVENT.M1 (PCA-X), Overview model
t[Comp. 1]/t[Comp. 2]



DV1	DV2	DV3
+	-	-
-	+	-
+	+	-
-	-	+
+	-	+
-	+	+
+	+	+



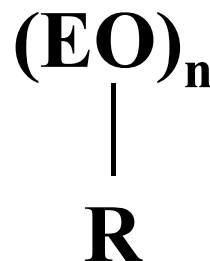
DV1	DV2	DV3
+	-	+
-	+	+
+	+	-



Example: Surfactants

- Non-ionic surfactants are increasingly used in commercially available detergent mixtures
- Lindgren/Uppgård studied non-ionic ethylene-oxide (EO) based surfactants and described these using 19 chemical variables
- 38 technical blends (distribution of EOs)

- R is hydrophobic part
 - straight chain
 - branching
 - aromatic
 - unsaturation
 - etc



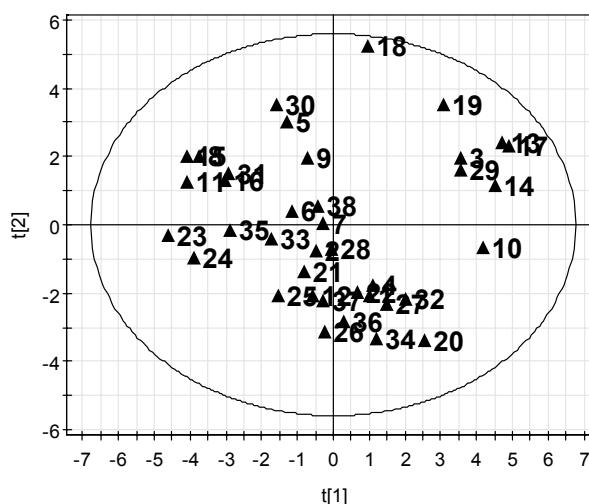
Objective of Surfactants study

- The objective of the study was to find a surfactant with good washing performance but without undesirable toxicity or biodegradability
- **Question 1:** Is it possible to quantitatively model the performance of technical blends as a function of chemical properties?
- **Question 2:** Which surfactants should be used as the basis for modelling?

Step 1: Multivariate characterisation

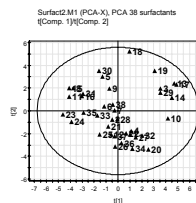
- K = 19 variables (Mw, C, redC, redC/C, EOw, HLBG, HLBD, CPP, redCPP, CP, dCP, chain, rmchain, f-alc, maxEO, w33EO, w66EO, CMC, logP)
- These reflect hydrophobicity, molecular weight, branching of R, technical blend properties, critical micellar conc., etc.
- PCA of 38 x 19 data matrix →
($R^2 = 0.78$, $Q^2 = 0.51$, A=3)

Surfact2.M1 (PCA-X), PCA 38 surfactants
t[Comp. 1]/t[Comp. 2]

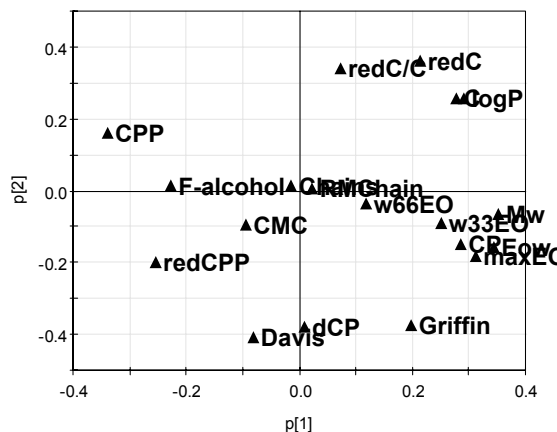


Interpretation of model

- PP1 describes lipophilicity and shape of EO-distribution chromatogram
- PP2 accounts for hydrophilic/lipophilic balance
- Surfactants in the upper right corner are too lipophilic to be interesting (these were excluded from further consideration)
- PC-score plot is a map of surfactants



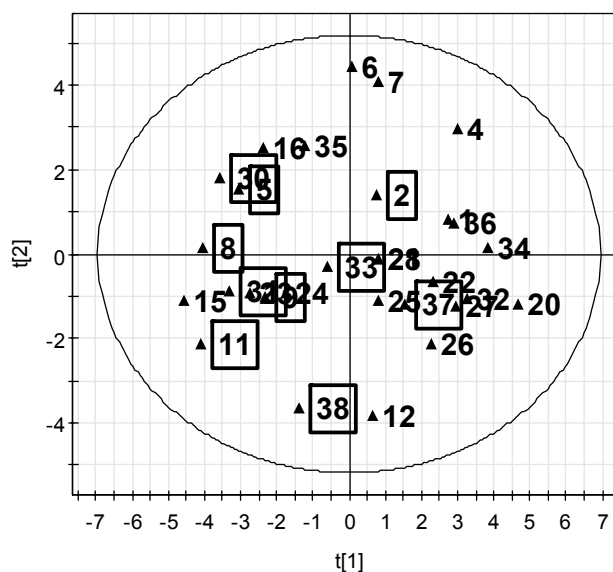
Surfact2.M1 (PCA-X), PCA 38 surfactants
p[Comp. 1]/p[Comp. 2]



Step 2: Selection of surfactants for further testing

- The 8 lipophilic surfactants were excluded, and an updated PC-model was computed
 - $R^2X = 0.76$
 - $Q2 = 0.52$
 - $A = 3$

Surfact2.M2 (PCA-X), PCA 30 surfactants
t[1]/t[2]



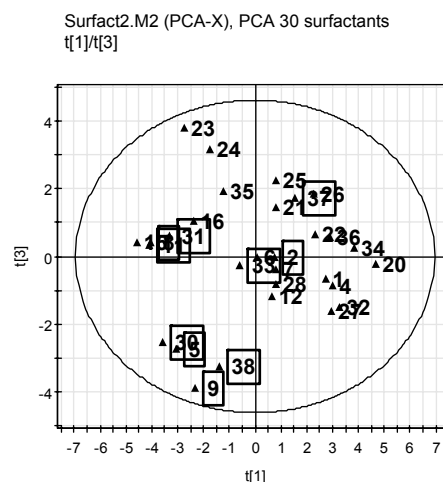
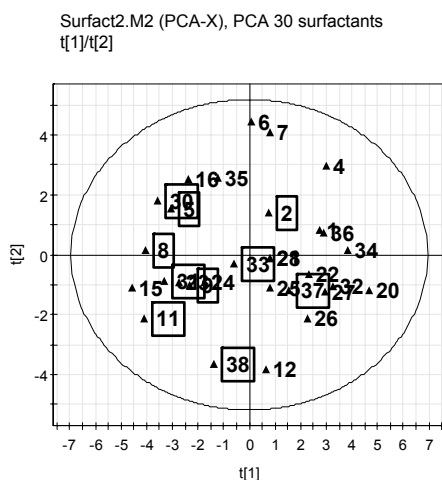
Step 2: Selection of surfactants for further testing

- The three PCs (or PPs) describe, quantitatively, the change in property profile when going from one surfactant to another
- Since PPs are mathematically independent of one another, they are useful as design variables in statistical experimental designs

	A	B	C	D	E
1	Obs ID (Primary)	Obs ID (OBSNAM)	M2.t[1]	M2.t[2]	M2.t[3]
2	1	1	2.75	0.81	-0.67
3	2	2	0.76	1.40	-0.04
4	4	4	2.99	2.96	-0.85
5	5	5	-3.07	1.56	-2.73
6	6	6	0.05	4.46	0.00
7	7	7	0.79	4.10	-0.39
8	8	8	-4.07	0.16	0.40
9	9	9	-2.31	-1.03	-3.90
10	11	11	-4.08	-2.10	0.33
11	12	12	0.65	-3.84	-1.19
12	15	15	-4.55	-1.11	0.42
13	16	16	-2.37	2.53	1.06
14	20	20	4.65	-1.20	-0.20
15	21	21	0.80	-0.12	1.46
16	22	22	2.32	-0.67	0.68
17	23	23	-2.75	-0.91	3.83
18	24	24	-1.75	-0.92	3.19
19	25	25	0.82	-1.10	2.24
20	26	26	2.28	-2.10	1.85
21	27	27	2.96	-1.23	-1.61
22	28	28	0.80	-0.13	-0.81
23	30	30	-3.60	1.81	-2.54
24	31	31	-3.30	-0.86	0.64
25	32	32	3.25	-1.04	-1.50
26	33	33	-0.58	-0.28	-0.27
27	34	34	3.85	0.16	0.26
28	35	35	-1.21	2.57	1.93
29	36	36	2.89	0.72	0.57
30	37	37	1.52	-1.20	1.74
31	38	38	-1.40	-3.66	-3.26

Step 2: Selection of surfactants for further testing

- Ten representative surfactants providing a reasonable coverage of an interesting area in the PP-space were selected
- Selected surfactants: 2,5,8,9,11,30,31,33,37,38



Step 3: Detergency measurements

- Detergency efficiency was measured as a function of washing conditions using design
- For each selected surfactant a CCC design in three factors was set up
- ^{14}C trioleine was used for soiling of cotton/polyester cloths
- Surfactant concentration, washing time and washing temperature were varied
- Three responses, optimal YDet, YTemp, and YConc, were determined in the 15 min washing experiment

Step 4: Toxicity measurements

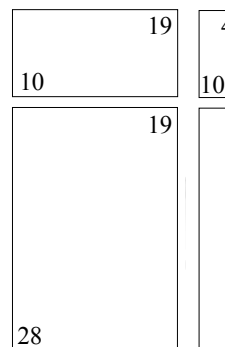
- Toxicity to a fairy shrimp (*Thamnocephalus platyurus*); log LC_{50}
- Summary of performance profiles: →

Surfact.	YDet	YConc	YTemp	YTox
2	86.3	1.4	60	1.29
5	81.8	1.6	48	0.52
8	82.4	1.9	44	0.96
9	81.8	1.5	82	0.55
11	85.4	2.25	54.5	1.25
30	78.8	1.7	48	0.59
31	84.6	2.8	50	1.77
33	88.9	1	74	1.11
37	85.9	2	67	1.52
38	86.5	1.1	61	

- Goals: →
 - YDet ↑ (important)
 - YConc ↓
 - YTemp ↓ (40-60 °C, important)
 - YTox ↑ (low toxicity, important)
- * Nos 2 or 37 look promising

Step 5: PLS modelling

- Develop model from selected training set
- Make predictions for test set and identify most promising surfactant(s)
- Strong model obtained



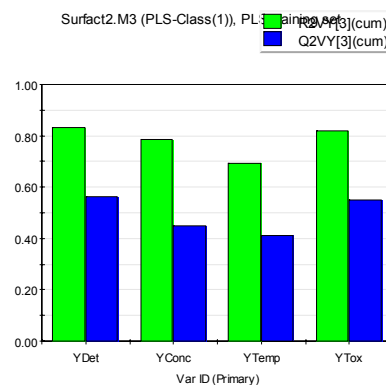
Surfact2 - M3

Workset... Options... Title: PLS training set

Type: PLS-Class(1) Observations (N)=10, Variables (K)=23 (X=19, Y=4)

Components:

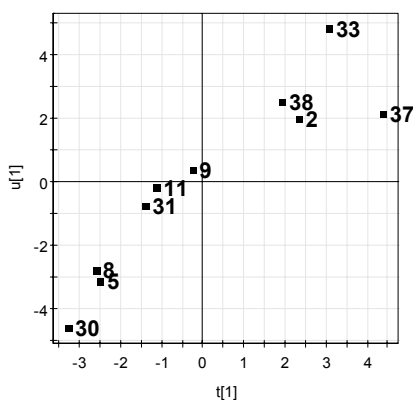
A	R2X	R2X(cum)	Eigen...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signifi...	Ite...
0	Cent.		3.97	Cent.						
1	0.397	0.397	3.97	0.377	0.377	0.143	0.05	0.143	R1	16
2	0.238	0.635	2.38	0.338	0.715	0.36	0.05	0.451	R1	5
3	0.148	0.783	1.48	0.0669	0.781	0.0657	0.05	0.487	R1	14



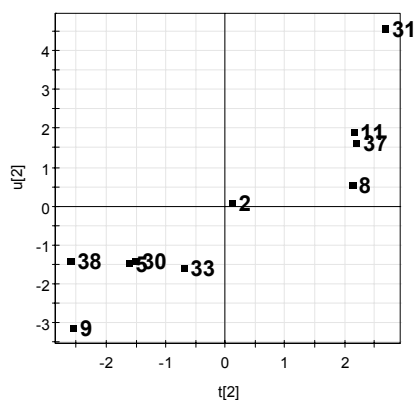
Model interpretation - use scores & loadings

- Strong correlation between chemical data and performance data

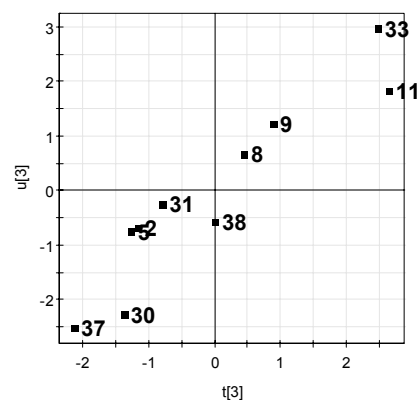
Surfact2.M3 (PLS-Class(1)), PLS training set
t[Comp. 1]/u[Comp. 1]



Surfact2.M3 (PLS-Class(1)), PLS training set
t[Comp. 2]/u[Comp. 2]

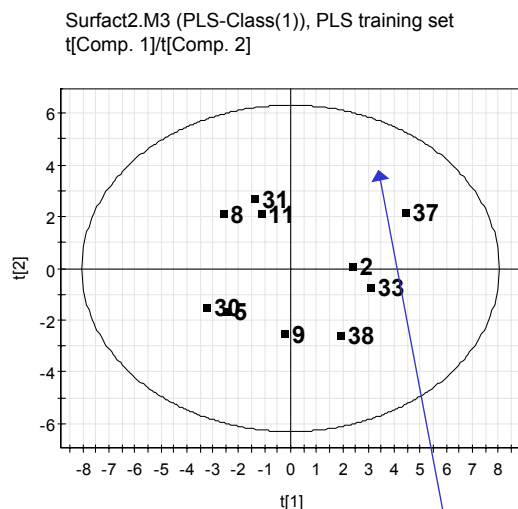


Surfact2.M3 (PLS-Class(1)), PLS training set
t[Comp. 3]/u[Comp. 3]

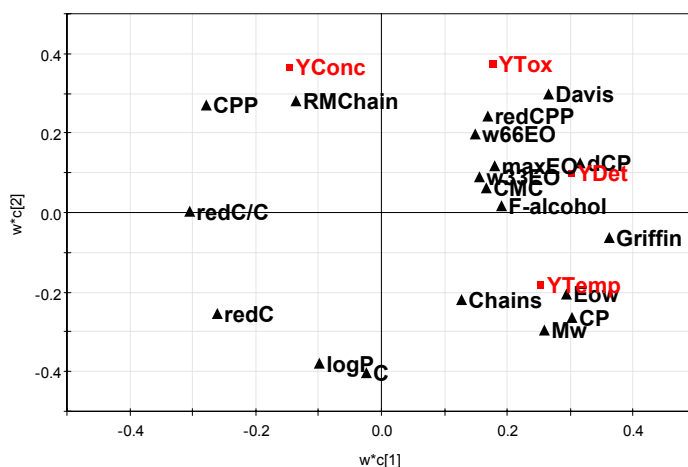


Model interpretation - use scores & loadings

- Strong correlation among responses



Surfact2.M3 (PLS-Class(1)), PLS training set
w*c[Comp. 1]/w*c[Comp. 2]



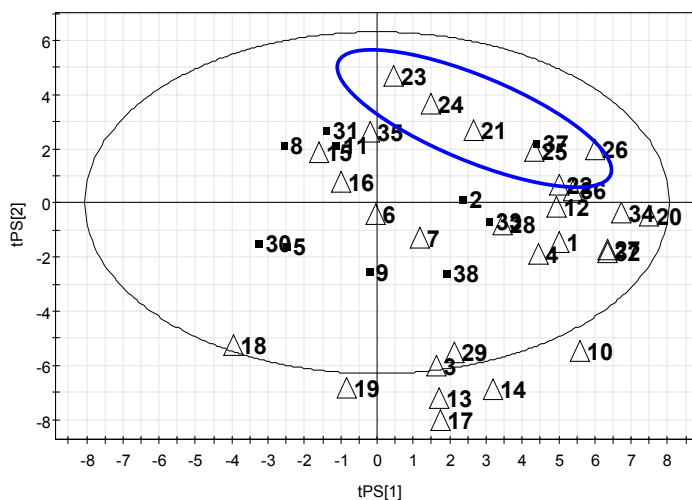
Target area (High YTox, High YDet and Low YTemp)

Model predictions

- Predictions can be used to identify promising surfactant structures, e.g., # 21, 24 & 25

Obs num	Ydet	Yconc	Ytemp	Ytox
37	85.9	2.0	67.0	1.5
15	84.6	2.1	54.2	1.2
21	86.3	2.2	56.1	1.8
22	88.1	1.5	70.1	1.6
23	86.7	2.6	51.9	1.9
24	87.8	2.2	59.2	1.8
25	89.7	1.7	71.5	1.7
27	89.6	0.9	83.1	1.3
28	87.1	1.3	71.5	1.2

Surfact2.M3 (PLS-Class(1)), PLS training set, PS-Surfact2 No Class
tPS[Comp. 1]/tPS[Comp. 2]



Conclusions - Surfactant example

- All surfactants cannot be tested due to lack of time and resources. Multivariate characterisation and design are useful to select a set of *representative* compounds.
- Surfactant performance is a multivariate property and must be treated as such.
- Strong relationships between measured physico-chemical properties of surfactants and their performance profiles.
- Predictions from PLS model identify interesting surfactants for further performance optimisation.

Conclusions - General

- Multivariate characterisation quantifies a discrete change, for instance, the shift from one surfactant or batch of raw material to another
- Critical aspect: the surfactant or raw material “effect” in the “real” system can only be mapped with a set of well chosen model systems
- Statistical experimental design is an excellent tool for selecting representative compounds/items/cases based on a multivariate characterisation

Multivariate Data Analysis and Modelling Basic Course

Chapter 6 Multivariate Calibration



Contents

- Six main steps of multivariate calibration
- Training, test & prediction sets
- Calibration – A short review
- Problems with traditional calibration
 - selectivity
 - precision
 - diagnosis
- Multivariate calibration
 - many signals
 - multivariate space
- Example: SUGAR
- Signal correction
 - Orthogonal Signal Correction, OSC
 - Multiplicative Signal Correction, MSC
 - Standard Normal Variate, SNV
 - Derivation (1st and 2nd derivatives)

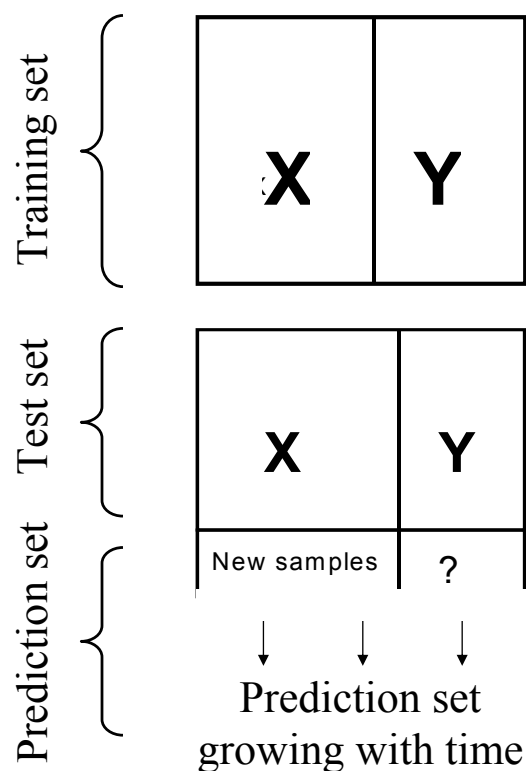
Six main steps of multivariate calibration

1. **Specification** of the analytes with concentration ranges
2. **Selection** of a representative calibration set (training set)
3. **Measurement** of spectra (X) – The analyte concentrations are measured with the reference method (Y)
4. **Evaluation of raw data** (outliers) and pre-processing (filtering and compression)
5. **Calculation** of the calibration model (review of fit, interpretation)
6. **Prediction** of analyte concentrations in new samples (prediction set)

Reference: Wold, S., and Josefson, M., Multivariate Calibration of Analytical Data, in: Meyers, R., A., Encyclopedia of Analytical Chemistry, John Wiley & Sons Ltd, 2000, pp. 9710-9736.

Training, test & prediction sets

- Steps 1- 5 comprise the training phase
- Step 6 is the prediction phase
- When predictive power is satisfactory the calibration model is applied to new samples



Two calibration situations

- Reference samples can be prepared with desired levels of analytes and interferences
 - DOE is useful to make reference samples of known composition
 - Typically a full or fractional grid design with many levels
- Samples are collected from the system, and are analysed by a reference method
 - Many samples (30 – 100) are needed to ensure proper spanning of all properties
 - Selection depends on situation (e.g., location or batch, see next slide)
 - Or use multivariate characterisation and design in principal properties to select representative spectra (see, e.g., Svensson, O., PhD Thesis, Gbg Univ., Sweden)

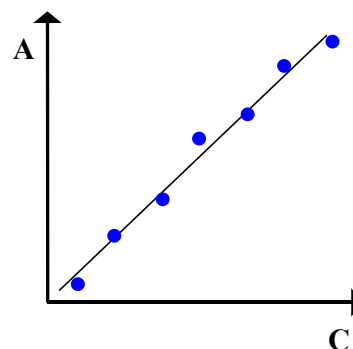
Selection of samples - continued

- Natural samples, e.g., wheat, lake-water, peat etc., *base on geography*
 - If prediction only within given locations:
 - Select training samples from all locations
 - Sort by Y; even = training, odd = test / with many Y make a design in Y
 - Prediction of results at new locations (most common case)
 - Training samples from locations A – O that cover the whole area
 - Test samples from locations P – Z
- Industrial samples, *base on batch*
 - If prediction only within given batches (or continuous process)
 - Select training samples from all batches
 - Sort by Y; even = training, odd = test / with many Y make a design in Y
 - Prediction of results of new batches (most common case)
 - Training samples from batches A – O, spread over time and other relevant properties
 - Test samples from batches P – Z

Calibration – A short review

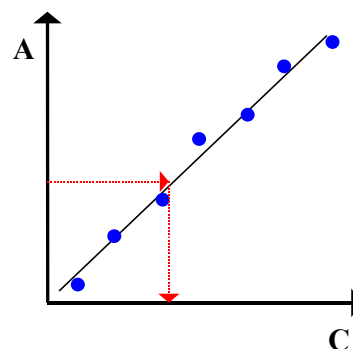
1) Samples with **known** concentrations (c_i) are measured on an instrument

- Resulting signal amplitudes (A_i)
- Standard curve



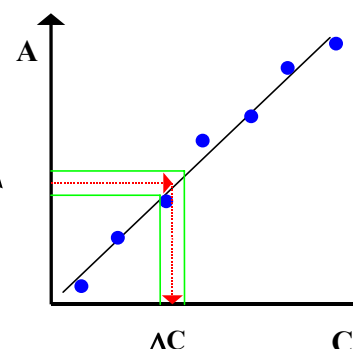
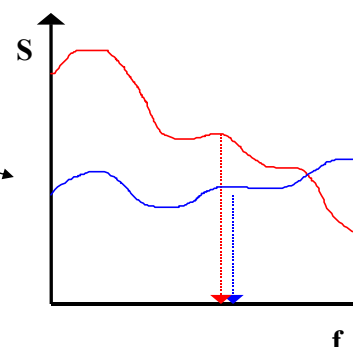
2) New samples with **unknown** concentrations

- Measurements \Rightarrow signal amplitudes, A_j
- \Rightarrow estimated concentration values, c_j (via standard curve)



Problems with traditional calibration

- **Selectivity:** there is NO frequency where ONLY the analyte absorbs
- **Precision:** noise in signal amplitude transmits to the estimated concentration of a new sample
- **Diagnosis:** standard curve valid ONLY for samples similar to the calibration samples



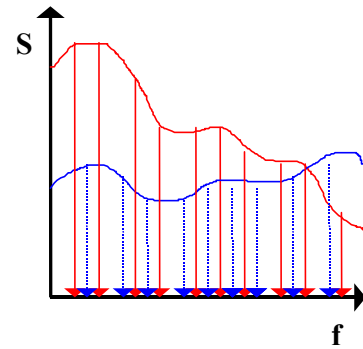
Multivariate calibration

- Many signals (spectrum digitised at K different wavelengths already in the instrument)

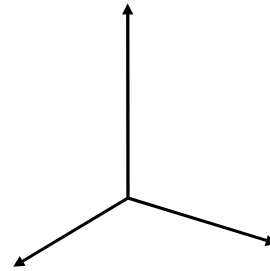
⇒

K variables

K signals

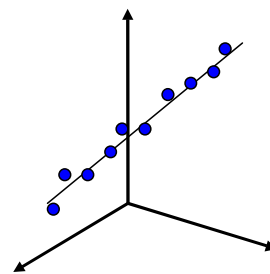


- Multivariate space
 - Each wavelength defines one co-ordinate axis
 - Space with K axes
 - Points, lines, distances, ..., have similar properties in K as well as in 2 and 3 dimensions

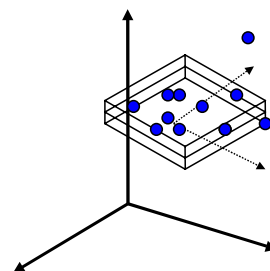


Multivariate calibration

- One analyte:
 - all points (digitised spectra) are situated on a line \pm noise (Lambert-Beer's "law")



- one analyte + interacting compounds, or several analytes + interacting compounds:
 - all points are situated on a hyper-plane \pm noise in K -space



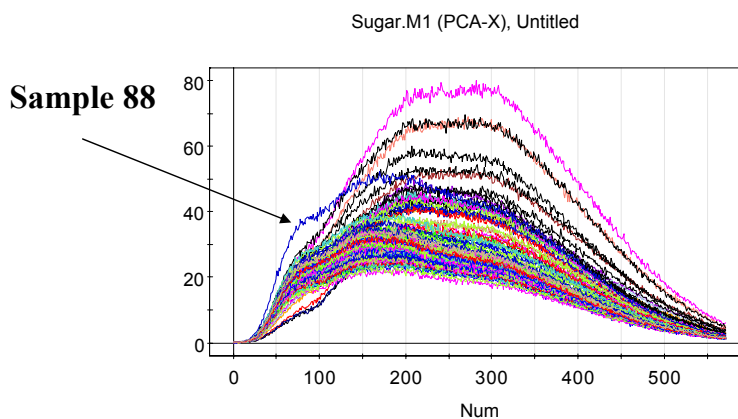
Application areas

- | | | |
|-------------------------------|-----------------------------------|--------------------|
| • Wheat, corn, ... | Protein, water, fat | NIR |
| • Peat, coal | Water, energy, sugars, C, N, S | NIR |
| • Lake water | Humic acids, lignin sulfonate | Fluorescence |
| • Beer, wine | alcohol, protein, sugars, etc. | NIR, IR |
| • Whisky, wines | Taste, smell, "quality" | GC, HPLC |
| • Cellulose, paper | Raw material, lignin, paper char. | NIR, UV, IR, NMR |
| • Pigs (living) | Fat, meat etc. | X-ray |
| • PAH | | UV, NMR |
| • Pharmaceutical apps. | Drug compounds & metabolites | UV-vis, FT-IR, NIR |
| • Process quality | | Sensors, NIR, NMR |
| • + many, many more | | |

Example: Sugar

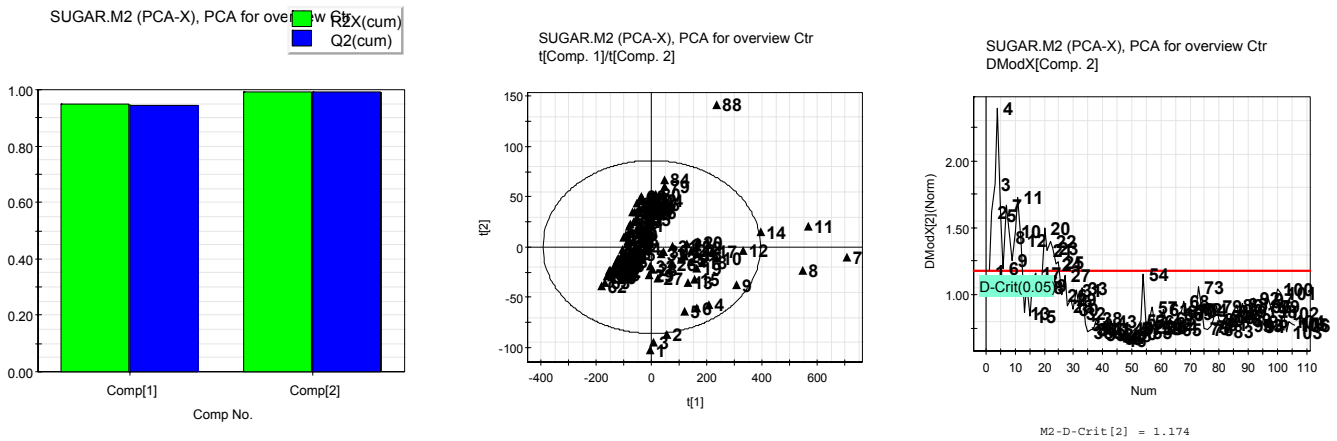
- Data: Fluorescence measurements on white sugar, the final product in the sugar production, dissolved in phosphate buffered distilled water
- 106 samples, 571 X-variables
- Excitation: 240 nm, Emission: 275-560 nm
- Response: Impurity ("ash content")
- Reference: Rasmus Bro, "Håndbog i Multivariabel Kalibrering"

- Plot of spectra
(the X-data):



PCA modelling

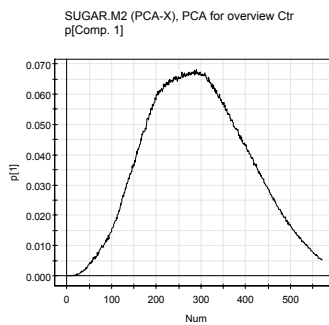
- PCA of unscaled and centered data shows problems in the beginning of the process; stabilisation from around sample 15; sample 88 probable outlier



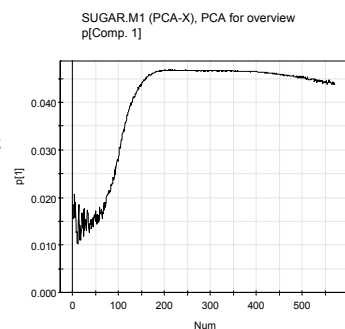
Scaled or unscaled data?

- Two independent spectral contributions are found in data – results of unscaled data are easier to interpret

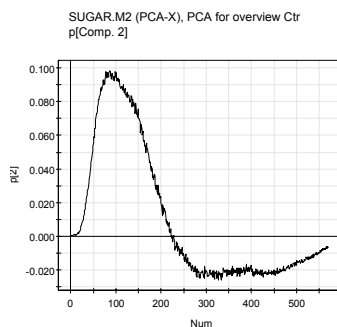
p_1 , unscaled data, has structure and resembles average spectrum



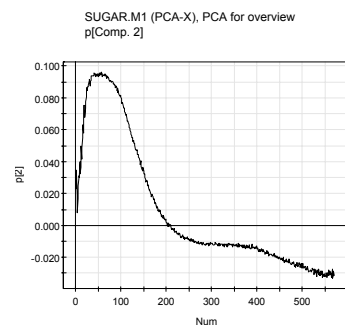
p_1 , scaled data, does not mimic the average spectrum



p_2 , unscaled data, also carries structure, peak-like shape

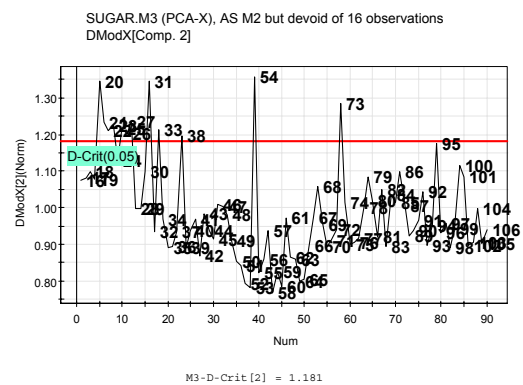
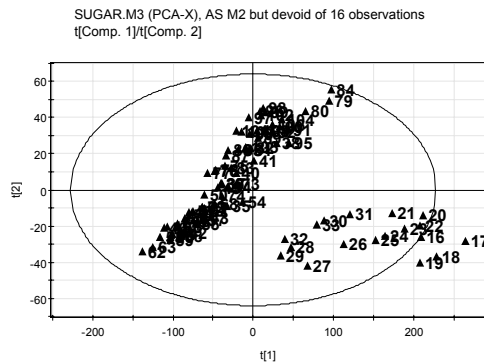


p_2 , scaled data, with a peak



PCA on 90 observations

- Start-up phase and the outlier have been eliminated
- The two main clusters are preserved
- No long-lasting period of consistently high DModX's
- No new sub-clusters or strong outliers emerge



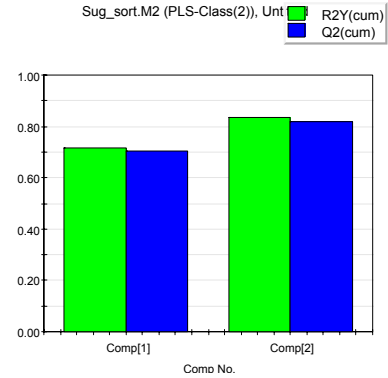
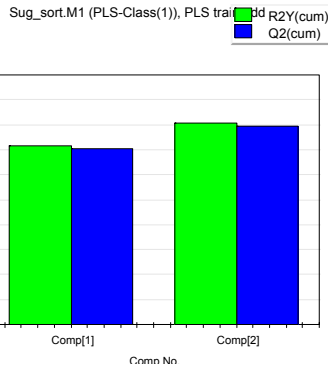
- Conclusion: It is reasonable to apply PLS to these 90 observations!
Gives a good spanning of Y.

PLS - modelling of reduced data set

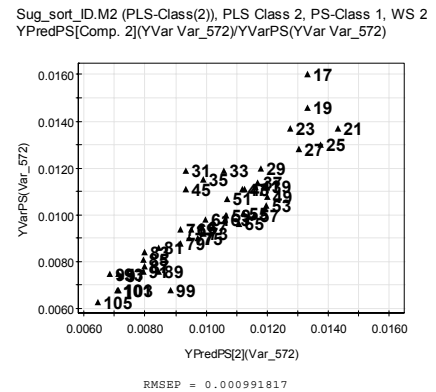
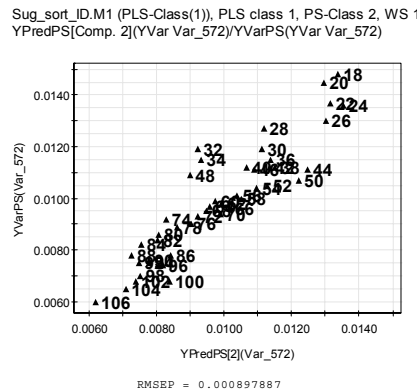
- 16 samples were removed (Nos 1 - 15 + 88)
- 90 samples remained, which were divided in two sub-groups
- Models were computed for each sub-set
- Model validation strategies:
 - Internal validation (cross-validation, Q^2_{int})
 - Response permutation testing ("Validate")
 - External validation (RMSEP, Q^2_{ext})

PLS - modelling of reduced data set

- Results of cross-validation indicate Q^2_{int} in the order of 0.8



- External predictions indicate RMSEP in the order of 0.00010 ($\leftrightarrow Q^2_{ext} \approx 0.8$)

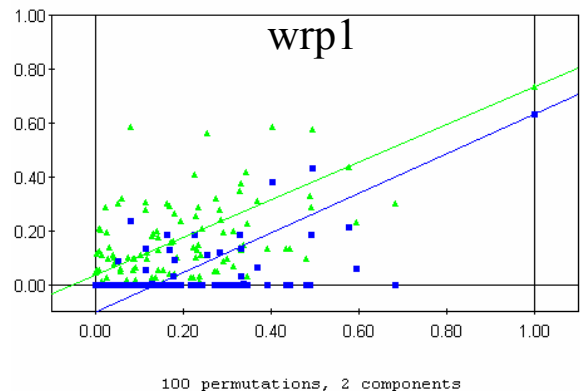
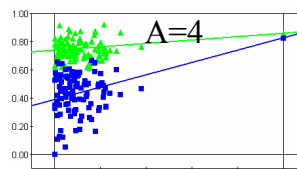
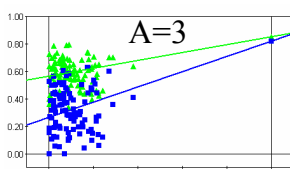
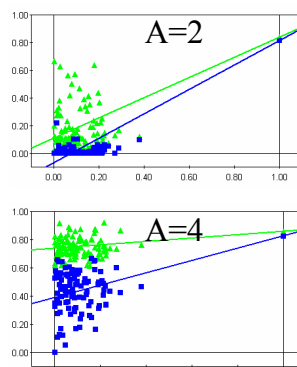
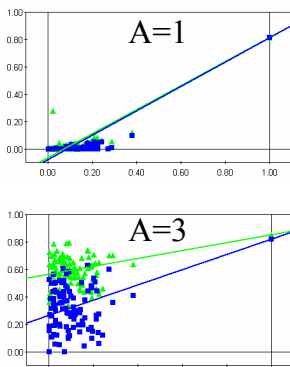


Response permutation test - "Validate"

- To check whether the existing model is the best predictive alternative and the degree of overfit
- Rules of thumb: Y-axis intercepts $R^2 < 0.3$, and $Q^2 < 0.05$
- If the R^2 -line is close to horizontal, this test indicates model overfit


/	Factors				Responses			
	1	2	3	4	Randomise wrp1 in new columns			
Onu m	glas	crtp	mic a	amp/	wrp1	Wrp1:1	Wrp1:2	Wrp1:3
1	40	10	10	40	0.9	3.7	0.6	0.3
2	20	20	0	60	3.7	0.6	3.6	0.6
3	40	20	0	40	3.6	0.3	1.2	1.2
4	20	20	20	40	0.6	1.2	0.3	3.7
5	20	10	20	50	0.3	0.9	0.9	3.6
6	40	0	20	40	1.2	3.6	3.7	0.9

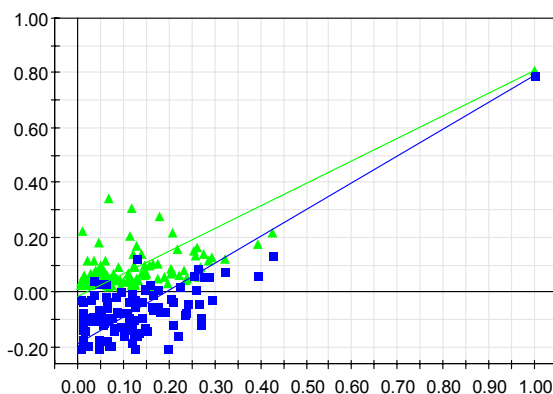
▲ R2
■ Q2




Validate applied to SUGAR models

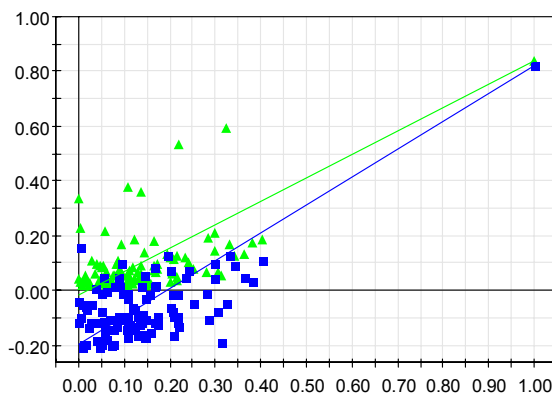
- Validate – excellent results, i.e., calibration models are valid

Sug_sort.M1 (PLS-Class(1)): Validate Model  R2
Var_572 Intercepts: R2=(0.0, -0.0138), Q2=(0.0, -0.188)



100 permutations 2 components

Sug_sort.M2 (PLS-Class(2)): Validate Model  R2
Var_572 Intercepts: R2=(0.0, -0.0163), Q2=(0.0, -0.198)



100 permutations 2 components

Summary of initial PCA/PLS modelling

- Scaled or unscaled spectral data?
 - Unscaled data ("centered but not scaled") easier to interpret
- Marked drift in the process \Leftrightarrow start-up variation
 - Unrepresentative samples should be removed (1 – 15 + 88)
 - 90 observations remained
- Cross-validation and dependency among adjacent observations
 - Data were sorted to break the auto-correlation structure (Note: good solution for our modelling efforts, but not necessarily for long-term process monitoring)
- External predictions
 - To enable external predictions data were split in two groups, odd- and even-numbered
 - External $Q^2 > 0.8$ for both groups
- Conclusion: Fluorescence data allow for reliable on-line prediction of ash content

Signal correction (“filtering”) of SUGAR data

- Signal correction can be used to pre-process data and remove systematic unwanted behaviour, such as baseline variation and multiplicative scatter effects
- Problem 1: risk of removal of variation in X that correlates with Y
- Problem 2: risk of over-training of model
- Spectral filters applied to SUGAR data:
 - **Multiplicative Signal Correction, MSC** (Geladi & Martens, 1985)
 - **Standard Normal Variate, SNV** (Barnes, 1989)
 - **Orthogonal Signal Correction, OSC** (Wold, 1997)
 - **Derivation** (1st and 2nd derivatives)

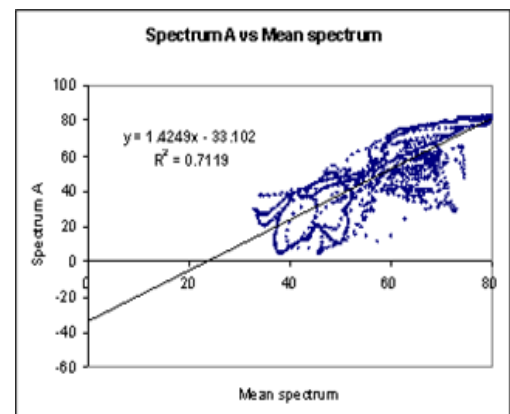
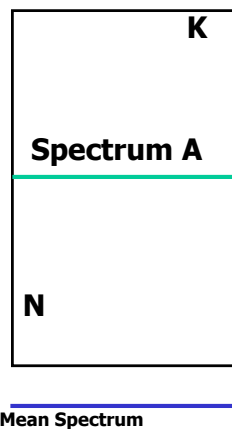
Spectral filters in SIMCA

- **MSC:** Each digitised spectrum (x_i' , row-vector in X) is regressed against the mean spectrum (m):

$$x_{ik} = a_i + b_i m_k + e_{ik}$$

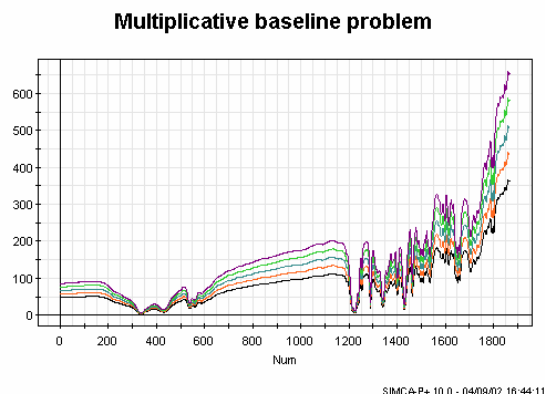
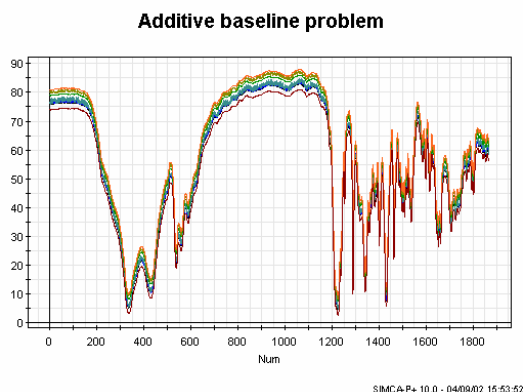
- From each spectrum one subtracts the intercept (a_i) and divides by the slope (b_i):

$$x_{i,corr}' = (x_i' - a_i) / b_i$$



Spectral filters in SIMCA

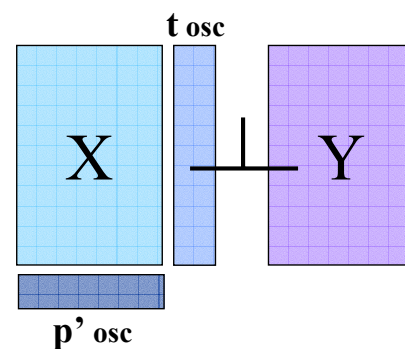
- **SNV**: Same mathematical form as MSC
 - Parameters a_i and b_i are calculated as the average and standard deviation of the i^{th} row of X ; Corresponds to row-centering and normalisation



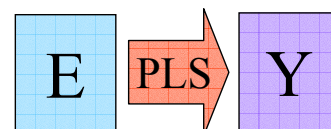
MSC and SNV are baseline corrections which remove additive and/or multiplicative effects!

Spectral filters in SIMCA

- Orthogonal Signal Correction, OSC
 - Calculate first PC of $X \Leftrightarrow$ score t
 - Orthogonalise t with regards to Y
 - $t_{\text{osc}} = (1 - Y(Y'Y)^{-1}Y')t$
 - Some NIPALS steps to give weights (w^*), and updates of t_{osc} and t , until convergence
 - Subtract correction $E = X - t_{\text{osc}}p'$
 - One or two OSC components recommended
 - Work with one response at a time
- OSC uses information in Y to construct a filter of X
- MSC and SNV work only with X , and may remove predictive information from X



$$E = X - t_{\text{osc}} * p'$$



Spectral filters in SIMCA

- 1st derivative spectrum
 - Provides the slope at each point of the original spectrum
 - Has peaks where the original spectrum has maximum slope and it crosses zero where the original has peaks
 - Removes additive baseline (“offset”)
- 2nd derivative spectrum
 - Measures curvature at each point in the original spectrum.
 - Is more similar to the original spectrum and has peaks approximately as the original spectrum, albeit with an inverse configuration
 - Removes a linear baseline
- Problem: May reduce the signal and increase the noise ⇒ noisy spectra
- Savitsky and Golay (SG) smoothing
 - SG-derivatives are based on fitting a low degree polynomial (quadratic or cubic degree) piece-wise to the data, followed by calculating the first and second derivatives

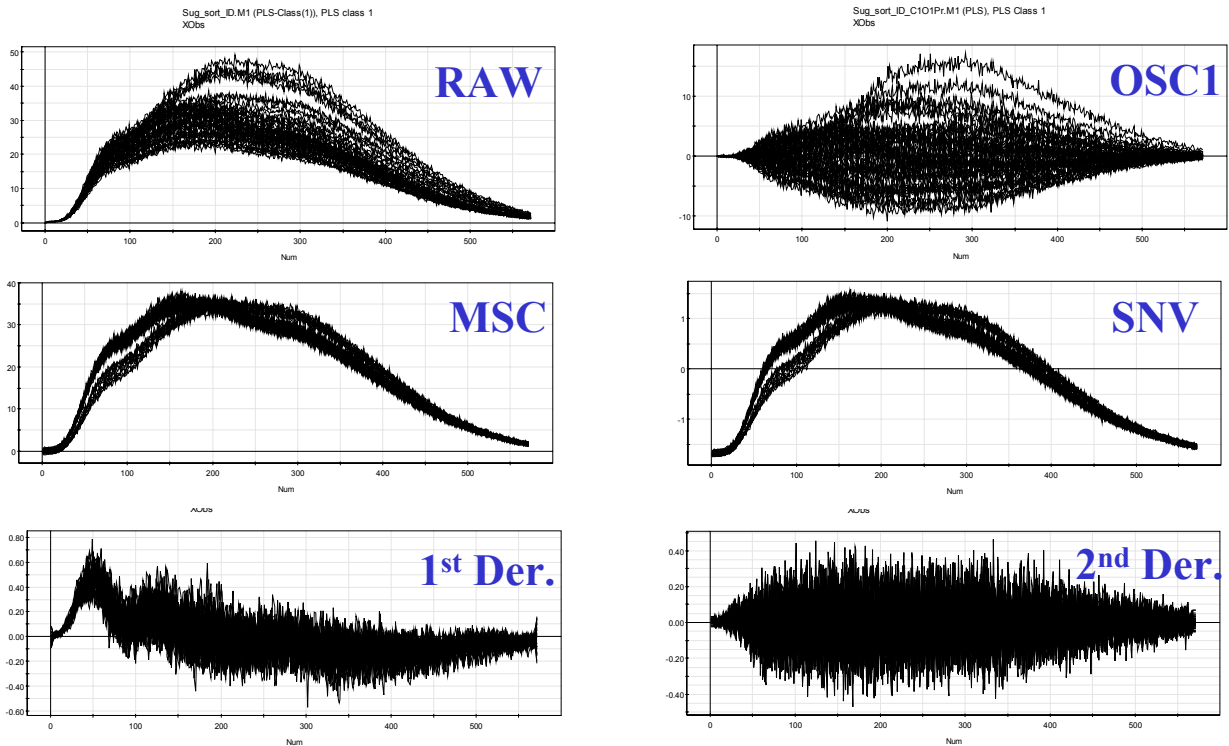
Results of signal correction (SUGAR)

- Procedure:
 - 1) Reduced data set (90 observations)
 - 2) Sorting and splitting of data preserved
 - 3) Standard PLS, and OSC, MSC, SNV & derivatives are compared
 - 4) Models trained on odd and tested on even, and vice versa
 - 5) External validation RMSEP and Q^2 used to evaluate predictive power

Model	RMSEP(o)	$Q^2_{\text{ext}}(o)$	A(o)	RMSEP(e)	$Q^2_{\text{ext}}(e)$	A(e)
PLS	0.00090	0.83	2	0.00099	0.80	2
OSC1	0.00084	0.85	1	0.00094	0.82	1
MSC	0.00174	0.36	5	0.00173	0.33	5
SNV	0.00163	0.43	5	0.00160	0.43	5
1st Der.	0.00111	0.81	2	0.00122	0.73	2
2nd Der.	0.00223	0.00	2	0.00258	0.00	2

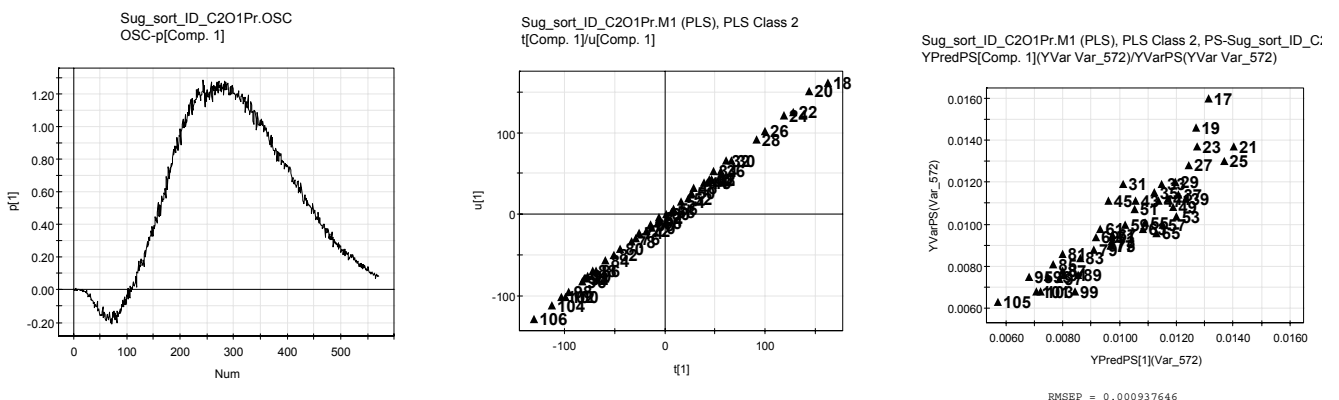
OSC1-PLS works equal to or better than PLS!

Results of signal correction (SUGAR)



Results of OSC1-PLS model (Train even/Test odd)

- 31% SS removed by OSC



Examine OSC-loading
(what was subtracted?)

Near perfect correlation
between X and Y for the
training set

Observed/predicted
for the test set

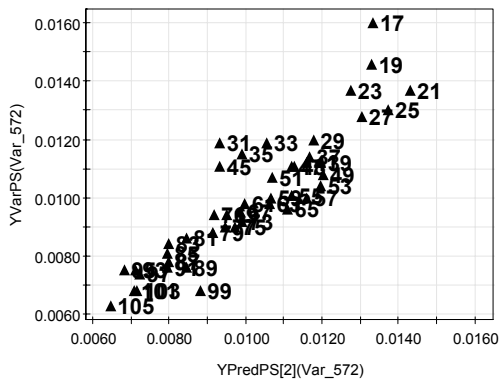
- OSC training set picture is exaggerated. OSC predictions still good.

Results of OSC1-PLS model (Train even/Test odd)

- Conventional PLS

- RMSEP = 0.00099
- $Q^2_{\text{ext}} = \mathbf{0.80}$ (A = 2)

Sug_sort.M2 (PLS-Class(2)), PLS train even, PS-Class 1, WS 2
YPredPS[Comp. 2](YVar Var_572)/YVarPS(YVar Var_572)

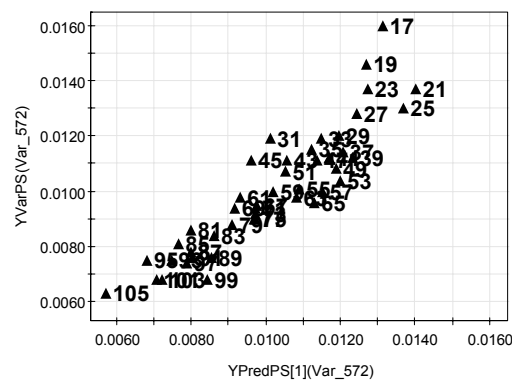


RMSEP = 0.000991817

- OSC1-PLS

- RMSEP = 0.00094
- RMSEP lowered by 5%
- $Q^2_{\text{ext}} = \mathbf{0.82}$ (A = 1)

Sug_sort_ID_C2O1Pr.M1 (PLS), PLS Class 2, PS-Sug_sort_ID_C:
YPredPS[Comp. 1](YVar Var_572)/YVarPS(YVar Var_572)



RMSEP = 0.000937646

Conclusions of SUGAR example

- We can make a predictively sound calibration model for impurity with predictive power above $Q^2 = 0.80$
- Fluorescence is a relevant technique in this application
- OSC gives a slight improvement in predictive power

Conclusions - Multivariate calibration

- PCA is an informative modelling tool prior to PLS analysis
- External validation of predictive power is extremely important (be aware of auto-correlation among time points in process data!!!!)
- Signal correction is useful to remove undesired systematic behavior in X-data
- SNV and MSC may remove variation from X that correlates with Y
- OSC removes variation from X that does not correlate with Y

Discussion - Representative calibration samples

- The uncertainty of predictions for a given signal increases rapidly outside the range spanned by the calibration samples. *Hence, the calibration model should not be applied far outside this range.*
- The best training set is obtained if a design is made in all factors relevant for the calibration, but this is often difficult.
- When DOE is not possible, a sampling where samples are selected to cover the major sources of variation is the best strategy

Multivariate Data Analysis and Modelling Basic Course

Chapter 7 Process Applications



Contents

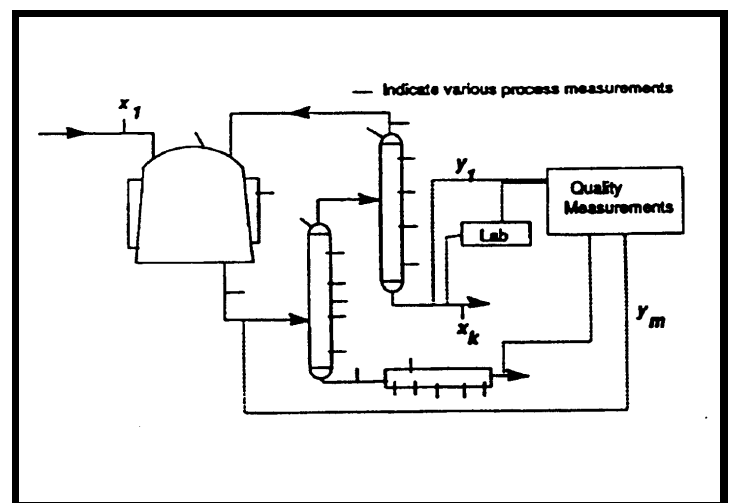
- Process modelling
 - Purpose
 - Data
- Monitoring a process
 - A chemical production plant
- Modelling a process output
 - A mineral sorting plant

Process Data Analysis - Purposes

- 1. Monitoring** the state of the process, statistical process control (SPC)
 - Early warning
 - Diagnostics - finding "assignable causes" (SPC jargon \Leftrightarrow interpret deviations)
- 2. Understanding** the relationship between
 - input variables, X (process data) and output variables, Y (product quality, cost, amount, ...)
- 3. Optimisation**
 - Use process models to improve process

The Problem

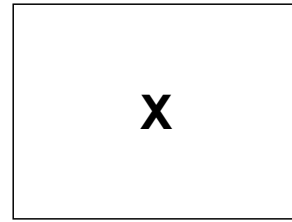
- 50 years ago
 - Few variables: T, P, flows
- Today
 - Many measurements and very often
 - Large data sets
 - The latent variable concept
- Process the same
- Data have changed
 - $K = 5 \rightarrow 500$
 - $N = 10 \rightarrow 1000$



Typical process (from MacGregor et al. 1991)

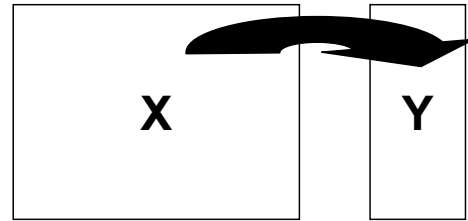
Process applications

- Monitoring a process
 - A chemical production plant
 - PROC1A



The process

- Modelling a process output
 - A mineral sorting plant
 - SOVR



The process

Quality

Example - Monitoring (PROC1A)

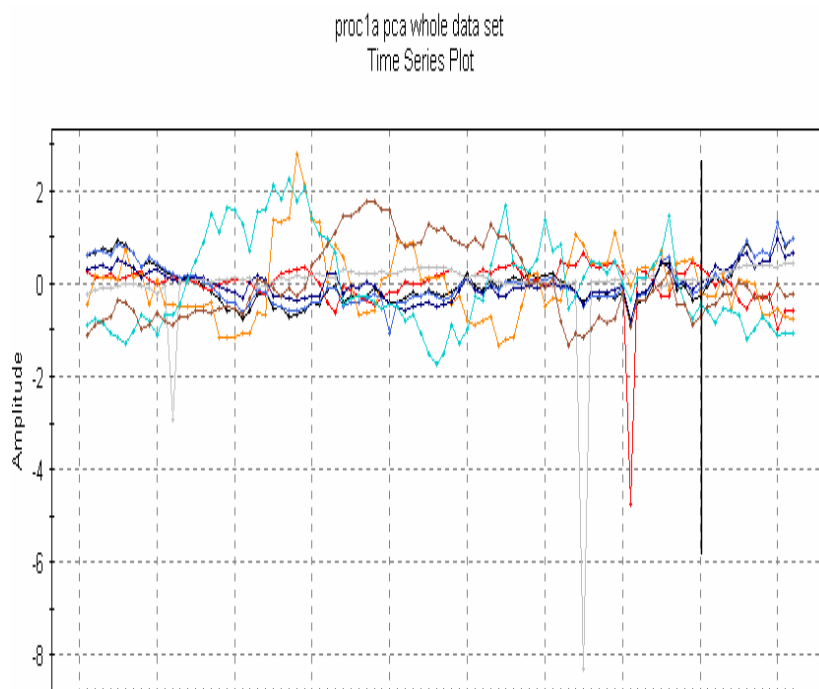
- A chemical production plant
 - A continuous steady state process
 - All data are coded, not to reveal any trade secrets
- The data - 33 variables, 92 observations
 - 7 controlled process variables (x1in-x7in)
 - 18 intermediate process variables (x8md-xpen)
 - 8 output variables (y1-y8)
 - Data sampled during 92 time units (e.g. hours or minutes)
- The process went out of control around time 80 and had to be shut down at time 92

Example - Monitoring a process (PROC1A)

- Questions:
 - How can we quality control the data?
 - How can we use these data to get an overview of the state of the process?
 - Do we see trends, groups of observations?
 - Can we detect changes in the process over time?
 - Can we detect sudden upsets in the process?
- We will use PCA modelling to address these questions

PROC1A – Time series plots of raw data

Just plotting the raw data does not reveal anything particular at time = 80

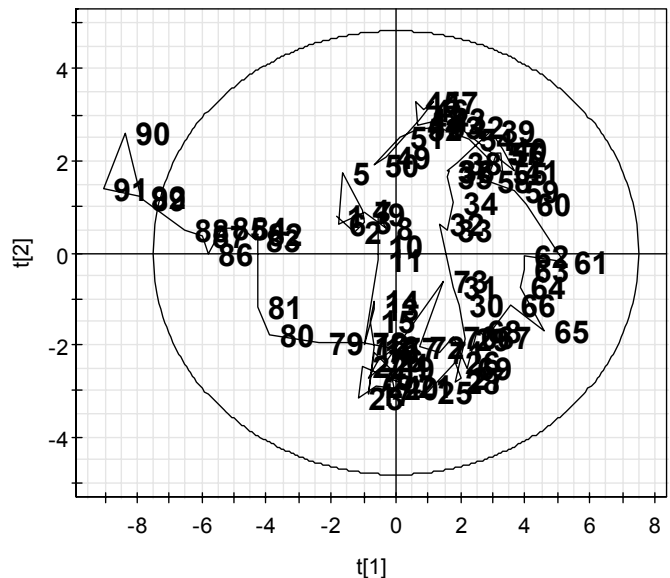


PROC1A – Overview PCA model

Model and understand the data

- Always start with an overview of the data
- Variable inspection indicated no need for transformation
- Calculate two first PCs
 - for overview this is usually sufficient
- The score plot reveals a clear trend in the data towards the end
 - All observations after process upset behave differently

proc1a.M2 (PCA-X), PCA entire data set
t[Comp. 1]/t[Comp. 2]



PROC1A – Modelling normal behaviour

Select observations representing normal operating conditions

- From the process engineer we know that the process was behaving normally until around time 70 and was definitely out of control at time 80
- Remove observations 70 - 92

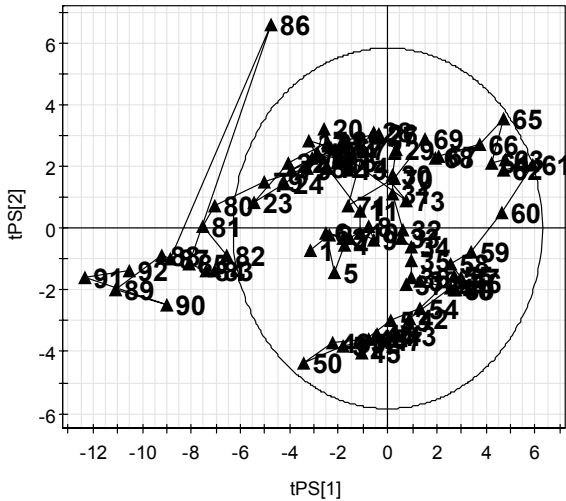
Model the normal process behaviour

- Make a PCA on the 69 first observations
 - Use cross-validation to indicate the number of PCs
 - Three significant components in PROC1A, modelling 51% of the variability
- Using the historical data (observations 1-69) one can define areas or intervals where process seems to be in control
 - These limits can then be used when monitoring the process
 - Multivariate Statistical Process Control, MSPC

PROC1A – Monitoring the process

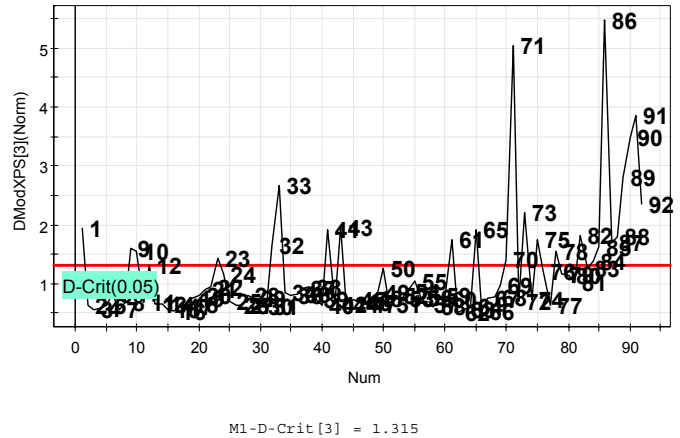
- Projecting the new data onto the model clearly indicates the process upset around time 80

proc1a.M1 (PCA-X), PCA on normal behavior 1-69, tPS[Comp. 1]/tPS[Comp. 2]



- In the DModX-plot we can see that most of the observations (before time 70) are below the critical distance – However, number 33 is somewhat high and will be investigated further

proc1a.M1 (PCA-X), PCA on normal behavior 1-69, PS-proc1a DModXPS[Comp. 3]

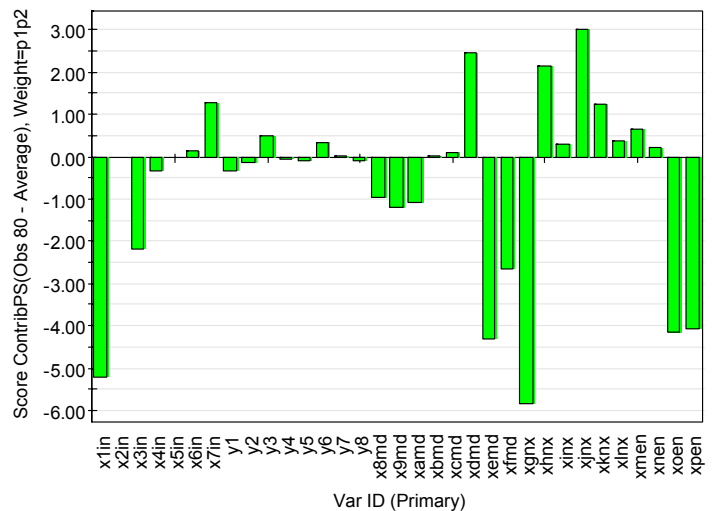


PROC1A – Contribution plots

Identifying contribution of variables

- Process changes detected in scores
 - Observation 80 and all thereafter are deviating from normal behaviour
- Contribution plot (Scores)
 - Identifies changes in variables, relative to the average or to a normal observation
 - Which variables are contributing to the **difference** between observation 80 and the average observation?
- Contribution = $\Delta X * \text{weight}$
 - weights: None, p, pp, RX

proc1a.M1 (PCA-X), PCA on normal behavior 1-69, PS-proc1a Score ContribPS(Obs 80 - Average), Weight=p[1]p[2]

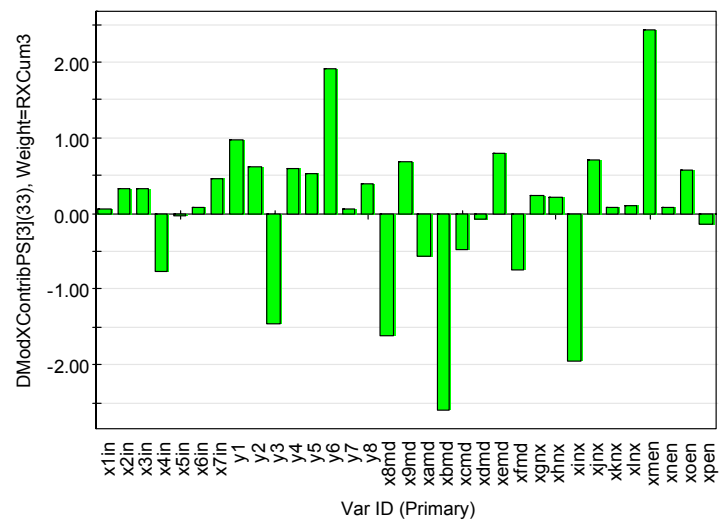


PROC1A – Contribution plots

Identifying contribution of variables

- Process changes detected in **DModX**
 - Observation 33 is deviating from normal behaviour
- Contribution plot (**DModX**)
 - Identifies the abnormal values or patterns in the variables causing the large residuals
 - Which variables are contributing to the **large residuals** in observation 33?
- Contribution = ResX * weight
 - weights: None, RX

proc1a.M1 (PCA-X), PCA on normal behavior 1-69, PS-proc1a
DModXPS Contrib(Obs 33), Weight=RX[3]



PROC1A – Summary

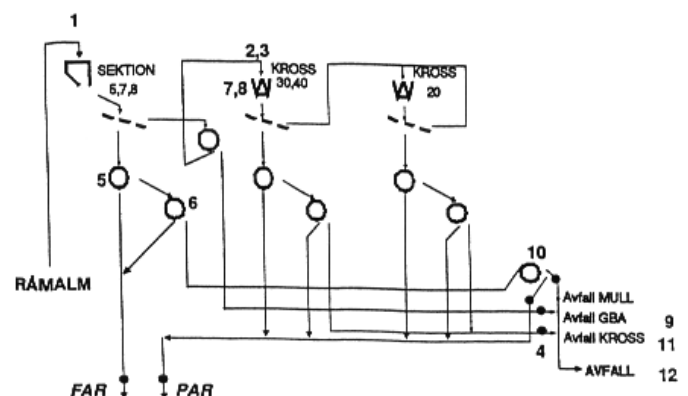
- Monitoring the new data
 - The process upset around time point 80 was easily detected
 - Observation 80 was projected outside the tolerance ellipse (Hotelling's T²)
- Interpreting the deviations
 - The contribution plot showed that the problem was to be found in a number of variables
 - e.g. x1ln, xemd, xgnx, xoem, and xpen were all too low
- Summary
 - The process upset was easily detected at an early stage by using PCA and MSPC
 - The possible variables related to the problem was identified
 - The problem might have been corrected by the operator if seen in time

Example - Modelling the process output (SOVR)

- A mineral sorting plant
 - A continuous dynamic process
 - Presence of feedback control
 - Many responses to consider
- Raw iron ore is divided into finer material by grinders. The material is sorted and concentrated by magnetic separators. Concentrated material is divided in two products
 - PAR that goes to a pelletization process
 - FAR (fines) that is sold as it is
- The primary goal was to identify the most important process factors and set up a prediction model to use on-line

Example - Modelling the process output (SOVR)

- The Data - 18 variables, 572 observations
 - 3 manipulated process (input) variables
 - 9 intermediate process (input) variables
 - 6 output variables
 - X-data were from process log (minutes)
- DOE was used
 - A CCC design in two levels
 - 14 runs + 3 centre points
- The output (or quality) variables were sampled and analysed in the laboratory once for each design setting
 - Each output measurement was then preserved for 20 minutes to capture normal process variation (variations in the X-block)

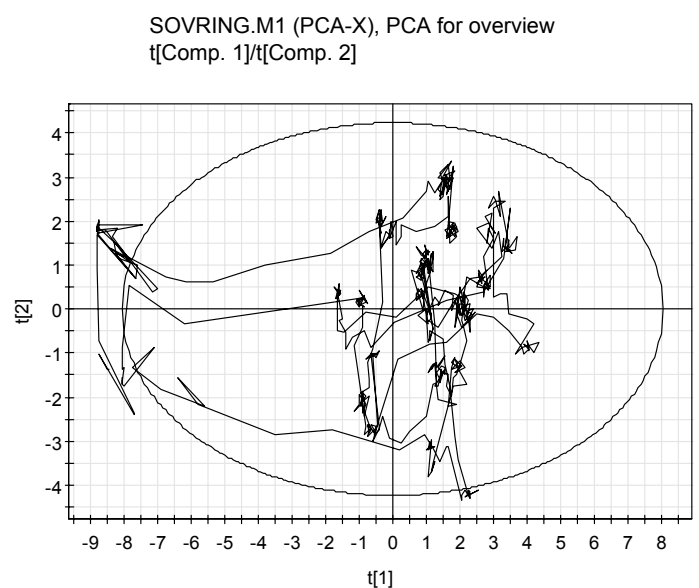


Example - Modelling the process output (SOVR)

- Questions
 - Can the data be used to model the process?
 - Is it possible to monitor and identify process upsets?
 - Are there trends, groups, different states of the process?
 - Can we understand the relationship between input and output variables?
 - Can we make predictions?
- We will use PCA and PLS modelling to address these questions

SOVR - The overview model

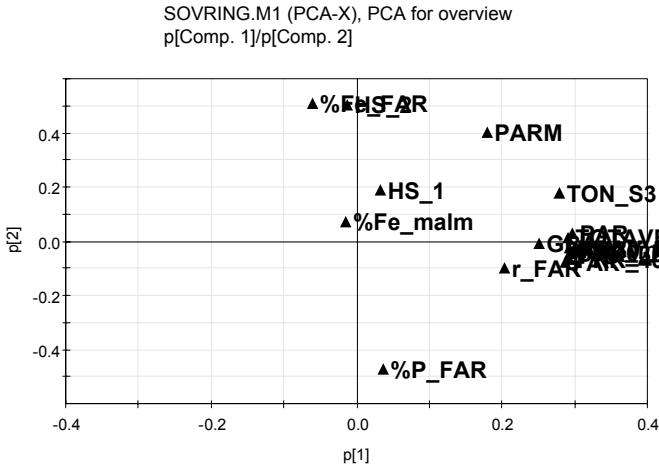
- The first two PCs were calculated using all data in order to get a good overview of the data
- The PC-model reveals clustering induced by the experimental design changes
- Two of the clusters deviate to the left in the score plot
 - Why?



SOVR - The overview model

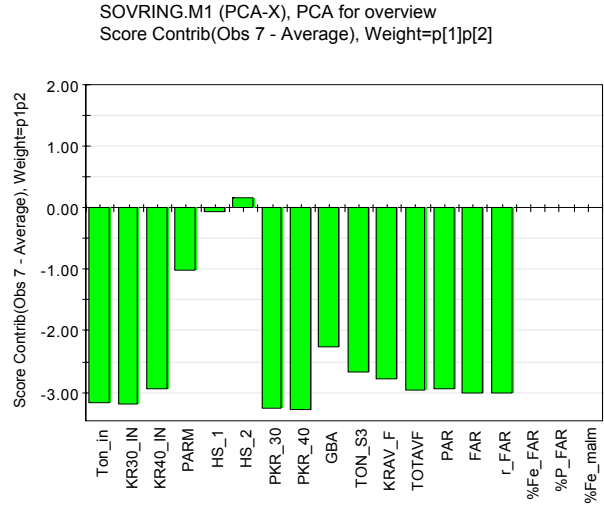
- The answer to why the two clusters deviate in the score plot can be found in two ways. First the more general information

– Loading plot



- For a more specific answer we ask the model: Which variables contribute to the change from an average sample to one of the samples in the clusters?

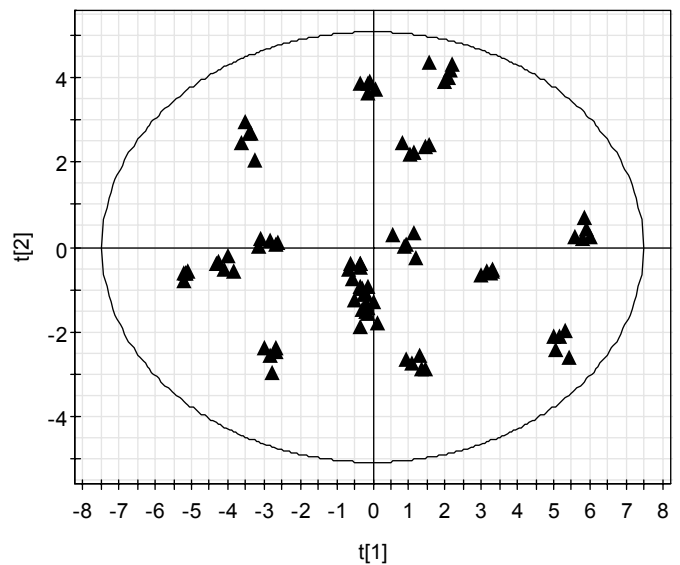
– Contribution plot



SOVR - Overview selected observations

- A new PC-model with selected observations was made
 - 5 observations from each of the 17 design points were selected
- The new PC-model made on the 85 selected observations and all variables clearly shows a clustering
 - the process is stable in all the design points
 - the centre points are well reproduced
- These 85 observations will be used in the following PLS-model
 - PLS-model with expanded terms

SOVRING.M6 (PCA-X&Y), PCA on 85 obs
t[Comp. 1]/t[Comp. 2]

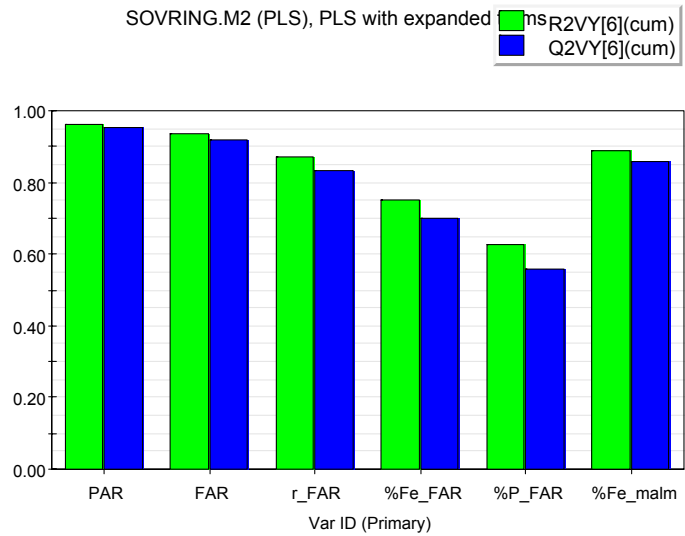
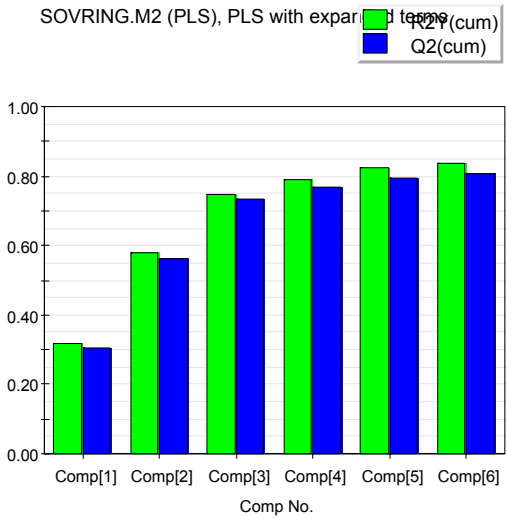


SOVR - The PLS-model

- Cross-validation indicates 6 PLS-components

– $R^2Y(\text{cum}) = 0.84$, $Q^2(\text{cum}) = 0.81$

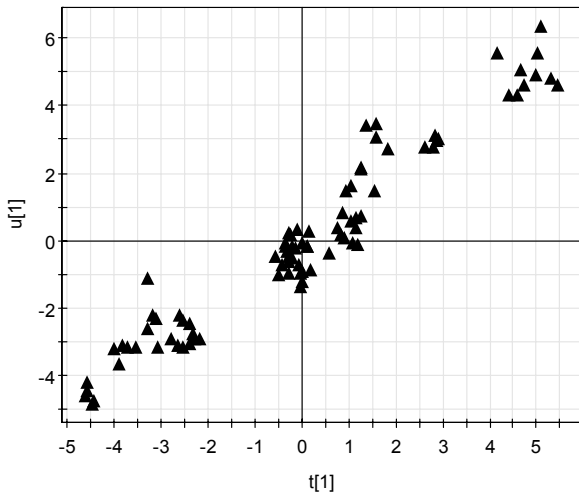
- Most of the output variables are well explained by the model



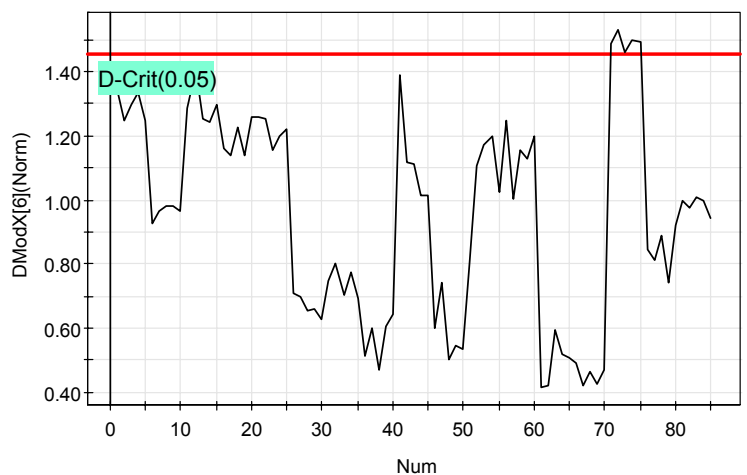
SOVR - Analysing the PLS-model

- The inner relationships and DModX are OK

SOVRING.M2 (PLS), PLS with expanded terms
t[Comp. 1]/u[Comp. 1]



SOVRING.M2 (PLS), PLS with expanded terms
DModX[Comp. 6]

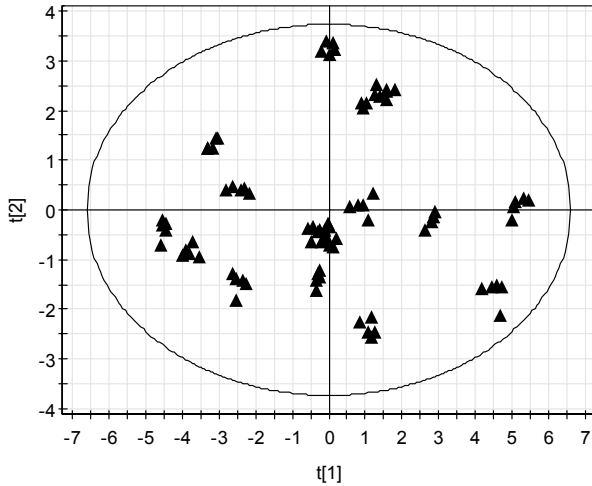


M2-D-Crit [6] = 1.455

SOVR - Interpreting the PLS-model

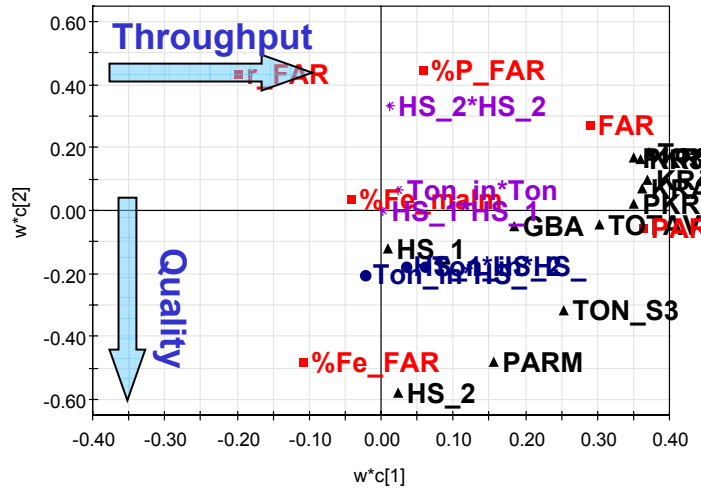
- The t_1/t_2 score plot looks as expected

SOVRING.M2 (PLS), PLS with expanded terms
t[Comp. 1]/t[Comp. 2]



- In the corresponding loading plot we can assign properties to the components

SOVRING.M2 (PLS), PLS with expanded terms
w*c[1]/w*c[2]

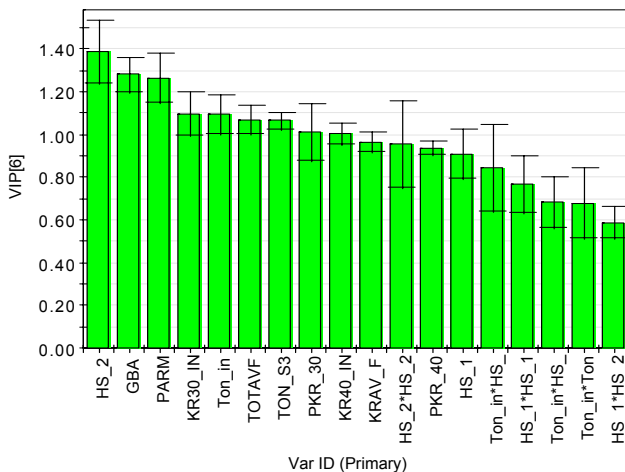


- Observations in the 4th quadrant combine high throughput and high quality

SOVR - Interpreting the PLS-model

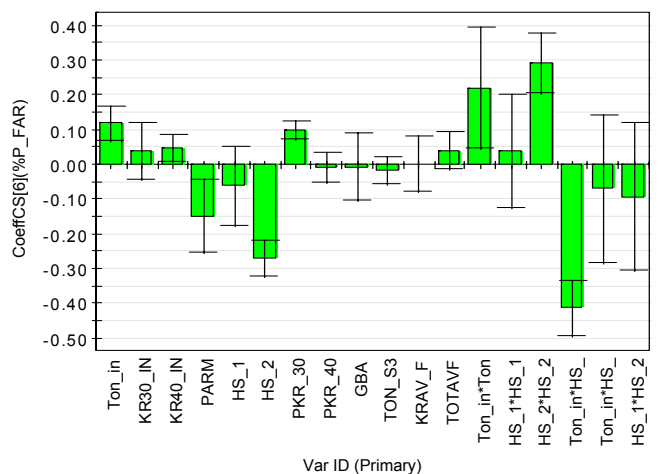
- The **VIP-plot** shows how important each input variable is to the total model
 - All outputs, all PLS-components

SOVRING.M2 (PLS), PLS with expanded terms
VIP[Comp. 6]



- The **coefficient plots** show how the input variables affect each individual output (here; %P_FAR)
 - Individual outputs, all PLS-components

SOVRING.M2 (PLS), PLS with expanded terms
CoeffCS[Comp. 6](YVar %P_FAR)

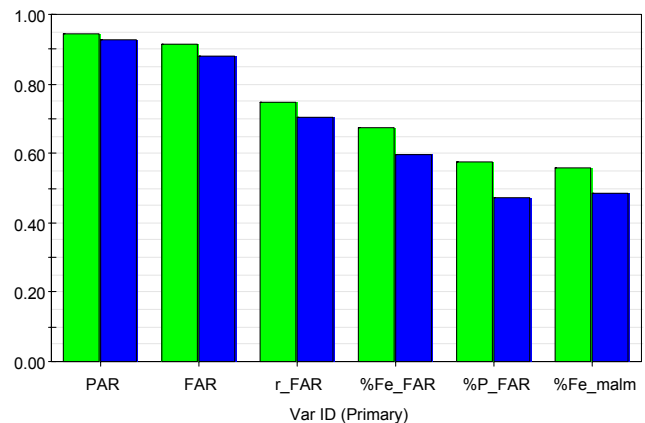


SOVR - Interpreting the PLS-model

- The influence of an input variable can be visualised in a contour plot
- Note: A correct contour plot can only be created for input variables varied according to DOE
 - Causality versus correlation
- A new PLS-model was calculated with only the 3 manipulated input variables
 - 3 manipulated input variables
 - 6 interaction & square terms
 - 6 output variables
- 4 significant PLS-components
 - $R^2Y(\text{cum}) = 0.74$, $Q^2(\text{cum}) = 0.68$

- The only output that was modelled worse by this reduced model was, as expected, %Fe_malm (% Fe in incoming ore)

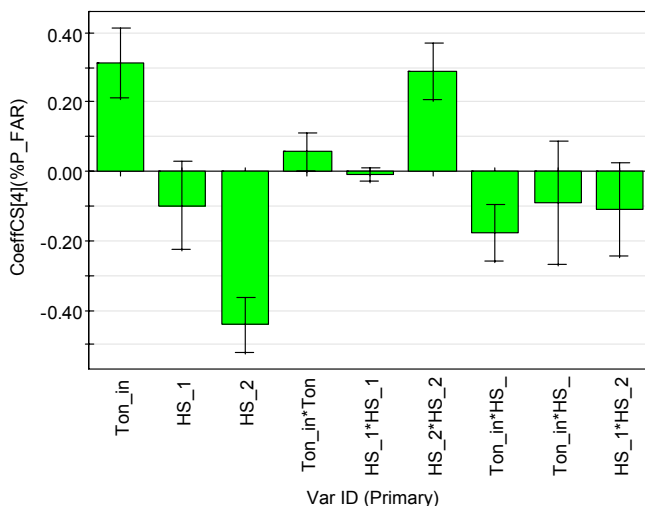
SOVRING.M7 (PLS), PLS only with designed variables



SOVR - Interpreting the PLS-model

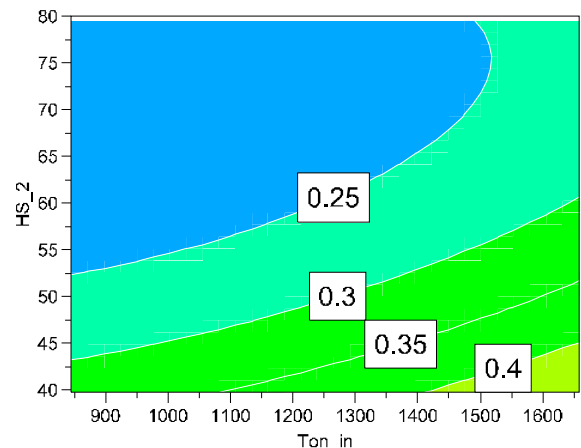
- The coefficient plot for %P_FAR shows almost identical relationships as the first PLS-model

SOVRING.M7 (PLS), pls only with designed X-variables
CoeffCS[4](%P_FAR)



- With the contour plot it is easier to interpret the effects of combined factors
 - here HS_2 and Ton_in (HS_1 kept at its centre value)

SOVRING.M7 (PLS), PLS only with designed X-variables
%P_FAR, Comp 4



For each Ton_in, there is an optimal HS_2

SOVR - Summary

- For prediction and monitoring purposes use the first PLS-model
 - The online measurements enable online predictions
 - The correlation structure stabilises the model and the predictions
 - For early fault detection it is also better to use more process parameters in the model, as more types of problems can be detected
- One interesting output to make online predictions for is the Iron content in the incoming ore (%Fe_malm)
 - Very difficult to get representative samples for analysis
 - When %Fe gets too low in the incoming ore, the process engineer must inform the miners

Modelling process output - Summary

- Collect data from process
 - Use experimental design if possible
- Use PCA to select data to use in PLS-model
 - Remove unwanted process situations from data
- Make PLS-model
- Use model for on-line predictions of final quality
- Use model for understanding and optimisation
 - Cause and effect relationships only with experimental design

Multivariate Data Analysis and Modelling Basic Course

Chapter 8 Conclusions



Combine DOE and multivariate data analysis

- In many process applications the system can be characterised by inherent, latent variables, which are few compared with the number of observed variables.
- Multivariate projection methods find the important LVs. These LVs become more stable the more relevant variables are included.
- Processes are monitored and responses are predicted by means of the latent variables.
- **The process world is multivariate - use multivariate methods! Capture the essentials in a few plots!**

Use DOE to improve/optimize products and processes

- High quality, efficiency and consistency are defining characteristics of a successful organisation
- DOE is the most efficient means of achieving these objectives
- DOE leads to savings through
 - shorter lead times
 - optimal raw material use/less waste
 - fewer product defects
 - less pollution
 - efficient operating conditions
 - ...
- DOE leads to earnings through
 - higher quality products
 - increased throughput
 - new and innovative products
 - ...

Key features of DOE

- How to make experiments efficiently
 - Span the experimental domain with the aid of a suitable experimental design
- How to analyse the data
 - Use good statistical tools to evaluate experimental results
- How to interpret the results
 - With the use of user-friendly PC-based graphical facilities
- How to convert modelling results into concrete actions/decisions
 - MODDE optimiser & verifying experiments

Umetrics has the Solution

- We can offer a solution to almost any multivariate problem.
- Umetrics has consultants with a wide range of experience and expertise, ready to advise, assist, or lead your projects.
- Umetrics is committed to staying ahead in the fast developing field of multivariate methods.
- Our knowledge and experience are continuously transferred into our software products and courses.

Multivariate Methods - a Total Concept from Umetrics

- Products
 - **MODDE**: is our state-of-the-art DOE software
 - **SIMCA-P**: for off-line modelling and execution
 - **SIMCA-4000**: for on-line execution in connection with your PCS
 - **SIMCA-Batch OnLine**: for on-line execution of batch processes
- Courses
 - **Umetrics Academy** offers a wide range of courses, from beginners to advanced users
- Consulting
 - **Profit on our knowledge** to speed up your projects
- On the Web
 - Get the latest news and links at our web-site: www.umetrics.com

Multivariate Data Analysis and Modelling Basic Course

Chapter 11 Additional Topics I - Pre-processing Methods



Contents

- **Scaling and Centering**
 - Unit variance, Pareto, and No scaling; With and without mean-centering
 - Set point centering & Control/Action limit scaling
 - Block-scaling, Double centering, ...
- **Transformation and Expansion of data**
- **Signal Correction and Compression**
 - Orthogonal signal correction, OSC
 - Multiplicative signal correction, MSC
 - Standard normal variate correction, SNV
 - Derivatives
 - Wavelet analysis

Scaling and Centering



Contents

- Mean-centering and scaling to unit variance
- Mean-centering but no scaling
- Block-scaling
- No centering, but scaling
- Set point centering & Control/Action limit scaling
- Double centering (“Correspondence analysis”)
- Scaling procedure in SIMCA

Mean-centering and scaling to unit variance

- The most common scaling type. All variables are scaled to unit variance
- Useful when variables are of different kinds and not directly comparable
- Problems:
 - If a large number of variables have low variation, UV-scaling will blow them up
 - typical example: Spectral data
 - The absolute variation will be lost
 - typical examples: Spectral data and COMFA

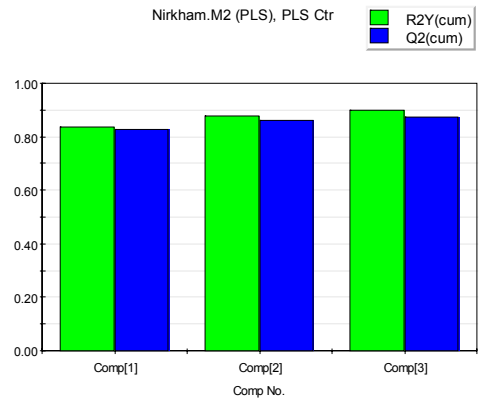
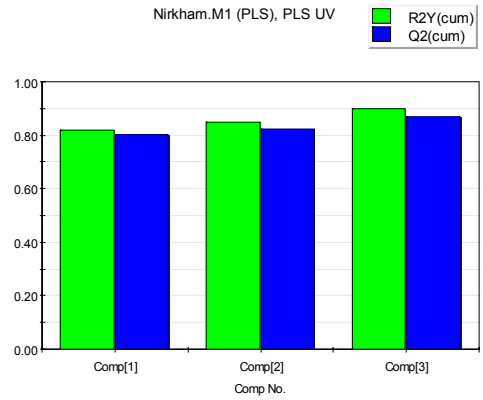
Example: Mean-centering and UV-scaling

- UV-scaling is easiest understood by comparing the results from scaled and unscaled data
- Example: NIRKHAM, using the sugar data as X in PLS
 - with UV-scaling all variables have equal chance to contribute to the model
 - without scaling the importance of Glu is increased
- The SD for Glu is about 6 times larger than the others

Rha_s	Fuc_s	Ara_s	Xyl_s	Man_s	Gal_s	Glu_s
0.84	0.18	1.04	0.92	0.64	0.80	6.79

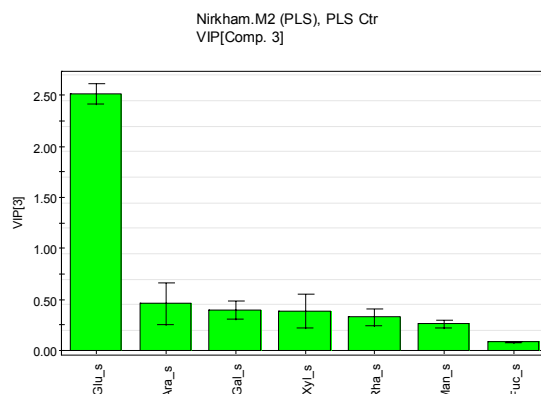
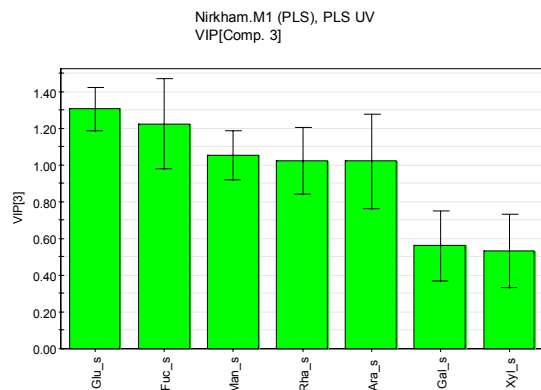
Example: Mean-centering and UV-scaling

- The effect of scaling on R^2 and Q^2 might be minor, but in principle a higher Q^2 is advantageous
- Above: UV-scaling (and mean-centering)
- Below: No scaling (but mean-centering)



Scaling influences variable importance

- UV-scaling, and mean-centering: All variables have had the same chance to influence the model
- No scaling, but mean-centering: Often, the variable with the highest SD will get too much influence



Mean-centering but no scaling

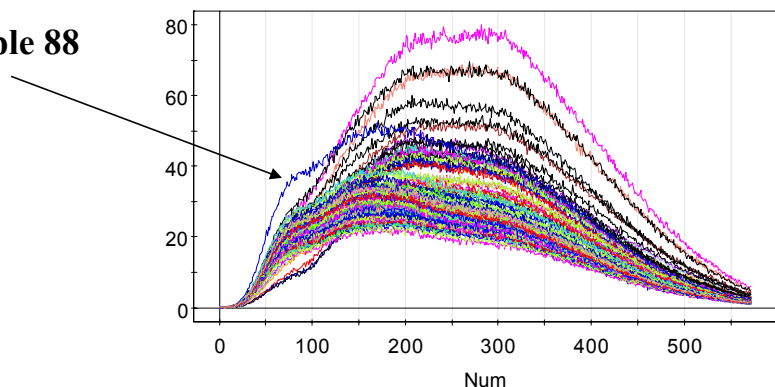
- This kind of pre-treatment is particularly useful when
 - all variables are of the same kind and their numerical size and intrinsic variation carry information
- Examples
 - spectral data, like UV, NIR, fluorescence
 - CoMFA
 - questionnaires

Example: Sugar (Mean-centering but no scaling)

- Data: Fluorescence measurements on white sugar, the final product in the sugar production, dissolved in phosphate buffered distilled water
- 106 samples, 571 X-variables
- Excitation: 240 nm, Emission: 275-560 nm
- Response: Impurity ("ash content")
- Reference: Rasmus Bro, "Håndbog i Multivariabel Kalibrering"

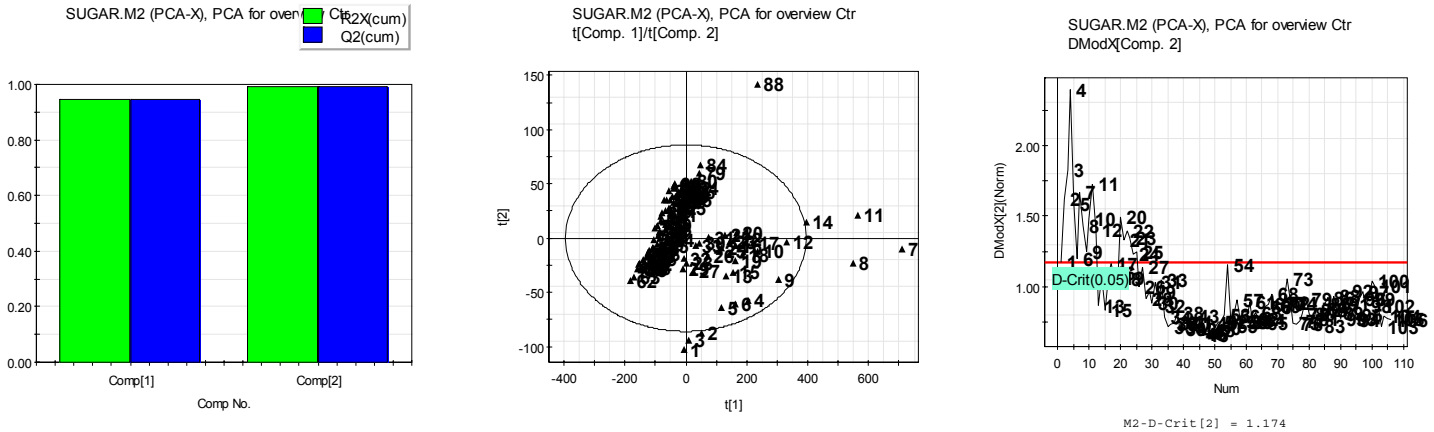
- Plot of spectra
(the X-data):

Sample 88



PCA modelling

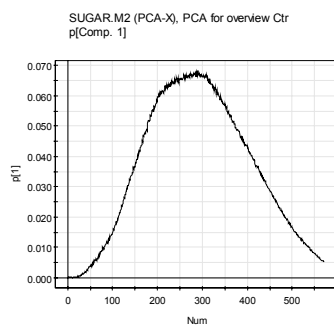
- PCA of unscaled and centered data shows problems in the beginning of the process; stabilisation from around sample 15; sample 88 probable outlier



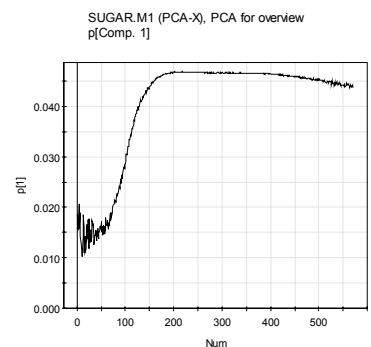
Scaled or unscaled data?

- Two independent spectral contributions are found in data – results of unscaled data are easier to interpret

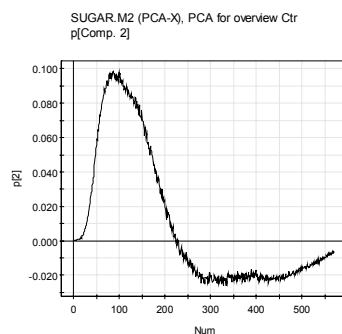
p_1 , unscaled data, has structure and resembles average spectrum



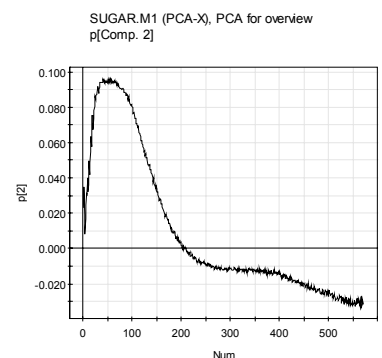
p_1 , scaled data, does not mimic the average spectrum



p_2 , unscaled data, also carries structure, peak-like shape



p_2 , scaled data, with a peak



Summary: No, Pareto, and UV-scaling

- No scaling (but mean-centering): Useful when all variables are expressed in the same unit, such as with spectroscopic data
- UV-scaling (and mean-centering): Useful when variables are of different kinds and not directly comparable
- Pareto scaling (and mean-centering): Intermediate between the extremes of no scaling and UV-scaling. Gives each variable a variance numerically equal to its initial standard deviation instead of unit variance

Block-scaling

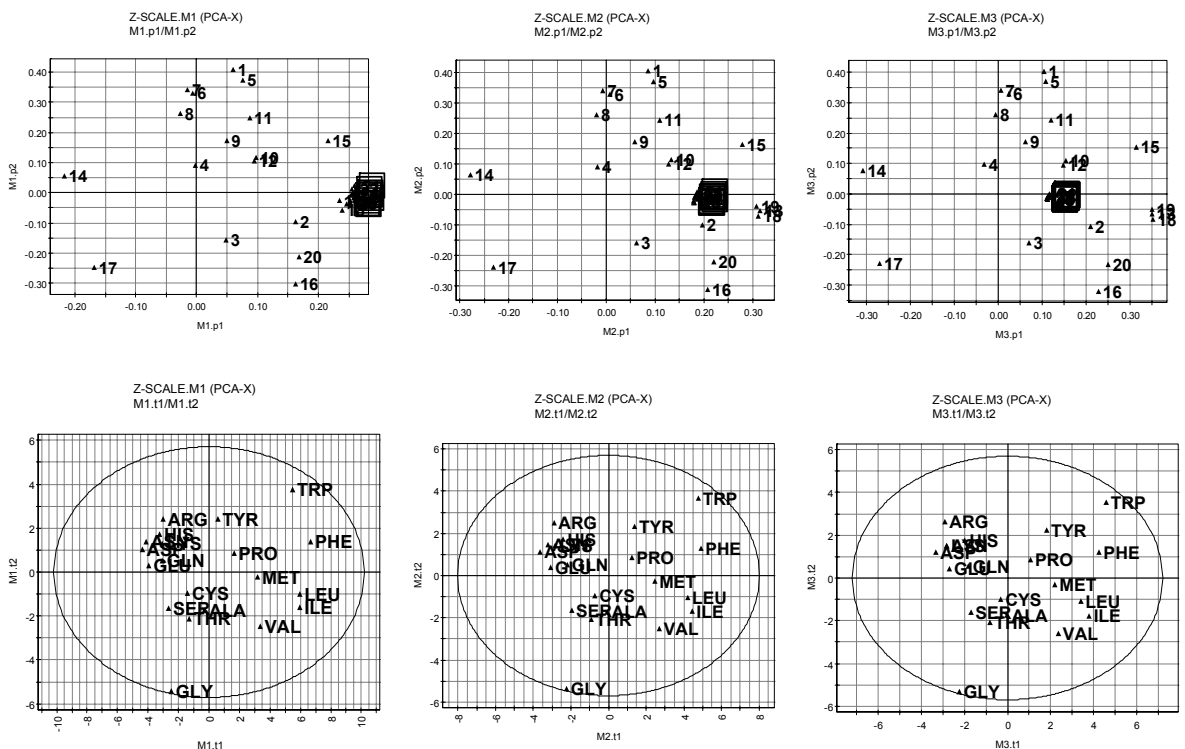
- Block-scaling is used with blocks of different kinds of variables,
Example: Questionnaire about health
 - 10 questions about physical exercise
 - 30 questions about food consumption
 - 5 questions about tobacco and alcohol
 - 3 measured variables: weight, blood pressure and cholesterol
- The three measured variables probably carry more information than the others, but will be masked by the variation of the others
- Problem: How to do block-scaling is both dependent on the number of variables in each block, the importance of the variables, and correlation between the variables

Example: Block-scaling (Z-Scales of amino acids)

- Hard block-scaling: Unit variance for all blocks
 - Multiply each variable weight by $1/(K_{\text{block}})^{0.5}$, where K_{block} denotes the number of variables in a block
- Soft block-scaling: Scale each block to have a variance equal to the square root of the number of variables in that block
 - Multiply each variable weight by $1/(K_{\text{block}})^{0.25}$
 - This is the recommended procedure
- In Z-scale we have
 - 9 HPLC variables $\Rightarrow \text{sqrt}(9)$
 - 20 other single variables $\Rightarrow \text{sqrt}(1)$ per variable

Example: Block-scaling (Amino acids)

- UV-scaling & soft and hard block-scaling

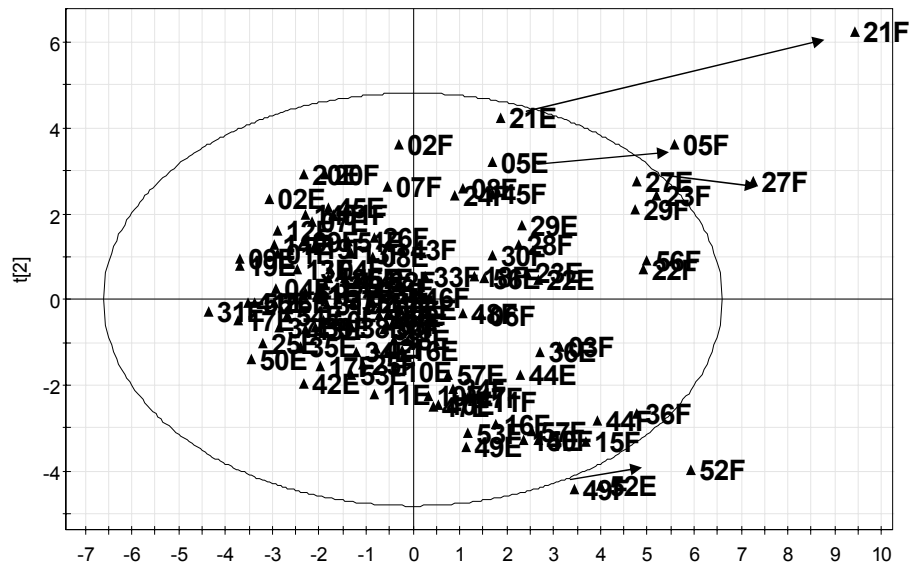


- Score plots do not change much!!

No centering, but scaling

- Why? Sometimes mean-centering will remove the interesting effect
- Example: KROPPAR with paired subjects
- PCA of scaled & centered data
 - Each subject appears twice, before (F) and after (E) treatment
 - One arrow per subject

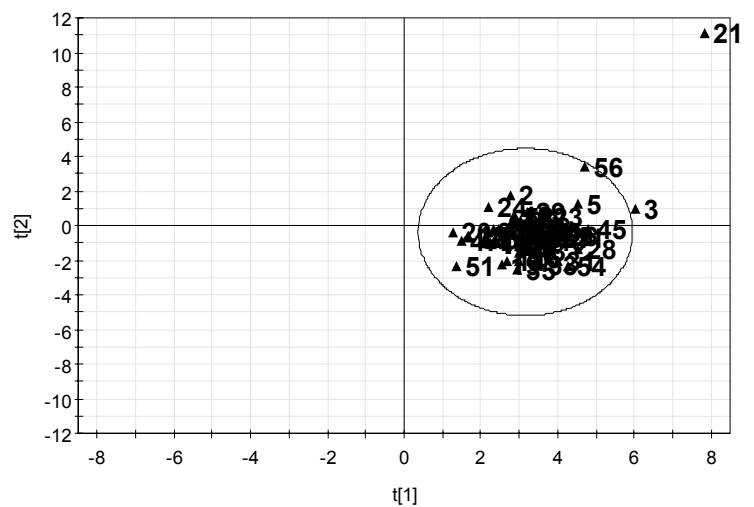
HEALTH.M1 (PCA-X), Overview entire data set
t[Comp. 1]/t[Comp. 2]



No centering, but scaling

- A better approach is to convert the original data-set to a data-set of differences
- Then each subject appears only once
- A PC model can be made with the variables UVN scaled (scaled relative to zero (no treatment effect) and not centered)
- If there were no treatment effect the subjects would appear as a cluster around origin

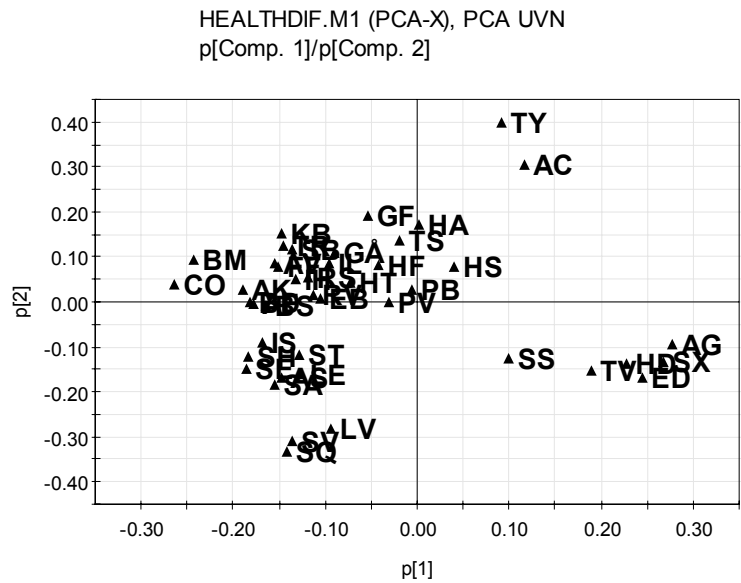
HEALTHDIF.M1 (PCA-X), PCA UVN
t[Comp. 1]/t[Comp. 2]



Subjects have moved to the right along the t_1 -axis.

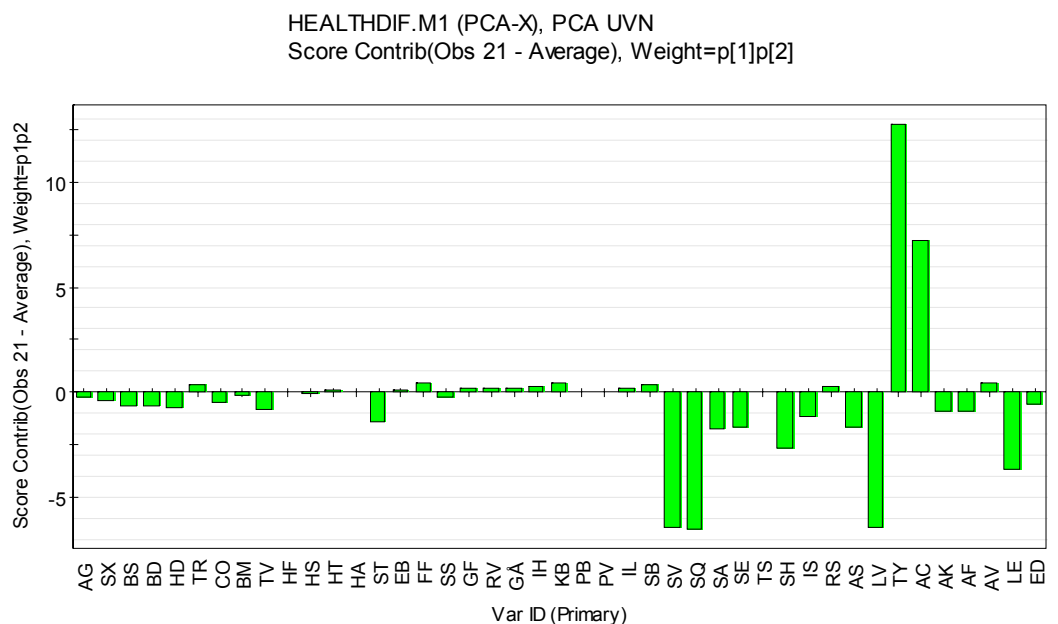
No centering, but scaling

- The loading plot shows that variables like cholesterol (CO) and body-mass-index (BM), have been reduced, and that variables like physical fitness (TV) and HDL blood lipids (HD), have been increased with respect to before and after the treatment



No centering, but scaling

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



No centering, but scaling

- Conclusions:
 - We can see that a reduction in blood pressure is correlated with reduction in body-mass-index (BM), cholesterol (CO), and increase in HDL blood lipids (HD)
 - If the variables were mean-centered, the main effect of the treatment would not be seen in the model plots. Then the model would not explain the changes themselves but only the variability in the changes.

Set point centering & Control/Action limit scaling

- In process modelling an alternative to mean-centering and UV-scaling involves
 - centering with regards to the set-point of the process and
 - scaling according to the limits defined as control and/or action limits.
- Set-point centering allows the process operator to focus on the variability around the set point and not around the average point.
- Control or action limit scaling might produce a more realistic model of the process, especially in the case when a number of variables happen to display a variation that is much smaller than their normal variation range.

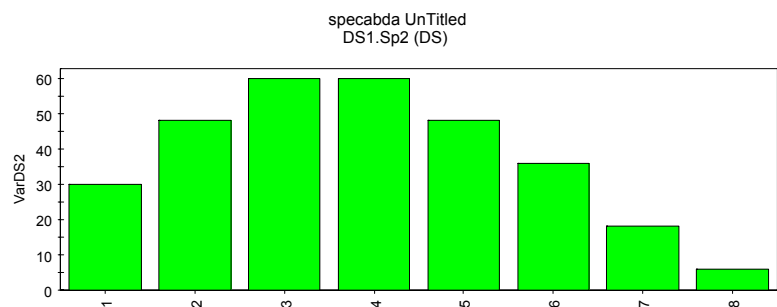
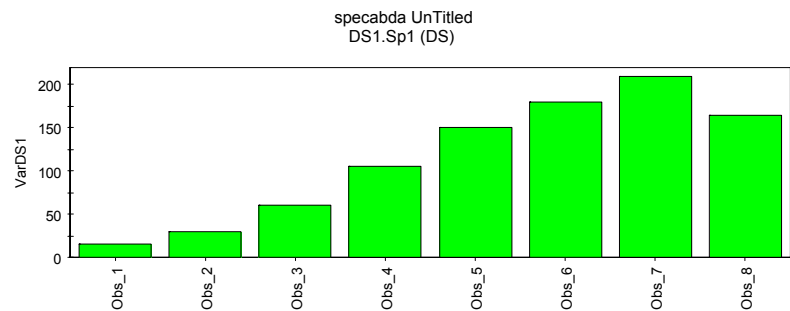
Double centering

- Often called correspondence analysis; CA is similar to PCA with a specific pre-treatment of X (row and column-centering)
- Typical example is to generate a map from a distance matrix

	ATA LANTA	CHI CAGO	DEN VER	HOUS TON	LA	MIA MI	NEW YORK	SAN FRAN
ATLANTA	0	587	1212	701	1936	604	748	2139
CHICAGO	587	0	920	940	1745	1188	713	1858
DENVER	1212	920	0	879	831	1726	1631	949
HOUSTON	701	940	879	0	1374	968	1420	1645
LA	1936	1745	831	1374	0	2339	2451	347
MIAMI	604	1188	1726	968	2339	0	1092	2549
NEWYORK	748	713	1631	1420	2451	1092	0	2571
SAN FRANCISCO	2139	1858	949	1645	347	2594	2571	0

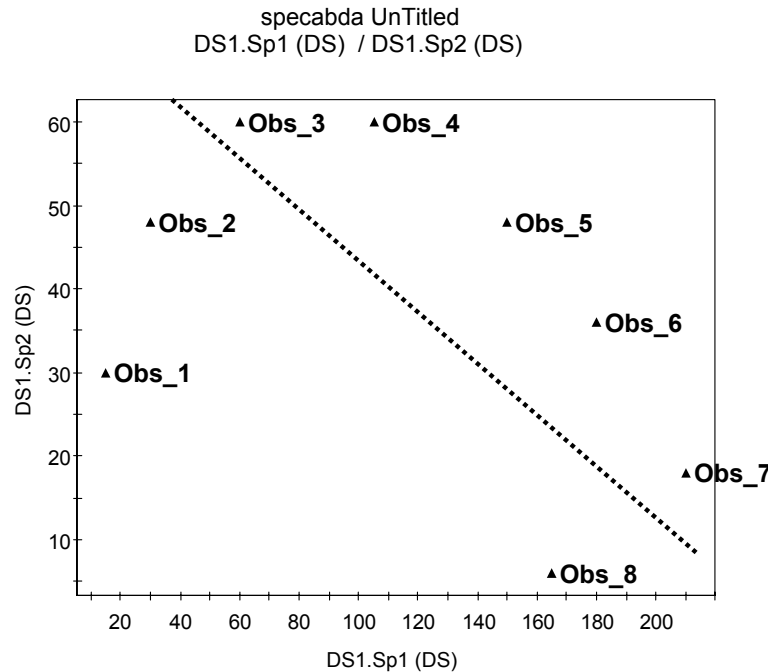
Example of Correspondence Analysis

- CA is mostly used for species abundance data in eco-toxicological monitoring
- Hypothetical example: Typical uni-modal evolution of two species, 1 and 2, for eight measuring stations along a pollution gradient, e.g., outside an offshore oil production site



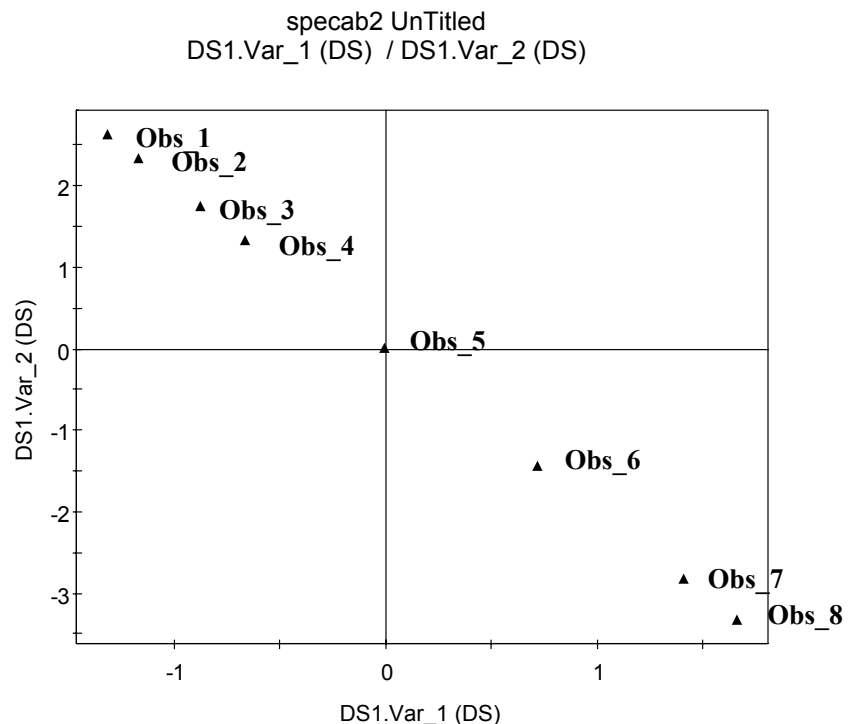
Example of Correspondence Analysis

- Abundances of the two species 1 and 2 plotted against each other for the eight measuring stations
- First PC is represented by the dotted line
- Problem: An inversion of the measuring stations occurs along the concentration gradient
- “Horseshoe effect”



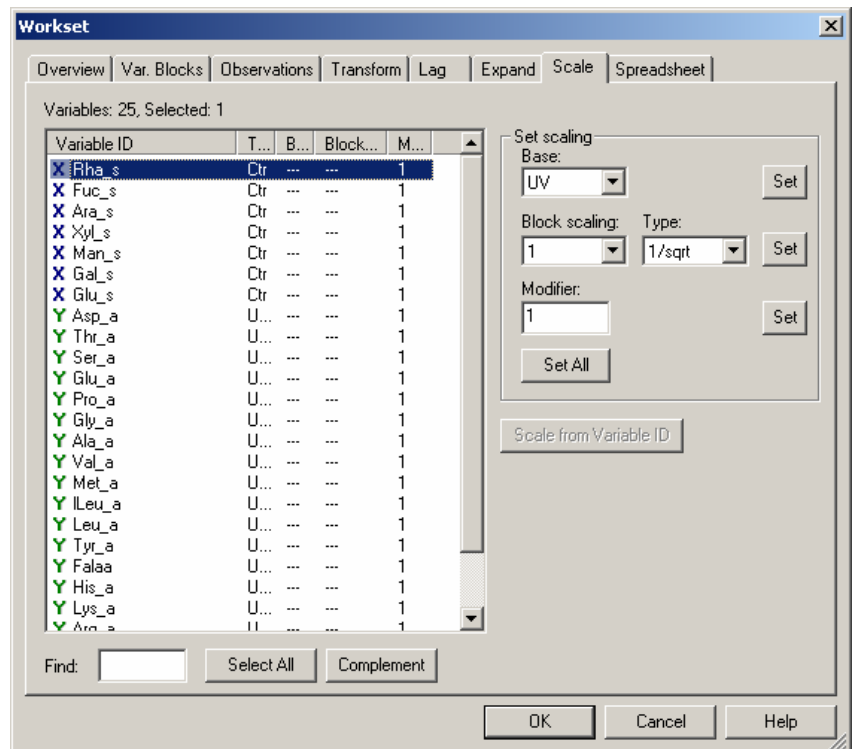
Example of Correspondence Analysis

- Plot of the transformed species counts against each other, which were obtained through CA
- No inversion of sites
- CA is particularly useful for “species abundance data”
- Reference: B. Massart, Ph.D. Thesis, University of Bergen, 1997.



Scaling procedure in SIMCA

- The scaling weight of a variable can be seen as the product of
 - A base weight
 - A block-scaling weight
 - A modifier



Scaling procedure in SIMCA - Base weight

- **UV:** Variable j is centered and scaled to "Unit Variance", i.e., the base weight is computed as $1/sd_j$, with sd_j being the standard deviation of variable j computed around the mean
- **UVN:** Same as UV, but the variable is not centered
- **Par:** Variable j is centered and scaled to "Pareto Variance", i.e., the base weight is computed as $1/\sqrt{sd_j}$. Pareto scaling is in between no scaling and UV scaling and gives the variable a variance equal to its standard deviation instead of 1.0
- **ParN:** Same as Par, but the variable is not centered
- **Ctr:** The variable is mean-centered but not scaled
- **None:** No mean-centering or scaling ($ws = 1$)
- **Freeze:** The variable is centered and the scaling weight of the variable is frozen (will not be re-computed when observations in the work-set change)

Scaling procedure in SIMCA - Block-scaling weight

- Hard block-scaling: $1/\sqrt{K_{\text{block}}}$:
 - This gives the whole block a variance equal to 1
- Soft block-scaling: $1/(\text{4th root}(K_{\text{block}}))$:
 - This gives the whole block a variance equal to the square root of K_{block}
- K_{block} = number of variables in the block

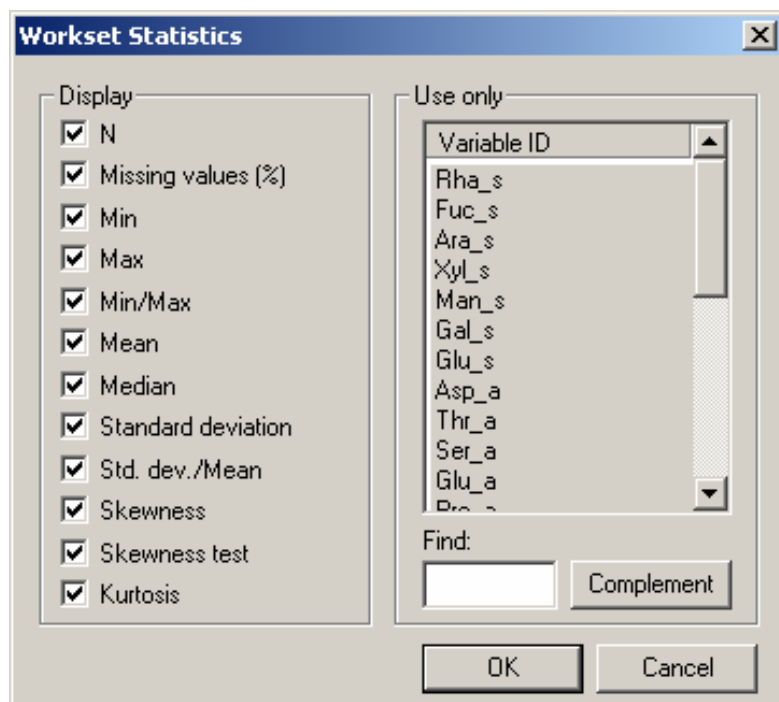
Scaling procedure in SIMCA - Modifier

- A modifier (default = 1) is utilized to scale variables up or down relative to the base scaling weight
- When the variable is not blocked (the block scaling weight is equal to 1) and when the modifier is equal to 1, the base weight is equal to the scaling weight

Transformations

Finding variables to transform in SIMCA

- Use Workset Statistics to create a table with min/max and skewness, or skewness test
- Coupling of min/max and skewness good for finding the need for transformation



Min/Max-ratio

- Look at ratio min/max.
 - When $|\text{Min/Max}| < 0.1$, SIMCA-P issues a warning (red colour)
- Make a histogram of a suspicious variable

Skewness and Skewness test

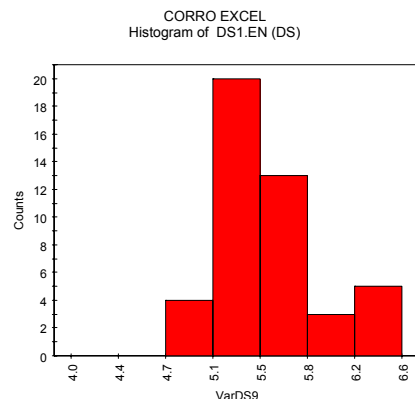
- SKEWNESS (- measures lack of symmetry of a distribution)
$$= N^{1/2} \Sigma(x_i - \bar{x})^3 / [\Sigma(x_i - \bar{x})^2]^{3/2}$$
- Skewness Test (- skewness statistic weighted for N)
$$= \text{Skewness} / (\text{sqrt}(6 * N * (N-1) / ((N-2) * (N+1) * (N+3))))$$
- When ABS (Skewness Test) ≥ 2 , SIMCA-P issues a warning (red colour)

When to transform and how - Rules of thumb

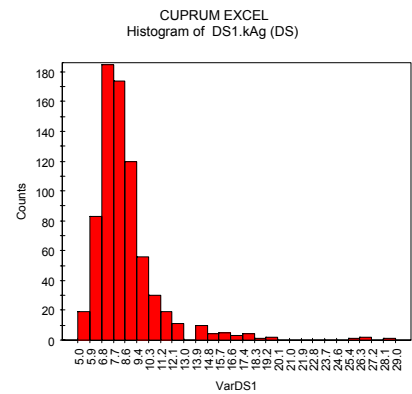
- Why transform?
 - Makes variables fairly symmetrically distributed
 - Gives simpler models with fewer dimensions
- Which transform?
 - A natural zero and min/max < 0.1, use log
 - Alternative to log is 4th root
- Percentages
 - < 15% use log
 - > 85% use log (100-y)
 - With values from 5-95 use logit
 - logit = $\log [(0+y)/(100-y)]$
- Known theory
 - Hydrogen ion concentration, use log (pH)
 - Size of rust-spot in mm², use square root
- How to evaluate effects of a transformation?
 - Answer: (R^2) Q^2
 - The Q^2 value should increase
 - The R^2 value should not decrease too much
 - Data and residuals should get more normally distributed

Distributions and proposed transformations

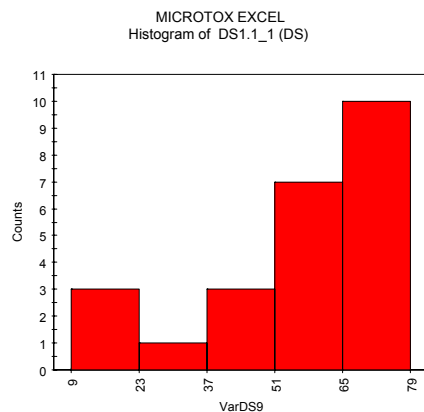
1. No Transf.
2. Log
3. NegLog
4. Logit



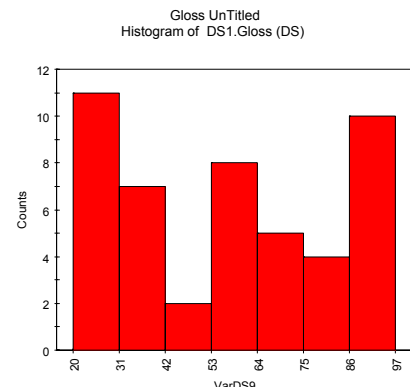
1



2



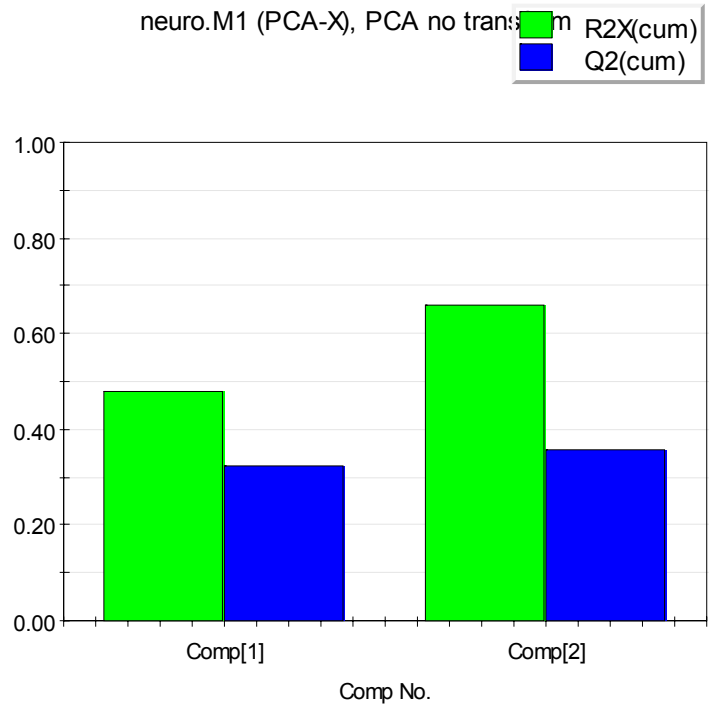
3



4

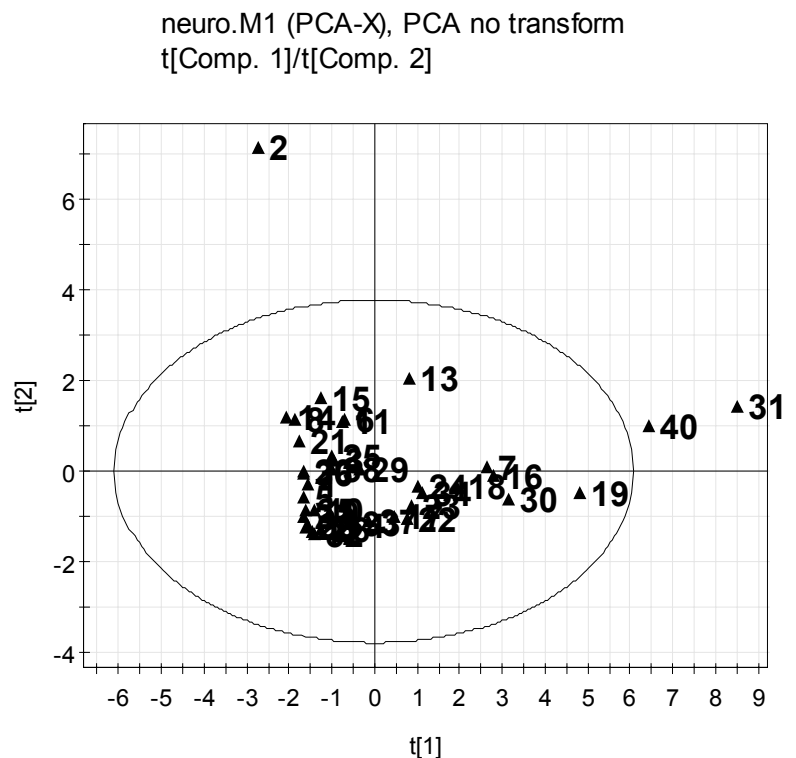
Example: Multivariate Biological Profiling of 40 Neuroleptics

- Neuroleptic compounds exhibit a variety of pharmacological activities
- 40 Neuroleptics were studied in twelve pharmacological tests in rat.
- Is it possible to classify these substances according to some kind of activity profile?
- PCA for overview of the 40x12 data table gave:



Example, 40 Neuroleptics

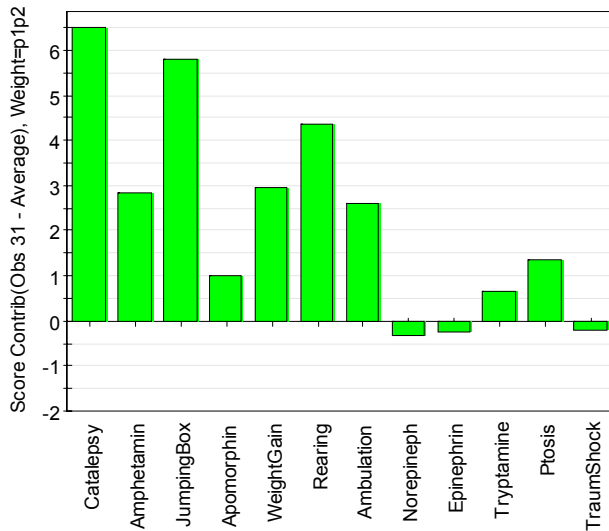
- Why do compounds # 2, 31 and 40 deviate so much?



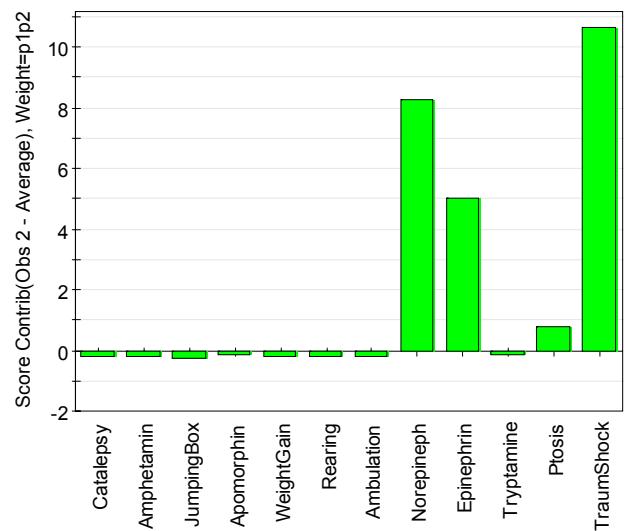
Contribution plots for observations 31 & 2

- Obs 31 is higher in variables 1-3, and 5-7 than the average obs.
- Obs 2 differs from the average in variables 8, 9, and 12.

neuro.M1 (PCA-X), PCA no transform
Score Contrib(Obs 31 - Average), Weight=p[1]p[2]



neuro.M1 (PCA-X), PCA no transform
Score Contrib(Obs 2 - Average), Weight=p[1]p[2]



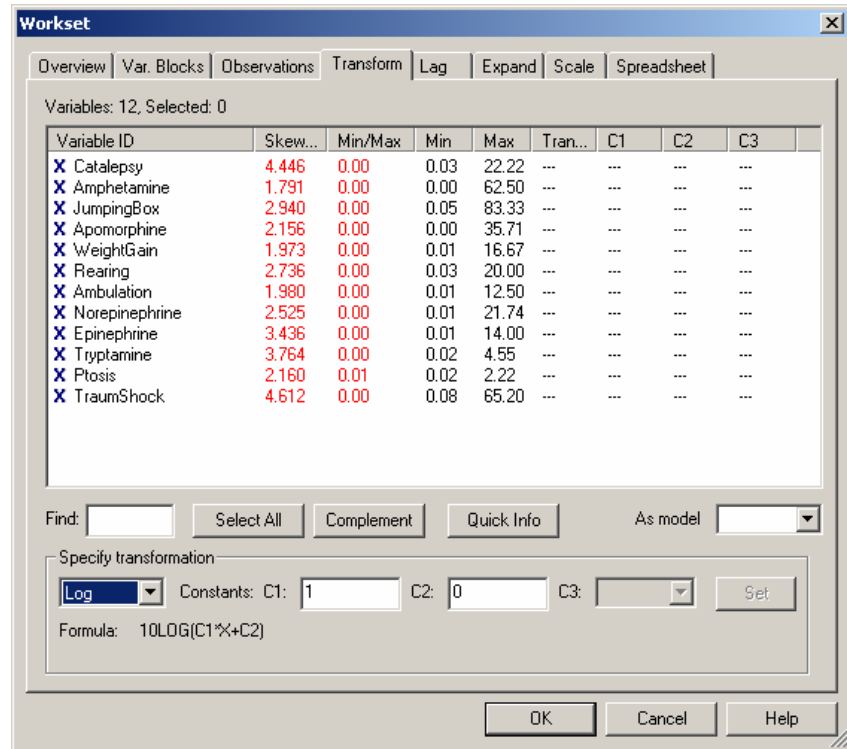
Example, 40 Neuroleptics

- Abs (Skewness Test) > 1.5 is an indication that a variable ought to be transformed. A related criterion is if | Min/Max | is < 0.1.
- All 12 variables are candidates for log-transformation

Workset Statistics - M1													
	1	2	3	4	5	6	7	8	9	10	11	12	13
	N	% MisVal	Min	Max	Min/Max	Mean	Median	Std.dev	Std.dev/m	Skewness	Skew.Test	Kurtosis	
2	Catalepsy	40	0	0.03	22.22	0.001350	1.887	0.56	3.72759	1.97541	4.44557	11.8934	23.3999
3	Amphetamine	40	0	0	62.5	0	10.656	3.005	16.0941	1.51033	1.79057	4.79041	2.36516
4	JumpingBox	40	0	0.05	83.33	0.000600	9.71	3.28	16.3075	1.67945	2.93997	7.86544	10.2751
5	Apomorphine	40	0	0	35.71	0	5.60378	0.515	9.70068	1.7311	2.15584	5.76763	8.91144
6	WeightGain	40	0	0.01	16.67	0.000599	2.80975	1.04	4.06511	1.44679	1.97336	5.27943	8.40059
7	Rearing	40	0	0.03	20	0.0015	2.7425	0.955	4.45391	1.62403	2.73581	7.31924	8.02778
8	Ambulation	40	0	0.01	12.5	0.0008	2.192	0.65	3.24609	1.48088	1.97982	5.29672	8.24041
9	Norepinephrine	40	0	0.01	21.74	0.000459	3.09475	1.42	4.44535	1.43642	2.52492	6.75504	7.64344
10	Epinephrine	40	0	0.01	14	0.000714	1.48325	0.61	2.72917	1.84	3.4364	9.19357	12.9585
11	Tryptamine	40	0	0.02	4.55	0.004395	0.5035	0.2	0.787666	1.56438	3.7642	10.0705	17.8901
12	Ptosis	40	0	0.02	2.22	0.009009	0.4285	0.235	0.439181	1.02493	2.16016	5.77918	6.21629
13	TraumShock	40	0	0.08	65.2	0.001226	4.92425	1.43	10.9243	2.21847	4.61178	12.3381	24.6267

Transformations in SIMCA

- Mark variables and select transform
- Add constants if necessary



Results after transformation

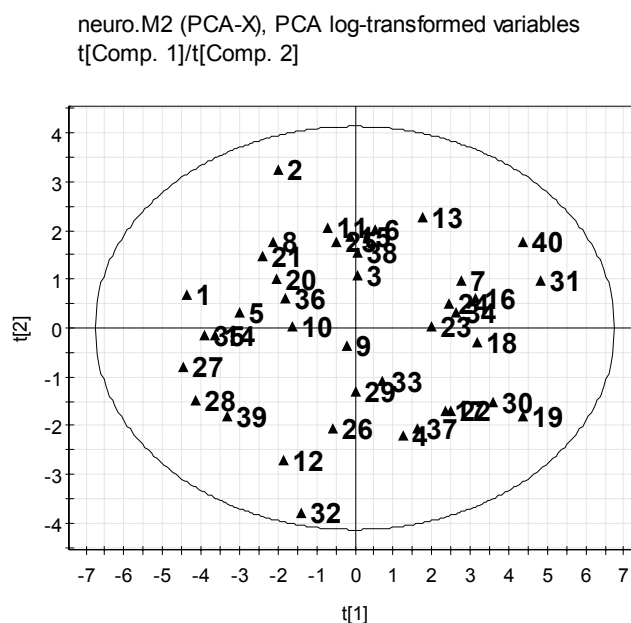
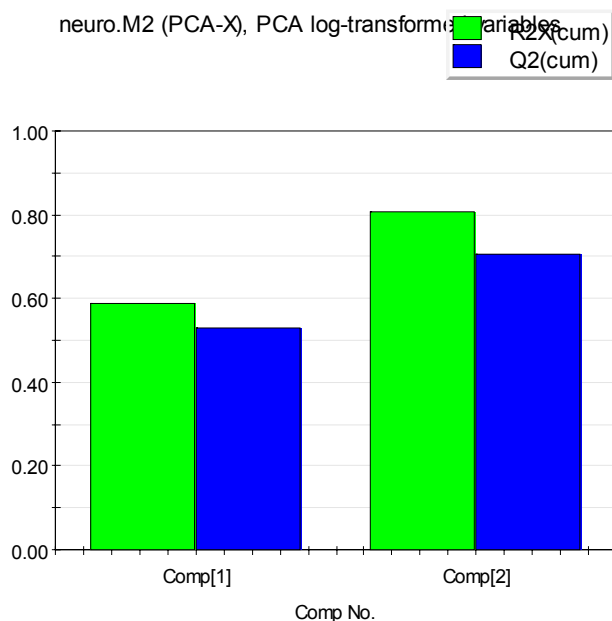
- Transforms used:
 - log (vars 1, 3, 5-12)
 - log y + 1 (vars 2, 4)

Workset Statistics - M2

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	N		% MisVal	Min	Max	Min/Max	Mean	Median	Std.dev	Std.dev/n	Skewness	Skew.Test	Kurtosis
2	Catalepsy	40	0	-1.52288	1.34674	-1.13079	-0.258552	-0.251812	0.741242	-2.86689	-0.030603	-0.0818735	-0.769237
3	Amphetamine	40	0	0	1.80277	0	0.686217	0.599905	0.582652	0.849078	0.469875	1.25708	1.13231
4	JumpingBox	40	0	-1.30103	1.9208	-0.677337	0.419039	0.515823	0.818267	1.95272	-0.29614	-0.792277	-0.578413
5	Apomorphine	40	0	0	1.56478	0	0.468314	0.180391	0.529298	1.13022	0.779487	2.0854	-0.780845
6	WeightGain	40	0	-2	1.22194	-1.63675	-0.064964	0.016953	0.780054	-12.0075	-0.329071	-0.880378	-0.408903
7	Rearing	40	0	-1.52288	1.30103	-1.17052	-0.095497	-0.020475	0.781992	-8.1886	-0.230661	-0.617099	-0.736007
8	Ambulation	40	0	-2	1.09691	-1.8233	-0.187071	-0.187292	0.787604	-4.2102	-0.303774	-0.812702	-0.463054
9	Norepinephrine	40	0	-2	1.33726	-1.4956	-0.012001	0.152278	0.801166	-66.7558	-0.484565	-1.29638	-0.481615
10	Epinephrine	40	0	-2	1.14613	-1.74501	-0.408607	-0.214904	0.82223	-2.01228	-0.263051	-0.703754	-0.787664
11	Tryptamine	40	0	-1.69897	0.658011	-2.58198	-0.646977	-0.69897	0.565803	-0.874534	0.158972	0.425305	-0.677251
12	Ptosis	40	0	-1.69897	0.346353	-4.90531	-0.575115	-0.629815	0.456421	-0.793618	-0.30598	-0.818603	-0.098954
13	TraumShock	40	0	-1.09691	1.81425	-0.604605	0.129671	0.155336	0.71096	5.4828	0.260262	0.696291	-0.526605

Model results after log-transformation

- The deviating compounds have "disappeared"! The model is better.



Transformations - Summary

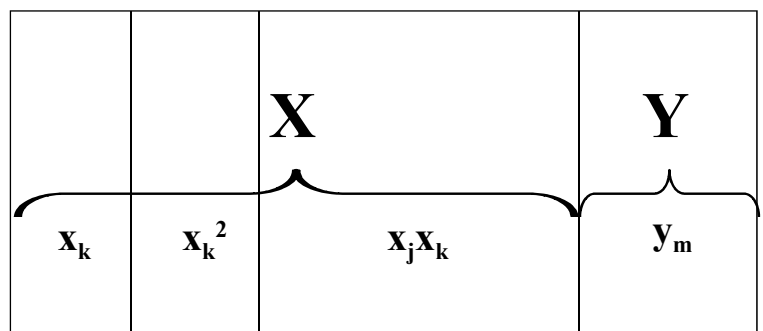
- The NEURO application shows that by carefully selecting a variable transformation data may be made more approximately normal.
- This, in turn, will enhance the efficiency of the data analysis.
- In regression analysis, the benefits of a *response* transformation are:
 - (i) a simplified response function by linearization of a non-linear response-factor relationship,
 - (ii) a stabilized variance of the residuals, and
 - (iii) a distribution of the residuals that is more nearly normal, which sometimes implies that outliers are eliminated.

Expansions



Expanding the X-matrix

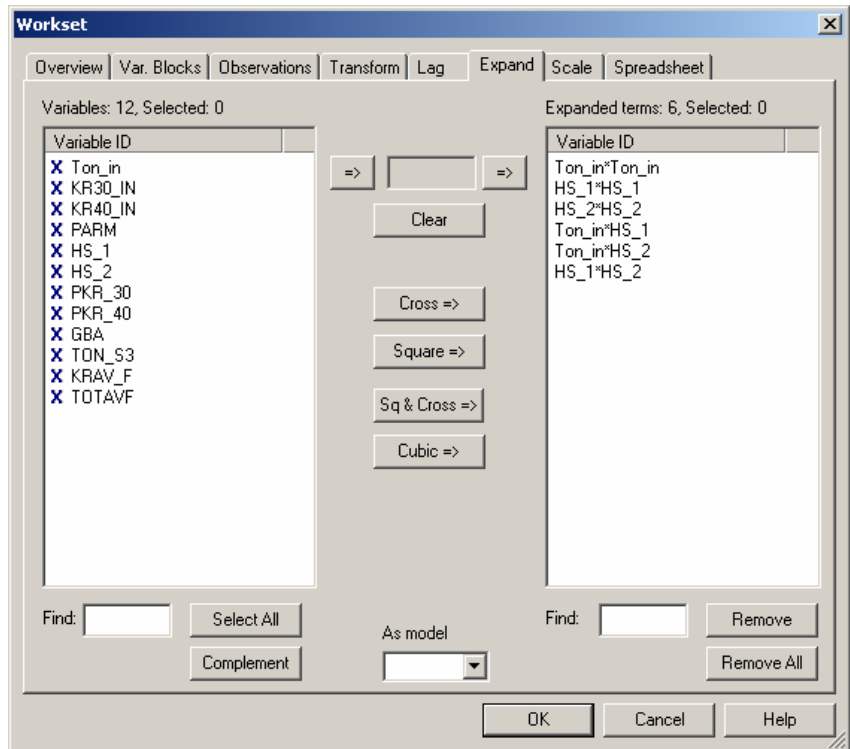
- In SIMCA, new variables can be constructed with the Expand function
- If a non-linear relationship is expected - expand with quadratic terms
- Do NOT expand with cross-terms unless supported by a design
- Do NOT expand simply to get the best fit (highest R^2)



- Example: Expanding 50 variables with square and cross terms gives $50 + 50 + 50 \cdot 49 / 2 = 1325$ variables \Rightarrow “redundancy problem” and the cross-terms dominate

Expansion in SIMCA

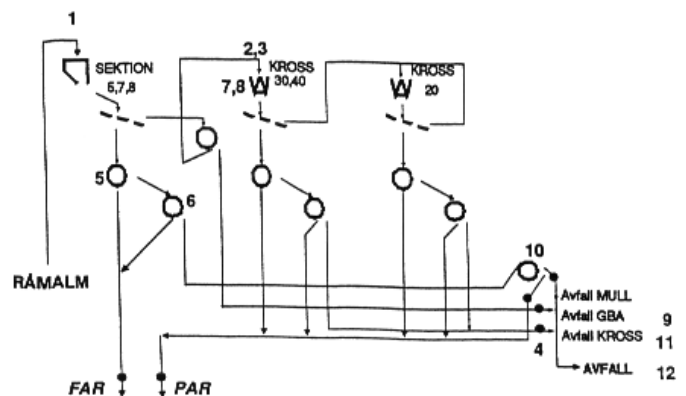
- Linear and interaction models are used for screening designs
- Quadratic and cubic models are employed in RSM phase
- Analyze data and refine model by deleting the unimportant expansion-terms



Example - Modelling the process output (SOVR)

- The Data - 18 variables, 572 observations
 - 3 manipulated process (input) variables
 - 9 intermediate process (input) variables
 - 6 output variables
 - X-data were from process log (minutes)

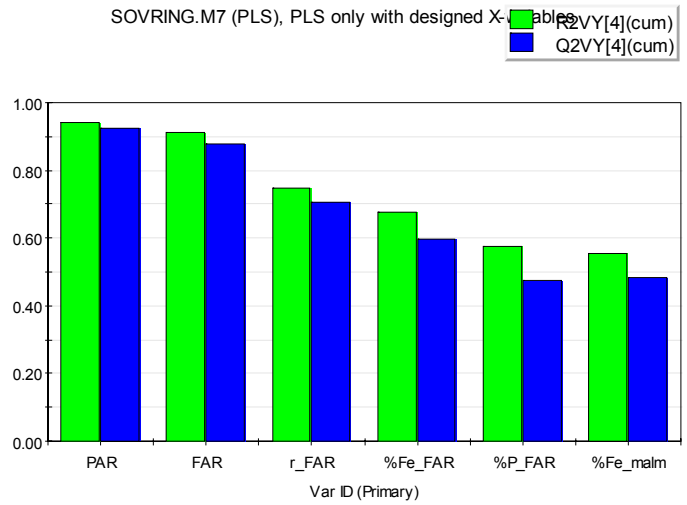
- DOE was used
 - A CCC design in two levels
 - 14 runs + 3 centre points



- The output (or quality) variables were sampled and analysed in the laboratory once for each design setting
 - Each output measurement was then preserved for 20 minutes to capture normal process variation (variations in the X-block)

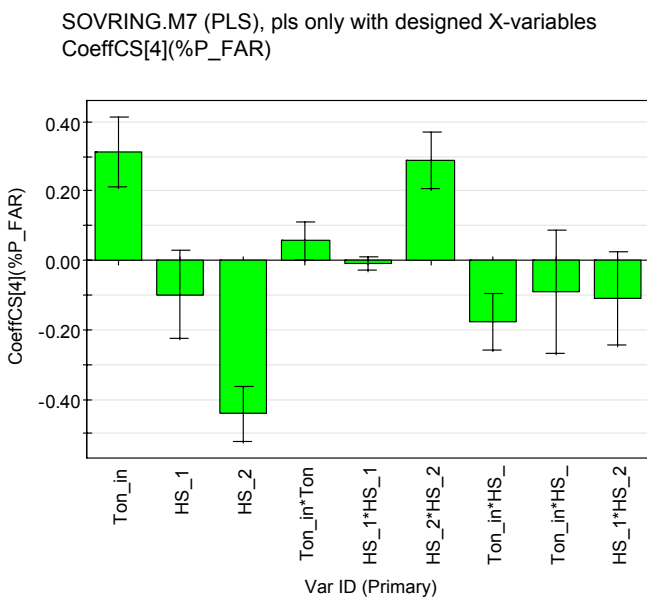
SOVR - Fitting the PLS-model

- A PLS-model was calculated with only the 3 manipulated input variables
 - 3 manipulated input variables
 - 6 interaction & square terms
 - 6 output variables
- 4 significant PLS-components
 - $R^2Y(\text{cum}) = 0.74$, $Q^2(\text{cum}) = 0.68$

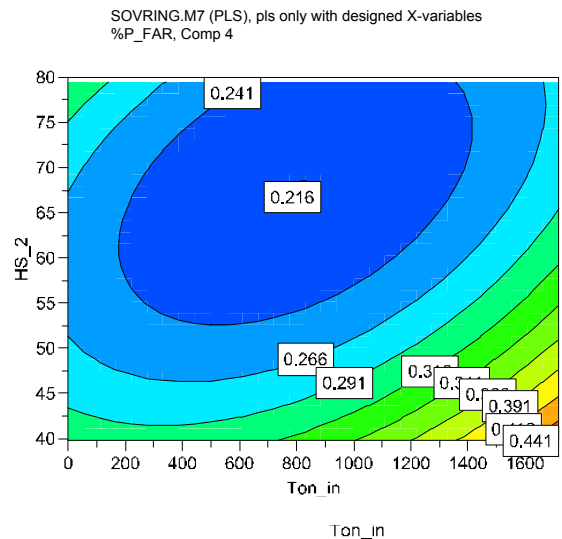


SOVR - Interpreting the PLS-model

- The coefficient plot for %P_FAR shows how the X-s influence Y



- With the contour plot it is easier to interpret the effects of combined factors
 - here HS_2 and Ton_in (HS_1 kept at its centre value)



For each Ton_in, there is an optimal HS_2

SOVR - Summary

- For prediction and monitoring purposes use also the non-designed X's
 - The online measurements enable online predictions
 - The correlation structure stabilises the model and the predictions
 - For early fault detection it is also better to use more process parameters in the model, as more types of problems can be detected
- One interesting output to make online predictions for is the Iron content in the incoming ore (%Fe_malm)
 - Very difficult to get representative samples for analysis
 - When %Fe gets too low in the incoming ore, the process engineer must inform the miners

Signal Correction and Compression

Contents

- Introduction
- Signal correction
 - de-noising with information scaling
 - multiplicative signal correction
 - standard normal variate correction
 - orthogonal signal correction
- Signal compression
 - wavelet analysis
 - steps in data compression
 - compression of rows
 - compression of columns

Introduction to signal correction

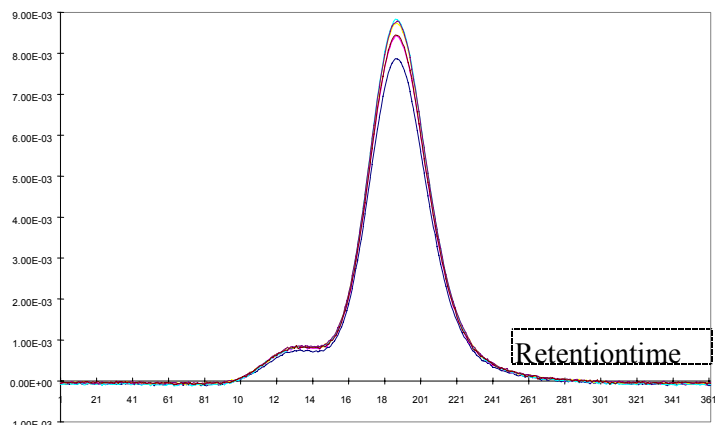
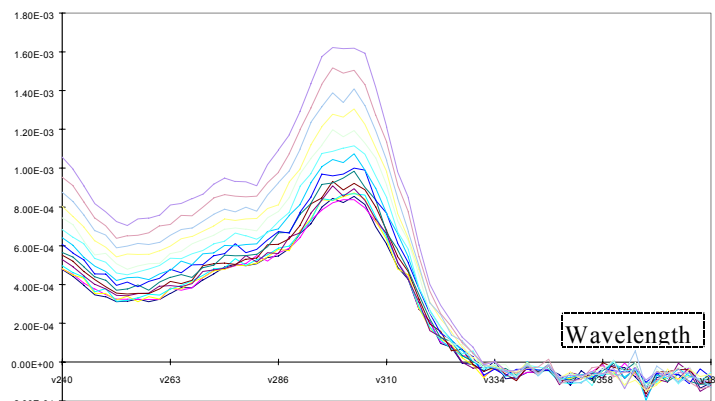
- The first step in multivariate calibration is often to pre-process spectral data to remove undesired systematic variation
 - base-line variation
 - multiplicative scatter effects
 - spectral regions of low information content
- Signal correction methods are interpretable as different cases of filtering
- A filter is a mathematical function through which a signal is passed to get “improved properties”
- Meaning of “improved properties” vary from case to case
 - signal should in some way be more pleasing to the eye
 - base-line removal
 - enhancement of predictive power of Y
 - ...

De-noising with information scaling

- Example: C3D7000
- UV-spectra (diode array) from an HPLC application in pharmaceutical industry
- Each row in the data table is a spectrum from 240-380 nm (studied in the loadings)
- Each column in the data table is comparable to a traditional chromatogram (seen in the scores)

Example C3D700 - Plots of raw data

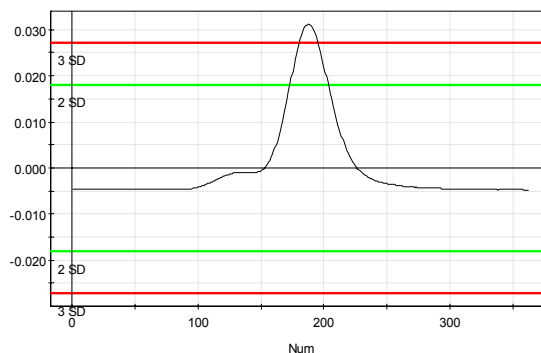
- High wavelengths (>340nm), variables 55-65, contain noise
- First 80 spectra contain no information



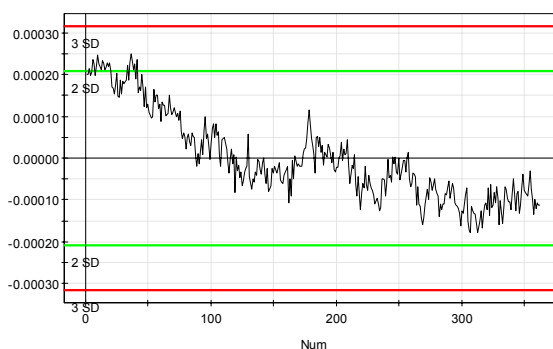
Example C3D700 - PCA

- PC-scores summarize HPLC chromatograms
- With all variables unscaled, t_1 is easy to understand, but t_2 is not smooth

C3d70000.M1 (PCA-X), pca ctr
t[Comp. 1]



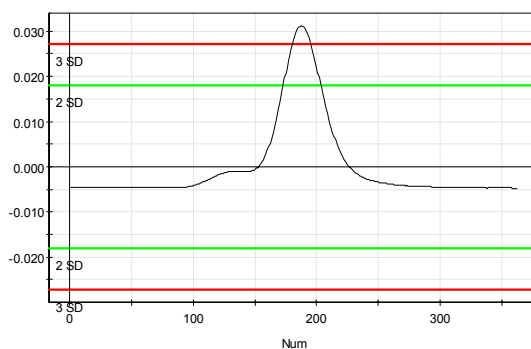
C3d70000.M1 (PCA-X), pca ctr
t[Comp. 2]



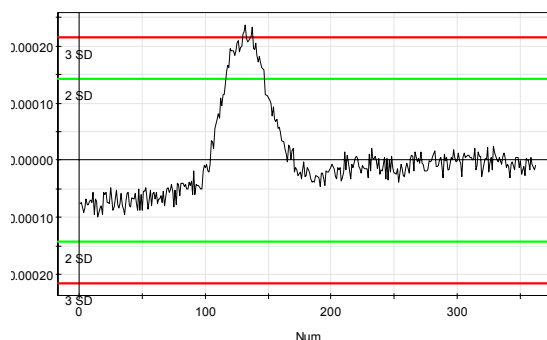
Example C3D700 - PCA with modified scaling weights

- The situation can be improved. By eliminating the noisy variables (zeroing); t_2 becomes more structured

C3d70000.M2 (PCA-X), as M1 but with zeroing
t[Comp. 1]



C3d70000.M2 (PCA-X), as M1 but with zeroing
t[Comp. 2]



Information scaling - Summary

- Example C3D700 demonstrates peak resolution of overlapping peaks
- Risk of information scaling: It is very easy to fulfill ones expectations; you can obtain any result desired
- Suggested procedure: use an external validation set
- Correctly used - with a cautious and careful attitude - information scaling can enhance modelling results

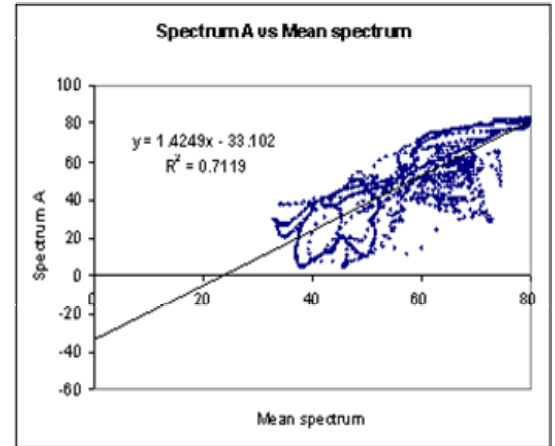
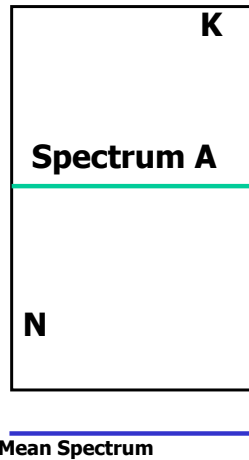
Signal correction techniques in SIMCA

- Signal correction can be used to pre-process data
 - Problem 1: risk of removal of variation from X that correlate with Y
 - Problem 2: risk for over-training of model
- Spectral filters available in SIMCA
 - Multiplicative Signal Correction, MSC
 - Standard Normal Variate Correction, SNV
 - Orthogonal Signal Correction, OSC
 - First and second order derivation
 - Wavelet Compression

Spectral filters in SIMCA

- **MSC:** Each digitised spectrum (x_i' , row-vector in X) is regressed against the mean spectrum (m):

$$x_{ik} = a_i + b_i m_k + e_{ik}$$

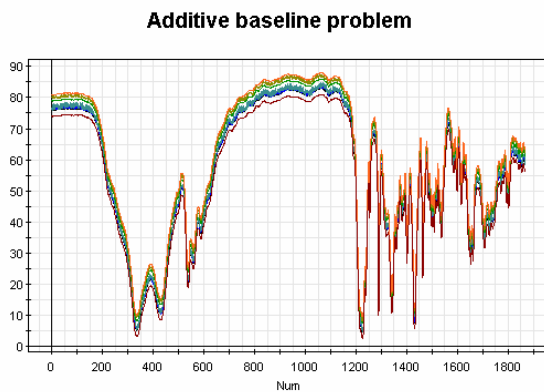


- From each spectrum one subtracts the intercept (a_i) and divides by the slope (b_i):

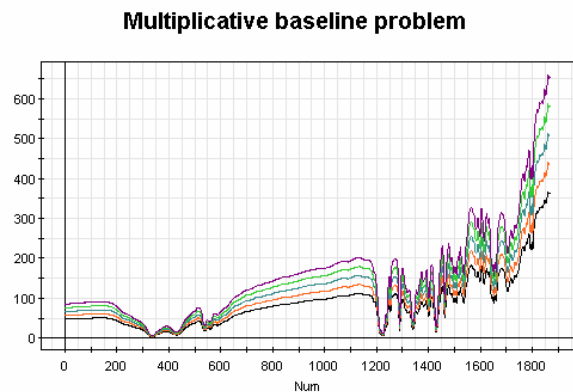
$$x_{i,corr}' = (x_i' - a_i) / b_i$$

Spectral filters in SIMCA

- **SNV:** Same mathematical form as MSC
 - Parameters a_i and b_i are calculated as the average and standard deviation of the i^{th} row of X ; Corresponds to row-centering and normalisation



SIMCA-P+ 10.0 - 04/09/02 15:53:52



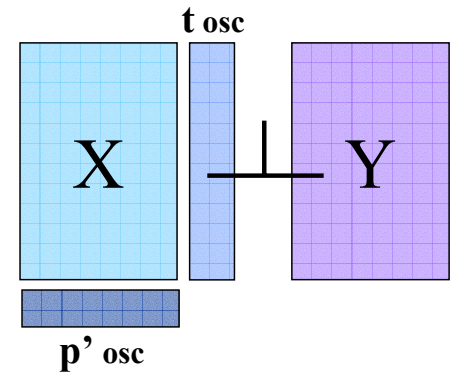
SIMCA-P+ 10.0 - 04/09/02 16:44:11

MSC and SNV are baseline corrections which remove additive and/or multiplicative effects!

Spectral filters in SIMCA

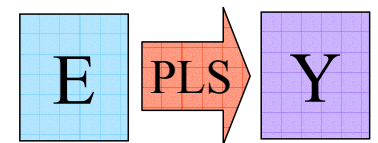
- Orthogonal Signal Correction, OSC

- Calculate first PC of $X \Leftrightarrow$ score t
- Orthogonalise t with regards to Y
- $t_{OSC} = (1 - Y(Y'Y)^{-1}Y')t$
- Some NIPALS steps to give weights (w^*), and updates of t_{OSC} and t , until convergence
- Subtract correction $E = X - t_{OSC}p'$
- One or two OSC components recommended
- Work with one response at a time



$$E = X - t_{OSC} * p'$$

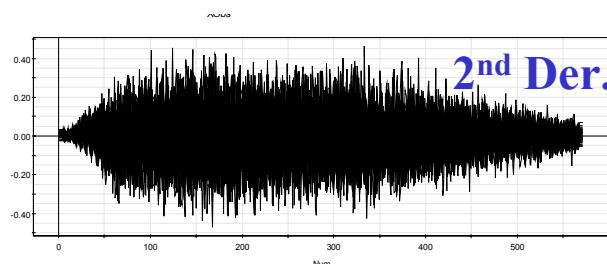
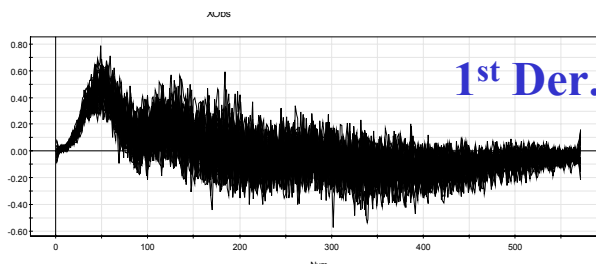
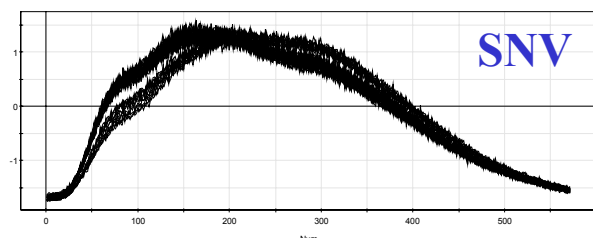
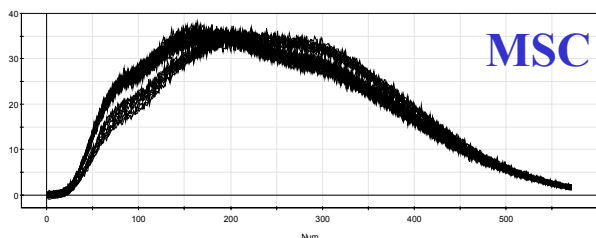
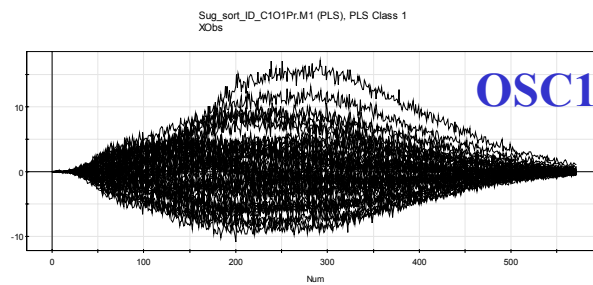
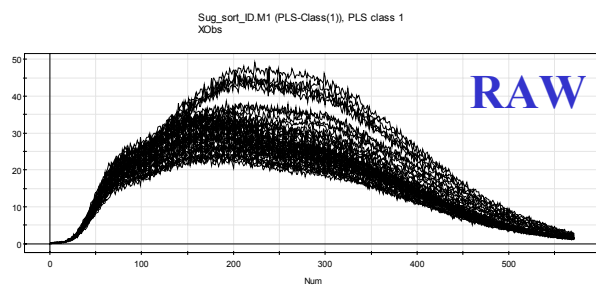
- OSC uses information in Y to construct a filter of X
- MSC and SNV work only with X , and may remove predictive information from X



Spectral filters in SIMCA

- 1st derivative spectrum
 - Provides the slope at each point of the original spectrum
 - Has peaks where the original spectrum has maximum slope and it crosses zero where the original has peaks
 - Removes additive baseline ("offset")
- 2nd derivative spectrum
 - Measures curvature at each point in the original spectrum.
 - Is more similar to the original spectrum and has peaks approximately as the original spectrum, albeit with an inverse configuration
 - Removes a linear baseline
- Problem: May reduce the signal and increase the noise \Rightarrow noisy spectra
- Savitsky and Golay (SG) smoothing
 - SG-derivatives are based on fitting a low degree polynomial (quadratic or cubic degree) piece-wise to the data, followed by calculating the first and second derivatives

Example: Effects of signal correction (SUGAR)



Introduction to signal compression

- Signal compression can reduce computational time and data storage, but also gives de-noising and smoothing
- In SIMCA, wavelet analysis is available
 - for row-wise compression (of spectral data)
 - for column-wise compression (of time-series data)
- User has to select (1) compression technique and (2) wavelet function and order

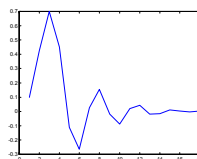
Introduction to wavelet analysis

- Wavelets look like small oscillating waves
- Wavelet analysis produces a linear transformation of the data
- The wavelet transform uses a mother wavelet (a basis function) with a certain “scale” (width of analyzing function window) to investigate the properties of a signal
- The mother wavelet is stretched or compressed. A narrow wavelet is used for detecting sharp signal features (high frequency) and a wider wavelet is used for uncovering more general signal properties (low frequency)

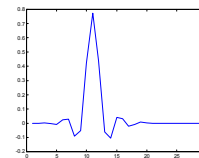
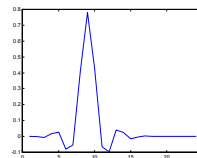
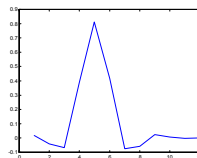
Wavelet filters

- The shape of the filter depends on type and order

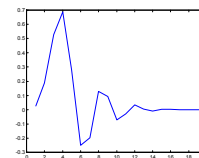
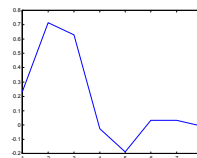
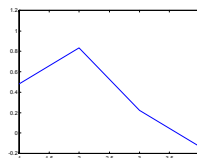
Beylkin
no order



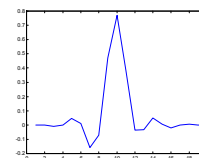
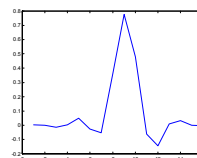
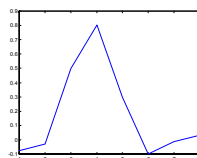
Coiflet
order 2, 4, and 5



Daubechies
order 4, 8, and 20



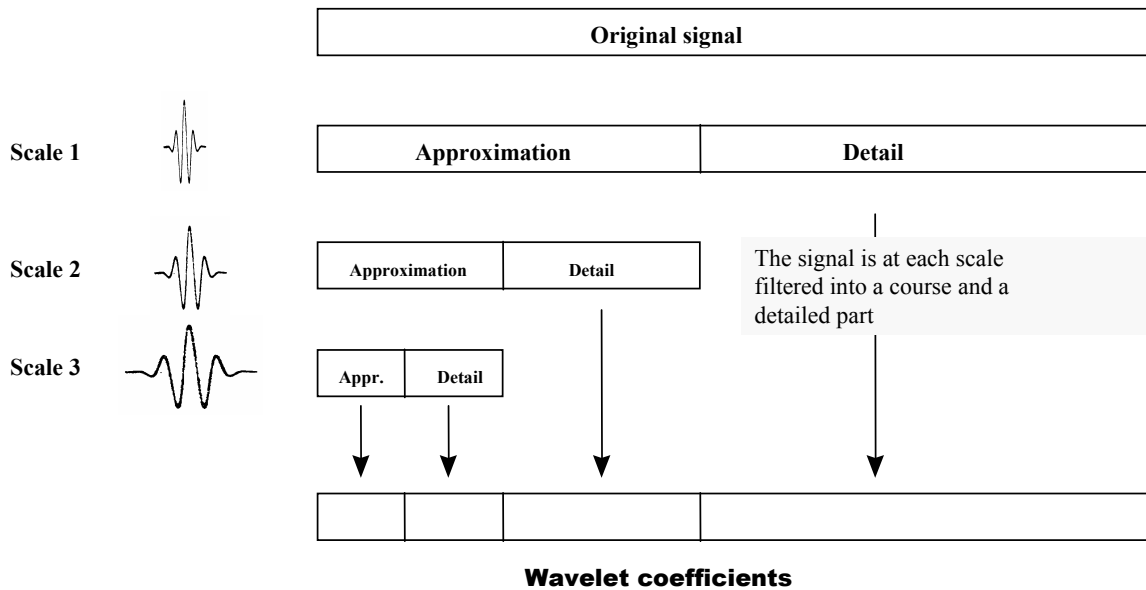
Symmlet
order 4, 8, and 10



Example of signal compression

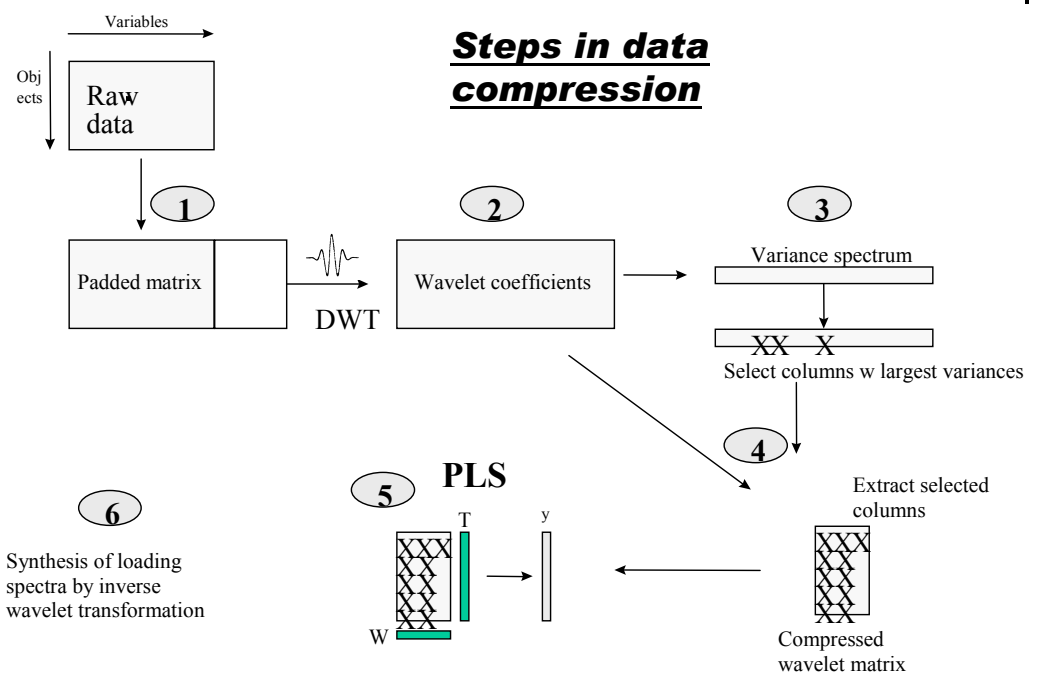
- Multi-resolution analysis, MRA, is a fast compression algorithm

Multiresolution analysis



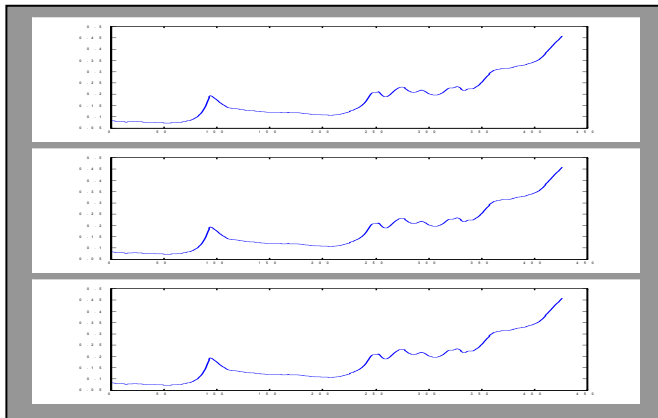
Wavelet compression - row-wise

The discrete wavelet transform (DWT) with MRA is a good approach to signal compression of spectral data

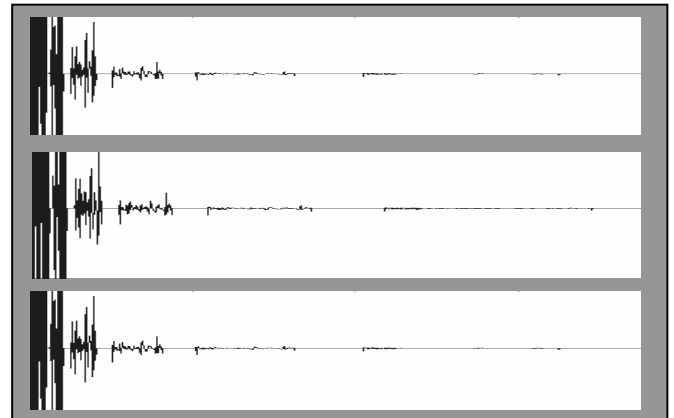


Each observation (spectrum) is wavelet transformed

Original data matrix

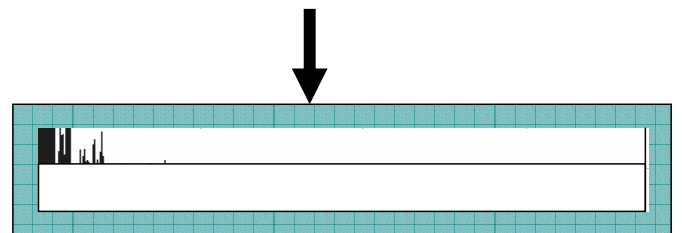


Wavelet data matrix



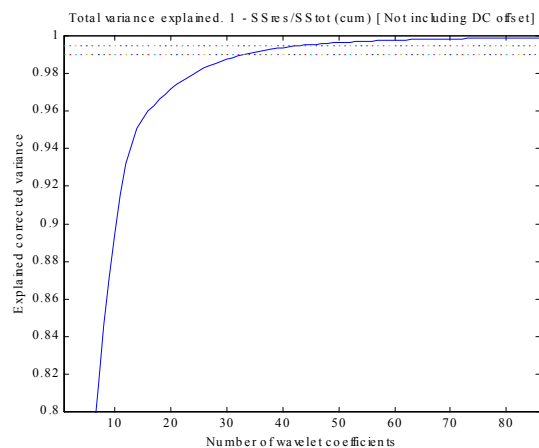
Largest coefficients selected from the variance spectrum

Wavelet variance spectrum



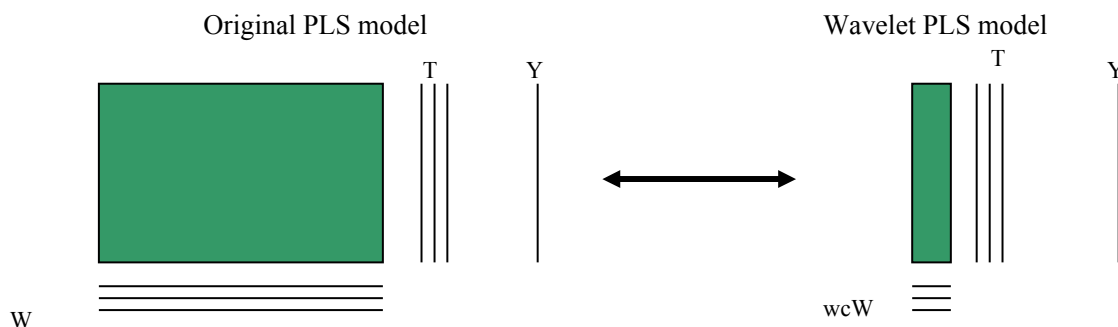
Number of wavelet coefficients?

- Optimal number of wavelet coefficients to use?
 - Problem dependent - depends on the data set
 - **X-block**, Explained variance (95 %, 99.5%)
 - 'trial and error' - stable wavelet PLS model
 - Visual inspection of the 'explained variance plot'
 - Prior knowledge
 - Default in SIMCA: 99.5%



Row compressed PLS models - Properties

- Scores - Original domain
- Loadings - Wavelet domain (can be back-transformed)
- Coefficients - Wavelet domain (can be back-transformed)
- Residuals - Original domain
- DModWX - distance to model in Wavelet space; new parameter
- Prediction set - Must be wavelet transformed in identical way as the training set



Column compressed PLS models - Properties

Scores

Wavelet domain (can be back-transformed)

Loadings & Coefficients

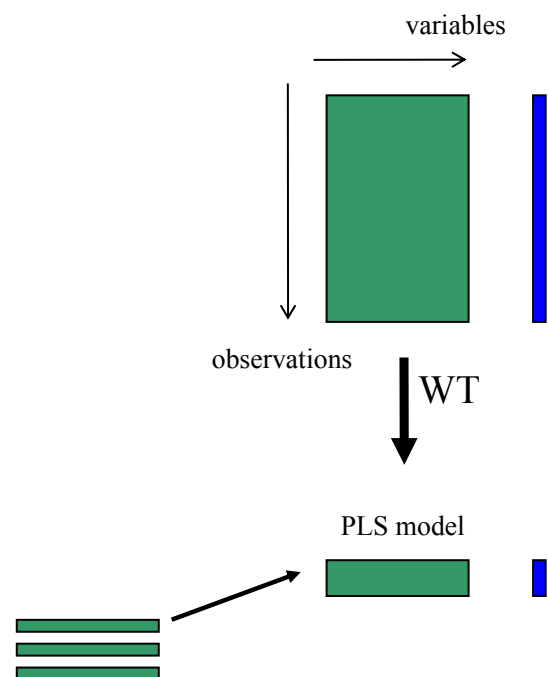
Original domain

Residuals

Wavelet domain (can be back-transformed)

Prediction set

Original test observation can be inserted, and give original predictions !

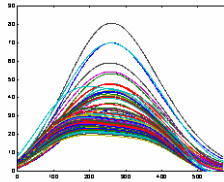


New **prediction set observation** inserted without transforming them !

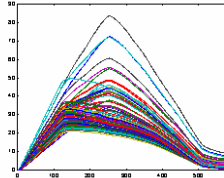
Wavelet compression - Sugar spectra

- With about 95% retained variability, the resulting shapes depend on the filter type and its order.

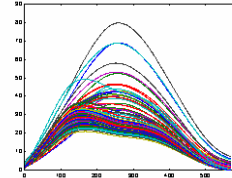
Beylkin
95.6% w. 5 coeff



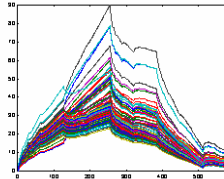
Coiflet-2
96.5% w. 5 coeff



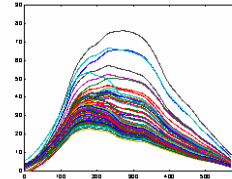
Coiflet-4
96.5% w 5 coeff



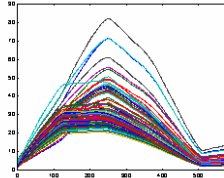
Daubechies-4
96.4% w. 7 coeff



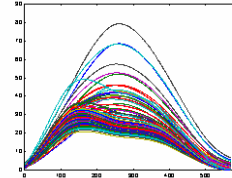
Daubechies-8
95.6% w. 6 coeff



Symmlet-4
96.1% w. 6 coeff



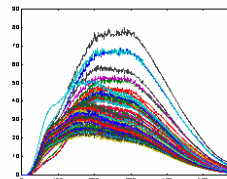
Symmlet-8
96.8% w. 5 coeff



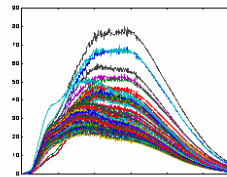
Wavelet compression - Sugar spectra

- With about 99.5% retained variability, the resulting shapes seem independent of the filter type and its order.

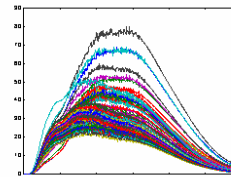
Beylkin
99.5% w. 150 coeff



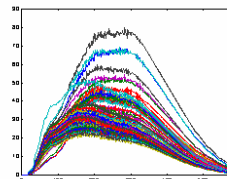
Coiflet-2
99.5% w. 150 coeff



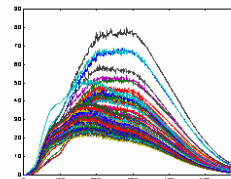
Coiflet-4
99.5% w. 151 coeff



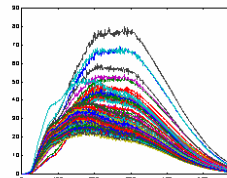
Daubechies-4
99.5% w. 152 coeff



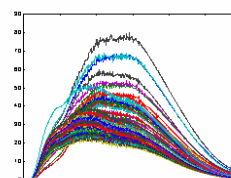
Daubechies-8
99.5% w. 151 coeff



Symmlet-4
99.5% w. 149 coeff



Symmlet-8
99.5% w. 151 coeff



Conclusions

- Signal correction may enhance properties of data
 - Potential of increasing the predictive ability
 - Risk for over-fitted model – always use external test data
- Signal compression
 - Spectra can be compressed (3-4% of original matrix size)
 - Time-series data can be compressed (of relevance for Batch-processes)
- Combination of signal correction and compression
 - Combination of OSC and DWT/MRA promising for future use in spectral on-line situations for quality monitoring
- Always use external prediction set for model verification

Multivariate Data Analysis and Modelling Basic Course

Chapter 12 Additional Topics II - Method Extensions and Special Cases



Contents

- PLS and one response variable
- PLS Discriminant Analysis (PLS-DA)

PLS and one response variable

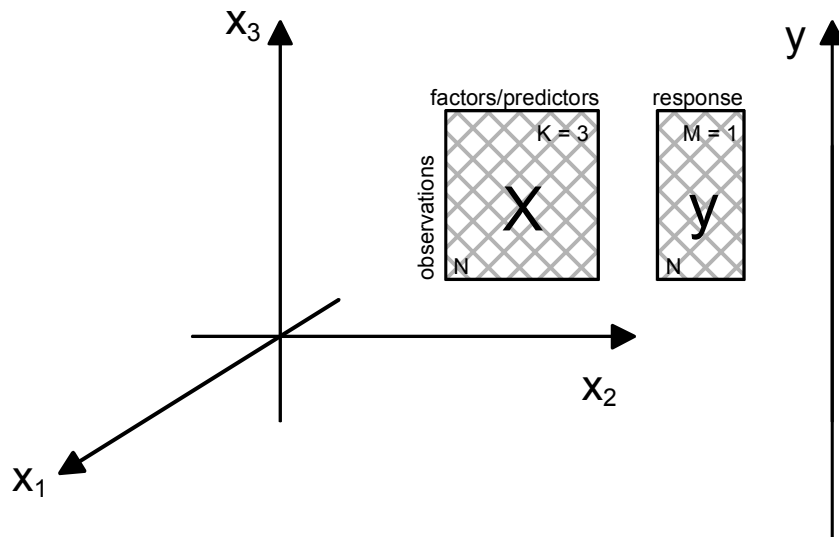


Contents

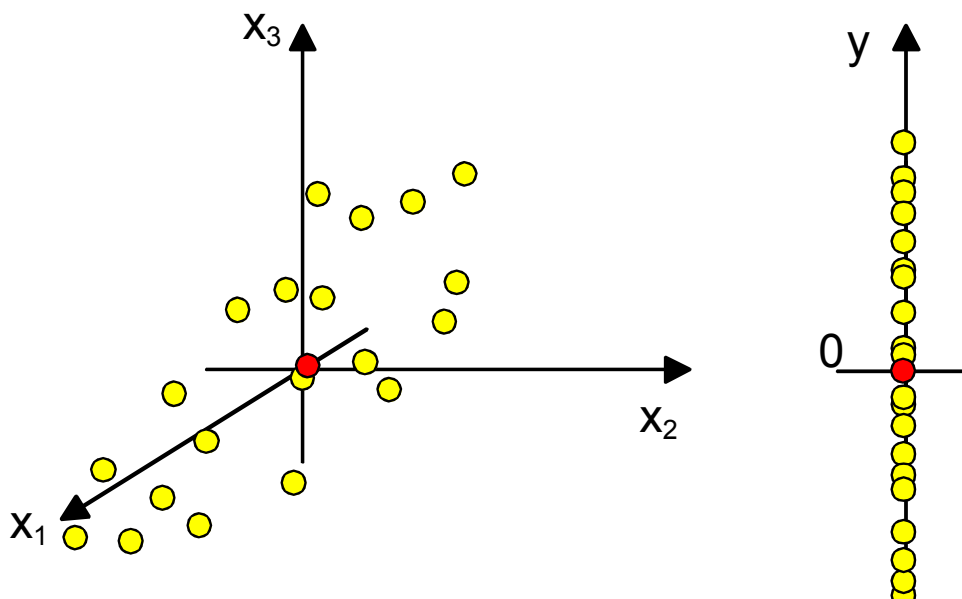
- Constructing the X and Y co-ordinate systems
- Each row in a data table corresponds to two points
- Finding the first PLS component ($M = 1$)
- Residual f_1 after the first component
- Extending the model with the second component ($M = 1$)
- An estimate of y after two components
- Residual f_2 after the second component
- Summary

Constructing the X and Y co-ordinate systems

- Consider a regression application with N observations, $K = 3$ X-variables, and $M = 1$ y-variable (here, we write lower case y to denote that a single variable is considered).

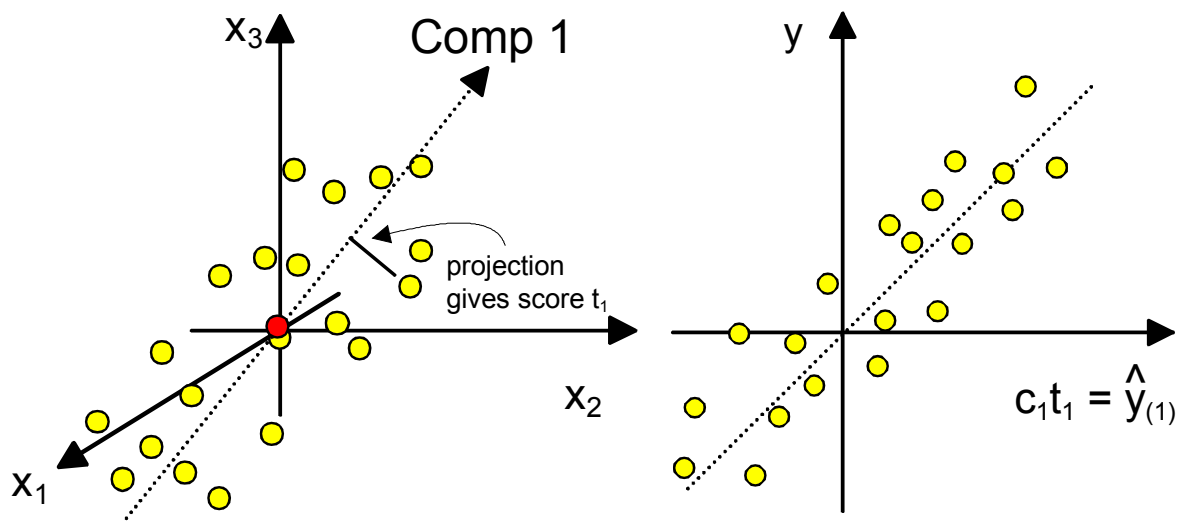


Each row in a data table corresponds to two points



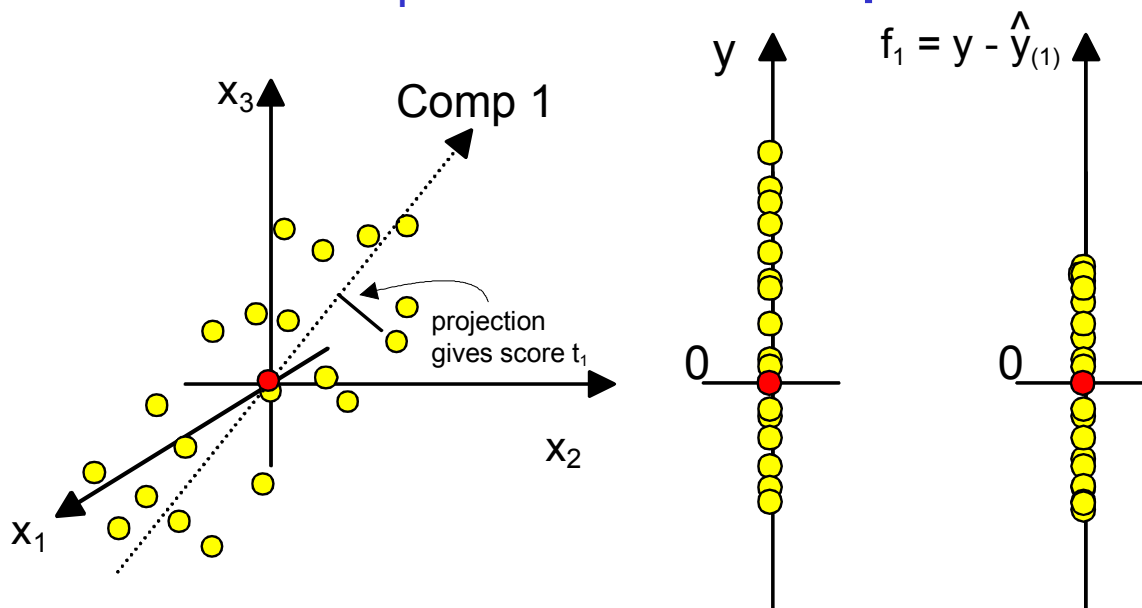
- In PLS each row of a data table corresponds to *two* points, one in the X-space and one in the Y-space. When considering only one y-variable, the Y-space reduces to a one-dimensional vector.

Finding the first PLS component (M = 1)



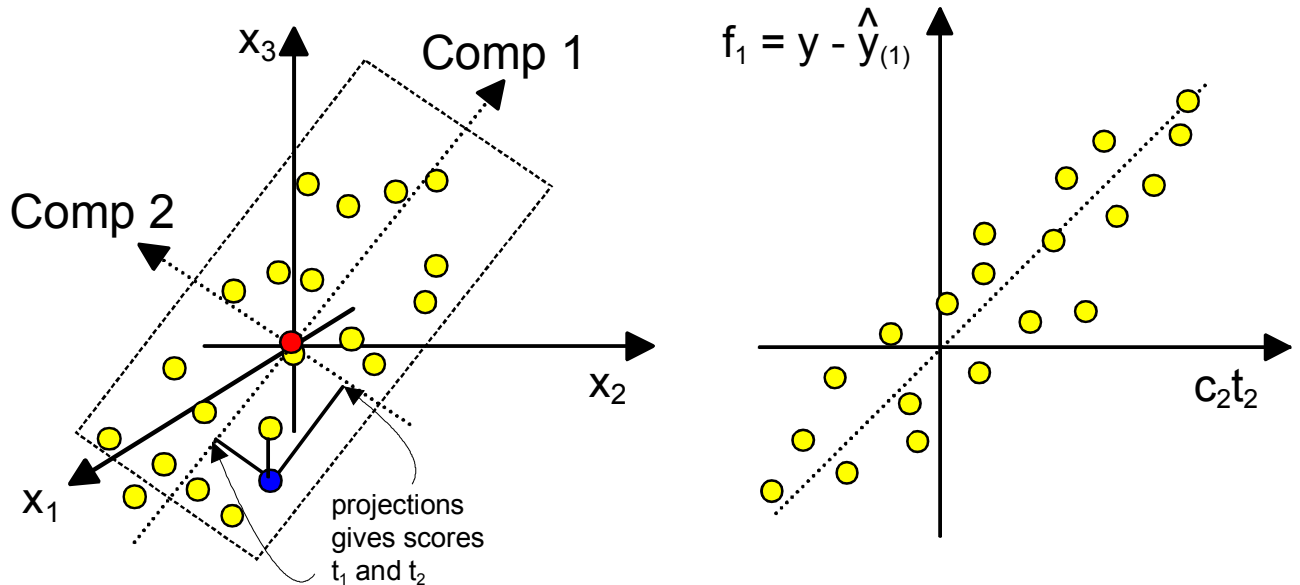
- The first component of the PLS model will orient itself so that it well describes the point-swarm in the X-space while at the same time accounting for a good correlation with the y-data. An estimate of y, $\hat{y}_{(1)}$, after the first PLS component is accomplished by multiplying t_1 by the weight of the y-data, c_1 .

Residual f_1 after the first component



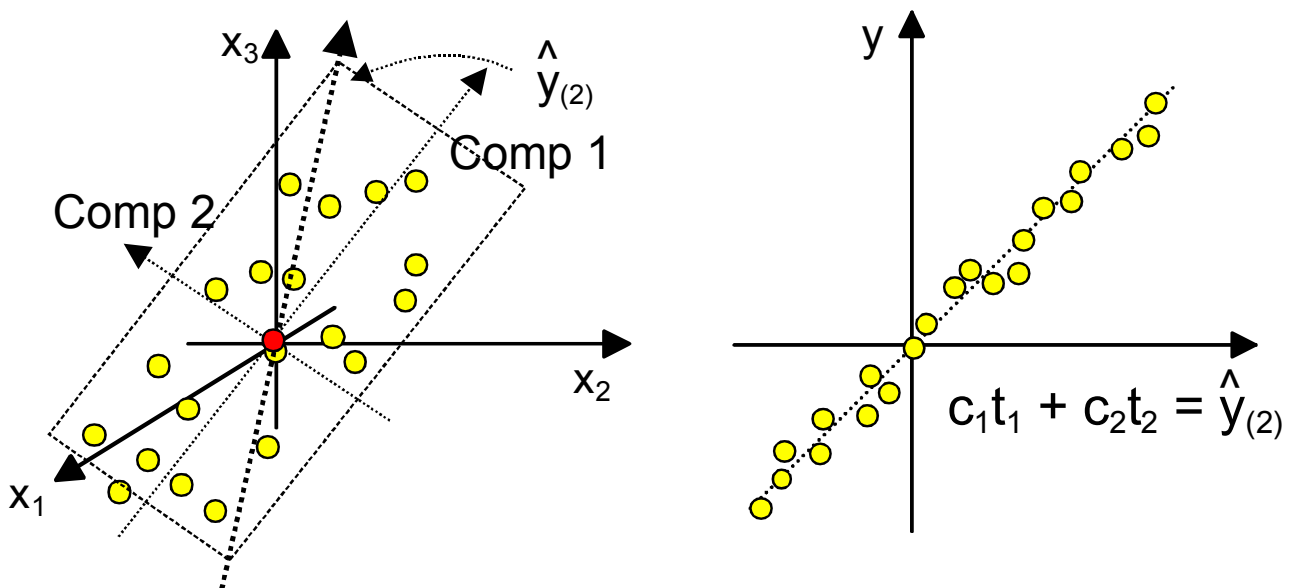
- The residual vector f_1 , obtained by subtracting $\hat{y}_{(1)}$ from y , is shorter than the vector consisting of the measured data. Therefore, the first component has explained a lot of response variation.

Extending the model with the second component (M = 1)



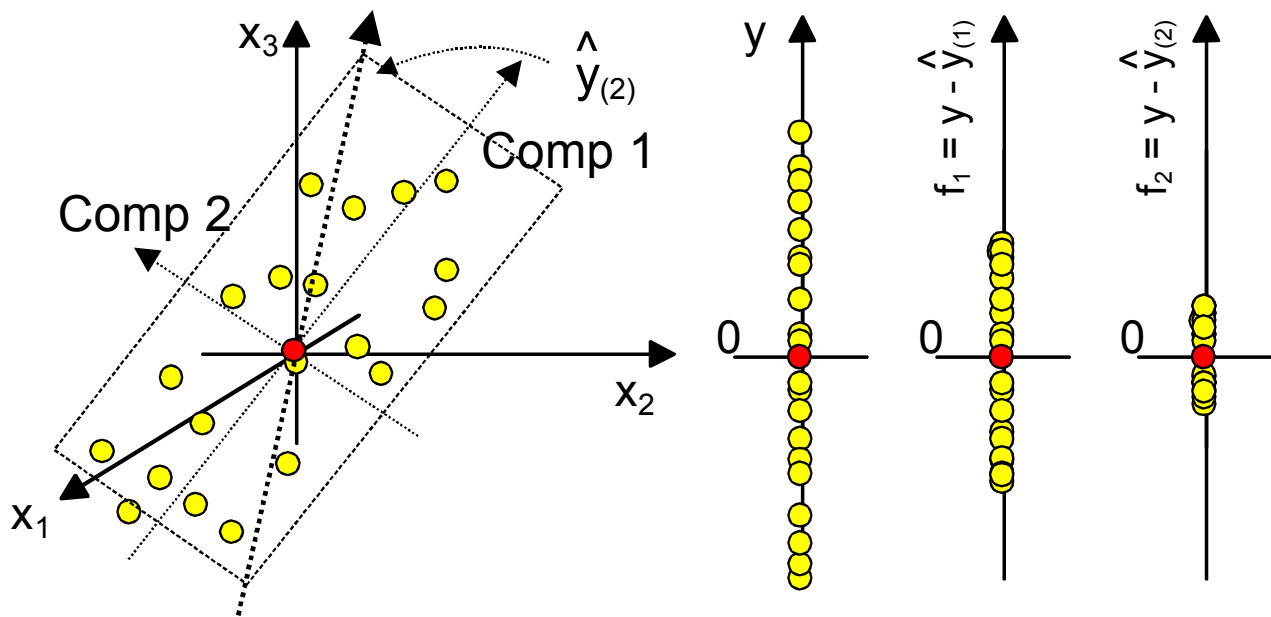
- The second projection co-ordinate in the X-space is orthogonal to the first one. By projecting the observations onto this line one obtains the score vector t_2 . In the right-hand part of the figure, we can see how the second score vector times the second weight of the y-data, c_2 , correlates with the y-residual, f_1 , after the first dimension.

An estimate of y after two components



- An estimate of y after two components is obtained by computing $c_1 t_1 + c_2 t_2$. Geometrically, $\hat{y}_{(2)}$ is interpretable as the resultant of the vector addition of component 1 and component 2 in the X-space.

Residual f_2 after the second component



- After one component, the y -residual (f_1) is significantly smaller than the spread in the measured variable. The situation is even better after the inclusion of the second component, as the residual f_2 is smaller than f_1 .

Summary, PLS1

- **One or several response variables?**
- PLS *has* the ability to model and analyse several Y -variables together, which has the advantage of giving a simpler picture than separate models for each response.
- In general, when the Y -variables are strongly correlated, one can recommend that they are analysed together, since the correlations stabilise the model.
- If the Y 's really measure different things, however, and hence are fairly independent, one gains little by analysing them in the same model.
- A simple geometric interpretation of PLS and one response variable is possible if the residual vector f_a is considered after each component a .

PLS Discriminant Analysis (PLS-DA)

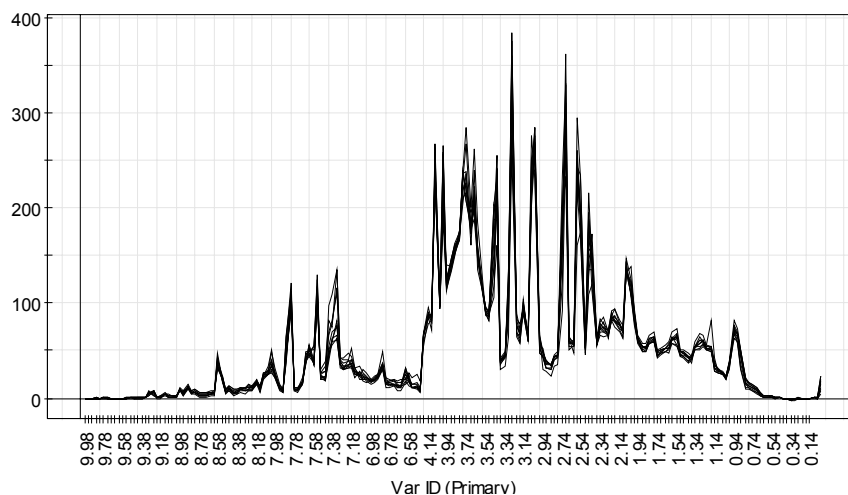


PLS - Discriminant analysis

- Assessment of drug exposure in metabonomics
- Metabonomics: monitoring of complex time-related metabolite profiles that are present in biofluids, e.g., urine, plasma, saliva, etc.

- Proton-NMR spectra of urinary profiles of drug-exposed rats

Metabonomics.DS1 Metabonomics
Observation

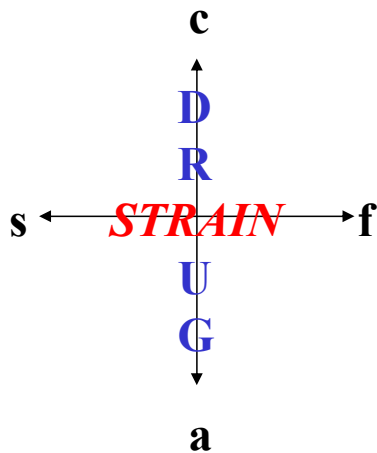


Metabonomics – The Data

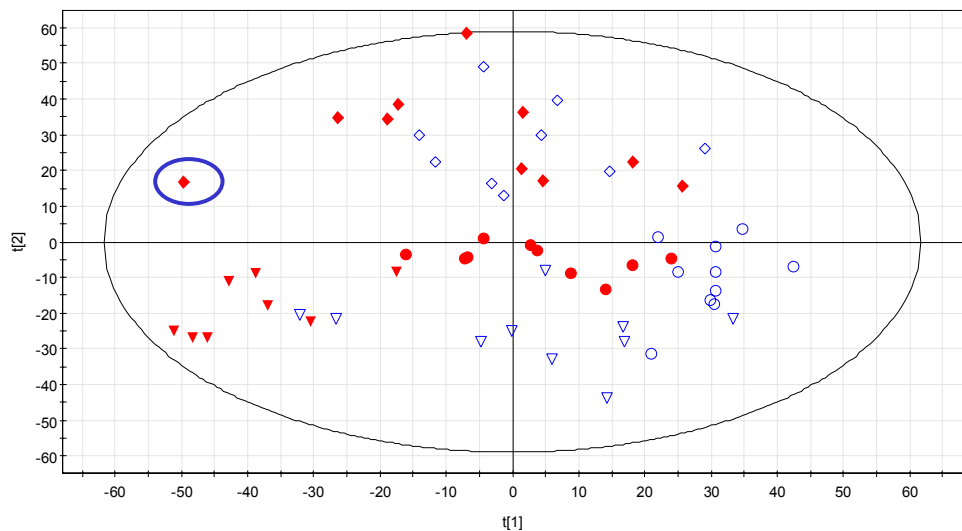
- Rats exposed to chloroquine (an antimalarial) or amiodarone (an antiarrhythmic)
- Observations: N = 57 rats
Variables: K = 194 variables (1H-NMR shift regions)
- Six groups (“classes”):
 - Control Sprague-Dawley, 10 rats, “s”
 - Sprague-Dawley treated with amiodarone, 8 rats, “sa”
 - Sprague-Dawley treated with chloroquine, 10 rats, “sc”
 - Control Fisher, 10 rats, “f”
 - Fisher treated with amiodarone, 10 rats, “fa”
 - Fisher treated with chloroquine, 9 rats, “fc”

Metabonomics – PCA to overview

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



Metabonomics.M1 (PCA-X), Overview with Pareto scaling
t[Comp. 1]/t[Comp. 2]

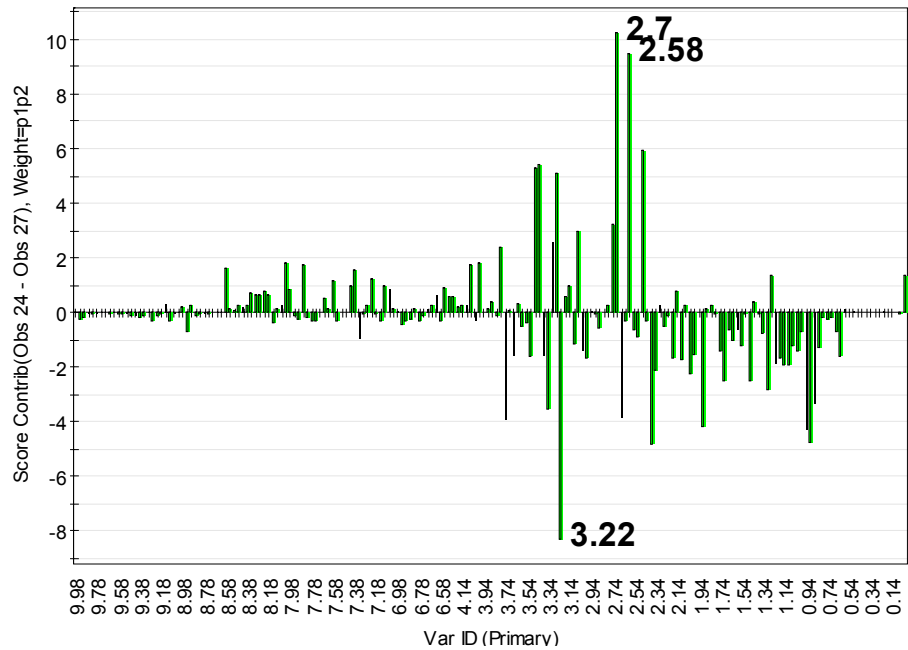


- One outlier, rat 27, encircled
 - Measurement error ?
 - Handling/environmental differences ?
 - Slow responder ?

Metabonomics – Contribution plot to reveal differences

Metabonomics.M1 (PCA-X), Overview with Pareto scaling
Score Contrib(Obs 24 - Obs 27), Weight=p[1]p[2]

- How is rat 27 different from a “normal” sc-rat?
- Chemical shift regions 2.58, 2.70 and 3.22



PLS - Discriminant analysis

- When clusters are found in PC score plots one can

- perform disjoint PCA (recall IRIS example), or
- carry out PLS-DA

- **What is PLS-DA ?**

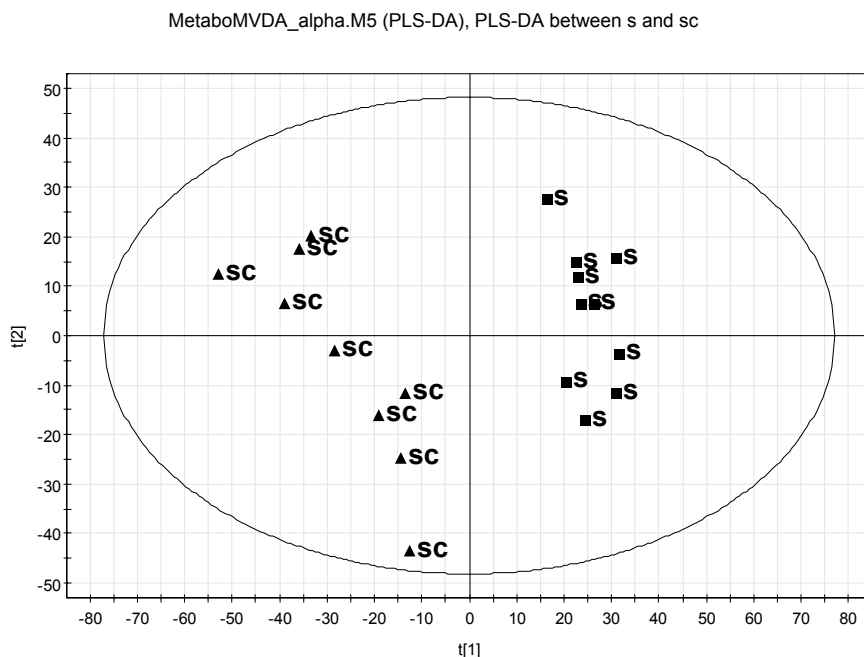
– A "dummy" variable is added for each category. PLS is then used to relate X and Y.

- **PCA:** Maximum variance projection
- **PLS-DA:** Maximum separation projection

	X		Y			
~200						
Class "s" (10)	1	0	0	0	0	0
Class "sc" (10)	1	0	0	0	0	0
Class "sa" (8)	0	1	0	0	0	0
Class "sa" (8)	0	1	0	0	0	0
Class "f" (10)	0	0	1	0	0	0
Class "f" (10)	0	0	0	1	0	0
Class "fc" (10)	0	0	0	0	1	0
Class "fc" (10)	0	0	0	0	1	0
Class "fa" (9)	0	0	0	0	0	1
Class "fa" (9)	0	0	0	0	0	1

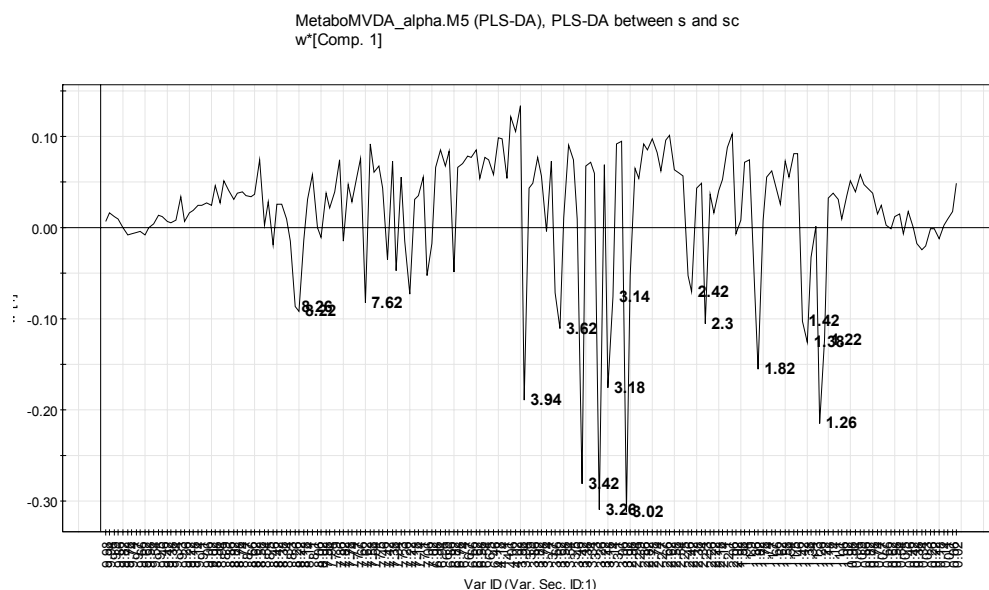
PLD-DA between "s" and "sc"

- A way to focus on drug effect; outlier removed
- The two classes are well resolved in component 1



PLD-DA between "s" and "sc"

- Line plot of w^*1
- Chemical shift regions influential for the separation of the two classes are indicated



Metabonomics – Conclusions

- Multivariate analysis of NMR-data creates one or several maps (i.e., score plots, loading plots) that show trajectories of biochemical changes in biofluids induced by toxin exposure or disease
- Through this technology it is possible
 - (i) to detect target organs or pathways of dysfunction
 - (ii) to uncover likely chemical mechanisms of toxicity, and
 - (iii) to identify useful biomarkers indicative of onset, development, and decay of abnormal animal health conditions.

Leading reference: Nicholson, J.K., Connelly, J., Lindon, J.C., and Holmes, E., Metabonomics: A Platform for Studying Drug Toxicity and Gene Function, Nature Reviews, 2002; 1:153-161.

Summary of PLS-DA

- Class memberships are explicitly given in PLS-DA; this gives a rotation of the latent variables, such that a maximum separation among the classes is obtained.
- PLS-DA works reliably when each class is "tight" and occupies a small and separate volume in the X-space.
- PLS-DA is useful with 2-4 classes; discrimination results may become incomprehensible with too many classes.
- When some of the classes are not homogeneous and spread significantly in X-space, the discriminant analysis does *not* work.

Multivariate Data Analysis and Modelling Basic Course

Chapter 13 Additional Topics III – Process Applications



Contents

- Multivariate Statistical Process Control (MSPC)
- Batch Statistical Process Control (BSPC)

Multivariate Statistical Process Control (MSPC)



Contents

Multivariate Statistical Process Control (MSPC)

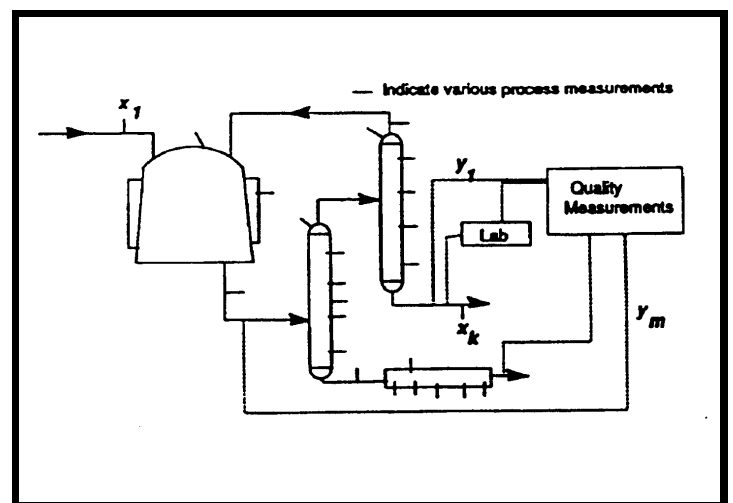
- The Problem
- SPC philosophy
- Control Charts
- MSPC
- Conclusions

Process Data Analysis - Purposes

- 1. Monitoring** the state of the process, statistical process control (SPC)
 - Early warning
 - Diagnostics - finding "assignable causes" (SPC jargon \Leftrightarrow interpret deviations)
- 2. Understanding** the relationship between
 - input variables, X (process data) and output variables, Y (product quality, cost, amount, ...)
- 3. Optimisation**
 - Use process models to improve process

The Problem

- 50 years ago
 - Few variables: T, P, flows
- Today
 - Many measurements and very often
 - Large data sets
 - The latent variable concept
- Process the same
- Data have changed
 - $K = 5 \rightarrow 500$
 - $N = 10 \rightarrow 1000$



Typical process (from MacGregor et al. 1991)

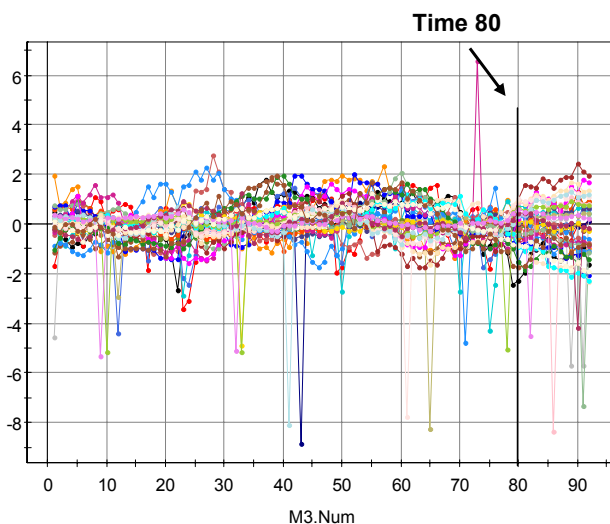
Example - Monitoring (PROC1A)

- A chemical production plant
 - A continuous steady state process
 - All data are coded, not to reveal any trade secrets
- The data - 33 variables, 92 observations
 - 7 controlled process variables (x1in-x7in)
 - 18 intermediate process variables (x8md-xpen)
 - 8 output variables (y1-y8)
 - Data sampled during 92 time units (e.g. hours or minutes)
- The process went out of control around time 80 and had to be shut down at time 92

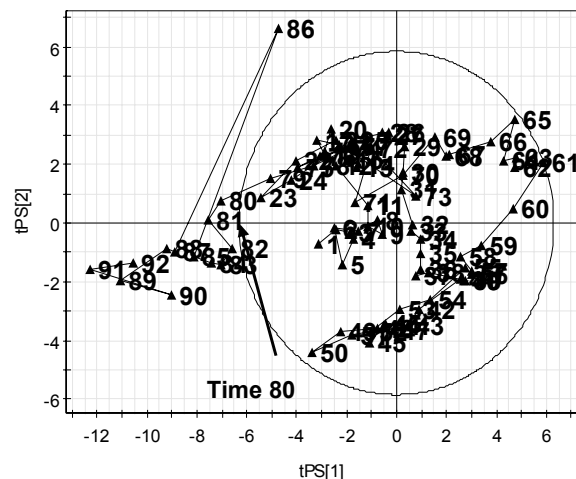
Example - Monitoring a process (PROC1A)

- Trend curves or traditional SPC on some important variables (Outputs)
 - In total, 33 trend curves to monitor
- PCA and multivariate SPC (MSPC) using all variables
 - The alarm limit (ellipse) is calculated from historical data

proc1a Pca on normal conditions 1-69



proc1a.M1 (PCA-X), PCA on normal behavior 1-69, PS-proc1a
tPS[Comp. 1]/tPS[Comp. 2]



Multivariate Statistical Approach to Monitoring

- Based on Statistical Process Control (SPC) philosophy
 - Based on historical process data
 - Future behaviour is referenced against a statistical model of good past behaviour
- Empirical models are easily built from the historical data base
 - First and second order models are often good approximations
- Non-directional
 - Detects any deviation from normal behaviour - needs to be complemented with tools for diagnostics (e.g., contribution plots)

Control Charts

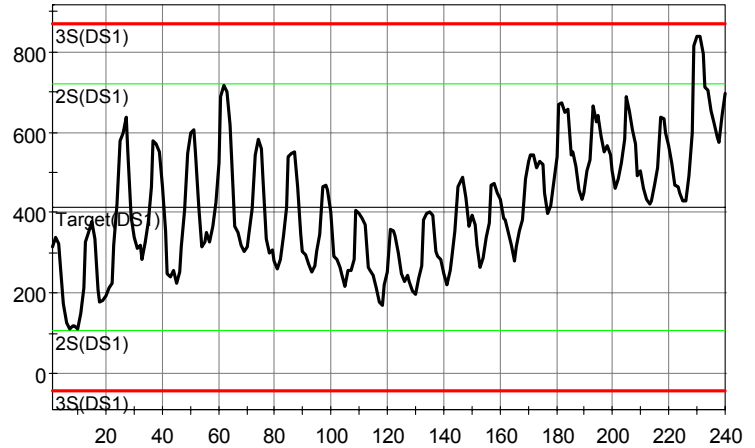
- A graphical method for evaluating whether a process is in a "state of statistical control"
- Measures:
 - current deviations Shewhart, Xbar-R, Xbar-s
 - cumulative deviations CuSum
 - filtered deviations EWMA
- Advantages:
 - Results from process displayed graphically and plotted against time
 - Feedback
 - Continuous improvements

The charts are illustrated with a data set with monthly Canadian unemployment numbers 1956-75.

Shewhart charts

- Started by Dr. Walter A. Shewhart
 - whose control chart approach remains the most widely used.
- Control limits are set so that, if the process remains in control, there is only a small probability of obtaining a point beyond the control limits.
- Normal distribution:
 - 95.5% within ± 2 standard dev.s (green)
 - 99.7% within ± 3 standard dev.s (red)

- Shewhart chart for monthly Canadian unemployment numbers 1956-75.
 - Shows current deviations



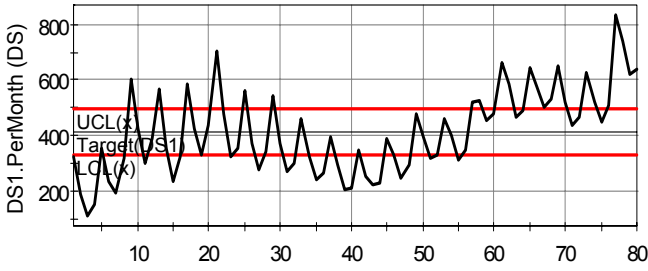
Shewhart charts, subgrouping

Subgrouping of the data is often used when there is a rational reason for doing so, e.g. over a batch, a day, or a week.

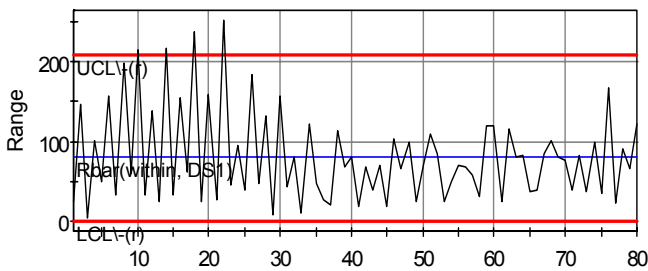
- Subgrouping has the advantage of making the subgroup averages more normally distributed
- We want to monitor the averages but still keep track of the variation between individual values. Two types of charts are common, both actually results in two charts
 - **Xbar-R**, subgroup averages and ranges
 - **Xbar-s**, subgroup averages and standard deviation
- The Xbar chart shows the variation *between* the subgroups (averages)
- The R or s chart shows the variation *within* the subgroups.
 - Use R when small subgroups
 - Use s when large subgroups (>10)
- Rational subgrouping maximises the between group variation and minimises the within group variation, e.g. per day, week, shift group, ...
- Subgrouping is also applicable for CuSum and EWMA charts

Shewhart Xbar-R chart

- Shewhart (subgroup 3) Xbar, quarterly averages



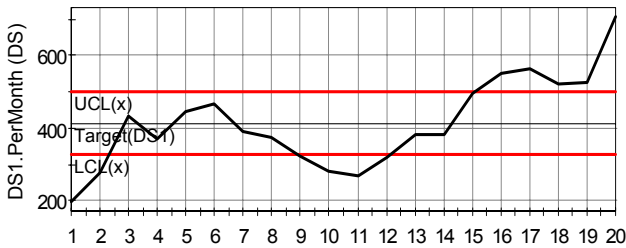
- Shewhart (subgroup 3) Range



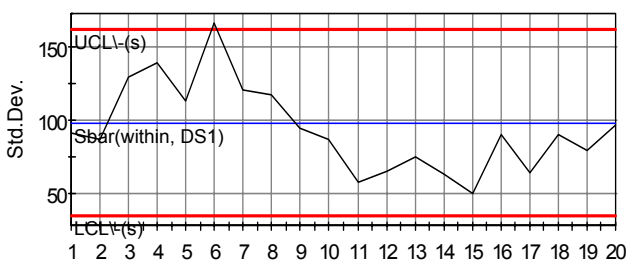
- The control limits for Xbar are given by
 - $UCL_x / LCL_x = \text{target} \pm A_2 Rbar$
 - A_2 is a control chart constant that depends on subgroup size
 - $Rbar$ is the average subgroup range.
- The control limits for R are given by
 - $UCL_R = D_4 Rbar, LCL_R = D_3 Rbar$
 - D_4 and D_3 are constants that depend on subgroup size

Shewhart Xbar-s chart

- Shewhart (subgroup 12) Xbar, yearly averages



- Shewhart (subgroup 12) standard deviation



- The control limits for Xbar are given by
 - $UCL_x / LCL_x = \text{target} \pm A_3 sbar$
 - A_3 is a control chart constant that depends on subgroup size
 - $sbar$ is the average subgroup range.
- The control limits for s are given by
 - $UCL_s = B_4 sbar, LCL_s = B_3 sbar$
 - B_4 and B_3 are constants that depend on subgroup size

CuSum charts

- Instead of plotting the individual observations, y_t , their cumulative deviations from the target value are plotted

$$S_T = \sum_{t=1}^T (y_t - \text{target})$$

- The CuSum chart shows small shifts in mean. It is suitable for autocorrelated data.
- This CuSum can wander remarkably far on the chart

- A more common way to present a CuSum is to subgroup the data and plot three points for each subgroup

- One point is the deviation from the target
- One point is the cumulative sum on the **high** side, SH
- One point is the cumulative sum on the **low** side, SL

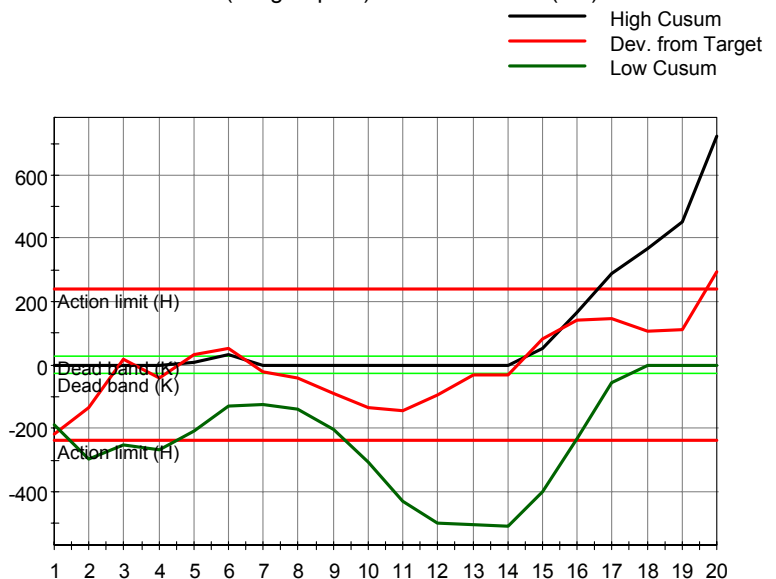
$$SH_T = \text{Max}(0.0, \sum_{t=1}^T (y_t - (\text{target} + K)))$$

- K = allowable variation (deadband), normally = $\sigma/2$
- SL_T is computed as SH_T , but with respect to target - K .

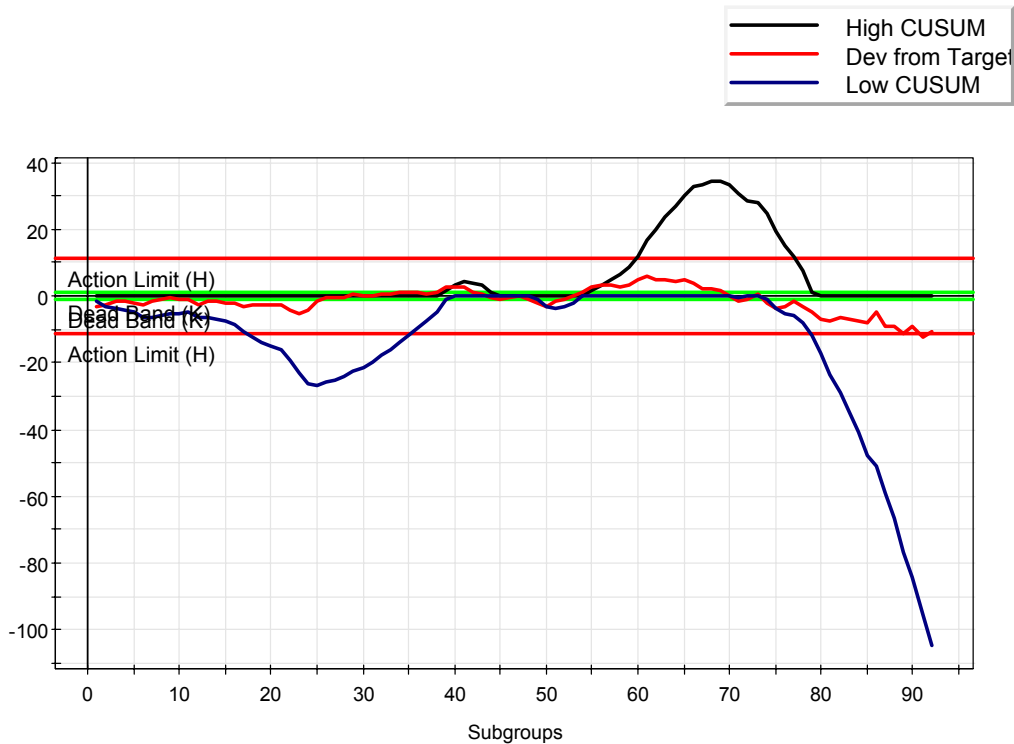
CuSum charts, continued

- CuSum chart (subgroup 12) for the unemployment data – cumulative deviation (shift from target)
- In the plot, the lines "K" represent the area where the process is operating on target. The lines "H" are the action limits, normally $H = 4.5\sigma$.

UNEMPLOY UnTitled
CuSum (subgroup 12): DS1.PerMonth (DS)



CuSum of t_1 of PROC1A



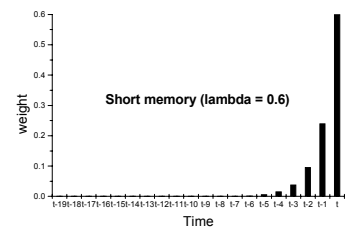
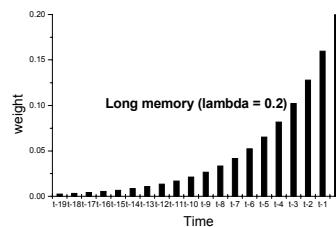
Exponentially Weighted Moving Average, EWMA, charts

- EWMA can be used to detect small process shifts. The choice of λ will affect how quickly. In general EWMA detects smaller process shifts ($< 2\sigma$) faster than the individuals control chart.

- Exponential weights (memory) can decrease
 - slowly, long memory $\lambda \Rightarrow 0.0$
 - or fast, short memory $\lambda \Rightarrow 1.0$

$$\hat{y}_{t+1} = \hat{y}_t + \lambda \varepsilon_t = \hat{y}_t + \lambda (y_t - \hat{y}_t) = \sum_t w_t y_t$$

- \hat{y}_{t+1} = predicted value at time t+1
- y_t = observed value at time t
- \hat{y}_t = predicted value at time t
- ε_t = prediction error at time t
- λ = weighting constant, determines the memory
- w_t = weight factor
- with $w_t = \lambda(1-\lambda)^{t-1}$ $\sum w_t = 1.0$



- EWMA can be seen as a compromise between the Shewhart and CuSum
- EWMA is also often used for forecasting.
 - Then value of λ is often set at 0.2 ± 0.1

EWMA charts, continued

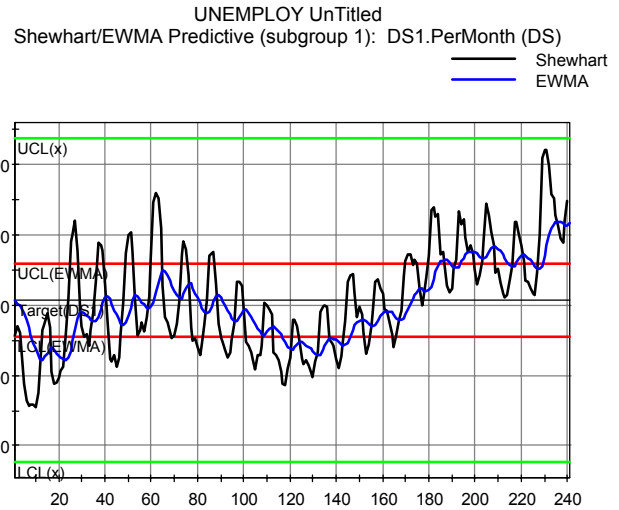
- The three-sigma limits for EWMA are given by

$$-UCL / LCL_{EWMA} = \text{target} \pm 3\sigma_{EWMA}$$

- $\sigma_{EWMA} = \sigma \sqrt{\frac{\lambda}{2-\lambda}}$

- σ is the standard deviation of the individual measurements.

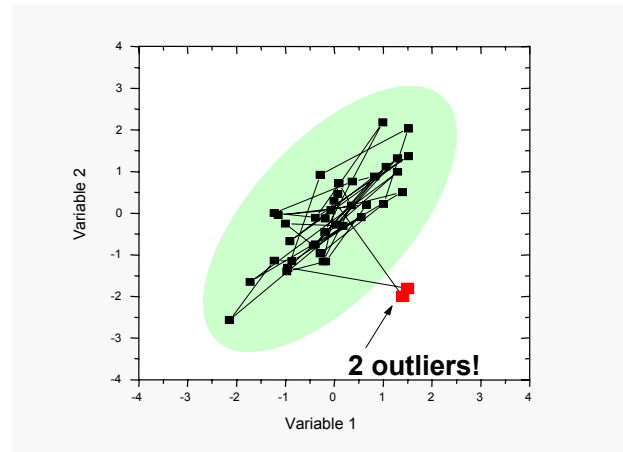
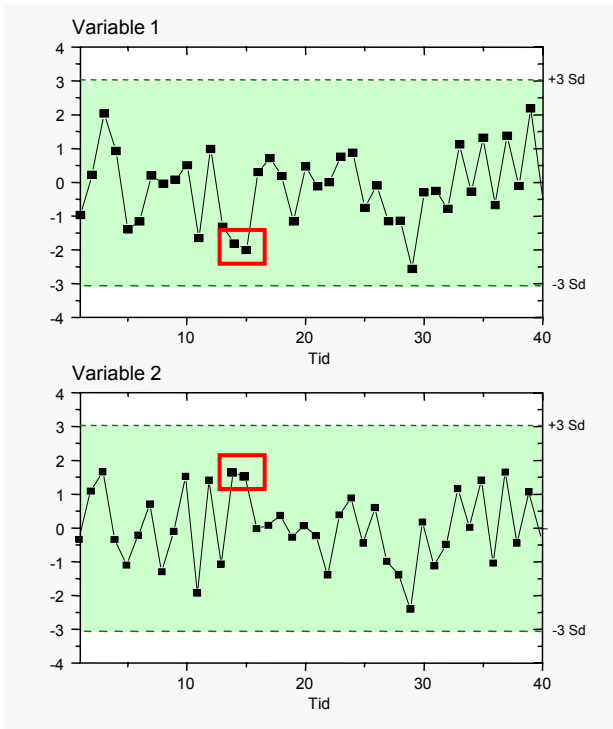
- The EWMA chart is normally used with individuals data
 - Here $\lambda = 0.1$



All data are needed

- In Shewhart's days (1930) process controllers were lucky to have one measurement of product quality
- Today we may get 10 or more quality measurements on each sample
- Most outliers remain undetected with the use of classical SPC techniques!
 - No covariance information

Fast and Correct Decision Making



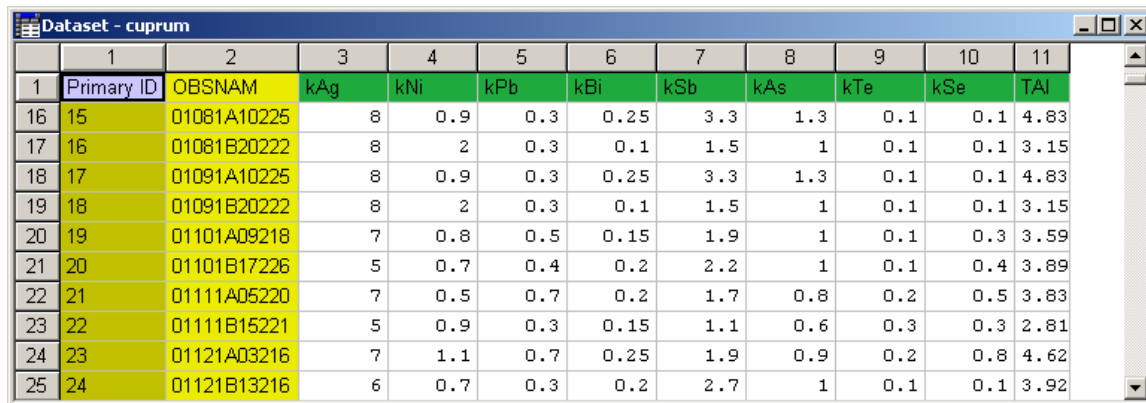
- The outliers are not detected until you look at the combination of variables
- The information is found in the correlation pattern - not in the individual variables!

Multivariate statistical process control (MSPC) charts

- Available charts in SIMCA:
 - Shewhart current value
 - XbarR mean and range of subgroup
 - XbarS mean and SD of subgroup
 - CuSum cumulative sum
 - EWMA exponentially weighted moving average
- MSPC = Control charting applied to multivariate parameters
 - Parameters which may be monitored in SIMCA
 - Scores score vectors, t
 - DModX distance to the model
 - Hotelling's T^2 multivariate generalisation of Student's t -test

Example – Quality control/MSPC (CUPRUM)

- Electrolytic production of Copper
 - Boliden AB produces approximately 300 tonnes of Copper every day
 - extremely pure (99.998 %) Copper
 - impurity testing twice a day to ensure quality (TAI, Total Analysis Index)
 - TAI is a weighted sum of 8 different impurities (PPM-level)



	1	2	3	4	5	6	7	8	9	10	11
	Primary ID	OBSNAM	kAg	kNi	kPb	kBi	kSb	kAs	kTe	kSe	TAI
16	15	01081A10225	8	0.9	0.3	0.25	3.3	1.3	0.1	0.1	4.83
17	16	01081B20222	8	2	0.3	0.1	1.5	1	0.1	0.1	3.15
18	17	01091A10225	8	0.9	0.3	0.25	3.3	1.3	0.1	0.1	4.83
19	18	01091B20222	8	2	0.3	0.1	1.5	1	0.1	0.1	3.15
20	19	01101A09218	7	0.8	0.5	0.15	1.9	1	0.1	0.3	3.59
21	20	01101B17226	5	0.7	0.4	0.2	2.2	1	0.1	0.4	3.89
22	21	01111A05220	7	0.5	0.7	0.2	1.7	0.8	0.2	0.5	3.83
23	22	01111B15221	5	0.9	0.3	0.15	1.1	0.6	0.3	0.3	2.81
24	23	01121A03216	7	1.1	0.7	0.25	1.9	0.9	0.2	0.8	4.62
25	24	01121B13216	6	0.7	0.3	0.2	2.7	1	0.1	0.1	3.92

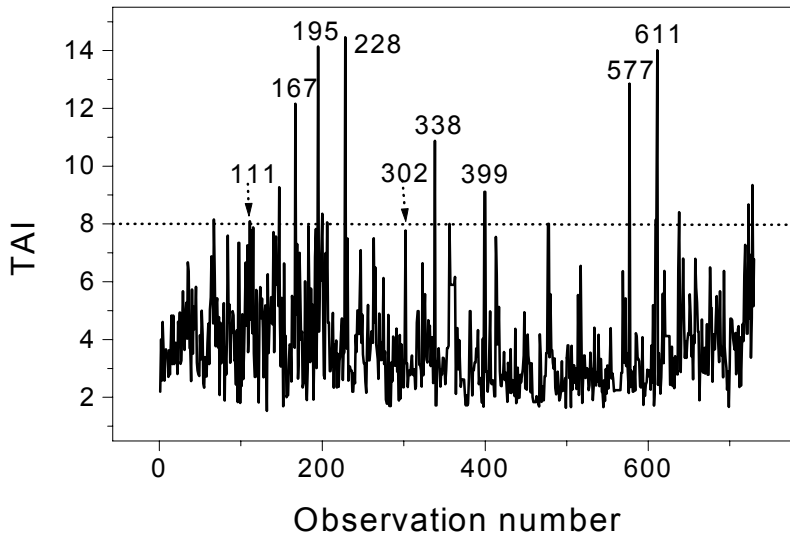
Example – CUPRUM

- The data - 9 variables, 730 observations
 - 8 measured variables (Ag-Se)
 - 1 calculated variable (TAI)
 - data sampled twice a day over one year giving 730 observations
 - all variables were log-transformed
- The Copper industry uses only the TAI value to determine the quality and thereby the price. Copper products with TAI over 8.0 are discarded.
- Question:
 - Can we do better with projection methods?

CUPRUM – Time series plot of TAI

- Quality control limit corresponding to $TAI = 8$
- Samples 111 ($TAI = 8.1$) and 302 ($TAI = 7.8$) have approximately the same TAI value

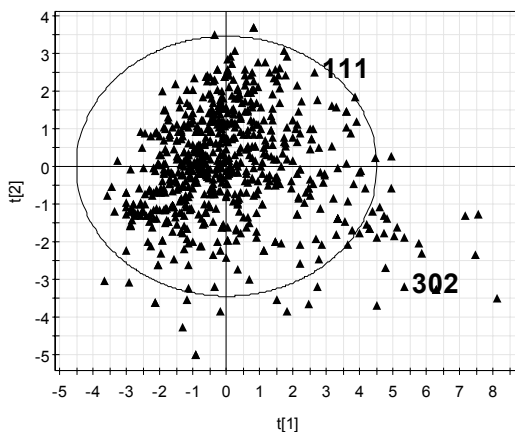
Time series plot of TAI



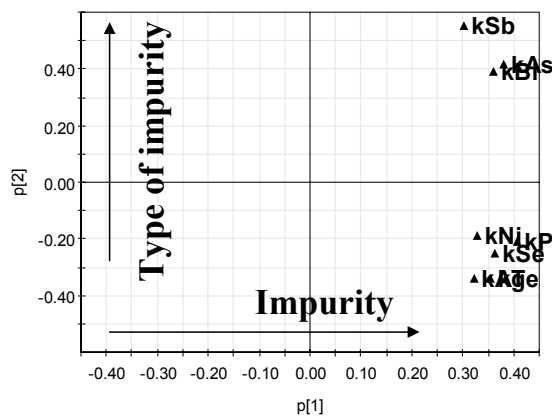
CUPRUM – Scores and loadings of PC-model

- A PC-projection of the table was made. The 8-dimensional table (the TAI variable excluded) was thus projected onto a two-dimensional plane, showing 67% of the variability in the data
- Samples 111 and 302 are situated far apart!
- The corresponding loading plot revealed two types of impurities

cuprum.M1 (PCA-X), pca for overview all vars log-transformed
t[Comp. 1]/t[Comp. 2]



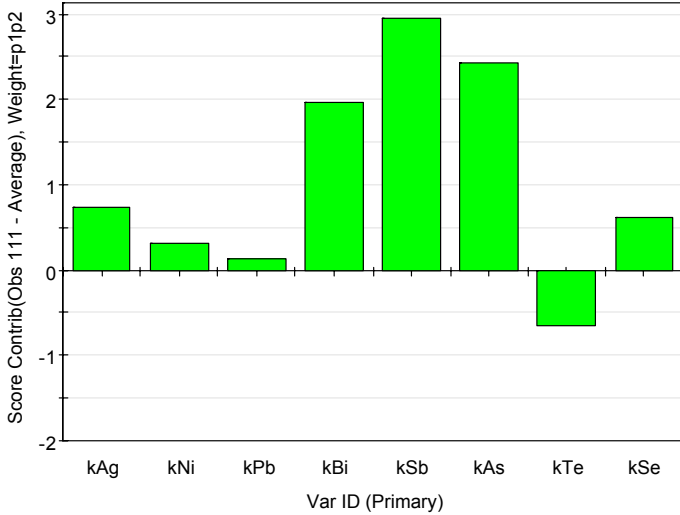
cuprum.M1 (PCA-X), pca for overview all vars log-transformed
p[Comp. 1]/p[Comp. 2]



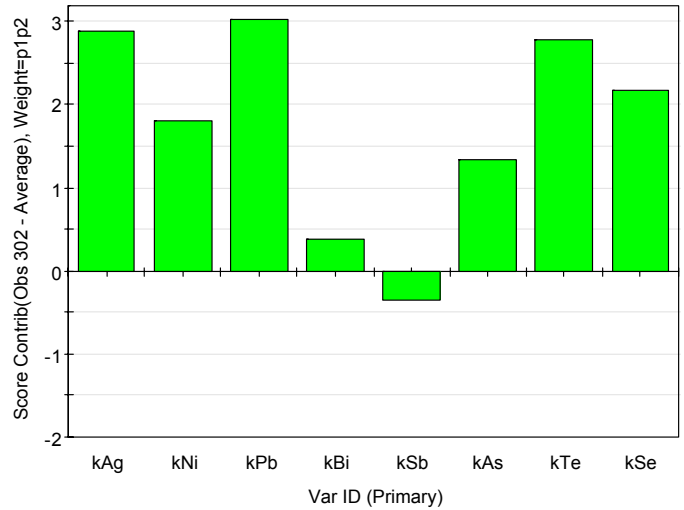
CUPRUM – Contribution plots

- Contribution plots “zoom in” on a single sample. Here, the variable profiles of samples 111 and 302 are shown.

cuprum.M1 (PCA-X), PCA for overview log-transform
Score Contrib(Obs 111 - Average), Weight=p1p2



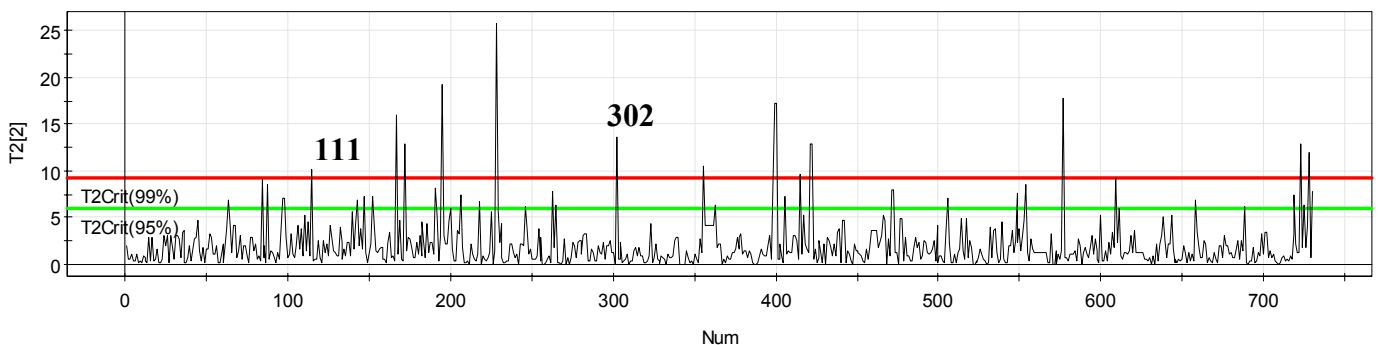
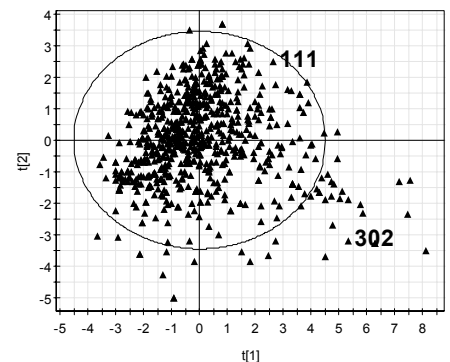
cuprum.M1 (PCA-X), PCA for overview log-transform
Score Contrib(Obs 302 - Average), Weight=p1p2



CUPRUM - MSPC monitoring

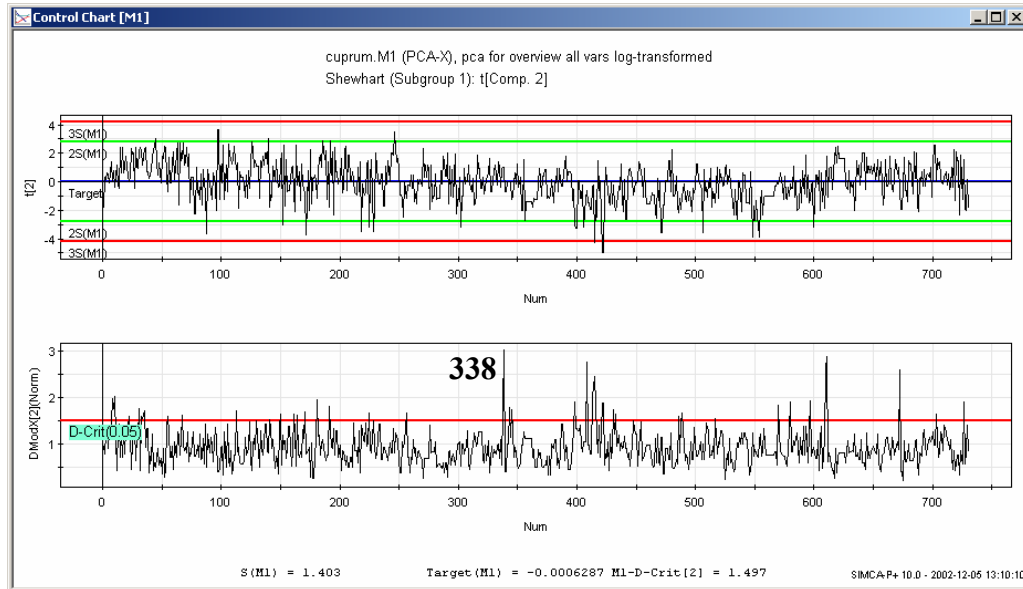
- Example: Hotelling's T^2
 - as tolerance limit in a scatter plot
 - “Normal operating conditions”
 - as control limit in conventional control chart

cuprum.M1 (PCA-X), pca for overview all vars log-transformed
t[Comp. 1]/t[Comp. 2]



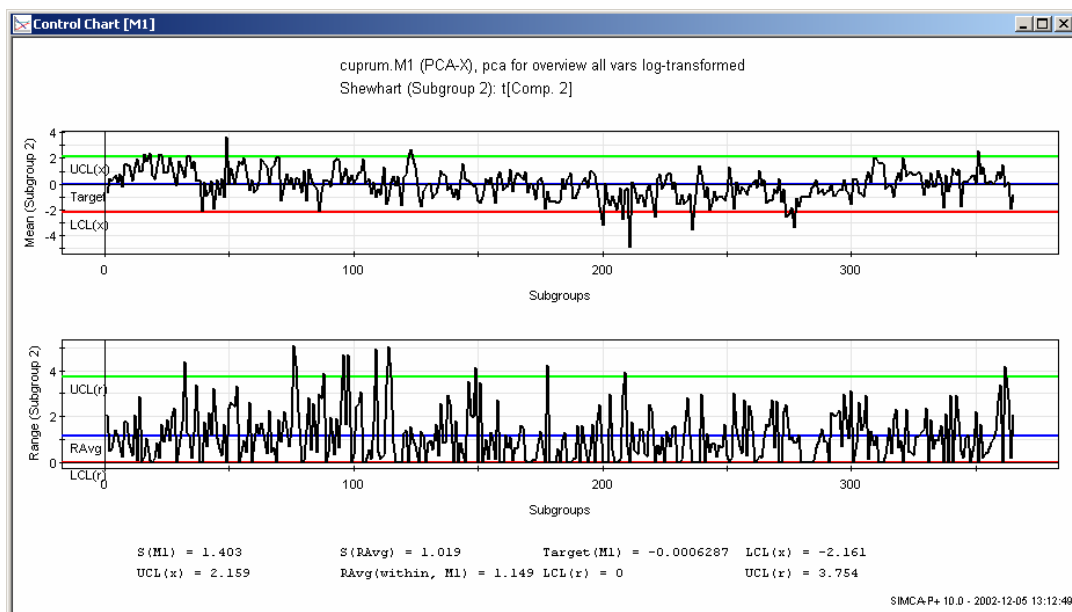
CUPRUM - MSPC monitoring

- The new, and independent, score vectors (\mathbf{t}) can be monitored in control charts. Here Shewhart chart on t_2 and DModX
 - used to detect trends and upsets (e. g. Observation 338, DModX)



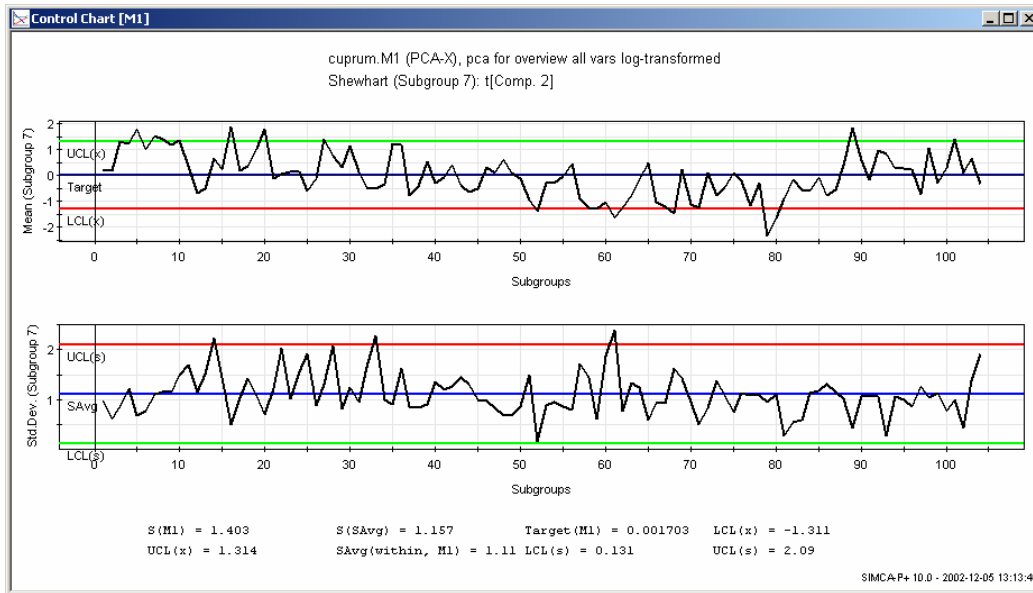
CUPRUM - MSPC monitoring

- The daily variation can be monitored in an XbarR chart; upper chart shows daily means, lower chart daily range



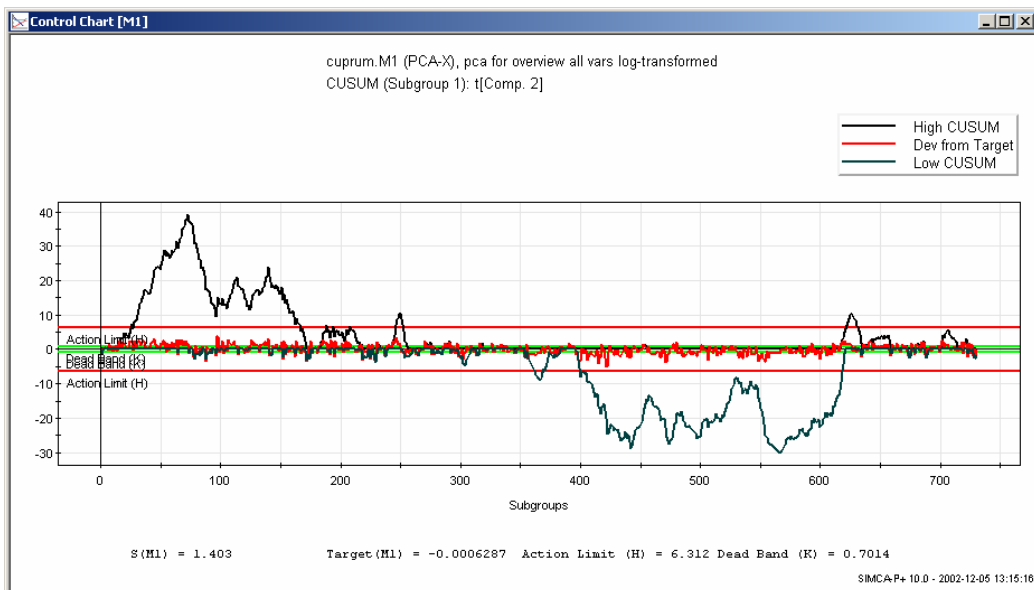
CUPRUM - MSPC monitoring

- The weekly variation can be monitored in an XbarS chart; upper chart shows weekly means, lower chart weekly standard deviation



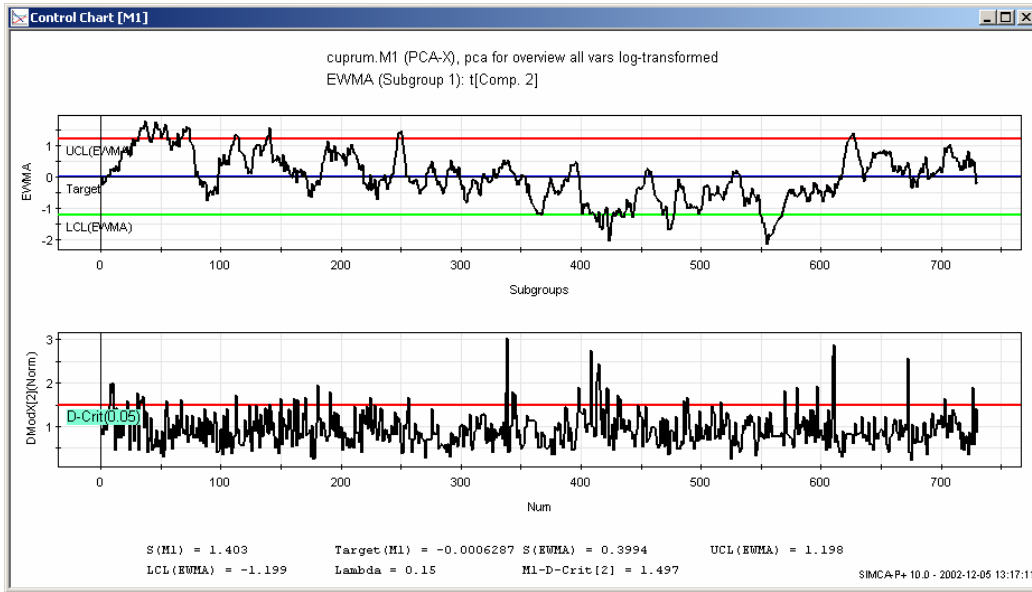
CUPRUM - MSPC monitoring

- Weak trends are best detected with CuSum charts on the score vectors
 - Here, two periods with deviations from the target can be detected



CUPRUM - MSPC monitoring

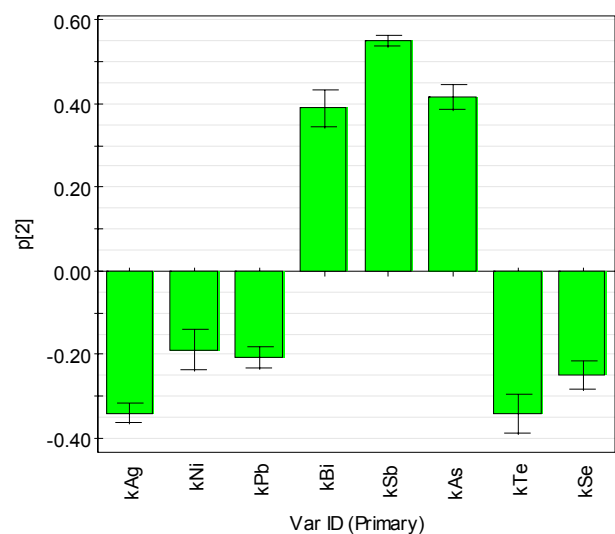
- The trend seen in the CuSum can also be seen in an EWMA chart, here $\lambda = 0.15$



CUPRUM - MSPC interpretation

- The next interesting part is to examine why the process behaves in a certain way
- The interpretation of drift in score vector t_2 is done by studying the loading vector 2
 - In this case the deviations were due to changes in contamination pattern. Probably due to changes in raw material.

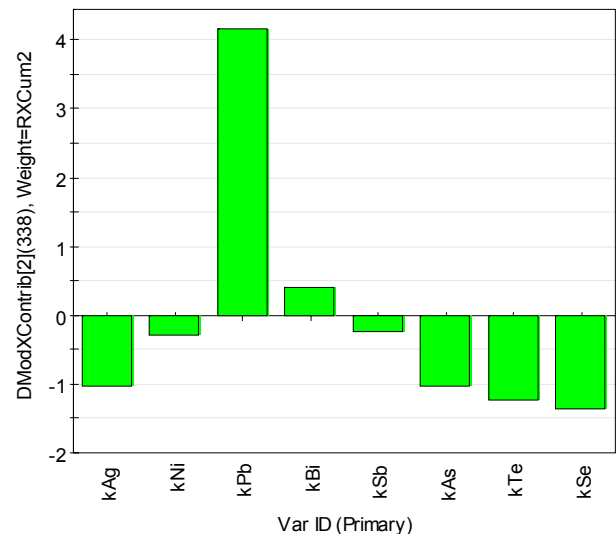
cuprum.M1 (PCA-X), pca for overview all vars log-transformed
p[Comp. 2]



CUPRUM - MSPC interpretation

- To explain why a single observation deviates from the model we use a contribution plot
- Contribution plots can be used both for deviations in scores and in residuals (DModX)
- In the DModX observation 338 seemed to be far away from the model
 - This sample has too high Pb-value in relation to the rest of the variables

cuprum.M1 (PCA-X), pca for overview all vars log-transformed
DModX Contrib(Obs 338), Weight=RX[2]



Multivariate Control Charts by PCA and PLS

Using t 's, T^2 , and DModX in control charts allows us to:

- Track the process over time
- Use all variables simultaneously
- Identify region where the process is operating normally
- Detect when the process starts to go out of control
- Identify anomalous process points
- Interpret the upsets

Conclusions - Advantages of MSPC

- Reduction of dimensionality
 - Few, new variables (scores) summarise all the information contained in the heap of variables describing the process over time, providing a model of the system
- Graphical display of the state of the process
 - Few plots (score plots) display the state of the process over time
 - Plots retain simplicity of interpretation and presentation
- MV control region
 - One can identify a multivariate region where the process is OK and under control
- Understanding
 - Loading and contribution plots give an identification of important variables

Batch Statistical Process Control, BSPC

Contents

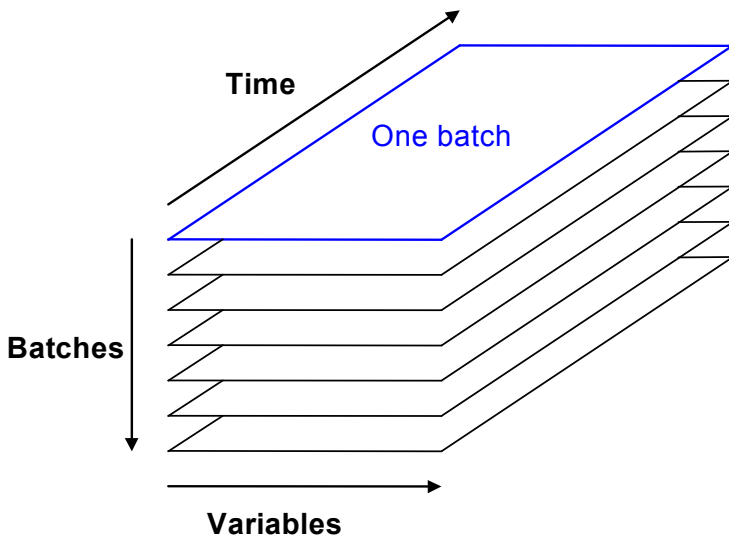
- Introduction to batch modelling
- Organisation of batch data
- Example: Baker's yeast production
- Two levels of batch modelling
 - observation level
 - batch level
- Stretching and shrinking "time"
- Tracing batch evolution
- Diagnosing upsets
- Batch modelling in practice
- Summary

Batch processes

- A batch process is a **finite duration process** (e.g., batch industry, metabonomics, QSAR, ...)
- The results depend on
 - the initial conditions
 - the evolution of the batch
 - interference during the batch evolution
- To model and monitor batches we need data concerning
 - initial conditions **Z** (sometimes absent)
 - data measured during their evolution **X**
 - data describing the interference
 - measurements of the results **Y** (sometimes absent)

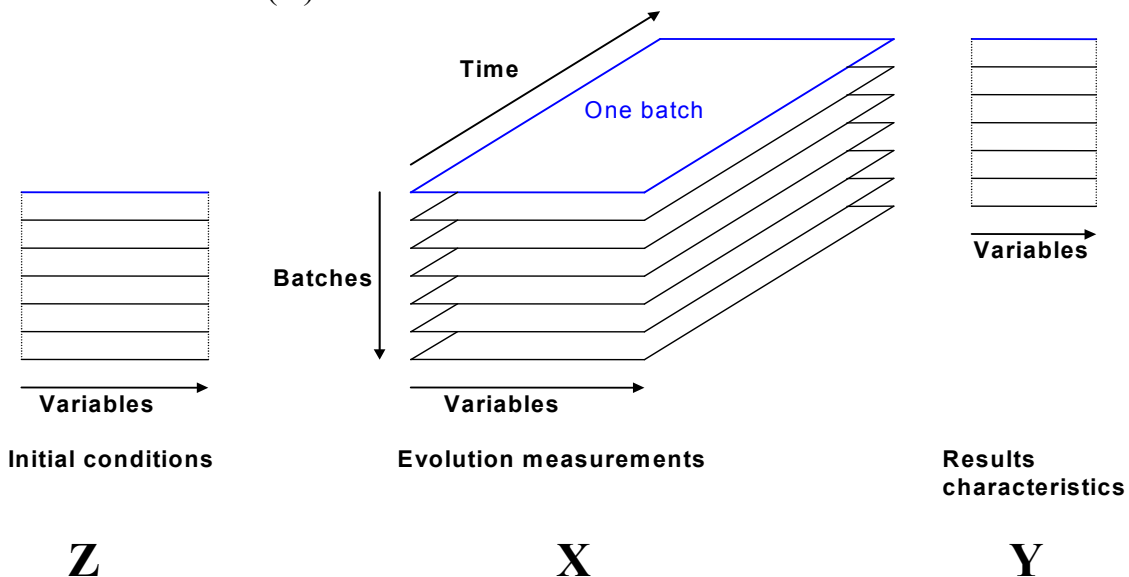
Batch X-data form 3-way tables

- We have a $K * N * B$ data table
 - K variables
 - N sampling times
 - B batches
- 3-way tables can be analysed with
 - PCA when only X data
 - PLS with response data Y



Three blocks of data

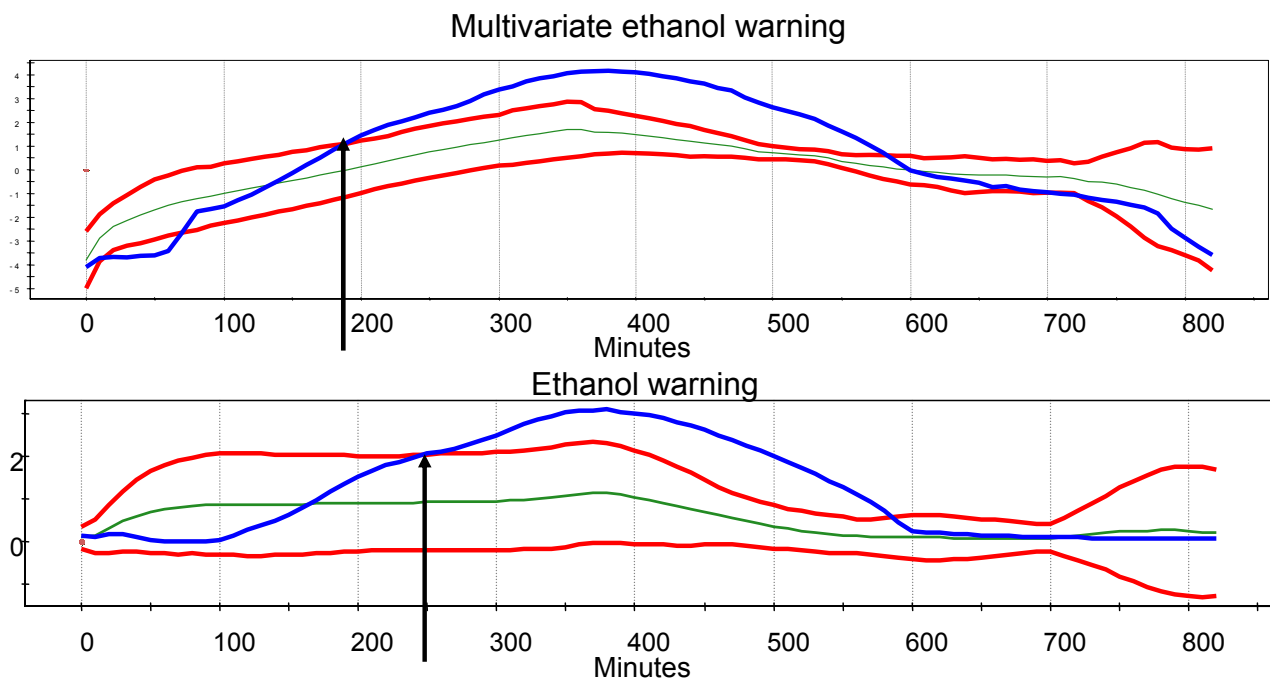
- In the general case, there are three blocks of data
 - Initial conditions data (Z)
 - Evolution data (X)
 - Results data (Y)



Baker's yeast production

- Data come from Jästbolaget AB in Sweden
 - The production of the final product took 14 hours
 - There were 33 batches, of which 20 were selected as reference batches
 - Each batch showed variability due to molasses used, temperature, pH etc.
-
- **Can the process be monitored efficiently by multivariate methods?**

Too much ethanol is a problem!



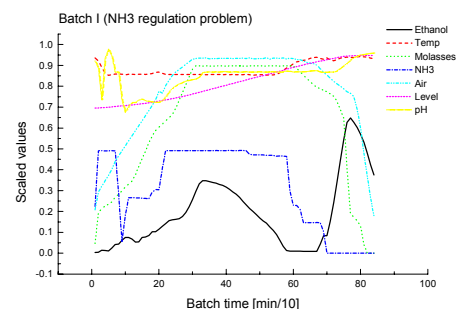
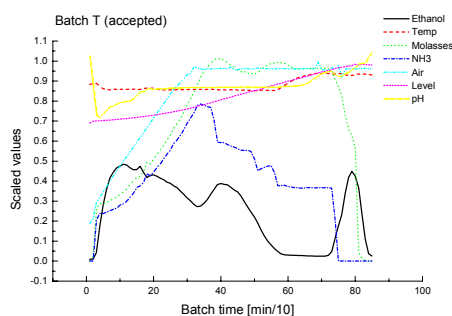
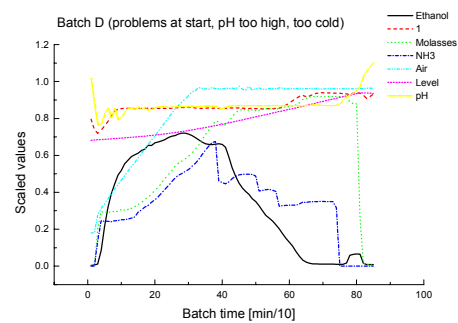
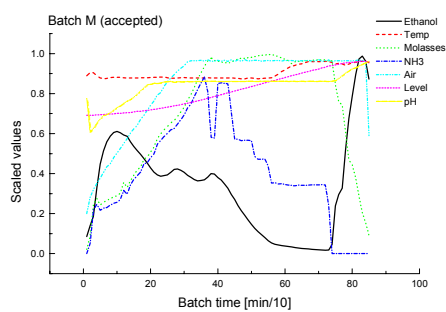
It is important to get an early warning

Multivariate warning occurs 1h before univariate warning !

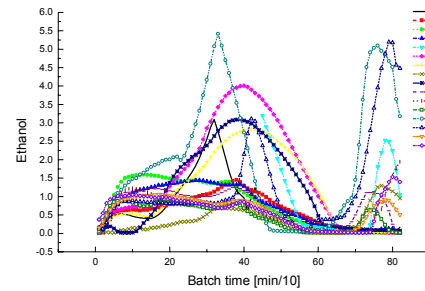
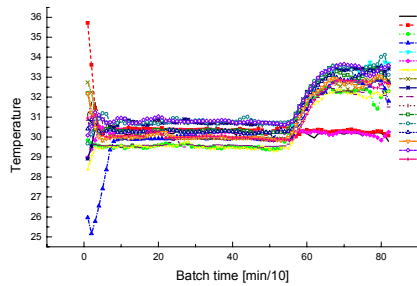
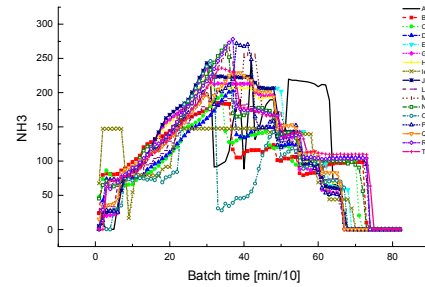
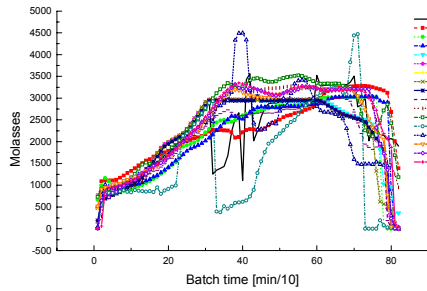
7 controlled/monitored variables

- Ethanol monitored
 - Temperature controlled
 - Feed of molasses controlled, $f(\text{quality of molasses})$
 - NH₃ feed controlled, $f(\text{feed of molasses})$
 - Air flow controlled
 - Level in tank monitored
 - pH controlled
- Data were sampled every 10 minutes. A batch took 14 hours, resulting in 84 data points per batch

Accepted batches (left) and not (right)



Variables from good and bad batches



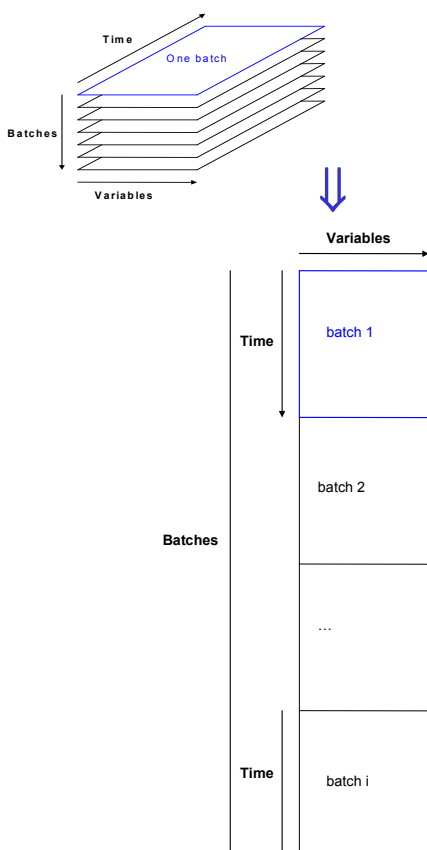
We want early fault detection and classification

- It is not easy to separate good and bad batches by means of the raw data
- We want to detect irregularities as soon as possible in order to have time to make corrections before it is too late
- **The solution: Combine multivariate modelling with SPC (Statistical Process Control), i.e., use MSPC/BSPC**

Two levels of batch modelling

- Observation level
 - looks at each individual observation
 - maturity prediction
 - progress monitoring
- Batch level
 - looks at all available data for the whole batch
 - results prediction

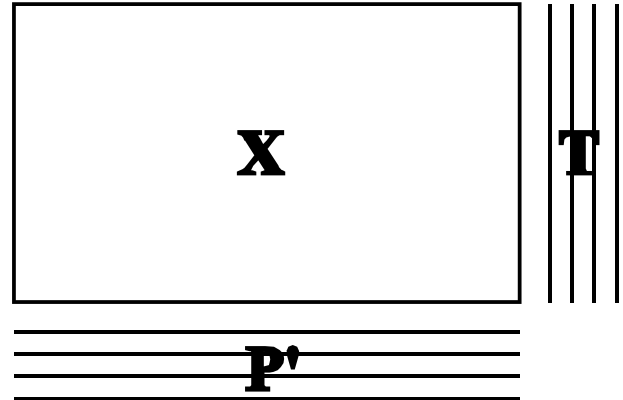
PCA - observation level



- The easiest way to analyse the 3-way table is to unfold the data to a 2-way table where the data from each batch follows the other, one below the other (variable direction preserved)
- PCA on such a table will show how the individual observations relate to each other

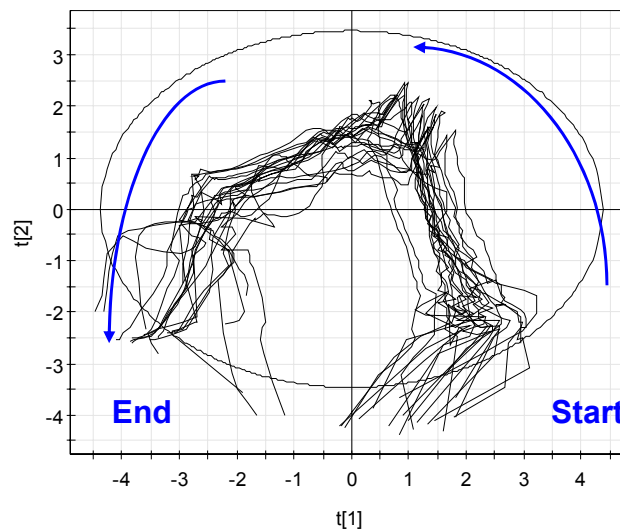
The data are modelled with PCA

- A model is computed that summarises X
- The resulting scores show how the batches evolve in the multivariate process space
- Deviating batches are easily detected in the score plots
- Accepts missing data
- Robust to noise, changes in observations and variables



Batches in multivariate process space

Bakers Yeast Primary.M2 (PCA-X), PCA 20 reference batches
t[Comp. 1]/t[Comp. 2]



We can see how the batches proceed in space as through a **curved tunnel**. Outliers are easily spotted

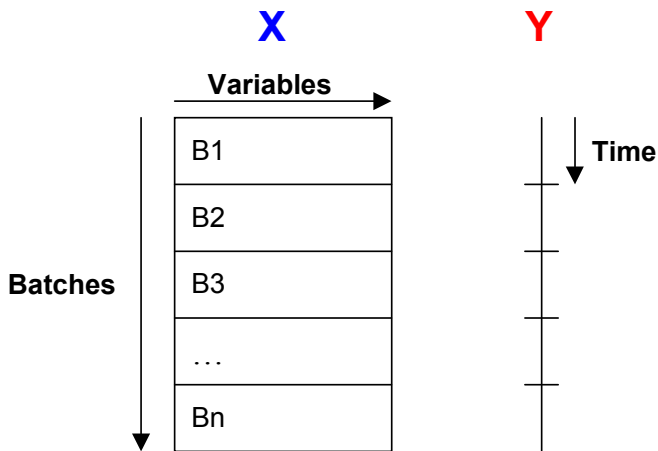
PLS is used at both levels

- Observation level
 - predict state of the process
 - predict phase (in lack of phase variable)
- Batch level
 - predict result variables

Stretching and shrinking of the “time”

- Batches may develop with different speed. Then time needs to be “normalised”, e.g., assigned according to the status of the process
- If a “maturity” variable, e.g., developed energy or volume, is present, a maturity index can be calculated and used instead of chronological time. The data can then be synchronised with respect to the maturity index
- If a “maturity” variable is lacking, the first score can often be used as a substitute maturity variable

PLS - observation level

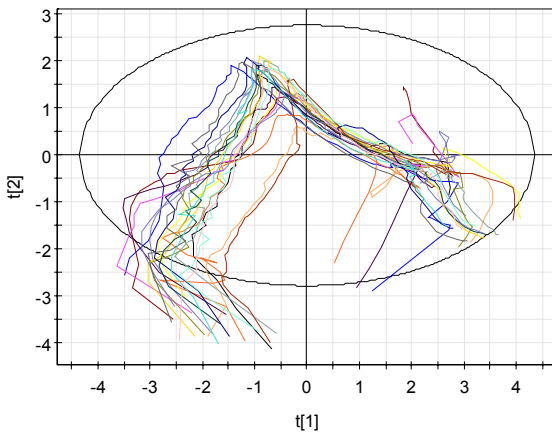


- Each row has the data from a single observation
- The batches follow each other
- Maturity (or time) is used as Y variable
- The resulting scores are new variables that capture
 - t_1 : linear relation to Y
 - t_2 : quadratic relation to Y
 - t_3 : cubic relation to Y

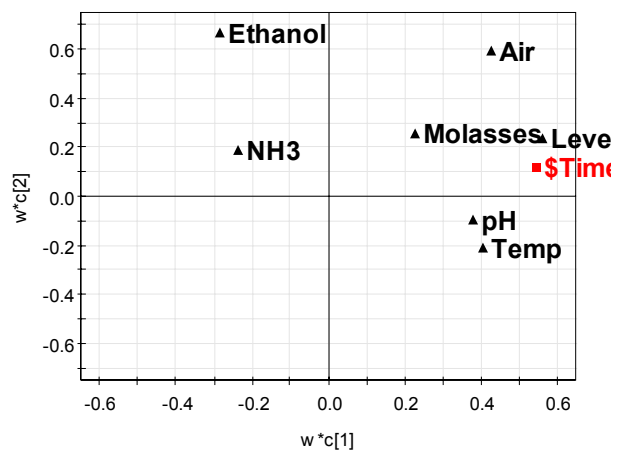
Scores and loadings of observation level PLS model

- Local batch time is positively correlated with level in tank, air flow, pH, and temperature. The response variable is little correlated with feed of molasses, ethanol content, and feed of NH_3 .

Bakers Yeast Primary.M1 (PLS), PLS 20 reference batches
t[Comp. 1]/t[Comp. 2]; Colored according to Batches

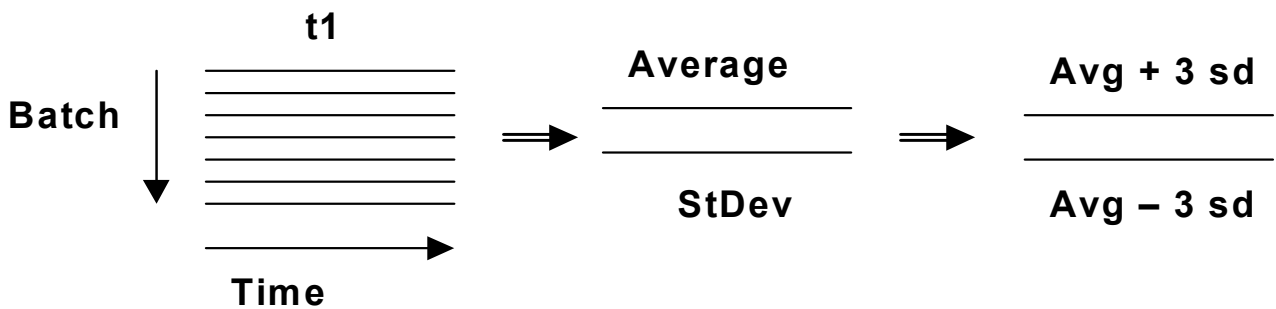


Bakers Yeast Primary.M1 (PLS), PLS 20 reference batches
w*c[Comp. 1]/w*c[Comp. 2]



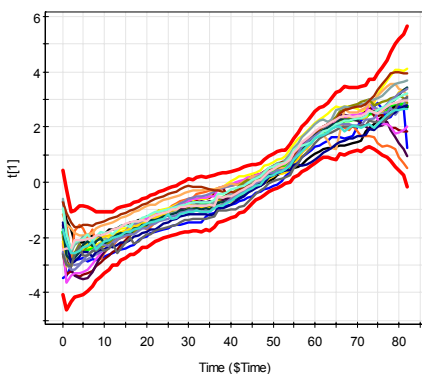
Re-arranging of scores and statistics computation

- Each score is re-arranged batch-wise. The averages and SD.s are calculated over the maturity index (time) of the batches

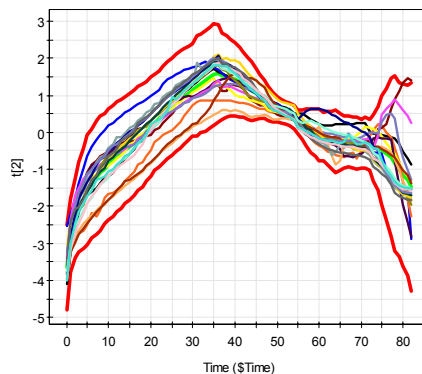


Batch control charts – scores $t_1 - t_3$

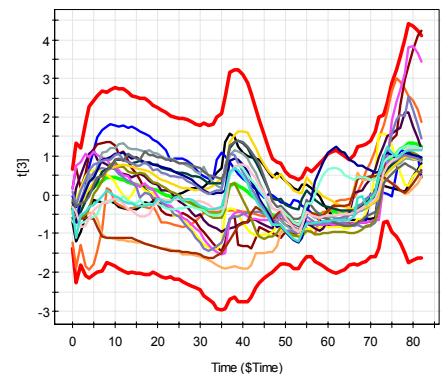
Bakers Yeast Primary.M1 - Scores [comp. 1] (Aligned)



Bakers Yeast Primary.M1 - Scores [comp. 2] (Aligned)



Bakers Yeast Primary.M1 - Scores [comp. 3] (Aligned)



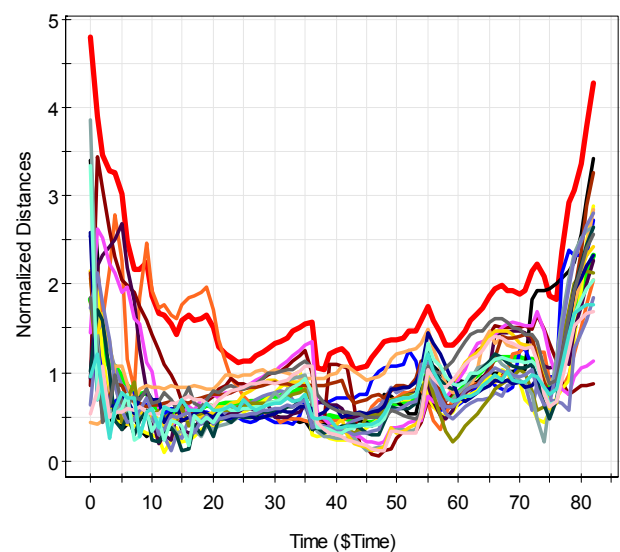
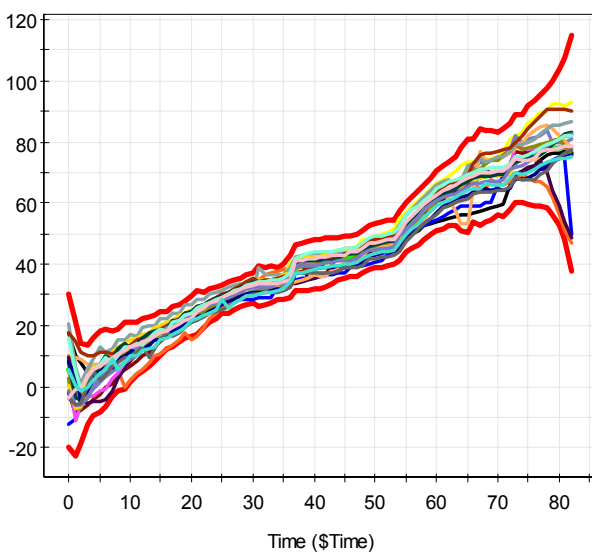
- Established using the 20 reference batches

Not only the scores can be monitored

- Apart from the scores the same procedure is made for
 - Hotelling's T^2 (summarizes all t.s)
 - Predicted time (maturity)
 - Distance to the model plane (residuals)
- If the predicted time (maturity) is higher/lower than the actual, the batch is progressing too fast/slow
- If the distance to the model is too high, the correlation structure in the data has changed. This is commonly the most sensitive indicator of process upsets

Batch control charts – Predicted time (maturity) & DModX

Bakers Yeast Primary.M1 - Observed vs. Predicted Time (\$Time) (Align) Bakers Yeast Primary.M1 - Distance to Model X (Aligned)

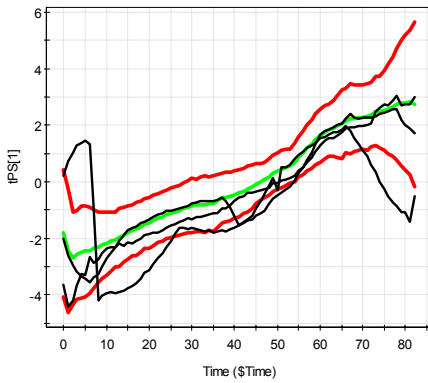


- Control charts of predicted time (left) and DModX (right) using the 23 reference batches

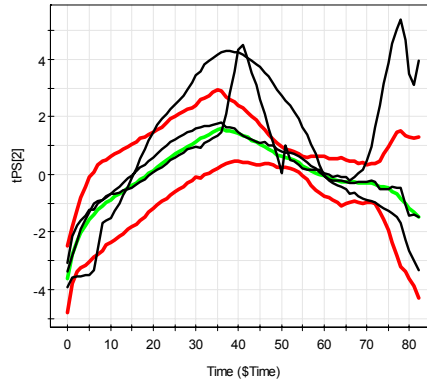
Monitoring the evolution of new batches - scores

- New batches can be monitored in these plots as they evolve, and deviations interpreted (contribution plots) – For clarity only three test batches are plotted

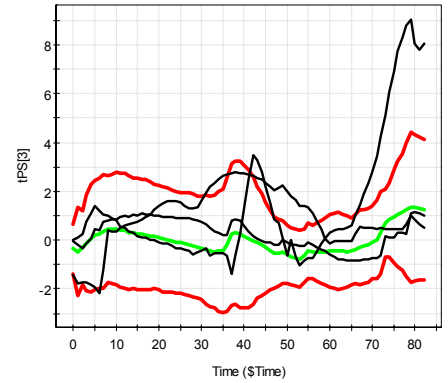
Bakers Yeast Primary.M1 - Predicted Scores [comp. 1]



Bakers Yeast Primary.M1 - Predicted Scores [comp. 2]



Bakers Yeast Primary.M1 - Predicted Scores [comp. 3]

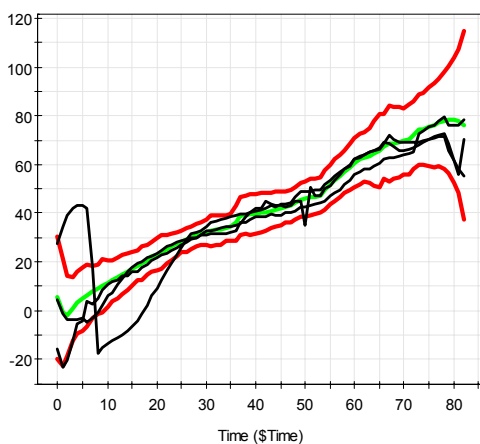


- Predictions for batches Da, Ja, and Pa, with entirely different evolutionary profiles

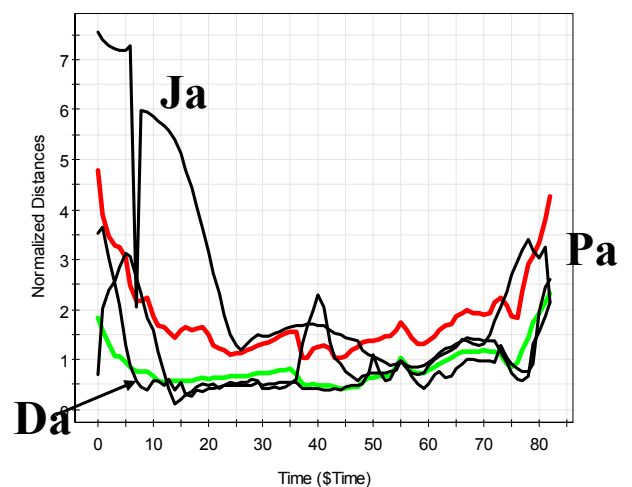
Monitoring the evolution of new batches – Maturity/DModX

- Except for a slight deviation of batch J in the beginning, all three batches behave well in the maturity control chart

Bakers Yeast Primary.M1 - Predicted Observed vs. Predicted Time (\$Time)



Bakers Yeast Primary.M1 - Predicted Distance to Model X

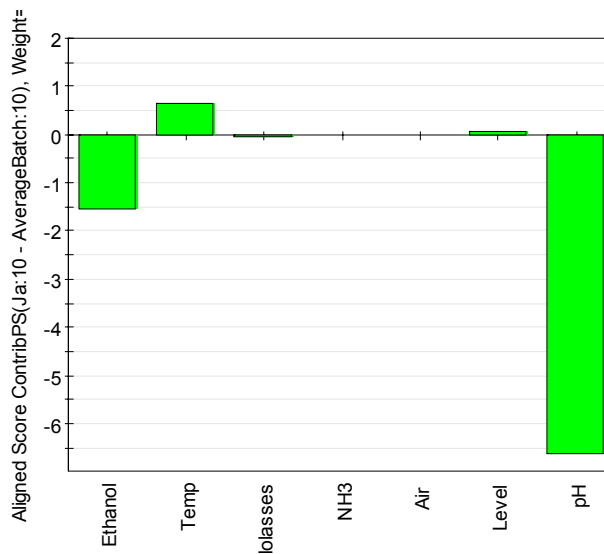


- The DModX chart shows that batch Ja deviates at the beginning, and batch Pa towards batch completion. Batch Da is basically OK all the time.

Finding variables contributing to deviation from "normality"

- Comparison of "bad" (Ja or Pa) and "good" (average) batches
- Early deviation of batch Ja (at local batch time 10)

Bakers Yeast Primary.M1 (PLS), PLS 20 reference batches, Score Contrib PS(Ja:10 - AverageBatch:10), Weight=p[1]

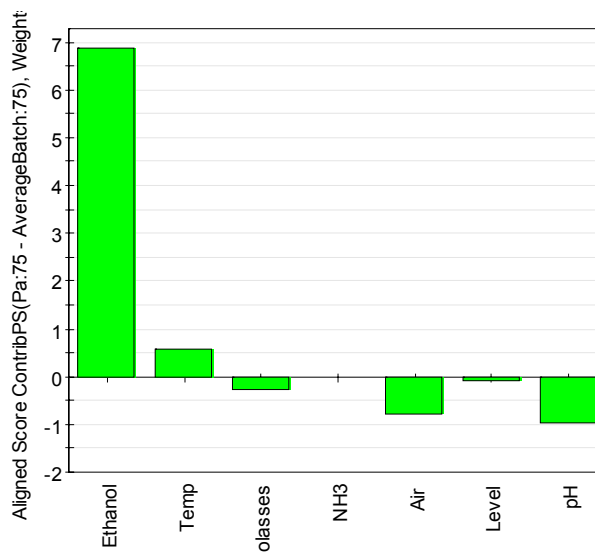


pH is low shortly after batch initiation for batch Ja

Finding variables contributing to deviation from "normality"

- Comparison of "bad" (Ja or Pa) and "good" (average) batches
- Late deviation of batch Pa (at local batch time 75)

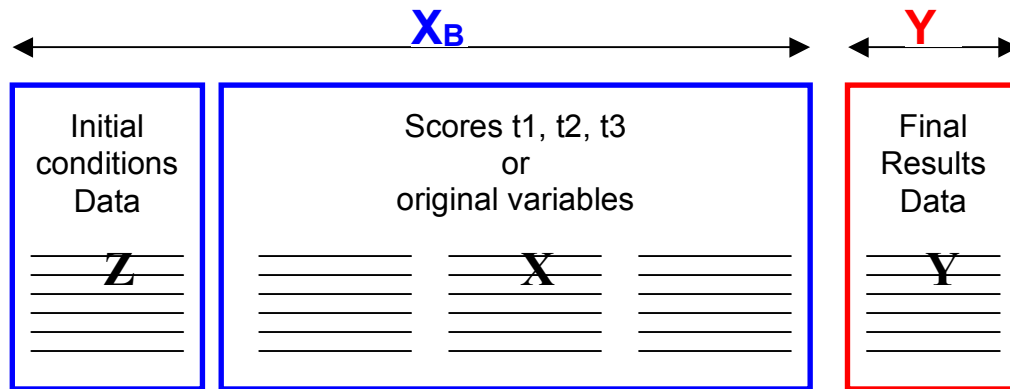
Bakers Yeast Primary.M1 (PLS), PLS 20 reference batches, Score Contrib PS(Pa:75 - AverageBatch:75), Weight=p[1]



Ethanol content higher than normal for batch Pa towards the end

PLS - batch level: Modelling final batch results

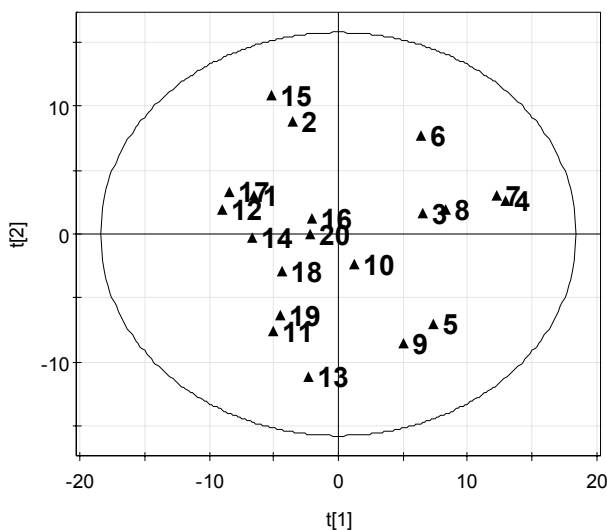
- Z = Initial conditions: Innoc (= total amount of dry substance added)
- X = Unfolded batch data: Score variables of t_1 , t_2 , and t_3
- Y = Final results: Amount of yeast & Yield (Amount corrected for amount of molasses used)
- Sub models can be made on initial data plus data from time 1, time 1-2, time 1-3, ... time 1-T. May be applied consecutively in the evolution of a batch as new data become available



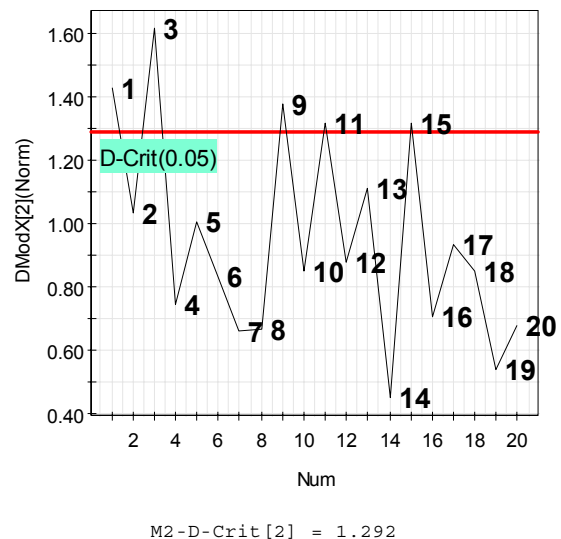
Scores and loadings of batch level PLS model

- PLS was used to relate X_B (initial conditions + obs level scores) to Y (Amount & Yield); $R^2X = 0.47$; $R^2Y = 0.61$, $A = 2$

Bakers Yeast Primary - batch level.M2 (PLS)
t[Comp. 1]/t[Comp. 2]



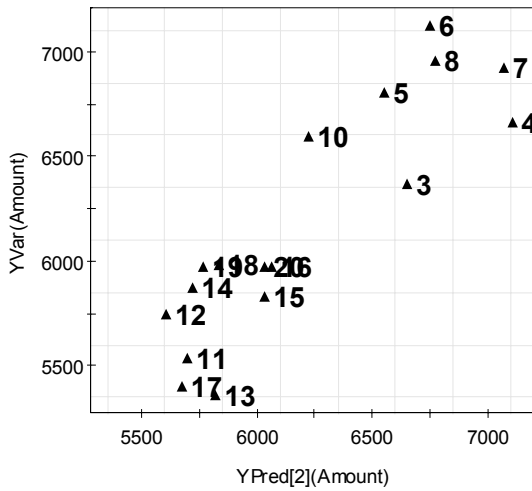
Bakers Yeast Primary - batch level.M2 (PLS)
DModX[2](Norm)



Y(obs) / Y (pred) of batch level PLS model

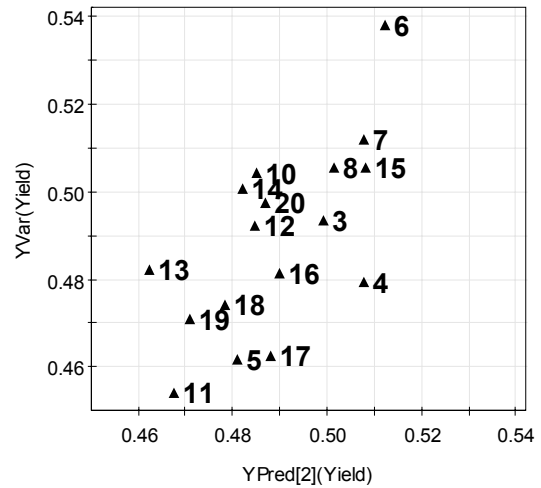
- PLS was used to relate X_B (initial conditions + obs level scores) to Y (Amount & Yield); $R^2X = 0.47$; $R^2Y = 0.61$, $A = 2$

Bakers Yeast Primary - batch level.M2 (PLS)
YPred[Comp. 2](YVar Amount)/YVar(YVar Amount)



RMSEE = 285.096

Bakers Yeast Primary - batch level.M2 (PLS)
YPred[Comp. 2](YVar Yield)/YVar(YVar Yield)

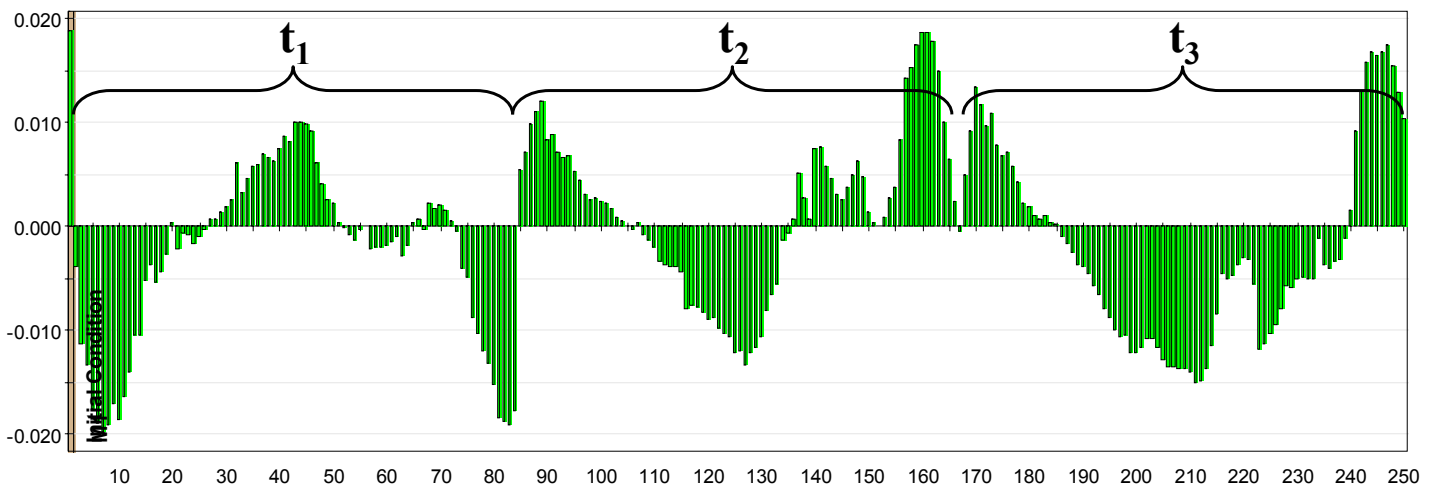


RMSEE = 0.0170927

Model interpretation of batch level model

- PLS regression coefficients

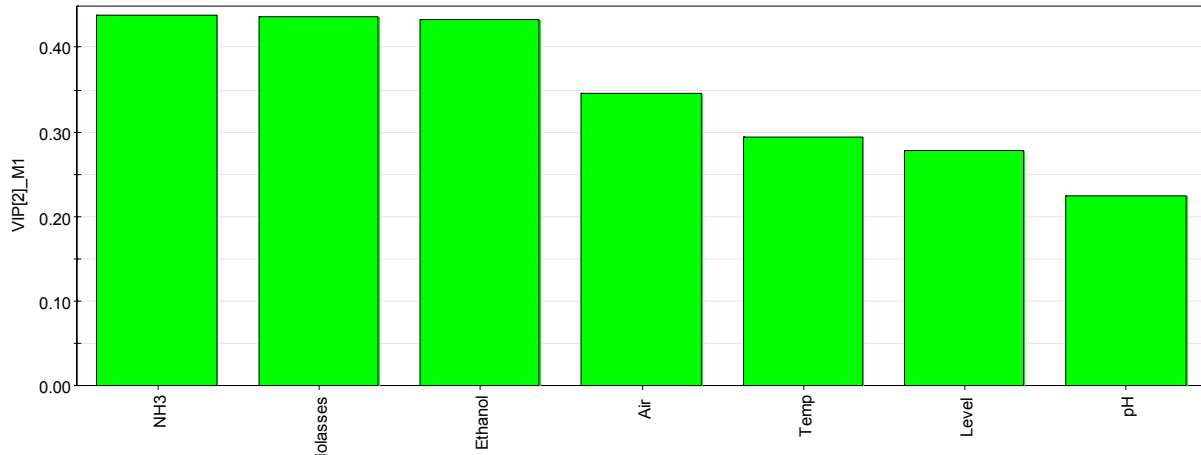
Bakers Yeast Primary - batch level.M2 (PLS), PLS 20 ref batches Amount and Yield
CoeffCS[Comp. 2](YVar Amount)



Model interpretation of batch level model

- Batch Variable Importance (provides an absolute measure)

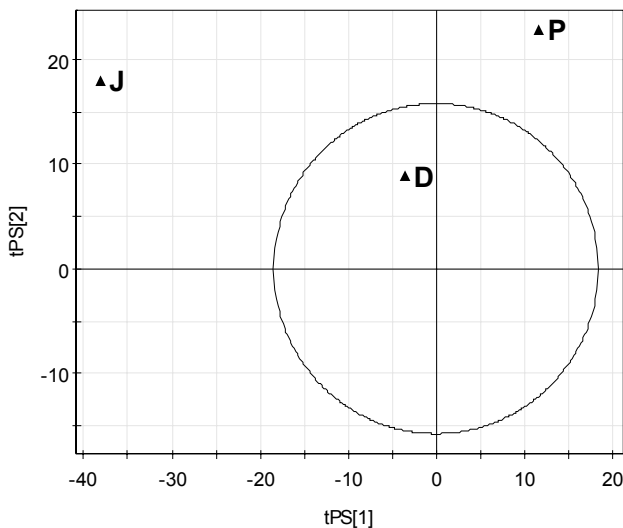
Bakers Yeast Primary - batch level.M2 (PLS), PLS 20 ref batches Amount and Yield
Batch Variable Importance[Comp. 2](YVar Amount), Phase M1



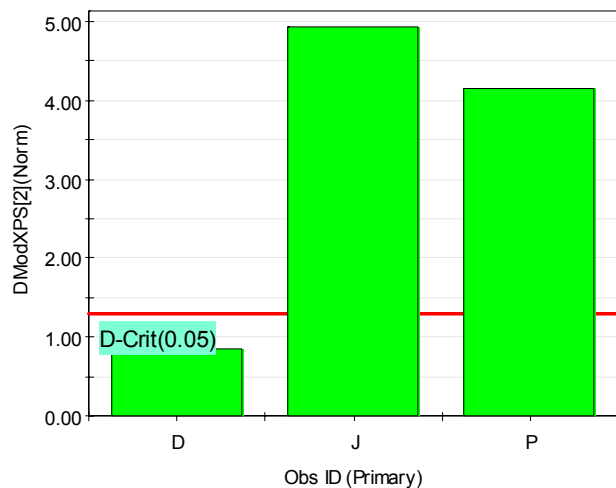
Predictions of new batches

- Predicted PLS scores and DModX; For clarity only three test batches are plotted

Bakers Yeast Primary - batch level.M2 (PLS)
tPS[Comp. 1]/tPS[Comp. 2]



Bakers Yeast Primary - batch level.M2 (PLS)
DModXPS[Comp. 2]

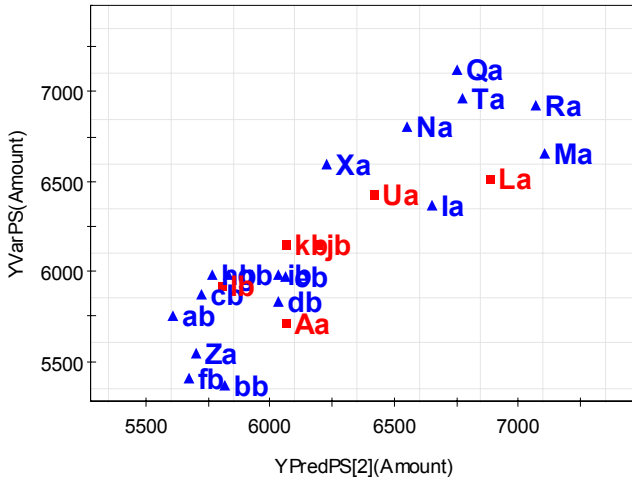
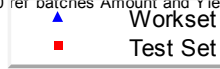


M2-D-Crit [2] = 1.292

Predictions of Y-data of new batches

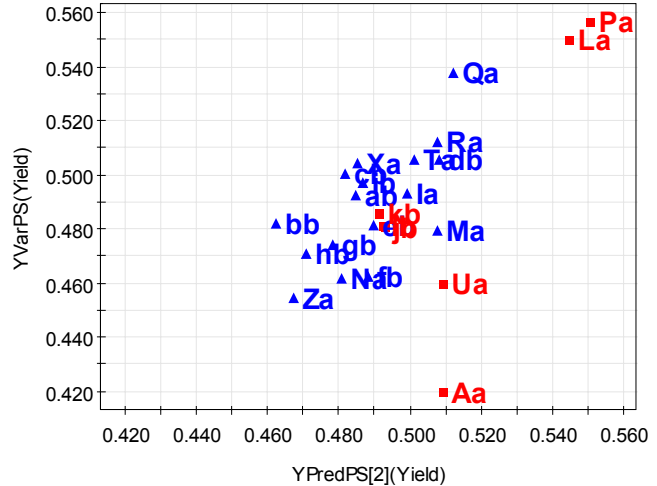
- Five new batches do not fit the model; Predicted Y-values are shown for the remaining eight batches.

Bakers Yeast Primary - batch level.M2 (PLS), PLS 20 ref batches Amount and Yield



RMSEP = 216.043 $Q^2_{ext} = 0.67$

Bakers Yeast Primary - batch level.M2 (PLS), PLS 20 ref batches Amount and Yield



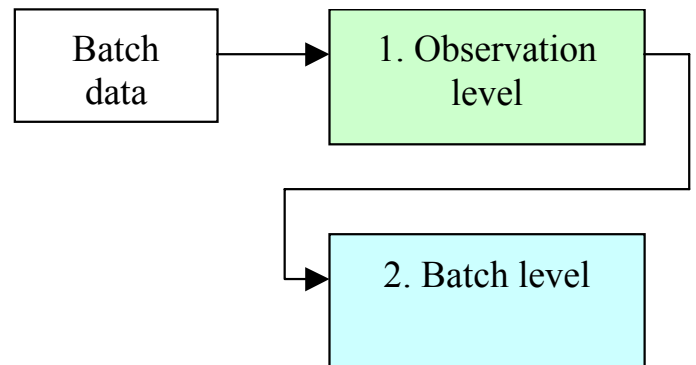
RMSEP = 0.0392013 $Q^2_{ext} = 0.52$

Summary - Batch modelling

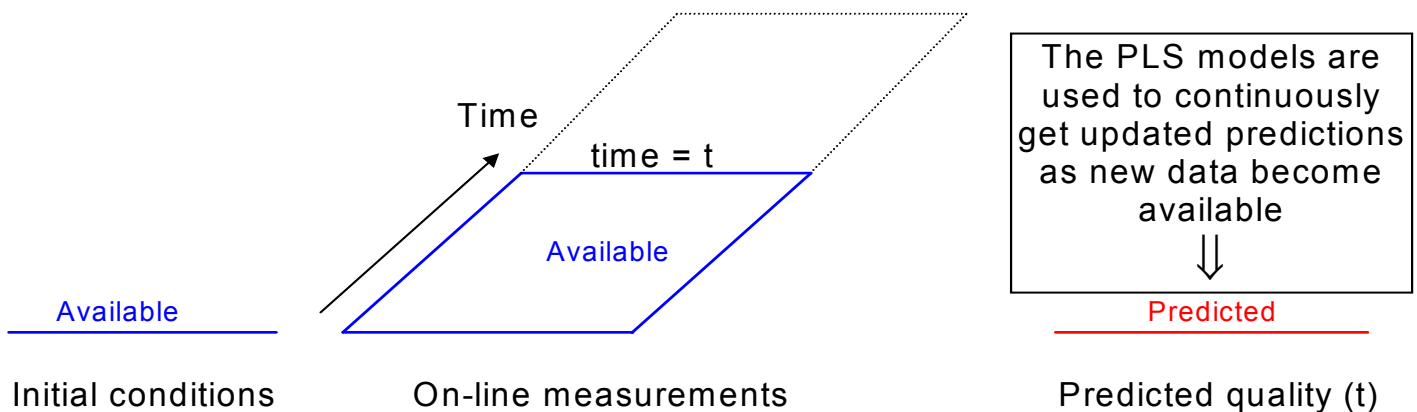
- Models are developed from a set of accepted batches
- These models provide a powerful tool to monitor new batches as well as to make on-line predictions
- The simplicity of presentation and interpretation of common SPC charts is retained despite the multitude of variables measured
- Diagnostic information is obtained with a mouse click

Summary - Batch modelling in practice

- Modelling and execution are made on two levels
 - **Observation level**
 - working with individual observations
 - monitoring the evolution of the batch
 - classifying current phase
 - **Batch level**
 - working with the whole of the batch
 - predicting the outcome of the batch



Summary - On-line results predictions



- In an on-line application new data are fed to the PLS batch level models as they become available
- Predictions with confidence intervals are computed and presented

Multivariate Data Analysis and Modelling Basic Course

Chapter 14: Exercises



Overview of Exercises Layout

- Each exercise contains the following headlines
 - *Background* (Why this investigation?)
 - *Objective* (What is the goal/objective with the exercise?)
 - *Data* (Description of X and Y and observations, originator(s) and literature source(s))
 - *Tasks* (What you are expected to do in this exercise)
 - *Solutions* (A proposed solution to the tasks given)
 - *Conclusions* (Emphasising main points of the exercise)
- Please do not hesitate to ask the course instructor(s) for help/advice
- Remember that our solutions are just proposals; other alternatives might exist...

Exercises – Part I

- Getting started
 - *FOODS*, overview of European food consumption profiles
 - *IRIS*, Classification of Iris flowers
- Easy PCA
 - *ARCHAEOLOGY*, Classification of soil samples
 - *METABONOMICS*, Investigation of Phospholipidosis
- Easy PLS
 - *LOWARP*, Polymer production using multiple responses
 - *USDVOLVO*, How to buy a second hand car!
- Quality Control
 - *THICKNESS*, Quality control of polymer disk manufacturing
 - *CUPRUM*, Multivariate quality monitoring of an electrolysis process

Exercises – Part II

- Multivariate Characterization
 - *SURFACTANT*, QSAR/QSPR modelling of surfactants
 - *PULP*, Modelling and prediction of pulp quality
- Multivariate Calibration
 - *SUGAR*, Multivariate calibration of sugar quality using fluorescence data
 - *NIR_CHIP*, Characterization and classification of wood chips using NIR-data
 - *CELLULOSE*, Modelling the viscosity of cellulose powder using NIR-data
- Process Applications
 - *SOVRING*, Process monitoring of a mineral sorting plant
 - *PROCIA*, MSPC on process data
 - *Baker's Yeast*, BSPC of a batch fermentation process

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

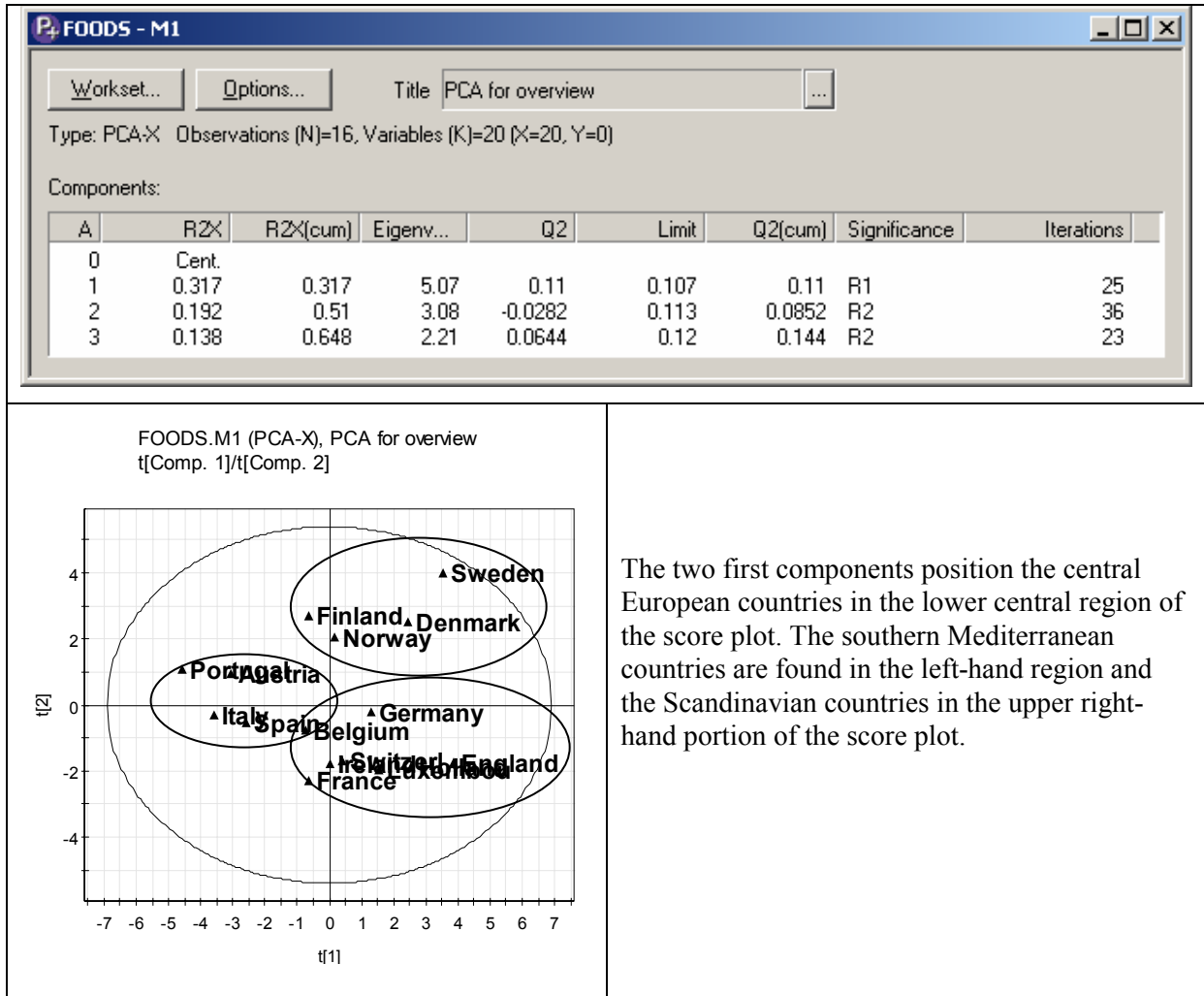
Solutions to FOODS

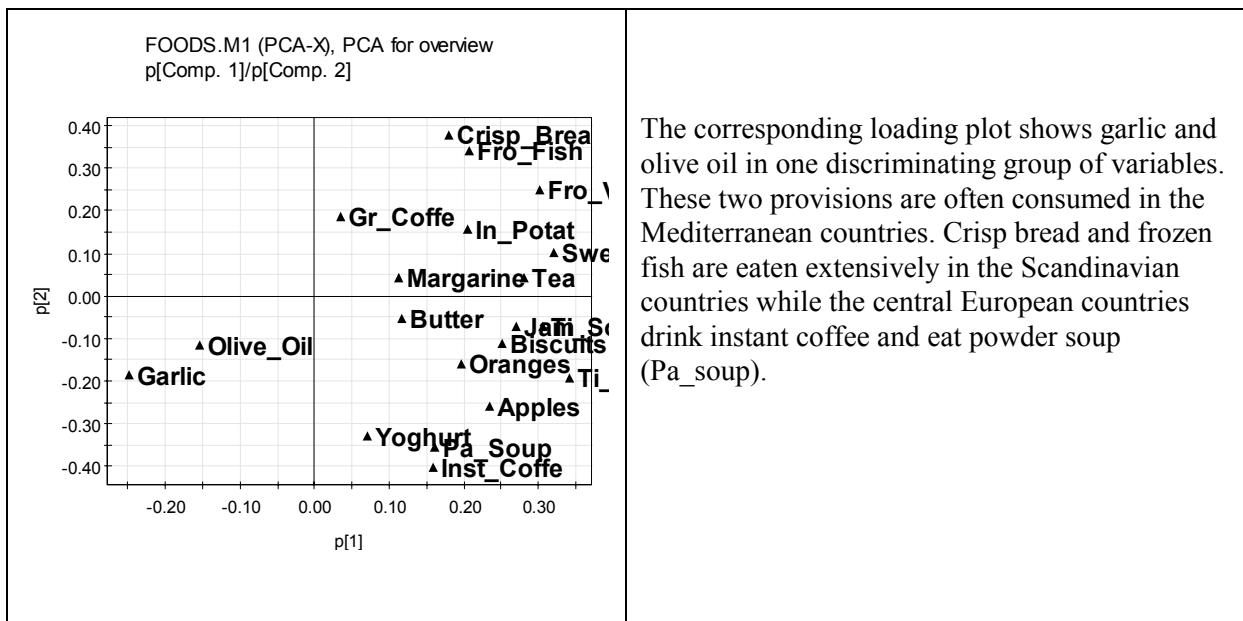
Task 1

There were 3 missing values.

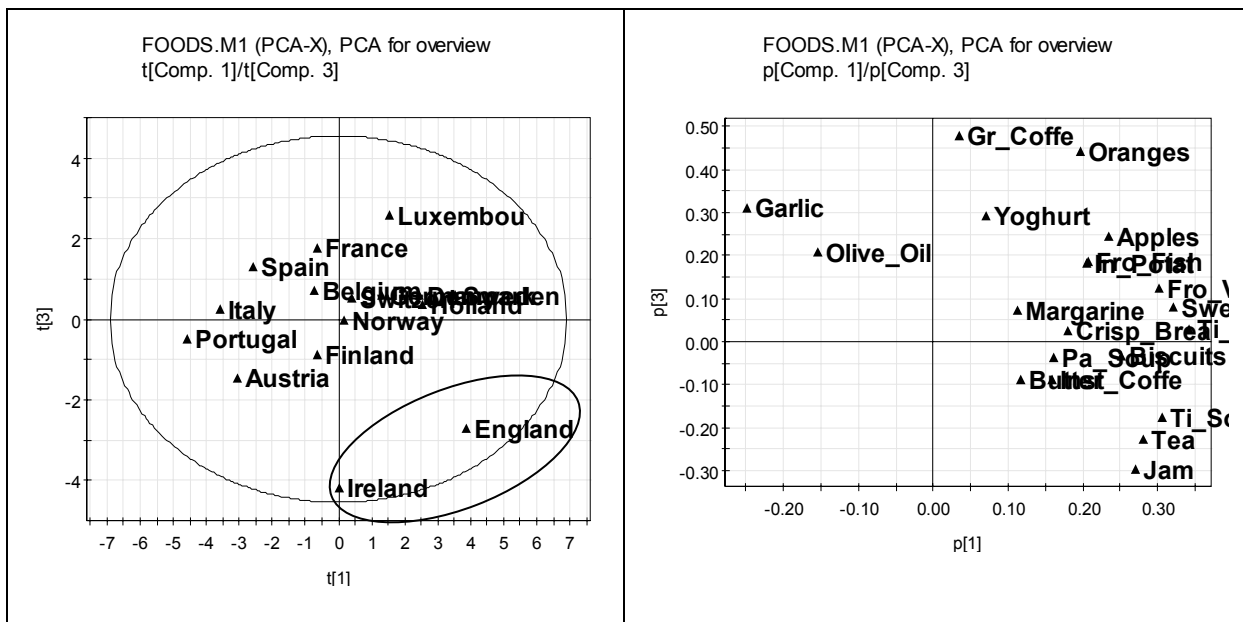
Task 2

A three component PC model was computed:



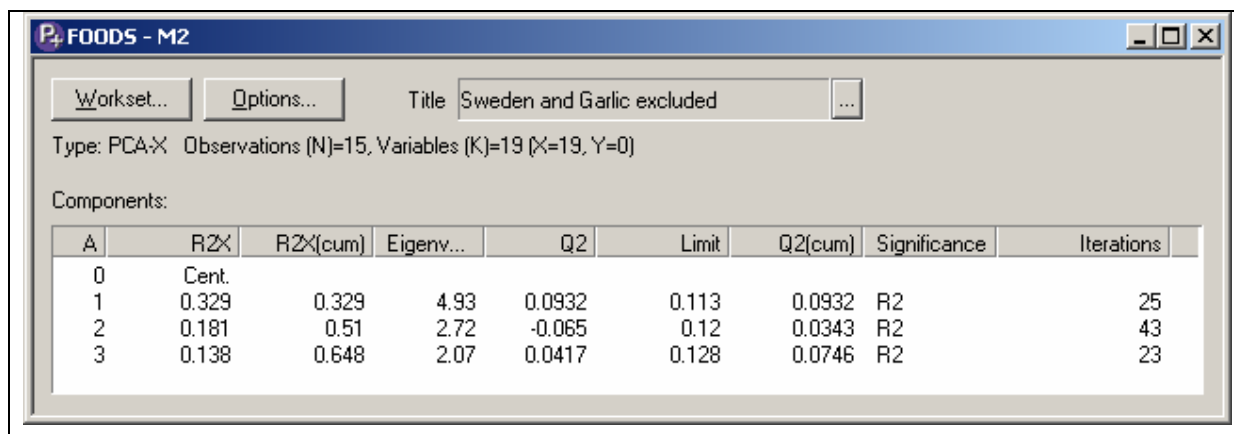


The third component separates England and Ireland from the rest of Europe. We can see the presence of the tea and jam habit, as well as the limited consumption of ground coffee, garlic, and olive oil on these islands.

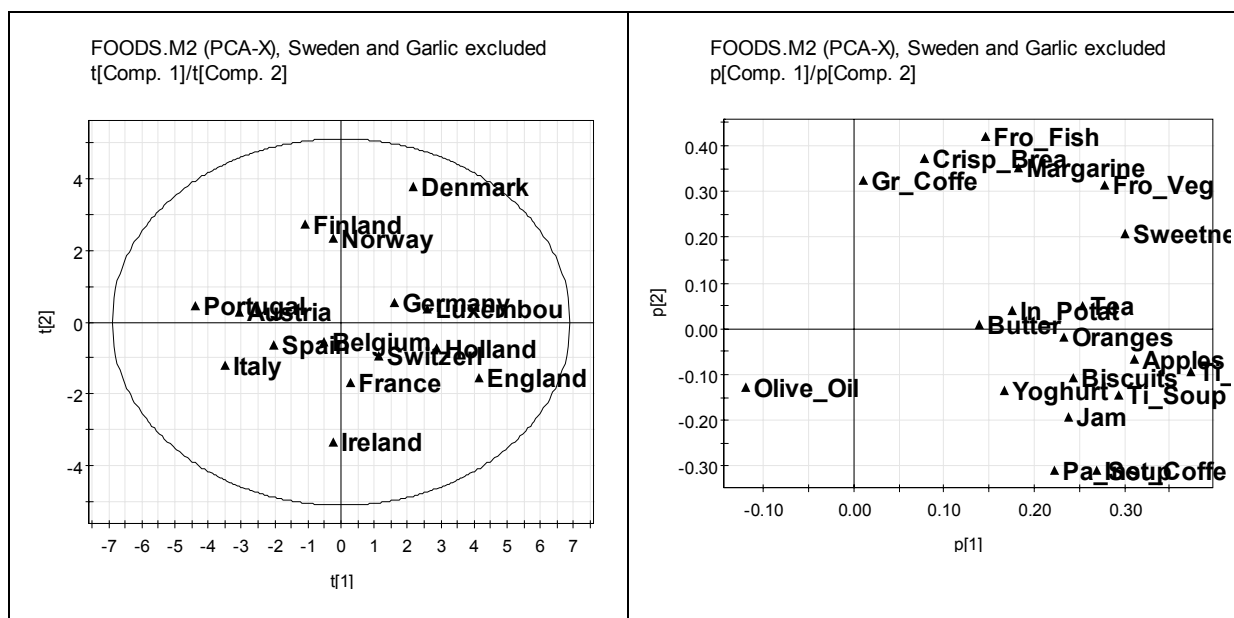


Task 3

A new model was made with Sweden and Garlic excluded.



We here show plots pertaining to the two first components.



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 IRIS

A classical data set in statistics

Background

In statistics there are a number of classical data sets which are used to test different methods and algorithms. IRIS is one of these historical data sets and has its origins in botany.

Objective

In this exercise you will make overview models of a full data set, and on parts of a data set. You should be able to interpret loading plots in terms of the variables that contribute to patterns in the corresponding score plots. You should be able to make separate models for each subgroup of observations and then see how new observations fit these models. This latter approach is called the SIMCA (Soft Independent Modelling of Class Analogy) method.

Data

The data set contains petal (sw: kronblad) and sepal (sw: foderblad) lengths and widths of 50 specimens each of *Iris setosa*, *Iris versicolor* and *Iris virginica*. The great statistician Fisher introduced this data set as early as 1936. It is commonly known as "The Fisher Iris Data" (see below for table header and the first ten observations). We will use 75 observations as training data (stored in IRIS training.xls) and 75 observations as prediction data (stored in IRIS prediction.xls).

Tasks

Task 1

Start a new project in SIMCA (*File|New*) with the Iris data (IRIS training.xls). Check the worksheet colouring and provide a project name. Check that all the training data have been imported; 75 observations and 4 variables.

Task 2

To define classes among the observations, use the command: *Workset|New*. Click on the Tab *Observations*, then click in the observation list, and finally click on the right mouse button and activate secondary observation name.

Observations: 75, Included: 75, Selected: 1

Observation ID	Class	Inc/Exc
1	---	Include
2	---	Include
3	---	Include
4	---	Include
5	---	Include
6	---	Include
8	---	Include
9	---	Include

Observations: 75, Included: 75, Selected: 1

Observation ID	OBSNAM	Class	Inc/Exc
1	Setosa__1E	---	Include
2	Setosa__1E	---	Include
3	Setosa__1E	---	Include
4	Setosa__1E	---	Include
5	Setosa__1E	---	Include
6	Setosa__1E	---	Include

Mark the 25 first observations, choose class 1, and click on *Set*. Continue with the rest of the observations according to the following:

Obs 1-25 inclusive, class 1 (*Setosa*)

Obs 26-50 inclusive, class 2 (*Versicolor*)

Obs 51-75, inclusive, class 3 (*Virginica*)

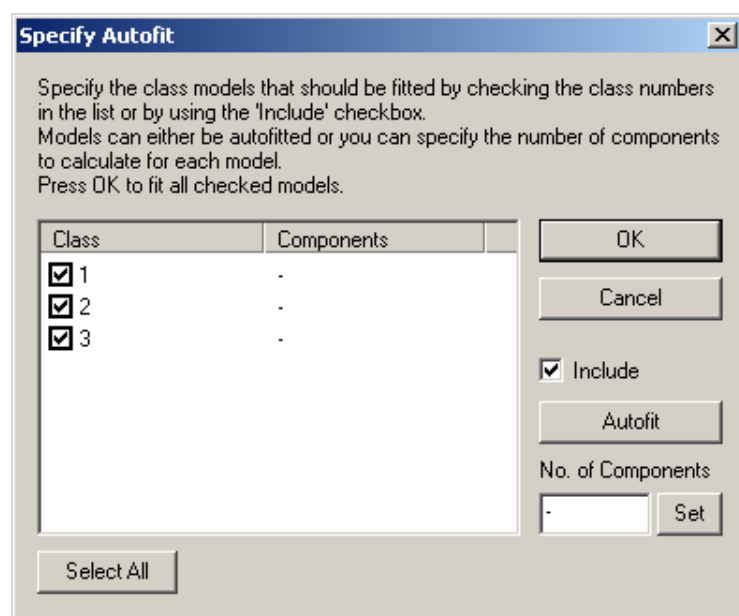
First we are going to make a PCA model on the **entire** training set, i.e. classes 1-3. Go to *Analysis | Change model type* and choose *PCA on X-block*. Make a two-component PCA overview of the data (Hint: *Analysis | Two First Components*). How are the 3 different species grouped? Which variables are responsible for this grouping? Are there any outliers?

Task 3

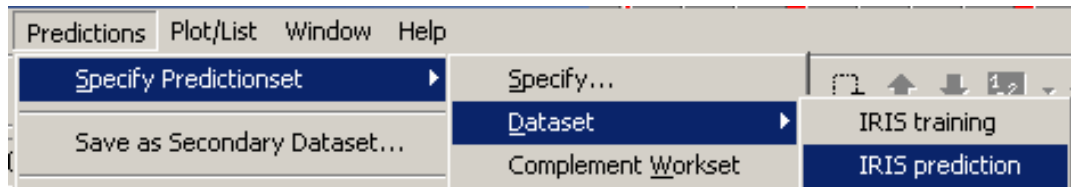
The difference between *Versicolor* and *Virginica* may be made more distinct by creating a PCA model where the *Setosa* observations have been omitted: Make a new PC model, *Work set|New as Model M1*. Include the *Versicolor* (class 2) and *Virginica* (class 3) observations, but exclude the *Setosa* observations (class 1). Use *Analysis|Autofit* to compute the model. You will get a one-component model. However, to be able to make plots we often add a second component to the model (use: *Analysis|Next component*). When we then interpret the plots we may disregard the structure along the 2nd component. Is there a separation between *Versicolor* and *Virginica*? How do they differ?

Task 4

We will now use the SIMCA method and compute separate models for each class of observations. Go to *WorkSet|New as Model M1* (this has to be done to activate class 1 observations again). Press *OK*. Go to *Analysis|Change model type* and choose *PCA Class* and the first class. Autofit the model. Save the model. Repeat this procedure for classes 2 and 3. Save these models. (Alternatively, the procedure specified above can be carried out directly in one step using *Analysis|Autofit Class Models*.)



We will now test the predictive ability of the three class models. For this purpose we have to import the prediction data (i.e. the last 75 observations). Use the command *File|Import Secondary Dataset* and select "IRIS prediction.xls". Press *Open*, *Next* and *Finish*. Then go to *Predictions* and select *Specify Predictionset|Dataset* and select as source the *IRIS prediction* data you just imported.



Plot DModX under *Predictions*|*Distance to Model* for each model and compare the distances. Any overlapping models? (The predictions are made on the model active in the project window). Produce a Coomans' plot (*Predictions*|*Coomans' plot*) for the *Versicolor* model against the *Virginica* model.

Explanation of data set

/ IRIS.XLS Last change 950418

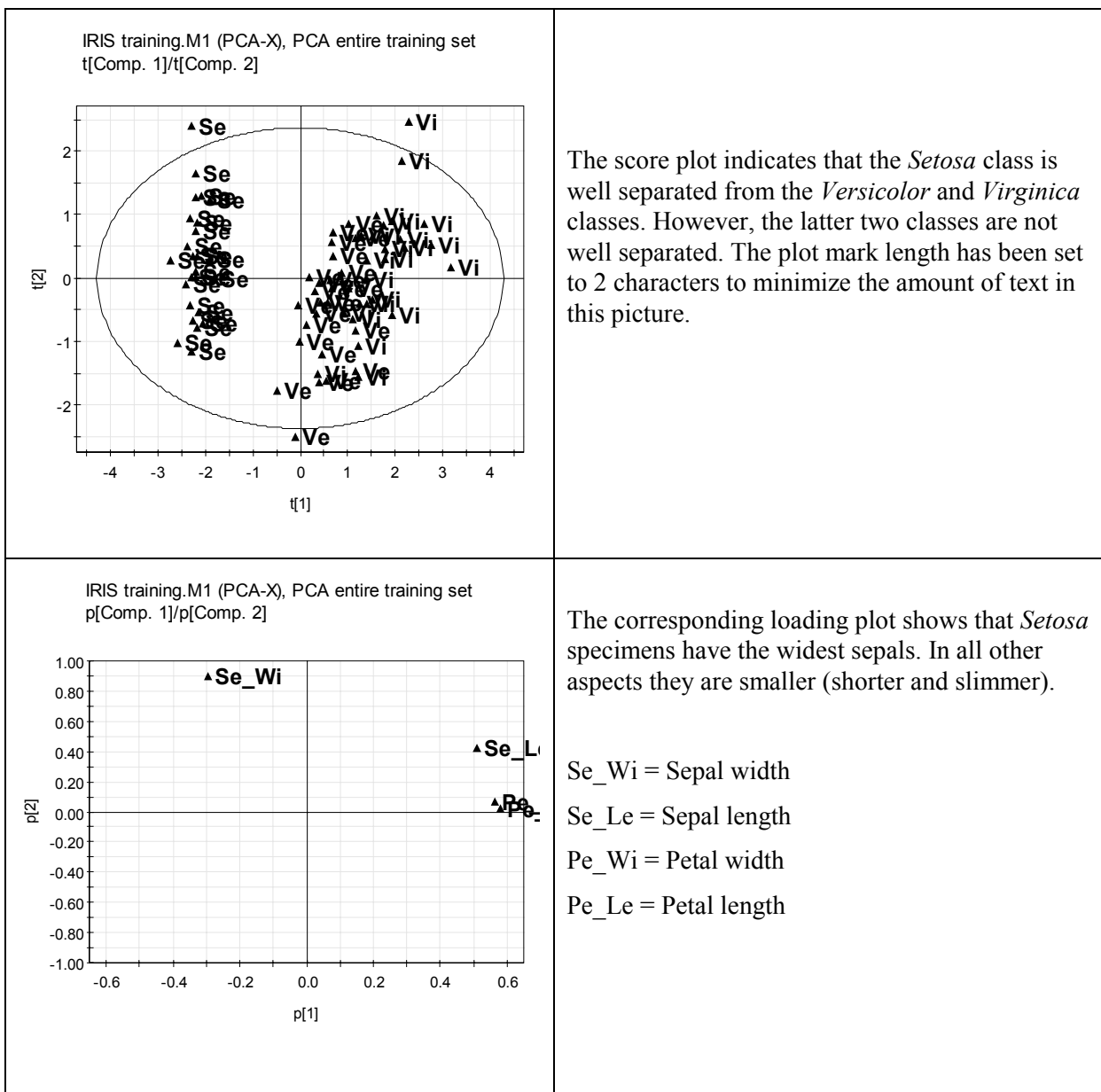
/ No	Name	Sepal Length	Sepal Width	Petal Length	Petal Width
/					
/ Whole data set					
/ Min		4.30	2.00	1.00	0.10
/ Max		7.90	4.40	6.90	2.50
/ Min/Max		0.54	0.45	0.14	0.04
/ Average		5.84	3.06	3.76	1.20
/ StDev		0.83	0.44	1.77	0.76
/					
/ Setosa					
/ Min		4.30	2.30	1.00	0.10
/ Max		5.80	4.40	1.90	0.60
/ Min/Max		0.74	0.52	0.53	0.17
/ Average		5.01	3.43	1.46	0.25
/ StDev		0.35	0.38	0.17	0.11
/					
/ Versicolor					
/ Min		4.90	2.00	3.00	1.00
/ Max		7.00	3.40	5.10	1.80
/ Min/Max		0.70	0.59	0.59	0.56
/ Average		5.92	2.77	4.28	1.33
/ StDev		0.51	0.31	0.48	0.20
/					
/ Virginica					
/ Min		4.90	2.20	4.50	1.40
/ Max		7.90	3.80	6.90	2.50
/ Min/Max		0.62	0.58	0.65	0.56
/ Average		6.59	2.97	5.55	2.03
/ StDev		0.64	0.32	0.55	0.27
/					
ONUM	ONAM	Se_Le	Se_Wi	Pe_Le	Pe_Wi
1	Setosa____1E	5.1	3.5	1.4	0.2
2	Setosa____1E	4.9	3.0	1.4	0.2
3	Setosa____1E	4.7	3.2	1.3	0.2
4	Setosa____1E	4.6	3.1	1.5	0.2
5	Setosa____1E	5.0	3.6	1.4	0.2
6	Setosa____1E	5.4	3.9	1.7	0.4
7	Setosa____1E	4.6	3.4	1.4	0.3
8	Setosa____1E	5.0	3.4	1.5	0.2
9	Setosa____1E	4.4	2.9	1.4	0.2
10	Setosa____1E	4.9	3.1	1.5	0.1

Solutions

Task 2

Two components were obtained. The model explains 96% of the variability in the data.

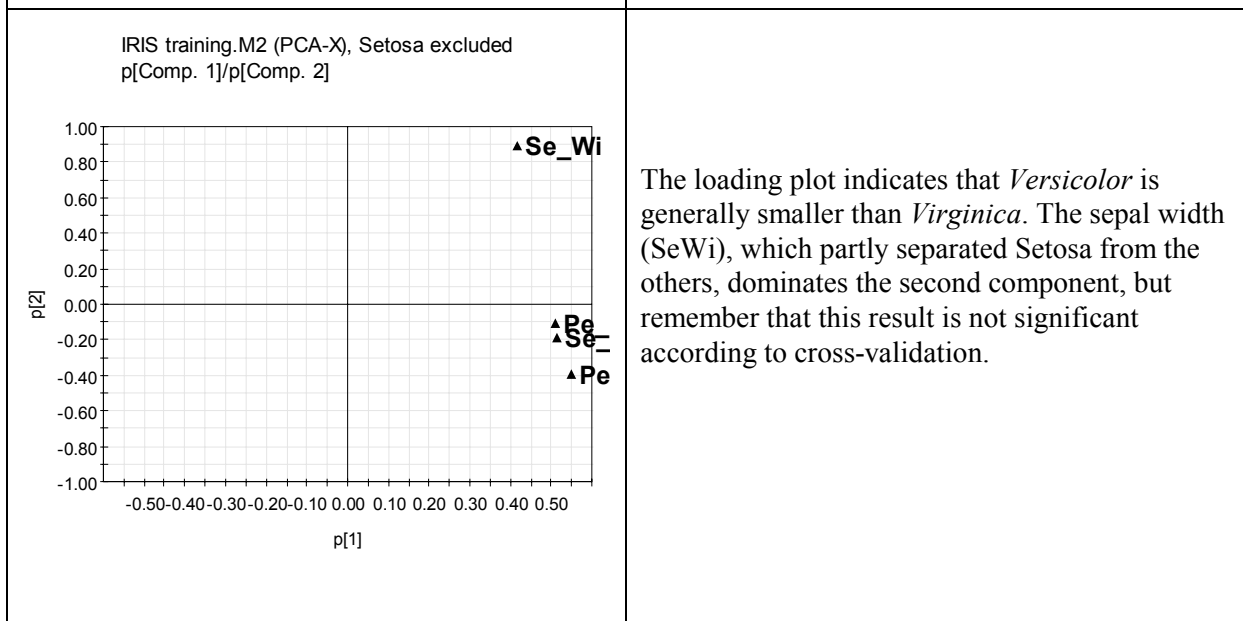
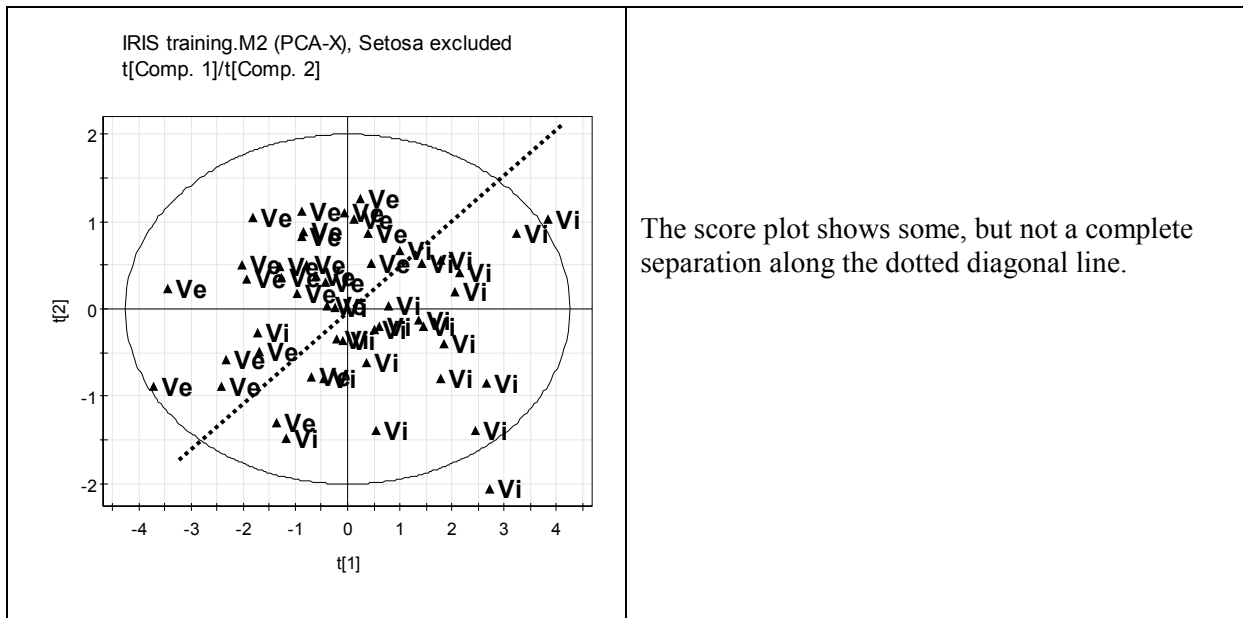
PCA entire training set									
Type: PCA-X Observations (N)=75, Variables (K)=4 (X=4, Y=0)									
Components:									
A	RZ	RZ(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations	
0	Cent.								
1	0.731	0.731	2.92	0.597	0.211	0.597	R1	10	
2	0.224	0.955	0.897	0.388	0.26	0.753	R1	7	



Task 3

All class 1 observations were removed. One component was obtained. For plotting purposes two components were calculated. The model describes 71% (87% after two components) of the variability in the data.

IRIS training - M2								
Workset...		Options...		Title: Setosa excluded				
Type: PCA-X Observations (N)=50, Variables (K)=4 (X=4, Y=0)								
Components:								
A	R2X	R2X[cum]	Eigen...	Q2	Limit	Q2[cum]	Significance	Iterations
0	Cent.							
1	0.713	0.713	2.85	0.441	0.216	0.441	R1	8
2	0.157	0.87	0.629	-0.354	0.265	0.385	NS	23

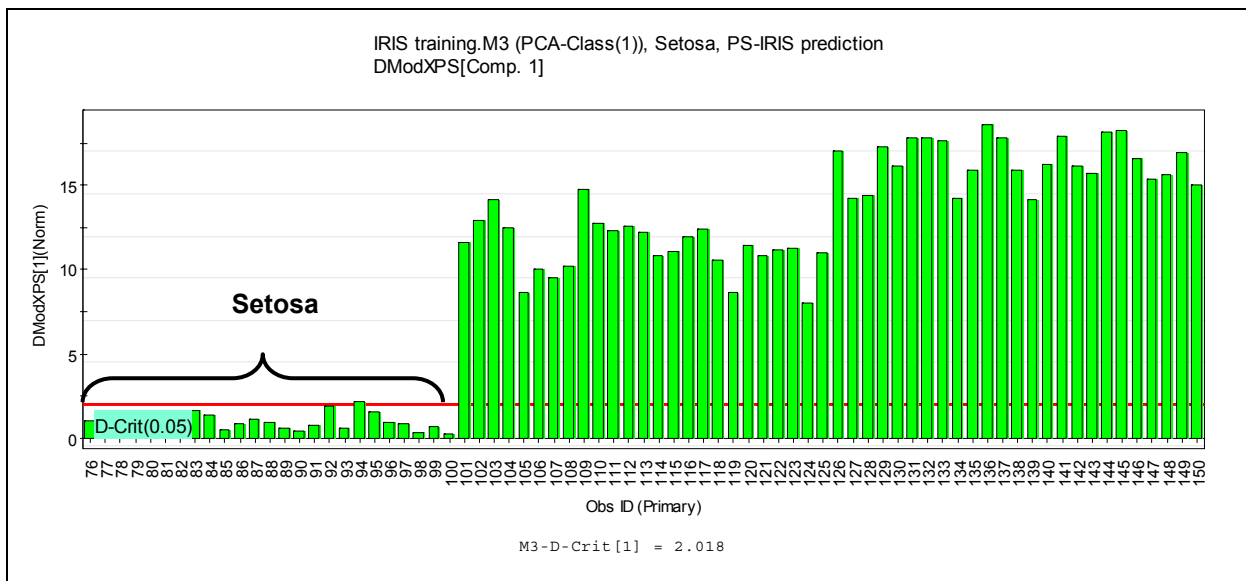


Task 4

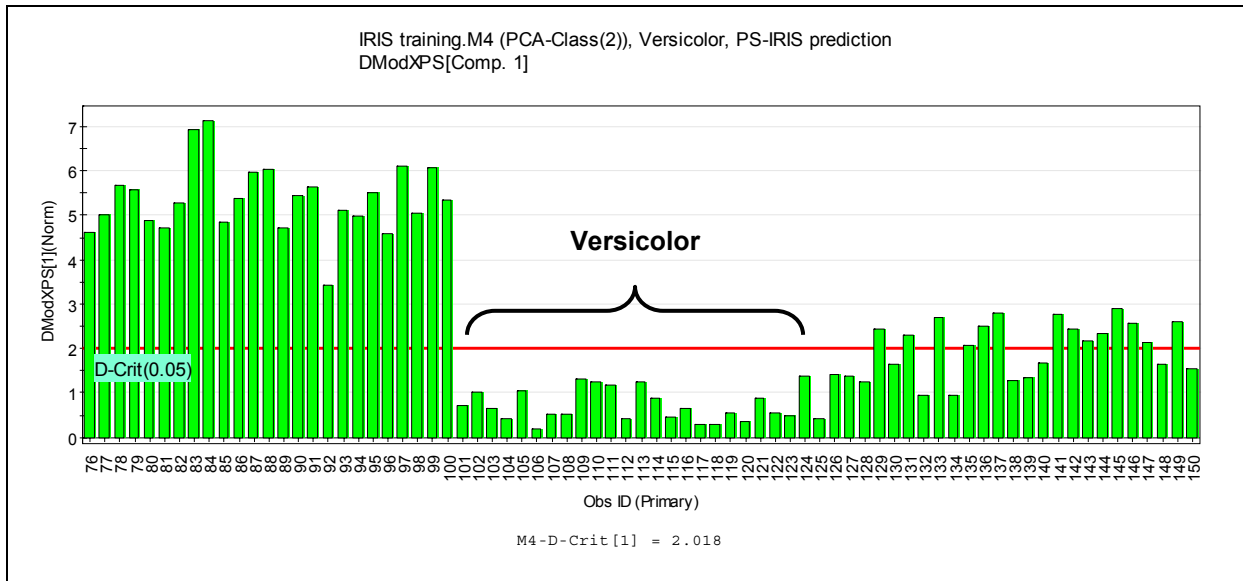
Three class models were made:

P4 IRIS training										
Observations [N] = 75, Variables [K] = 4										
Models:										
No.	Model	Type	A	RZX	RZY	Q2(cum)	Date	Title	Hiera...	
3	M3	PCA-Class(1)	1	0.589		0.214	2002-10-14	Setosa		
4	M4	PCA-Class(2)	1	0.671		0.387	2002-10-14	Versicolor		
5	M5	PCA-Class(3)	2	0.896		0.452	2002-10-14	Virginica		

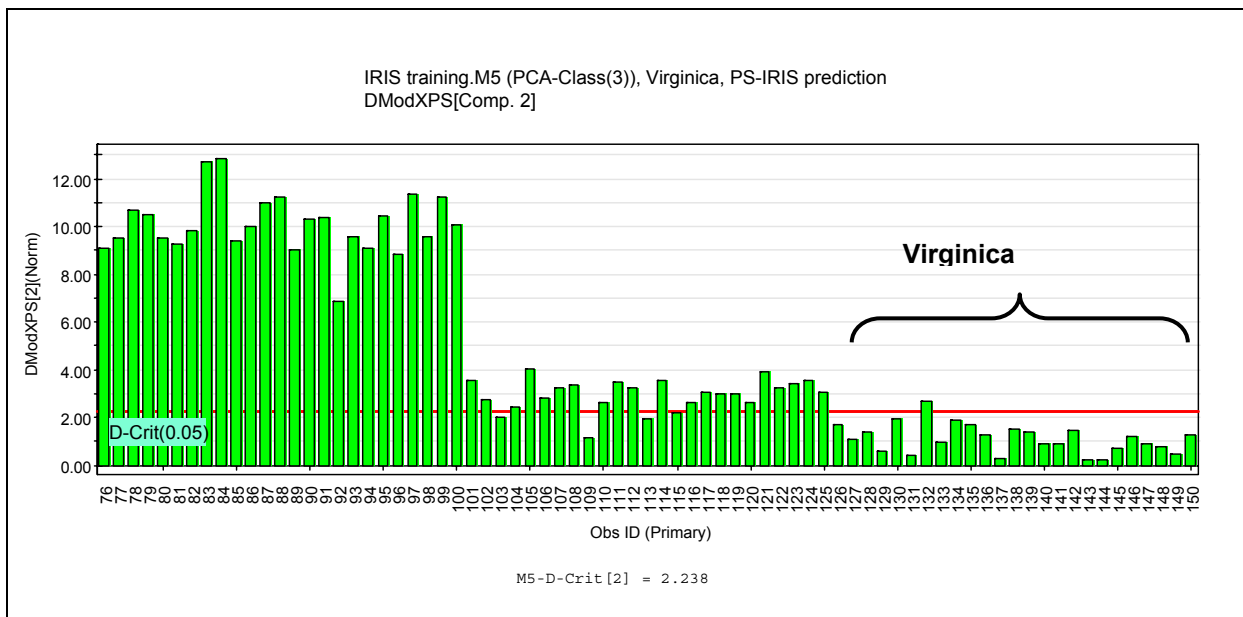
The plot below shows DModX for the *Setosa* model. We find that 24 out of 25 Setosas in the prediction set are correctly assessed. (Recall that on the 0.05 level 1 out of 20 are expected to lie outside Dcrit.) This fact justifies keeping one component for this model. We can also see that the other two species are far away from the model region for Setosa.



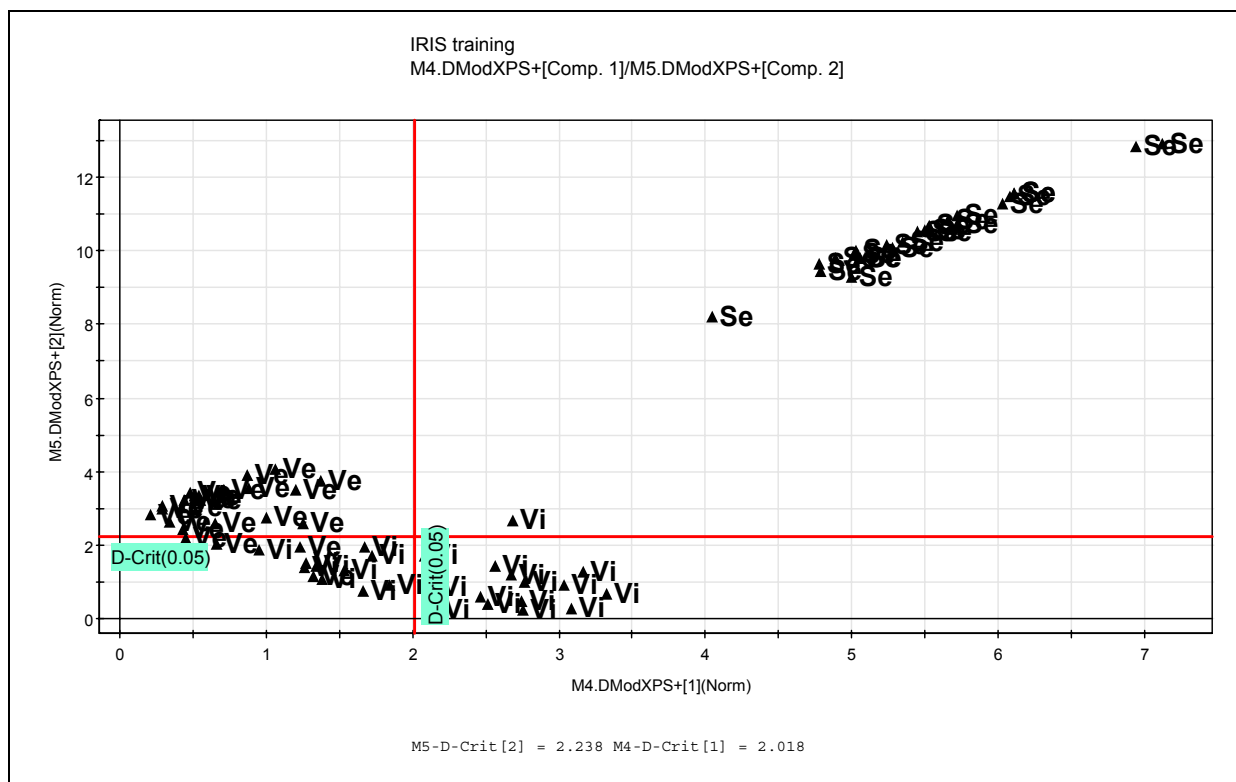
The next plot provides DModX for the *Versicolor* class. Here the validation observations are also very well predicted. However, we can see that some of the *Virginicas* are false positives, i.e. are found below DCrit. This means that these two classes have some common area in the multivariate space.



The *Virginica* model (below) has a similar resolution to the *Versicolor* class model. The two classes are not completely separated.



An alternative way of displaying class distances is to plot the distances for two models against each other in a Coomans' plot (below). By also plotting the critical distance, DCrit, for each model in the Coomans' plot, four areas of diagnostic interest are created. This plot is given below, which represents a scatter plot of DModX to the *Virginica* model against DModX for the *Versicolor* model. In the lower left-hand part of the plot there is a region where prediction set samples that fit both models are found. In the lower right-hand part and the upper left-hand part observations predicted to fit the *Virginica* model or the *Versicolor* model are located, respectively. Finally, we have the upper right-hand area where we find observations that do not conform with either of the models.



Conclusions

An overview PCA was made on the training set of 75 observations. This model revealed that *Setosa* flowers were different from the other two species. In an attempt to resolve the *Versicolor* and *Virginica* samples, a new model was founded on these two groups. It was found that some, but not complete, separation was the case. Finally, the SIMCA approach was tested on the IRIS data set. By constructing three local PCA models pertaining to each separate class of Iris species, and predicting probable class membership of 75 prediction set observations, it was corroborated that *Setosa* observations were markedly different from the others. Through the construction of a Coomans' plot, it was also confirmed that the classes of *Iris versicolor* and *Iris virginica* overlapped in multivariate space.

The conclusions that may be drawn are thus the following:

- (i) *Setosa* specimens are quite different from *Versicolor* and *Virginica* observations.
- (ii) There is an overlap between the *Versicolor* and *Virginica* classes, and they cannot be completely separated but it is possible to predict if an unknown sample was (a) definitely *Versicolor*; (b) definitely *Virginica*; (c) definitely neither; or (d) *Virginica* or *Versicolor* using this methodology.

MVDA-Exercise ARCHAEOLOGY

Classification of archaeological soil samples

Background

In archaeology it is important to confidently classify soil samples in order to determine their origin and age. Since undisturbed soil and soil modified by cultivation and dwelling exhibit quite different patterns of trace element composition, this objective might be accomplished if soil samples are characterised appropriately. Characterisation of soil samples is usually accomplished by measuring the level of phosphate. However, this is a univariate description that does not necessarily relate perfectly to prehistoric activity. Realising the need for multivariate characterisation of soil samples using a multitude of descriptors, Linderholm *et al* (J Arch Sci 21:303-314, 1994) carried out a feasibility classification study involving 18 chemical descriptors. This exercise is based on the data of Linderholm *et al*.

Objective

The objective of this study was to classify soil samples to determine their origin, and to gain a better understanding of how various soil samples differ.

Data

Linderholm *et al*. collected soil samples from an excavation area outside Mjölby in Sweden. The collected samples represent three categories: samples from recognisable features (F) of the site, samples from the occupation layer of the site (S) and off-site control samples (C). In total, 22 samples were assembled representing a varying degree and intensity of human cultural influence. To carry out multivariate chemical characterisation of these soil samples, levels of nine elements (Fe, Cu, P, Mn, V, Co, Zn, Cr, and Ca) were registered. These measurements were made using ICP-AES and testing two kinds of pre-treatments, viz. total dissolution (TD) and nitric acid (NA). Thus, the chemical characterisation embraced 18 descriptors. The data set is explained in more detail below.

Explanation of the data set:

Characters in observation names:

F = Feature (refuse pit or hearth) sample

S = Site sample

C = Control sample taken off-site

All other characters refer to specific details at the excavation area, such as geographic location (north, west, etc.). For more details reference is made to the original work.

Characters in variable names:

TD = total dissolution, first pre-treatment method

NA = nitric acid, second pre-treatment method.

		Fe_TD	Cu_TD	P_TD	Mn_TD	V_TD	Co_TD	Zn_TD	Cr_TD	Ca_TD
1	F_A100B	2.21	36	0.914	950	50	7	130	34	2.26
2	F_A100M	2.09	29	1.053	857	46	7	150.5	29	2.49
3	F_A100T	1.98	27	0.337	786	45	7	110	29	1.38
4	F_A52B	2.14	44	0.389	1044	52	7	160	31	1.6
5	F_A52M	1.99	81	2.926	3200	45.5	7	475	28.5	8.08
6	F_A52T	2.08	64	1.436	1998	45	7	352	29	3.33
7	F_A40	2.87	47	0.61	1593	63.7	9.3	311	41	3.36
8	S_1	2.45	24.5	0.136	870	64	8.5	81.5	31.5	1.21
9	S_2	2.21	26	0.147	873	54	8	94	31	1.3
10	S_3	2.42	28	0.133	903	61	8	97	32	1.28
11	S_4	2.28	31	0.133	708	58	7	98	32	1.01
12	S_5	2.28	29	0.136	679	62	7	84	31	0.95
13	C_N10B	1.78	7	0.027	263	41	5	37	27	0.93
14	C_N4B	1.24	4	0.028	222	30	4	24	20	0.87
15	C_W10B	3.38	12	0.023	1581	72	12	73	51	0.92
16	C_W8B	2.63	15	0.033	666	61	8	52	36	0.89
17	C_W6B	1.99	9	0.036	273	47.5	5.5	43	30	0.77
18	C_N10C	3.21	18	0.033	553	68	9	57	45	0.98
19	C_N4C	4.21	25	0.067	569	90	12	80	63	1.16
20	C_W10C	5.29	28	0.045	1315	116	16	89	87	1
21	C_W8C	4.54	27	0.05	620	102	13	82	72	0.8
22	C_W6C	2.52	13	0.03	324	56.5	7	45.5	35	0.73
		Fe_NA	Cu_NA	P_NA	Mn_NA	V_NA	Co_NA	Zn_NA	Cr_NA	Ca_NA
1	F_A100B	1.89		0.838	933	32	4	141	20	1.93
2	F_A100M	1.77		1.002	962	29	4	161	17	2.34
3	F_A100T	1.67		0.342	788	27	3	118	16	1.04
4	F_A52B	2.03	55	0.43	1208	37	5	185	20	1.49
5	F_A52M	1.77	114	2.679	3160	32	5	561	21	5.11
6	F_A52T	1.88	81	1.284	2098	33	4	396	20	3.24
7	F_A40	2.315	42	0.6	1522	36.5	6	300	23.3	2.79
8	S_1	2.34	30	0.147	1034	44	6	98	19	0.97
9	S_2	2.15	38	0.158	970	40	6	124	22	1.07
10	S_3	2.17	33	0.143	904	55	5	102	25	0.97
11	S_4	2.02	33	0.134	743	50	4	102	22	0.87
12	S_5	2.05	31	0.129	660	53	4	87	22	0.82
13	C_N10B	1.34		0.023	158	21	2	33	14	0.33
14	C_N4B	0.87		0.026	107	17	1	21	11	0.28
15	C_W10B	2.82		0.019	1499	45	8	66	31	0.5
16	C_W8B	2.375	13	0.03	656	39	4.5	50	23.5	0.465
17	C_W6B	1.73	8	0.031	198	30	3	39	18	0.47
18	C_N10C	2.65	16	0.029	505	38	6	68	28	0.42
19	C_N4C	3.56	27	0.059	483	53	7	72	41	0.73
20	C_W10C	4.91	20	0.034	1367	71	11	82	64	0.92
21	C_W8C	4.065	32	0.044	608	53	8	74	44	0.73
22	C_W6C	2.21	20	0.024	262	35	5	39	21	0.39

Tasks

Task 1

Open SIMCA and import ARCHAЕ.SIM (or ARCHAЕ.XLS). This file has 22 observations (soil samples) and 18 variables. The data set is incomplete and contains some missing values in variable 11 (Cu_NA).

Task 2

Select all variables and log-transform these (*WorkSet/New/Transform*). Run PCA to make an overview of the soil samples. Create the necessary score- and loading-plots. Review and interpret the model. Can you see any groupings in the data?

Task 3

When PCA results in clustering of observations, it is sometimes worthwhile to resolve these groupings by means of PLS discriminant analysis (PLS-DA). Define three classes of observations according to:

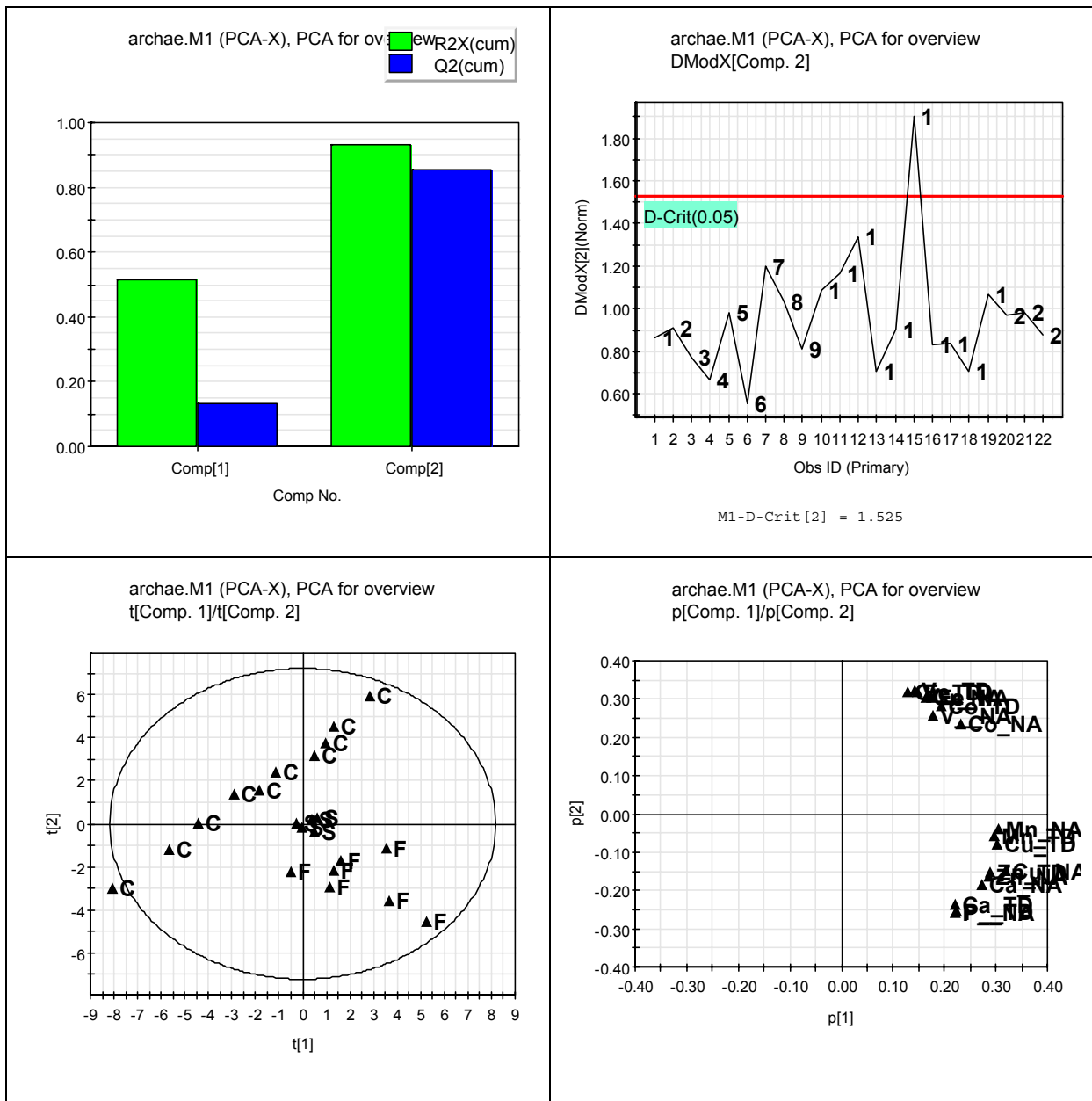
- class 1, F-observations
- class 2, S-observations
- class 3, C-observations

Then run PLS-DA and interpret the PLS model. Which variables have the highest discriminatory power?

SOLUTIONS to ARCHAEOLOGY

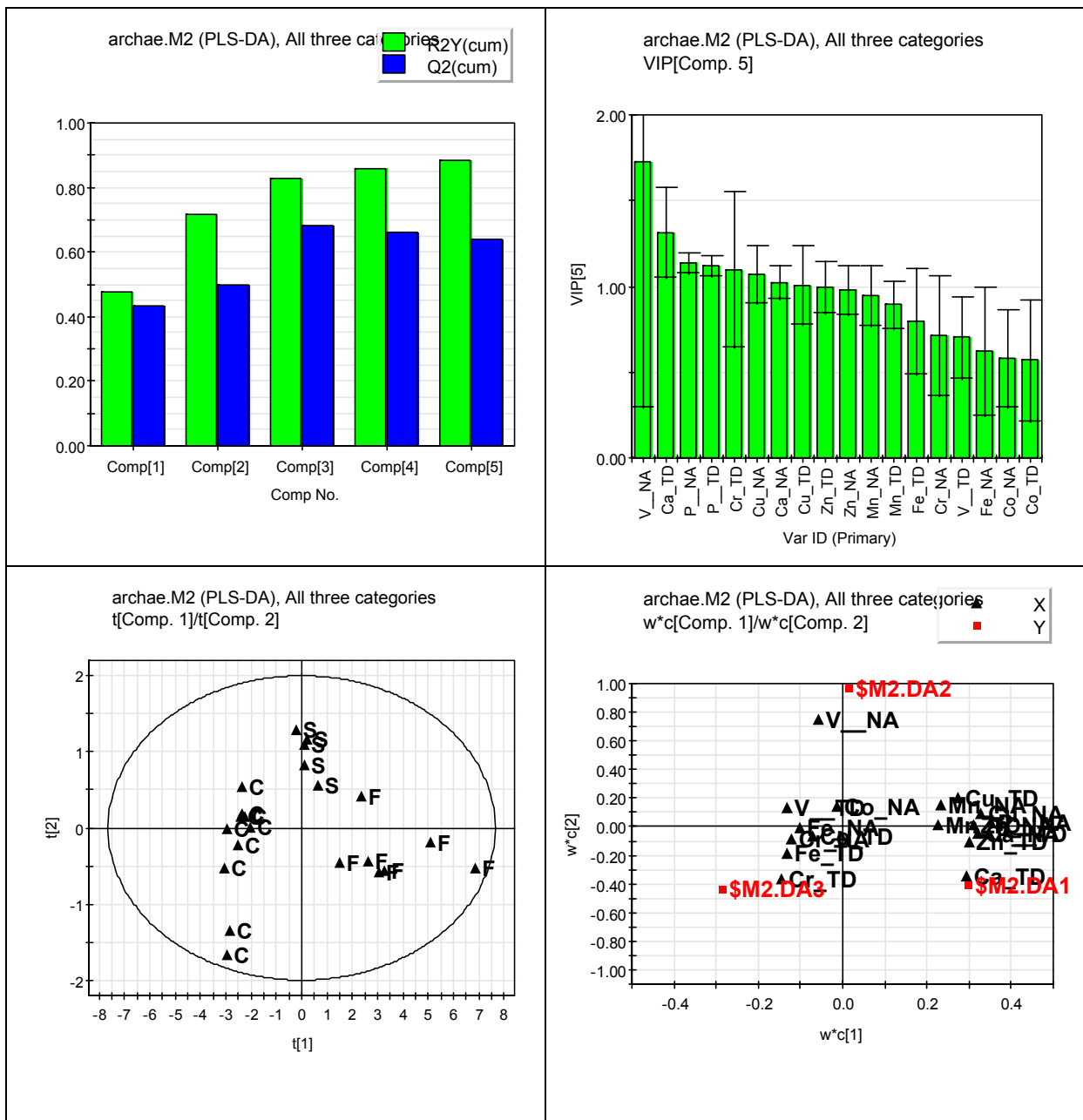
Task 2

PCA yielded a two-component model. In the t_1/t_2 score plot the three categories of samples are well separated. The diagonal going from the upper left-hand corner to the lower right-hand corner might be interpreted as the direction of cultural activity. The corresponding loading plot reveals that the trace elements cluster in two groups. One group contains Mn, Cu, Ca, Zn and P and the other Co, V, Fe, and Cr. There is a larger difference between these two groups of variables than between the TD and NA pre-treatments. Finally, the DModX graph tells us that observation 15 (C_W10B) is not explained well by this PC model.



Task 3

PLS discriminant analysis resulted in a five-component model. The t_1/t_2 score plot shows that the three classes of archaeological soil samples are clearly separated. To interpret this pattern we can look in the PLS loading plot. Here, the triangular distribution of the three dummy variables (denoted §M2.DA1 - 3) is expected because of their 0/1-nature. Chemical descriptors that are close to these dummy variables contribute strongly to the separation of the classes. Thus, the Control observations have comparatively high levels of Cr, Fe, V, and Co, and low levels of Mn, Cu, Zn, Ca, and P. The Feature group of observations has a reversed pattern, and the Site samples are between these two extremes. More quantitative estimates of the discriminatory power can be obtained from the VIP-plot. Obviously, the vanadium content measured with the NA pre-treatment method displays the strongest discriminatory power. However, according to the jack-knife estimates this is the most uncertain X-variable.



Conclusions

The initial PCA model revealed strong groupings among the three categories of archaeological soil samples. It was found that the scores t_1 and t_2 reflected jointly the level of cultural disturbance in the soil samples. Furthermore, it was concluded that there was a larger difference between the two main groups of elemental profiles than between the NA and TD pre-treatment techniques. In the last modelling stage, PLS-DA was attempted. A strongly significant PLS model was acquired, indicating that the 18 chemical variables contained class separating information. The separation of the three classes was slightly superior compared with previous modelling attempts. As a quantitative measure of the discriminating power of the 18 chemical descriptors, the VIP parameter was used. It was inferred that the vanadium content, monitored after NA pre-treatment, was the descriptor that carried most class discriminating information.

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 $^1\text{H-NMR}$ 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 $^1\text{H-NMR}$ 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 and PLS-DA may be used in state-of-the-art Metabonomics.

Data

In total, the data set contains $N = 57$ observations (rats) and $K = 194$ variables ($^1\text{H-NMR}$ 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 $^1\text{H NMR}$ spectra were reduced by summation of all the data points over a 0.4 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.

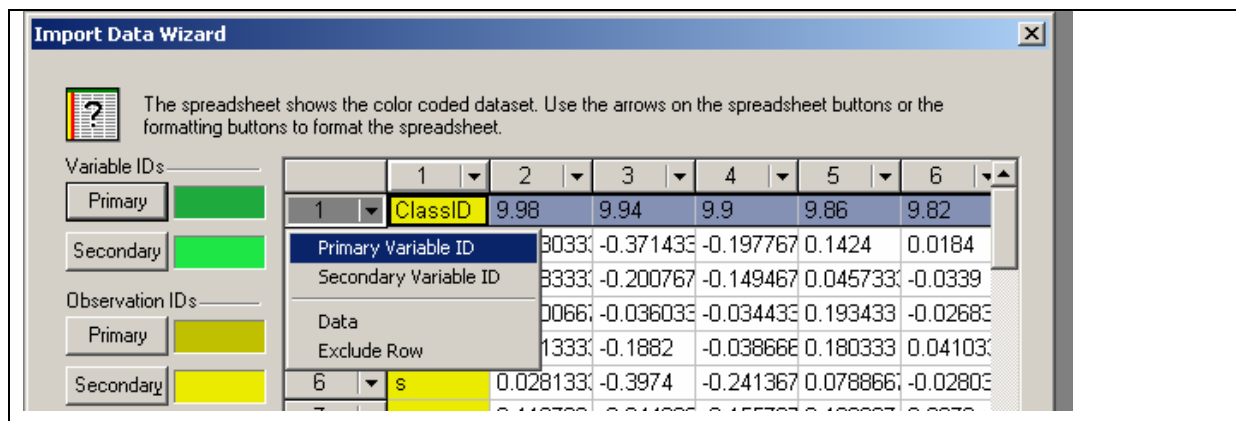
1) Nicholson, J.K., Connelly, J., Lindon, J.C., and Holmes, E., *Metabonomics: A Platform for Studying Drug Toxicity and Gene Function*, *Nature Review*, 2002; 1:153-161. 2) J.R. Espina, W.J. Herron, J.P. Shockcor, B.D. Car, N.R. Contel, P.J. Ciaccio, J.C. Lindon, E. Holmes and J.K. Nicholson. *Detection 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 *METABONOMICS.DIF* (*File/New*). The first column in the data set is labelled *ClassID*. SIMCA assigns this column to a Secondary Observation ID. Accept this. SIMCA will later auto-generate a Primary Observation ID.

To define a Primary Variable ID, first press the arrow as indicated in the picture below, then select Primary Variable ID. This first row is equivalent to the chemical shift regions in the NMR-spectra.



Press *Next*, and verify the entire data set has been imported: 57 observations (rats) and 194 variables (chemical shift regions). Are there any missing values in the data set? Press *Finish*.

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 overviewing the information in the data set using PCA.

To Pareto-scale and mean-center the data, follow these steps: *Workset/Edit*, select the *Scale* tab, and mark all the variables. Under *Set Scaling/Base* select "Par". Press *Set*. (By default "Par" scaling automatically mean-centres the data). Simply press *OK* and you will be ready to fit the principal component model.

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

In order to illustrate the utility of PLS-DA we are going to focus on the difference between group "s" (controls of Sprague-Dawley) and "sc" (SD rats treated with chloroquine). However, in so doing we must first eliminate the outlying "sc"-rat. PLS-DA requires homogenous groups devoid of outliers, otherwise inconsistent patterns may result.

To specify the two classes, do the following: *Workset/New as Model 1*, press the *Observations* tab. Right-click to get the following picture:

Observation ID	Class	Inc/Exc
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Select ObsID (Class ID) as the only displayed observation ID (i.e., deselect Obs ID Primary). Press *Select All* and *Exclude*. (Now all observations are excluded). Then go to *Find* and enter a lower case s to find the first class. Press *Include* and assign class number 1 to this class. Repeat this with the sc-group and assign class number 2 to this group of rats. Finally remove (exclude) the penultimate sc-rat, which was diagnosed as an outlier in the foregoing task. You should now have 19 rats selected.

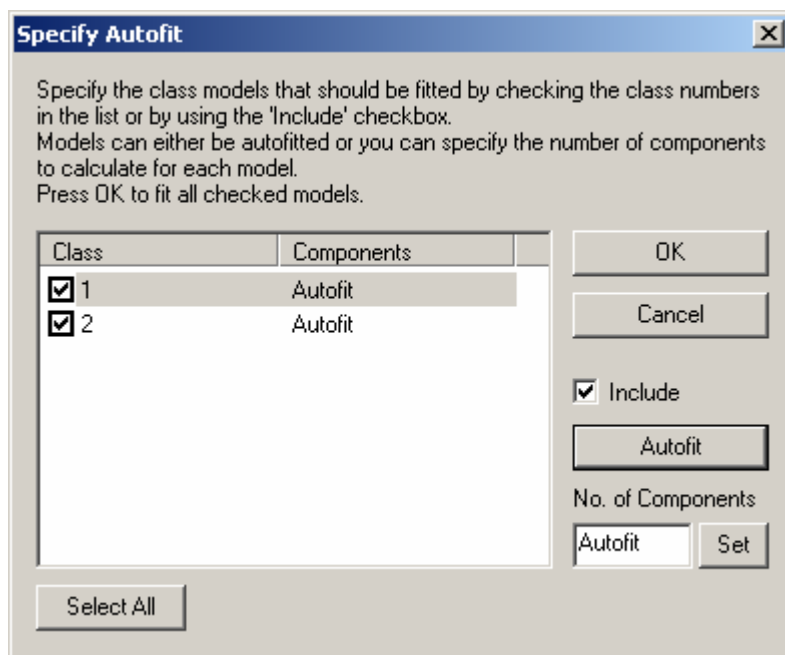
Fit the PLS-DA model (*Analysis/Change Model Type/PLS-DA/AutoFit*). Review the fit and interpret the model. Is it possible to use NMR-data to discriminate between these two groups?

Task 5

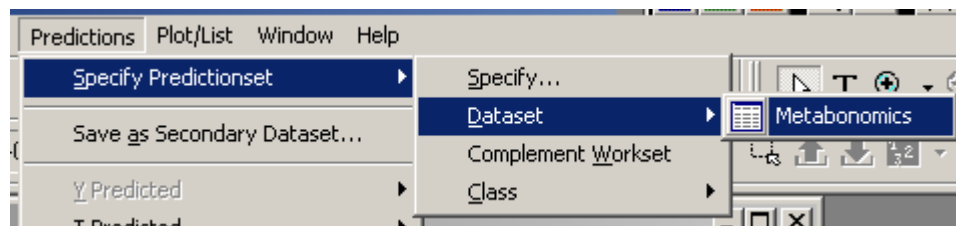
We will now apply the SIMCA method to the s and sc-groups.

Go to *Analysis|Change Model Type* and choose *PCA Class* and the first class. AutoFit the model and save. Repeat this procedure for the second class and save. Review and interpret these models.

Alternatively, the procedure specified above can be carried out directly in one step using *Analysis|AutoFit Class Models*. You may experiment with this facility if you wish, or continue with the exercise as described below.



We will now test the predictive ability of the two class models. For this purpose we define a prediction data set. Use *Predictions* and select *Specify Predictionset|Dataset* and select as source the entire *Metabonomics* data set.



Produce a Coomans' plot (*Predictions|Coomans' plot*) for the two class models. What can you say about classification ability of these two models?

Task 6

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

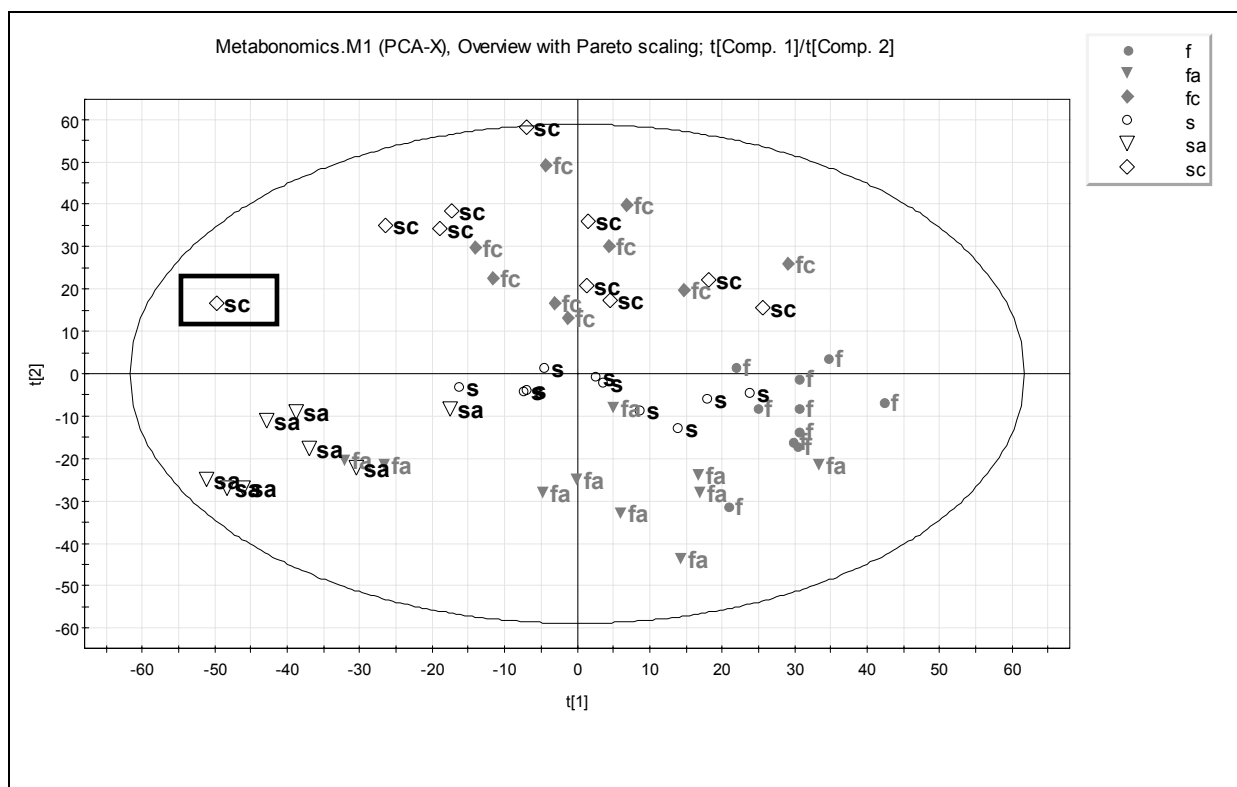
-

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 $Q^2X = 0.38$.

P Metabonomics - M1								
Workset...		Options...		Title Overview with Pareto scaling				
Type: PCA-X Observations (N)=57, Variables (K)=194 (X=194, Y=0)								
Components:								
A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	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.

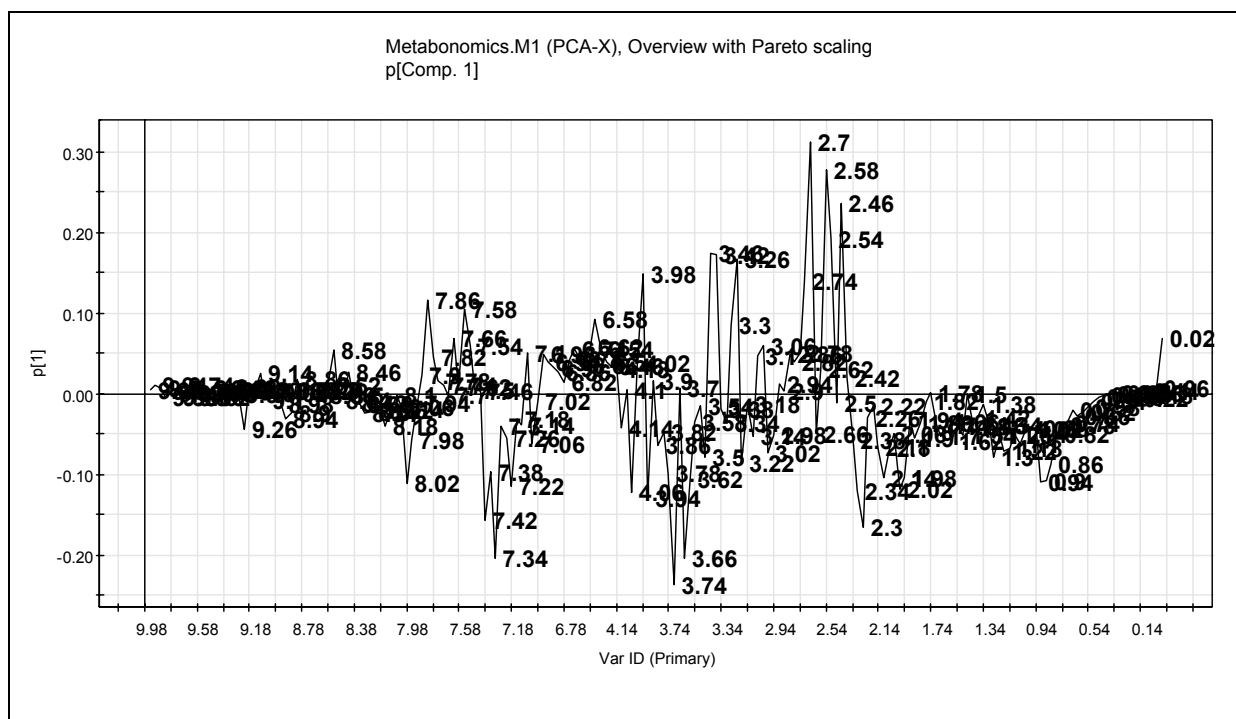


We can see that all the chloroquine-treated animals are positioned in the top part of the plot, whereas the majority of the amiodarone-treated rats are found in the bottom part. All controls are located in the central, predominantly right-hand, part of the plot. Hence, the second principal component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

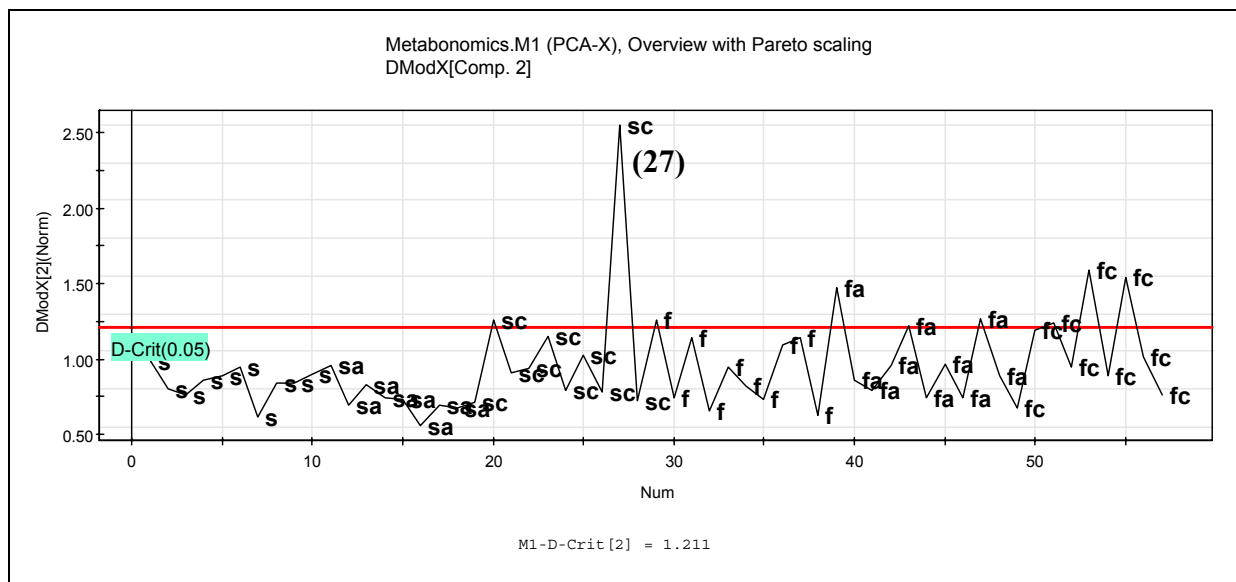
As seen, this score plot is not identical to the original one. We may take advantage of the Secondary Observation ID (here called ClassID) to modify this plot regarding colour, markers, etc. To accomplish this, right-click in the plot and choose *Label Types tab, Use Identifier, ClassID*, and press *Apply*. Next you select the *Colour tab, Colouring type, By Identifier, Choose the ID to colour by, ClassID*. Then you can assign any colour to the classes. However, since we only use black&white printing we decided to use only black and grey. Additionally, we modified the plot marks a little to get the plot seen above.

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 based technique, called PLS discriminant analysis (PLS-DA), becomes interesting. PLS-DA makes 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 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.

In PLS-DA, the X-matrix consists of the multivariate characterization data of the observations. In order to encode a class identity, one uses as Y-data a matrix of dummy variables, which describes the class membership of each observation in the training set. A dummy variable is an artificial variable that assumes a discrete numerical value in the class description. The dummy matrix Y has G columns (for G classes) with ones and zeros, such that the entry in the g.th column is one and the entries in other columns are zero for observations of class g.

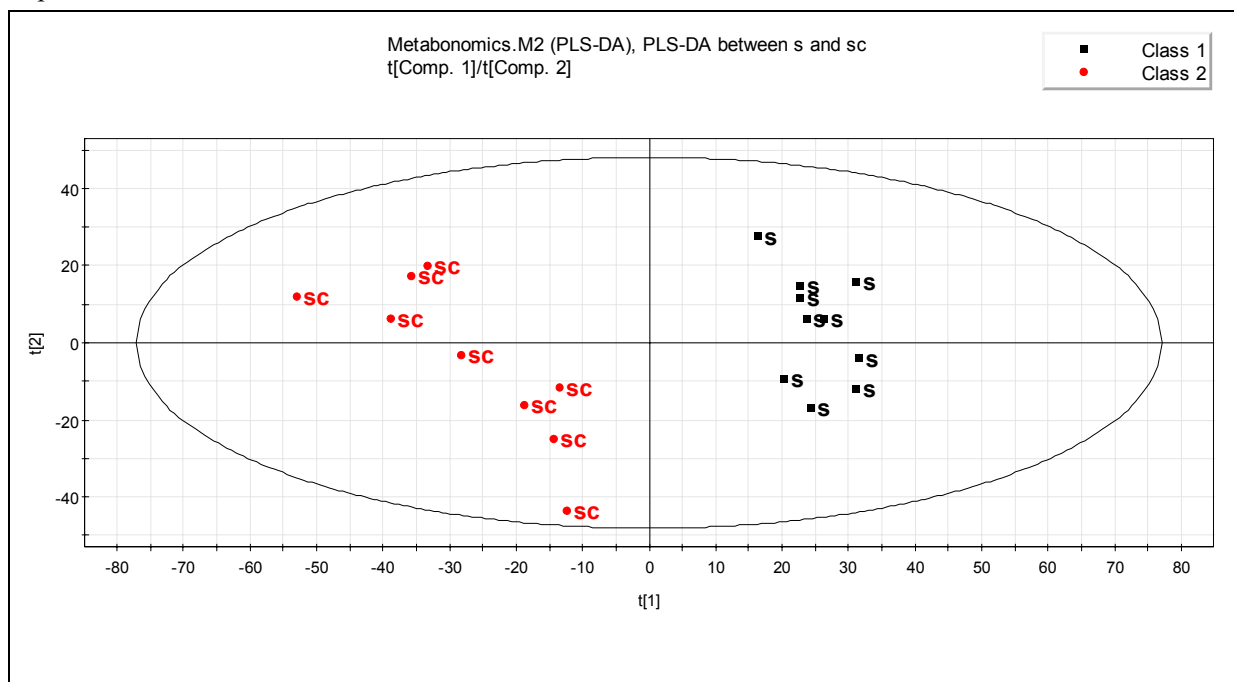
In the next task (Task 4) you will be asked to do a PLS-DA between two classes of rats, the “s” and “sc” classes.

Task 4

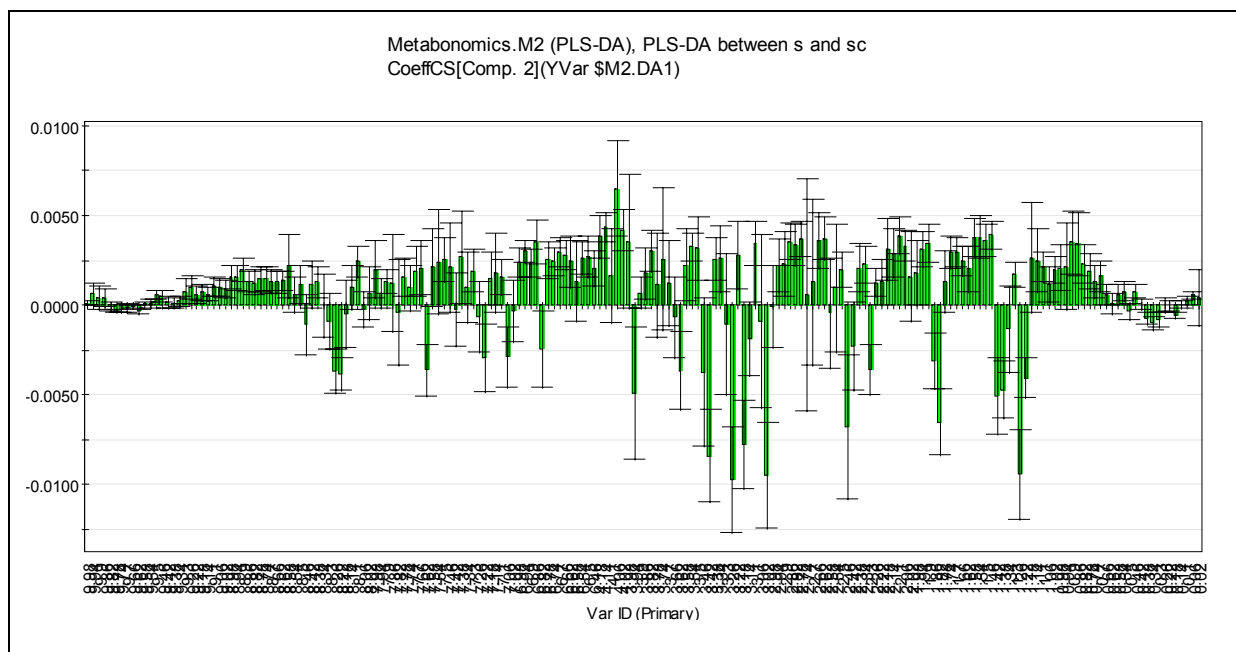
A PLS-DA model was calculated based on the 19 rats in the “s” and “sc”-groups. All variables were mean-centered and Pareto-scaled. This model contained two very strong components, showing the performance statistics $R^2X = 0.69$, $R^2Y = 0.94$ and $Q^2Y = 0.90$. We will neglect the two last components due to their minor importance.

Metabonomics - M2											
Workset...		Options...		Title PLS-DA between s and sc							
Type: PLS-DA Observations (N)=19, Variables (K)=196 (X=194, Y=2)											
Components:											
A	R2X	R2X(cum)	Eigenv...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signifi...	Ite...	
0	Cent.			Cent.							
1	0.479	0.479	9.1	0.88	0.88	0.84	0.05	0.84	R1	2	
2	0.207	0.686	3.93	0.063	0.943	0.384	0.05	0.901	R1	2	
3	0.0532	0.739	1.01	0.041	0.984	0.241	0.05	0.925	R1	2	
4	0.064	0.803	1.22	0.0105	0.995	0.483	0.05	0.961	R1	2	

The X-score plot of t_1 and t_2 of this model is displayed in the next figure. Evidently, there is strong separation (“discrimination”) between the “s” and “sc”-groups. It is mainly the first component that is responsible for separating the two groups of rat from each other. The second model component picks up within-class variation.



Thus, there is really no doubt that the chemical treatment of the rats induces a substantial and characteristic change in their NMR-profiles. The next coefficient plot shows which chemical shift regions contribute to the separation of the two classes.



An alternative to PLS-DA is SIMCA, short for “Soft Independent Modelling of Class Analogy”, or SIMCA for short [4,5]. SIMCA is a graphically oriented technique, and is applicable when clear groupings exist in the data, such as those seen above.

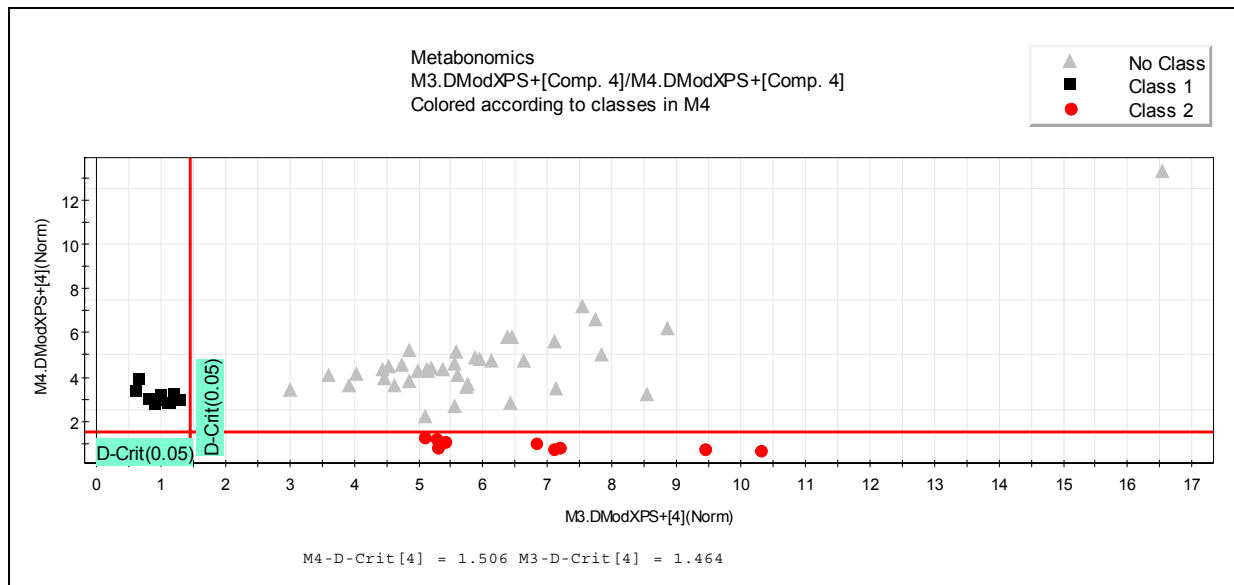
In SIMCA one separate PCA-model for each class is computed. After the separate modelling of each class, the models are used to predict a likely class membership (“classification”) for new observations. An observation is classified in SIMCA according to the tolerance intervals of the different classes. Observations that do not fit any class are then considered as outliers, or perhaps as founders of a new, hitherto unseen, class. Furthermore, in regions where tolerance intervals overlap, the observations cannot be unequivocally assigned.

These local PCA-models can be interpreted by inspecting loadings, scores, residuals, and contribution plots. Among other things, this will indicate which variables contribute to modelling class similarity (loading plot), and which variables do not (contribution plot in residuals of non-fitting observations).

In the next task (Task 5) you will be asked to carry out a SIMCA-analysis for the s- and sc-groups.

Task 5

One local PCA-model was fitted to the “s”-group and another to the “sc”-group. These two models were four-dimensional and were used to infer class membership of all the other rats. The results from the classification phase are summarized in the Coomans’ plot below.



The essence of the Coomans’ plot is that class distances (DModX’s) for two classes are plotted against each other in a scatter plot. By plotting also the critical distance, DCrit, for each model in the Coomans plot, four areas of diagnostic interest are created. In the lower left-hand part of the plot a region where prediction set samples (rats) that fit both models are found (no rats in this case). In the lower right-hand part and the upper left-hand part there are regions where those observations predicted to fit the “sc”-model or the “s”-model are found, respectively. Finally, we have the upper right-hand area where we find observations that do not conform to either of the models. These are all the “sa”, “f”, “fc”, and “fa”, which consistently are found to be different from the “s”- and “sc”-rats.

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 PLS-DA and SIMCA are very useful. A necessary condition for PLS-DA to work reliably is that each class preferably is "tight" and occupies a small and separate volume in the X-space. Also, the number of modelled classes must not be too high. Experience shows that PLS-DA is useful with 2-4 classes, but when the number of classes exceeds four, it is usually more tractable to switch to SIMCA.

In this exercise, we have focused on the differences between two classes, i.e. the “s” and “sc”-rats. This is an analysis that will 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 or PLS-loadings, or contribution plots. A few of these possibilities were hinted at throughout the exercise.

MVDA-Exercise LOWARP

Production of a polymer with desired properties

Background

The development of a polymer with a certain profile of properties was desired, i.e. low warp and high strength. To obtain this a polymer formulation was made with the following (coded) constituents:

1. Glas 20 to 40 %
2. Crtp 0 to 20 %
3. Mica 0 to 20 %
4. Amtp 40 to 60 %

A quadratic model was selected and a mixture extreme vertices design with 14 runs + 3 centre points was made.

Objective

The objective was to identify the most important constituents and to understand how to modify the polymer recipe in order to maximise strength and minimise warp.

Data

14 responses relating to both warp and strength were measured on the product (see data table on the next page).

Tasks

Task 1

Create a new project from LOWARP.SIM. Check that the worksheet colouring agrees with the description above. The first two columns contain observations ID.s. The X-data starts in the third column. Give the project a unique name.

Set the first 4 variables to X and the rest to Y (*WorkSet /New /Var. Blocks*). Make an overview of the responses using PCA (*Analysis /Change Model Type /PCA on Y-block*). Interpret the model (scores, loadings, residuals, ...). How are the responses related?

Task 2

Change model type to PLS and Autofit. Investigate the relevant score and loading plots. Which predictors are influential for the responses?

Task 3

When working with many response variables (14 in this case) the loading plot is convenient for model interpretation. However, it is possible to conduct model interpretation with the help of regression coefficient plots. What message(s) do coefficient plots convey in this case? What is the difference between a coefficient plot and a loading plot?

Data:

Dataset - lowarp																				
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	Primary	Obs. Se	glas	crtp	mica	amtp	wrp1	wrp2	wrp3	wrp4	wrp5	wrp6	st1	st2	wrp7	st3	st4	wrp8	st5	st6
2	1	1	40	10	10	40	0.9	5	0.2	1	0.3	4.2	232	15120	1.2	2190	26390	1.3	2400	0.7
3	2	2	20	20	0	60	3.7	7.3	0.7	1.8	2.5	5.4	150	12230	1.8	905	20270	2.1	1020	0.6
4	3	3	40	20	0	40	3.6	6.9	0.9	2.1	4.8	9.4	243	15550	1.2	1740	21180	1.4	1640	0.5
5	4	4	20	20	20	40	0.6	3.1	0.3	0.4	0.4	1.1	188	11080	1	1700	17630	1	1860	0.5
6	5	5	20	10	20	50	0.3	2.1	0.3	0.3	0.8	1.1	172	11960	1.2	1810	21070	1.3	1970	0.5
7	6	6	40	0	20	40	1.2	5					245	15600	1.1	2590	25310	1.3	2490	0.6
8	7	7	20	0	20	60	2.3	3.9	0.3	0.4	0.7	1.4	242	13900	1.5	1890	21370	1.6	1780	0.7
9	8	8	40	0	10	50	2.6	5.9	0.4	0.2	0.7	1.2	243	17290	1.6	2130	30530	1.6	2320	0.7
10	9	9	30	20	10	40	2.2	5.3	0.2	0.7	0.6	2	204	11170	1	1670	19070	1.1	1890	0.6
11	10	10	40	0	0	60	5.8	7	0.9	1	5.6	11.8	262	20160	1.6	1930	29830	1.8	1890	0.6
12	11	11	30	0	20	50	0.8	2.9	0.5	0.6	1.1	2	225	14140	1.3	2140	22850	1.3	2110	0.7
13	12	12	30	10	0	60	2.8	5.1	1	1.2	2.7	6.1	184	15170	1.9	1230	23400	2.1	1250	0.6
14	13	13	30	10	10	50	1.1	4.7	0.6	0.9	1.3	3.5	198	13420	1.4	1750	23790	1.4	1930	0.7
15	14	14	30	10	10	50	1.9	4.7	1	1	2.8	5.4	234	16970	1.5	1920	25010	1.6	1790	0.7
16	15	15	30	10	10	50	2.9	5.9	0.5	0.6	1	6.6	239	15480	1.5	1800	23140	1.6	1730	0.6
17	16	16	40	10	0	50	5.5	7.9	0.8	2.4	5.5	9.3	256	18870	1.5	1880	28440	1.8	1790	0.6
18	17	17	30	0	10	60	3.2	6	0.3	0.5	1.5	5.2	249	16310	1.5	1860	24710	1.7	1780	0.6

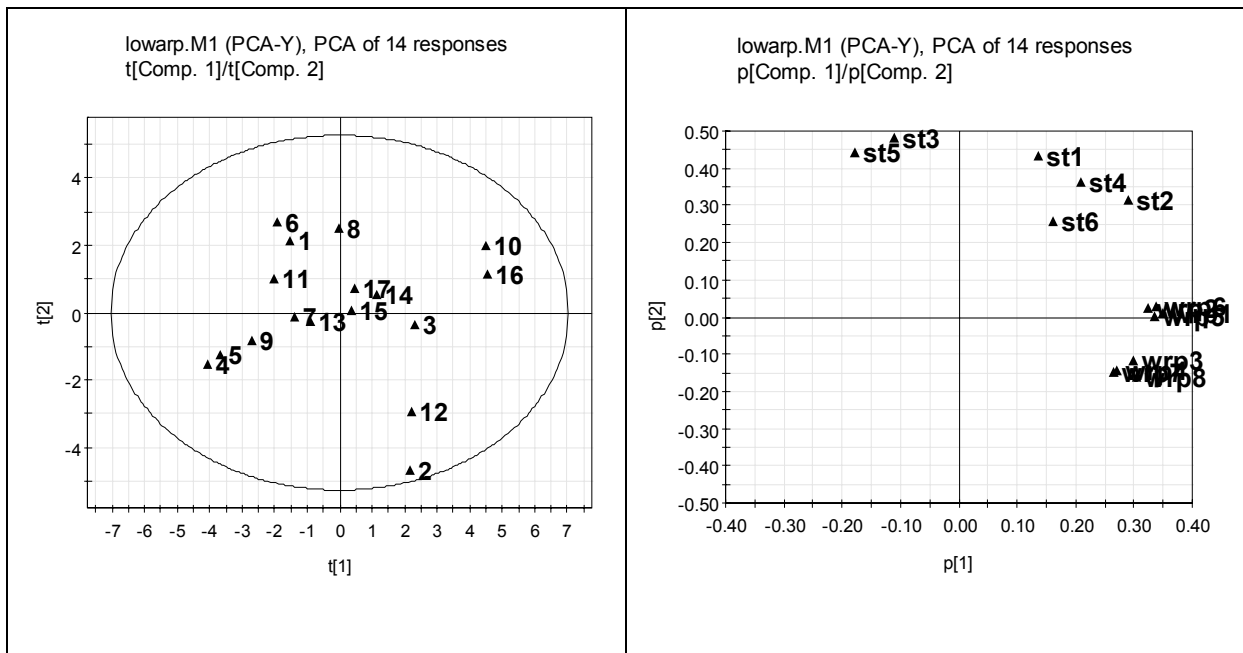
Solutions to LOWARP

Task 1

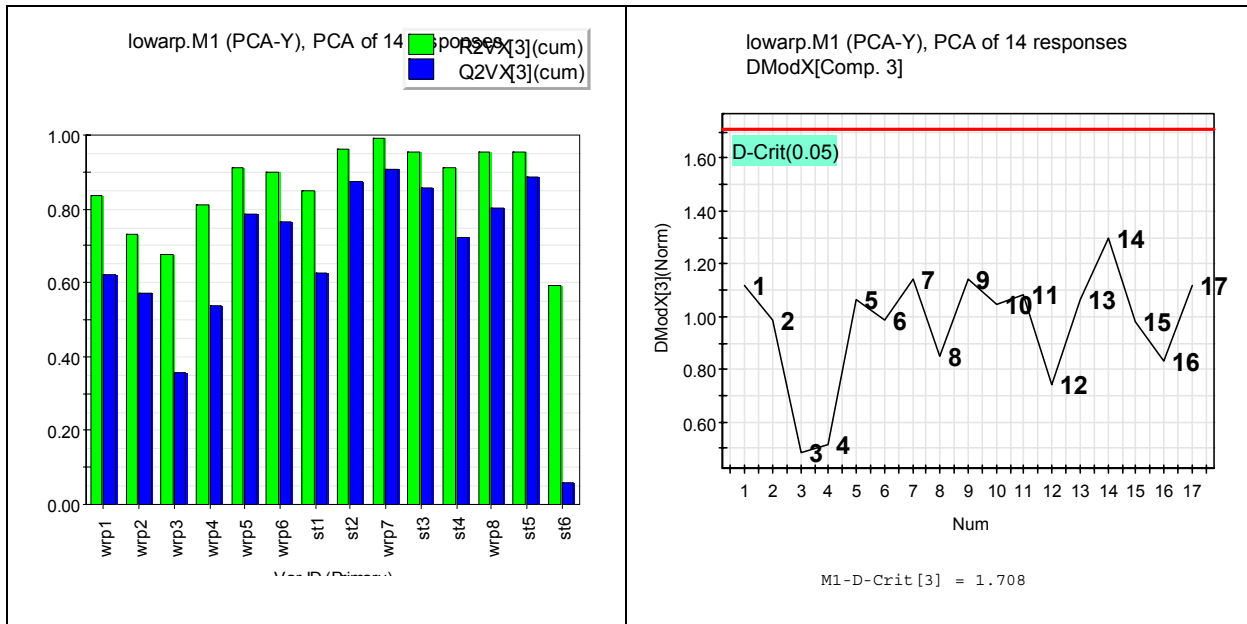
Three components were extracted with PCA.

lowarp - M1											
Workset...		Options...		Title PCA of 14 responses							
Type: PCA-Y Observations (N)=17, Variables (K)=14 (X=14, Y=0)											
Components:											
A	RZ	RZ(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations			
0	Cent.										
1	0.49	0.49	6.86	0.31	0.122	0.31	R1	13			
2	0.278	0.768	3.89	0.357	0.129	0.557	R1	10			
3	0.0989	0.867	1.38	0.21	0.138	0.65	R1	13			

The score plot indicates that the 17 observations are rather well distributed. This is a consequence of the underlying mixture design. According to the loading plot, the first component reflects warp and the second component describes variation in strength. In addition, we can see that the warp responses are clustered in two distinct groups.

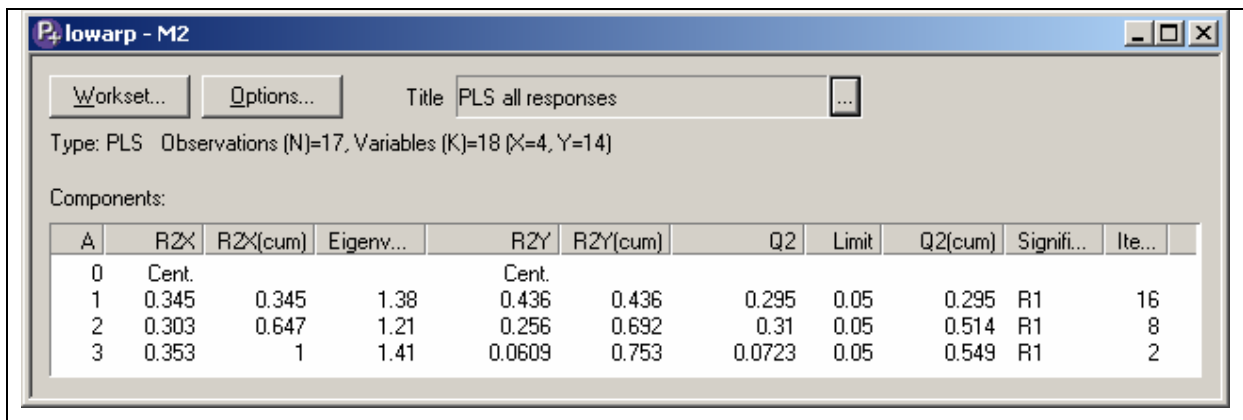


In the left plot below the variance explained for each variable is displayed. Obviously, strength is modelled better than warp. The right plot shows the distance to the model for each observation. All observations are well inside the critical distance.

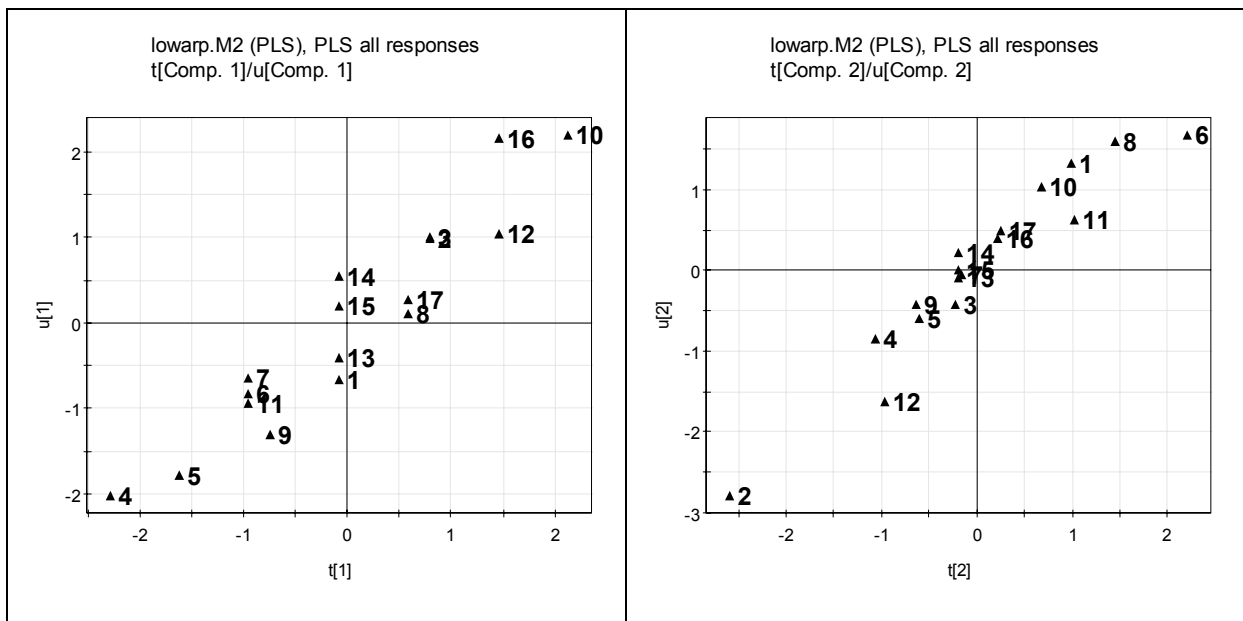


Task 2

A PLS model with three significant components was obtained. Three components will thus give optimal predictive power. Here, however, two components will be used, as our focus primarily lies on interpretation. The increase in Q^2 when going from two to three components is marginal.

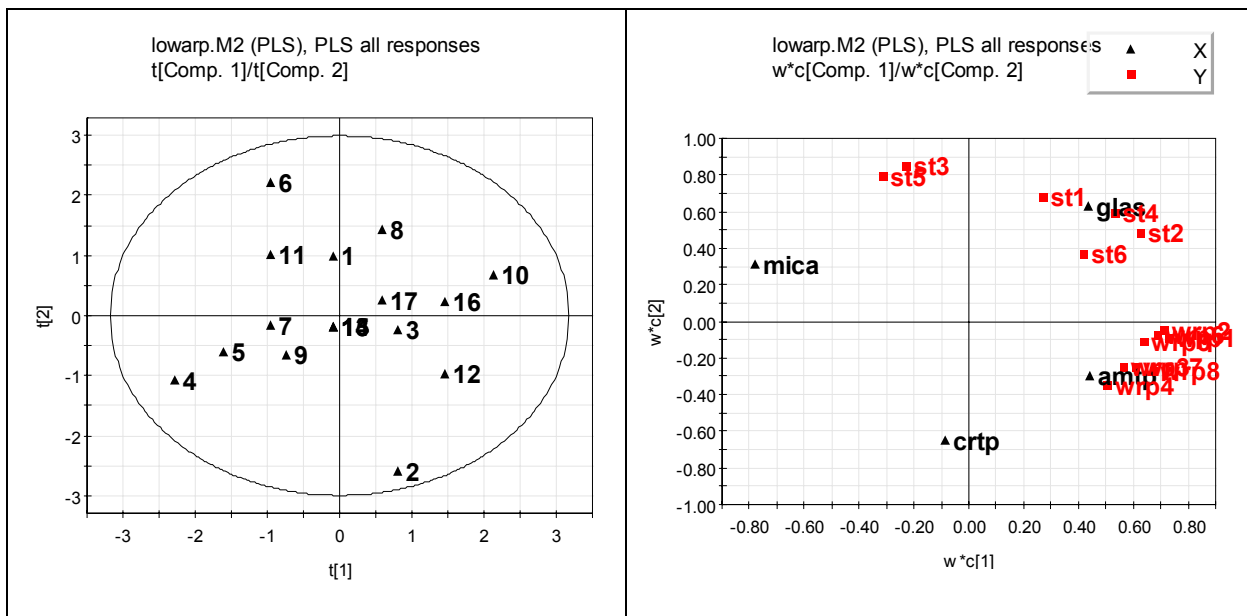


The score plots below show a good correlation between t and u, which means that there is information in X describing Y.



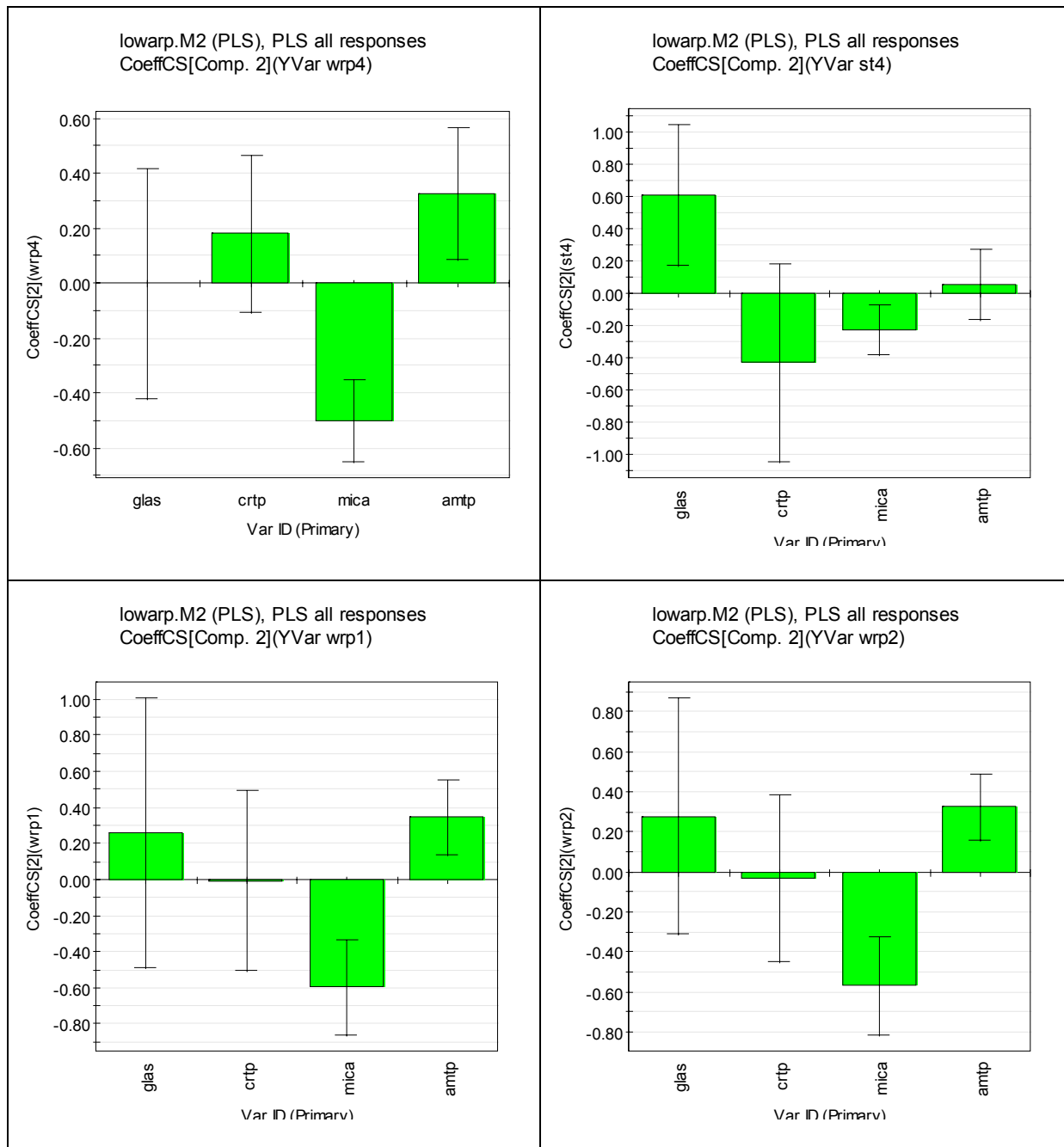
The loading plot below highlights the same grouping of responses that was discovered in the previous task. Obviously, the eight warp responses are strongly correlated, as they are situated tightly together. The six strength responses, however, are less correlated internally and therefore more spread in the loading plot. Two strength responses, st3 and st5, are weakly correlated with the warp responses, whereas the other four strength responses (st1, st2, st4 and st6) are partially correlated with the warp responses. We can see that **glas** is not a good factor for accomplishing the experimental objective, as it influences both strength and warp in the same way. Raising the amount of **glas** in the formulation, for example, will cause an increase in the values of all responses. In order to increase strength one should primarily focus on lowering the amount of **crtp** in the recipe, as this factor does not influence warp to any appreciable extent. And to decrease warp one should increase **mica** and decrease **amtp**.

Thus, observations likely to meet the demands placed on the manufactured polymer ought to be positioned in the upper left-hand quadrant (and as far away from the origin as possible) of the score plot. Because there are no observations in the peripheral part of the upper left-hand area, the conclusion is that no experimental run in the design perfectly matches the desired response profile. Consequently, the PLS model must be scrutinised to identify experimental conditions that provide a reasonable compromise between low warp and high strength. We can see in the score plot that observation numbers 6 and 11 might be appropriate.



Task 3

The loading plot above is an overview showing the correlations between all factors and all responses at the same time. It is possible to “zoom-in” and look at model details of a separate response using a coefficient plot. Responses situated together in the loading plot should have similar regression coefficient profiles, whereas uncorrelated responses should not have similar coefficient profiles. Below, regression coefficients are plotted for two uncorrelated responses and for two correlated responses.



Conclusions

There is a strong association between the four ingredients of the polymer and the measured property variables. **Glas** is not a good factor to change in order to accomplish the experimental objective, as it influences both strength and warp in the same way. In order to increase strength one should primarily focus on lowering the amount of **crtp** in the recipe. In order to decrease warp one should increase **mica** and decrease **amtp**.

MVDA-Exercise USDVOLVO

The optimum way to buy a second hand car

Background

Second hand car dealers often have large advertisements in the daily press. The challenge is to find the best car among all the possible candidates. These data come from one full-page advertisement where each car is described according to 6 different criteria. In total, data were listed for 111 cars. We thank Sven Ahlinder at Volvo Technical Development, Gothenburg, Sweden, for his ideas and suggestions for this exercise.

Objective

The objective of this exercise is to decide which second hand car is the best buy.

Data

The data set has both quantitative and qualitative variables. Some variable information is coded into the observation names to make the plots more interpretable.

Variable	Type	NamePos	in ObsName
• Selling Site	Qual	1-2	Sisjön, Sävedalen, Hisingen, Kungsbacka, Kungälv, Stenungsund, Lilla Edet
• /Type	Not to be used (used to compute Model and Combi)		
• Car Model	Qual	3-5	240, 340, 440, 460, 740, 760, 850, 940, 960
• Combi	Qual	0 or 1	
• Equipment	Qual	6-8	Turbo, S, SE, GL, GLT, GLE, E
• Engine	Quant	-	litres
• Year	Quant	9-10	
• Mileage	Quant	-	km/10
• Price	Quant	-	SEK

Tasks

Task 1

Make a new project by importing the data file UsdVolvo.xls. In the *Import Data Wizard* mark Site, Model, Combi and Equip as qualitative variables. **The final data set should have 110 observations and 8 variables.** Keep the variable “Engine” despite 60% missing values. Make a PC-model to obtain an overview of the data. Use the selling site, car model, equipment, and year as plot markers in the score plots to find any data related structure. For example, to have “Equipment” as the score plot mark, make the score plot, click on the right mouse button. Now, the Properties dialogue will open. Set Start = 6 and Length = 3 (in Plot Labels/Use Identifier/Secondary Observation ID).

Are there any outliers, groups or trends among the cars?

Which variables dominate the correlation structure? Which variables correlate with Price?

Are there any cars with large residuals? If so, why?

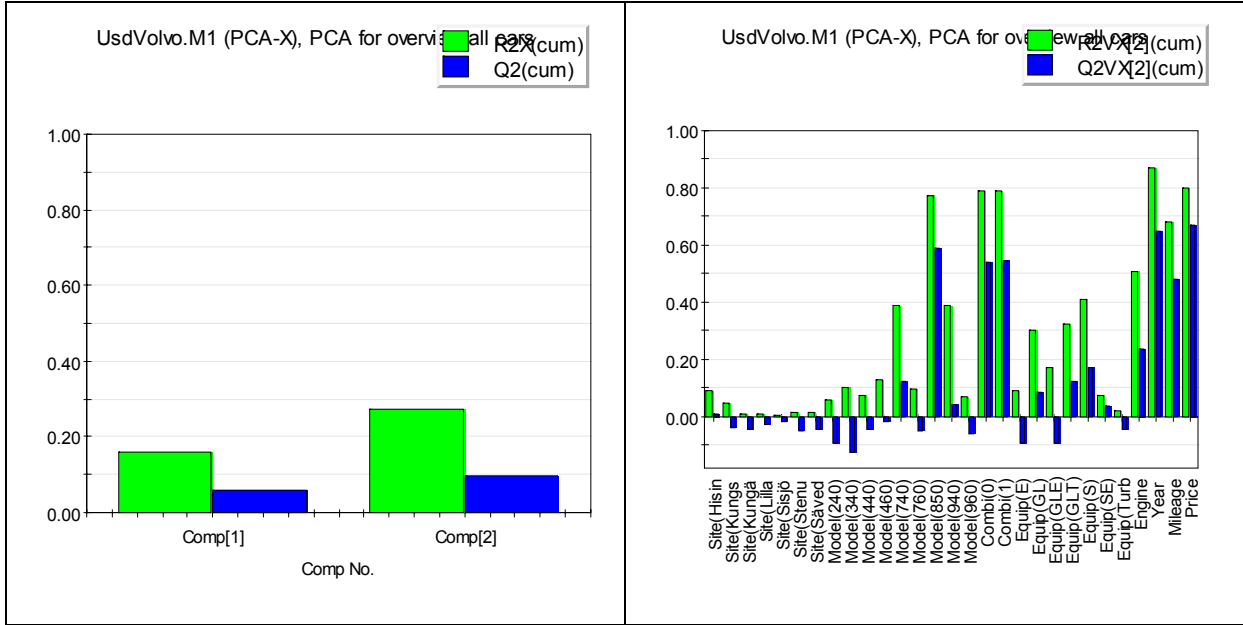
Task 2

Autofit a PLS model with Price as the Y-variable. Exclude any clear outliers after having looked at the raw data. Make a second model. Find the car that is most worth buying. What are the estimated prices of the excluded cars, if any? Can you say something about the different selling sites? Should any dealers be avoided?

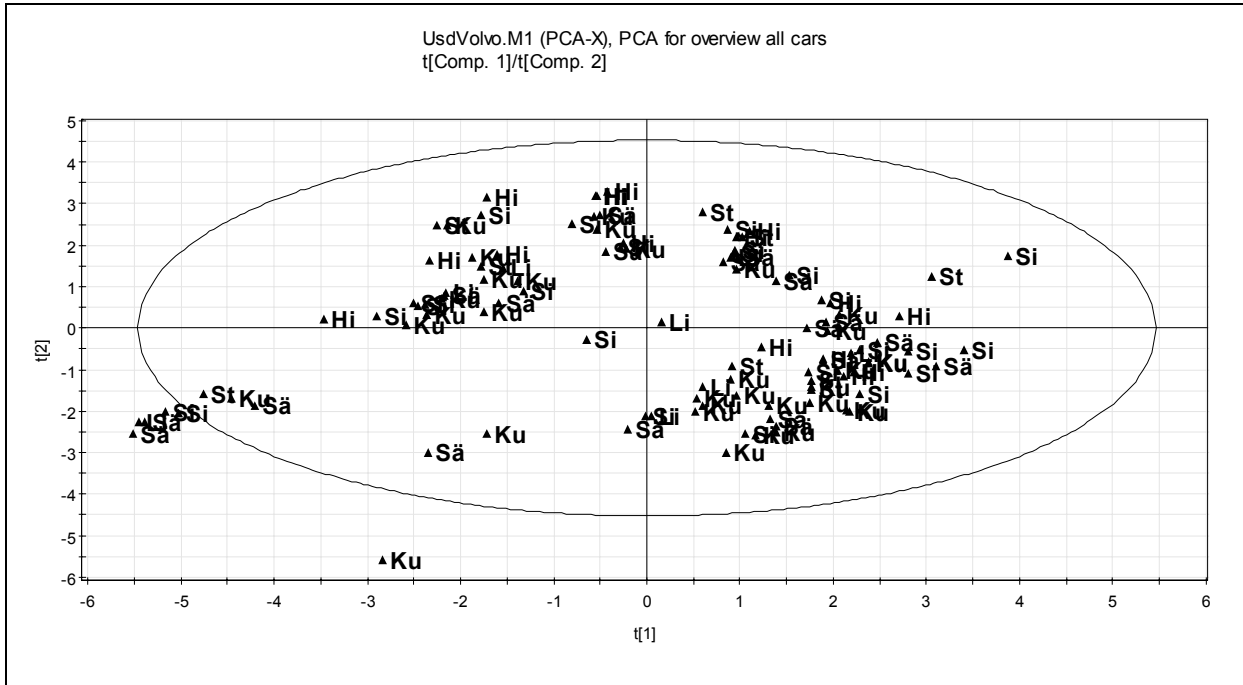
Solutions to USDVOLVO

Task 1

PCA based on default cross-validation suggests two PCs to be significant. The comparatively low values of R^2 and Q^2 is characteristic for data tables with a large portion of qualitative data. In fact, a few of the variables like Engine, Mileage, and Year are well explained and predicted by the model.

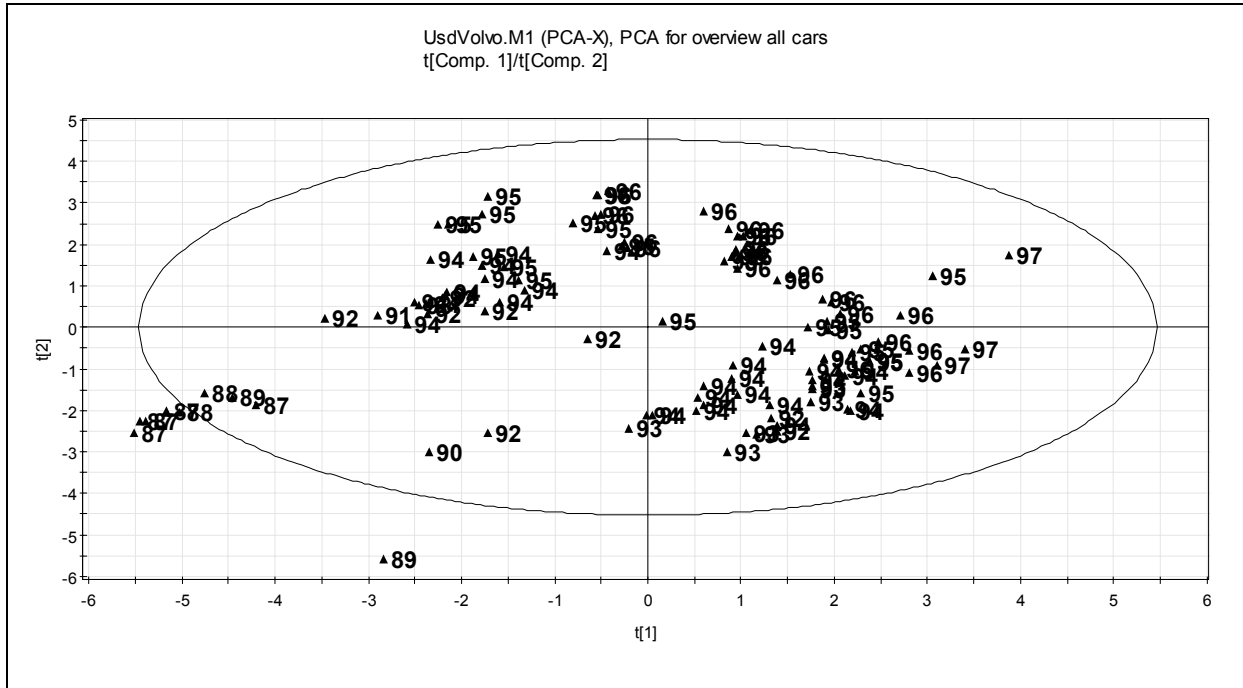


We used positions 1-2 in the name to display Site:



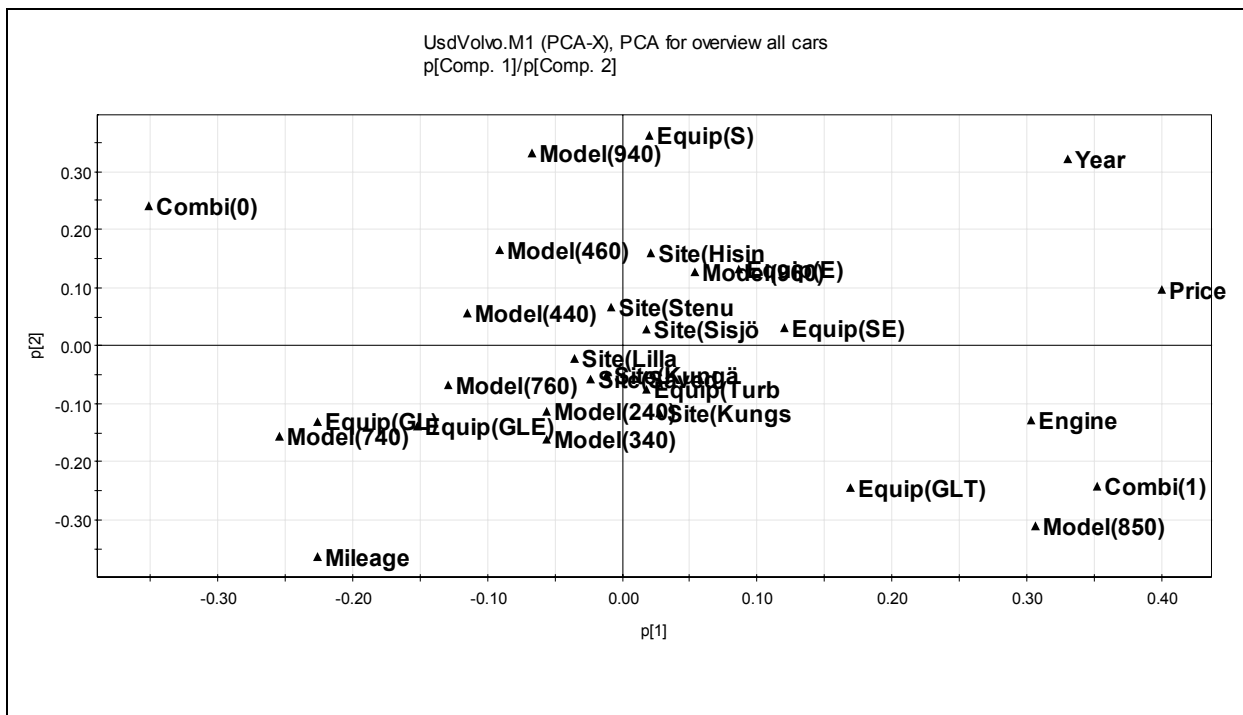
No visible structure correlated to Site.

We use positions 9-10 in the name to display Year:



There is a trend in Year from left to right.

The loading plot:

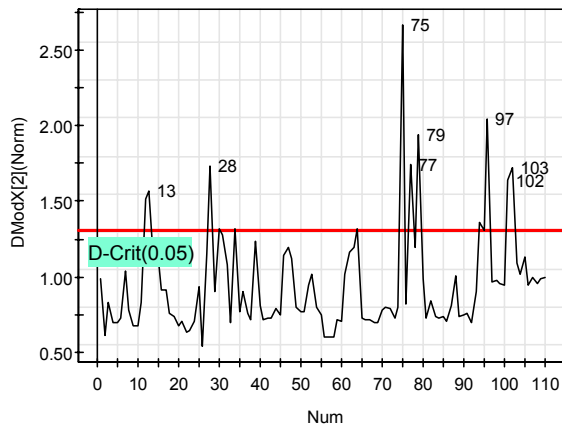


Price, Year, Mileage, and Model dominate the loading plot. Price and Year were positively correlated, and both were negatively correlated with the variables Mileage, 740, and GL.

When examining the residuals we can see that observation 75 is an outlier, also observations 79 and 97 are suspected outliers.

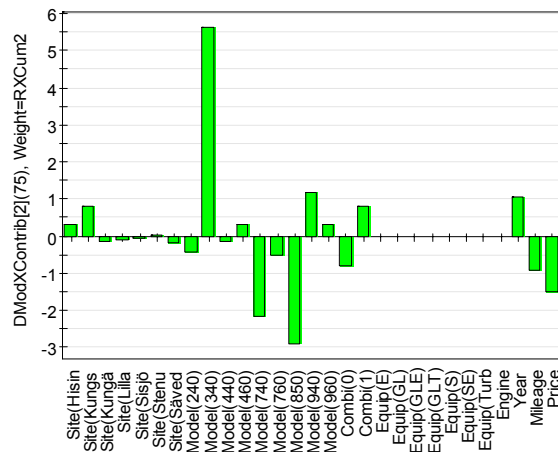
By making a contribution plot (DModX-mode) for observation 75 we find that variable 340 dominates the residual. The reason for this is that this car is the only 340 car in the data set. (Similarly, observations 79 and 97 are the only ones of model 760.)

UsdVolvo.M1 (PCA-X), PCA for overview all cars
DModX[Comp. 2]



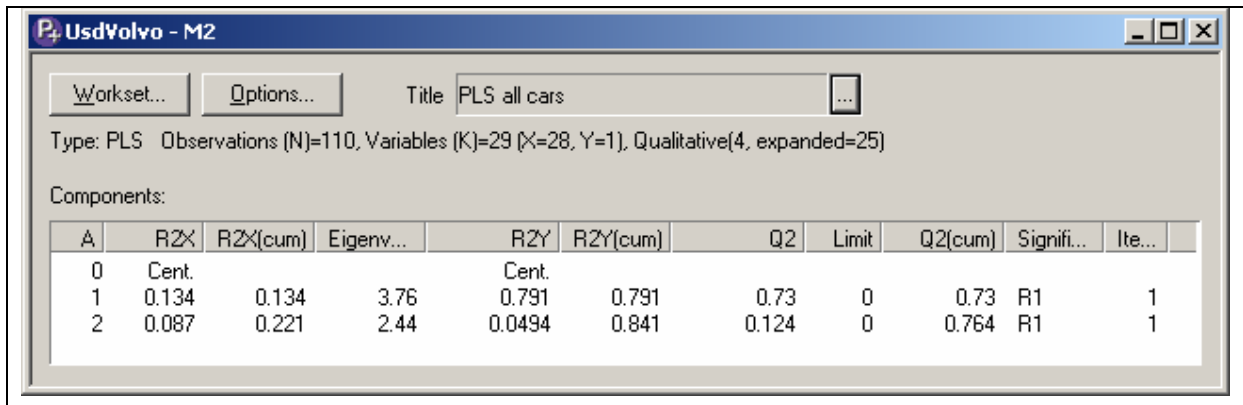
M1-D-Crit [2] = 1.306

UsdVolvo.M1 (PCA-X), PCA for overview all cars
DModX Contrib(Obs 75), Weight=RX[2]

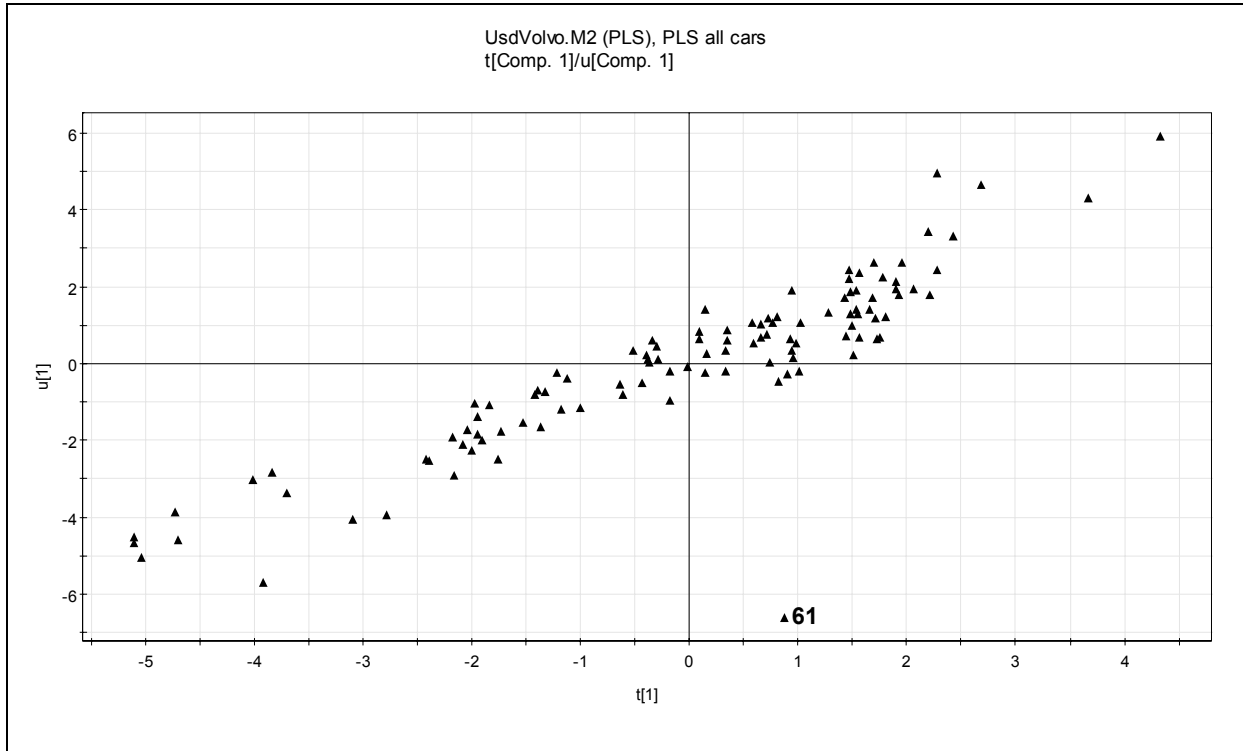


Task 2

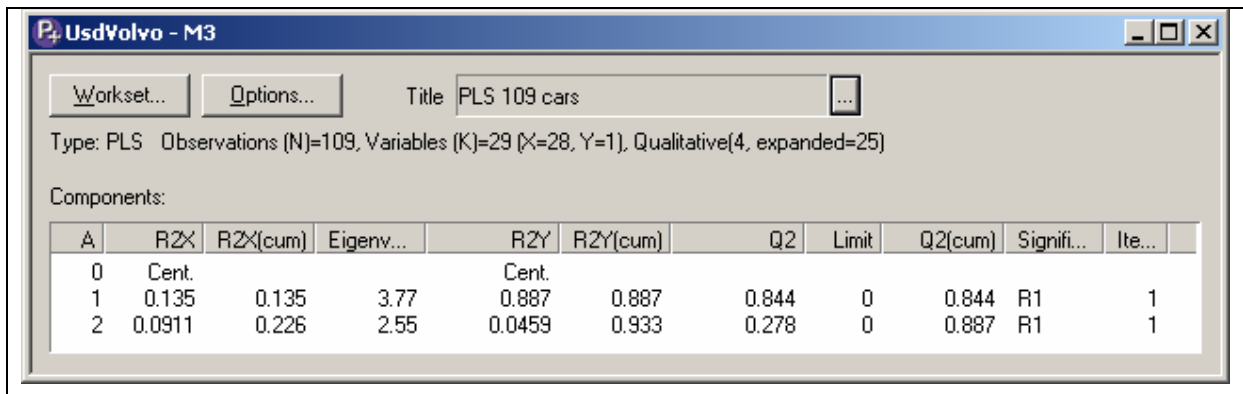
A two-component PLS model was obtained. It is a good model, which models 84% of the price with 76% validity.



One useful plot is the inner relationship t_1/u_1 .

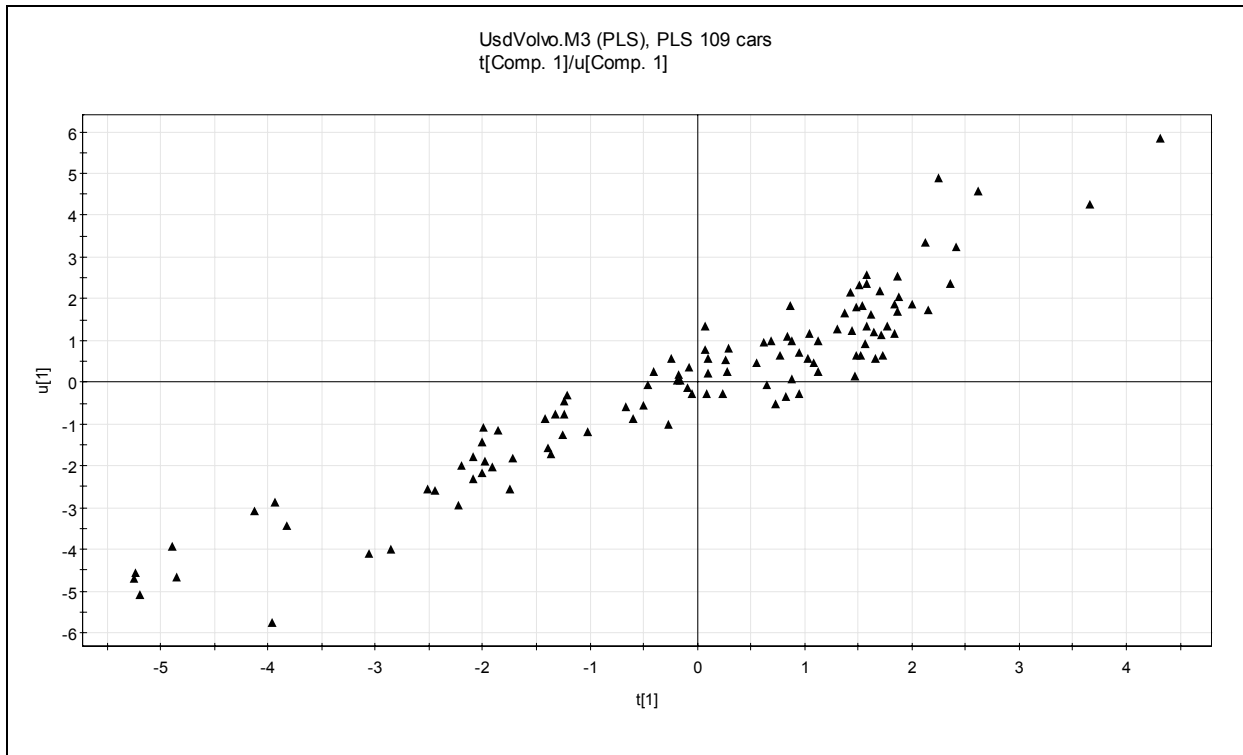


A clear outlier, observation 61, Hi945S_ is detected. It has the unrealistic price of 17.100 (due to an input error, was originally 171.000). Not knowing the reason we excluded this observation and computed an updated PLS model:

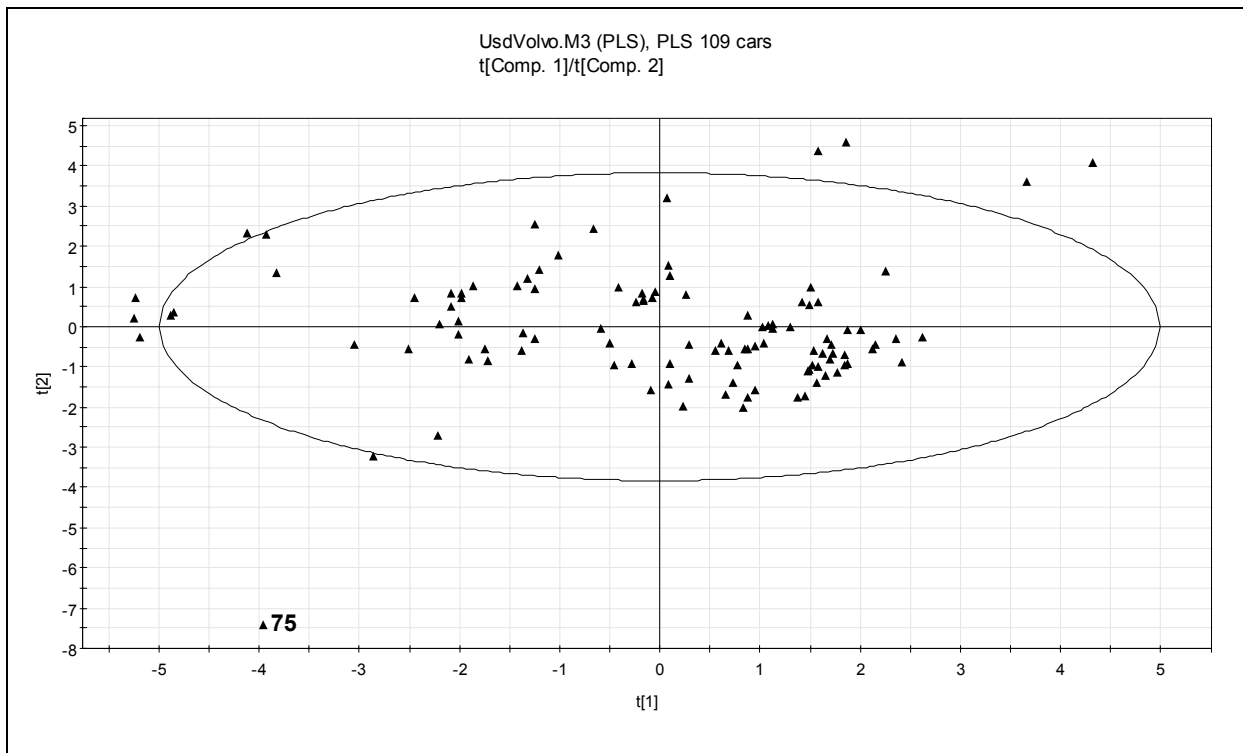


This model explains, surprisingly, 93% of the variation in price with a validity of 89%.

Now, the inner relation shows strong correlation.

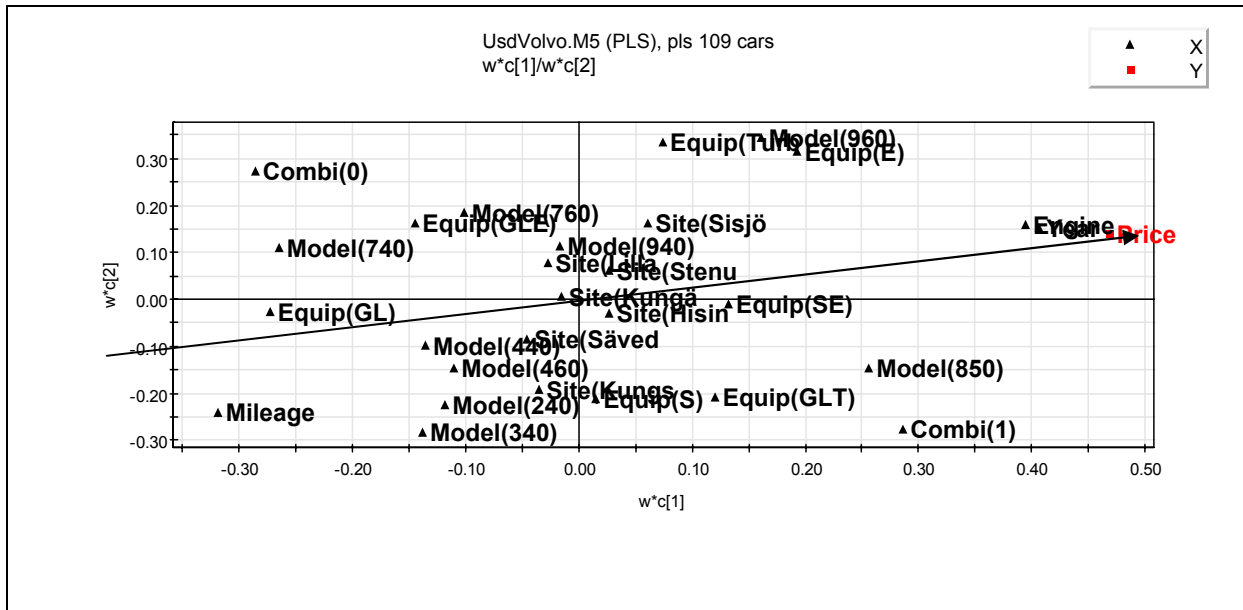


The t_1/t_2 plot shows some grouping by Model and one outlier, observation 75 (Ku345).

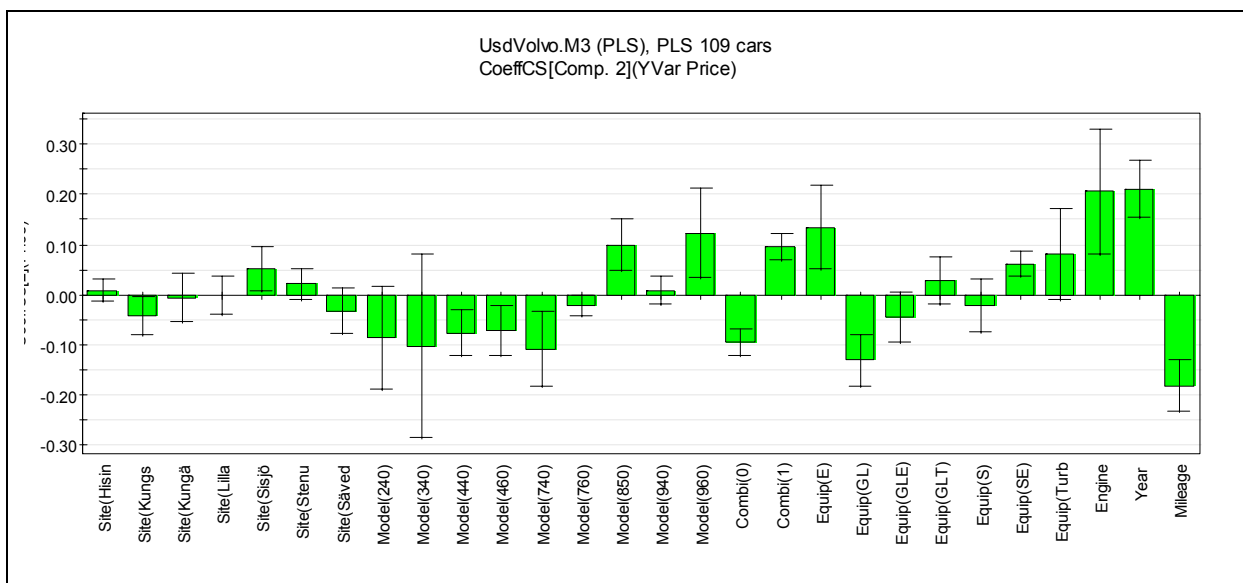


Having looked at the data we found that there is only one car of model 340. We might choose to exclude this car as this car design is not well represented in the data, or, alternatively, we could include more data for 340 cars.

The w_{c1}/w_{c2} loading plot shows the correlation structure among the variables.

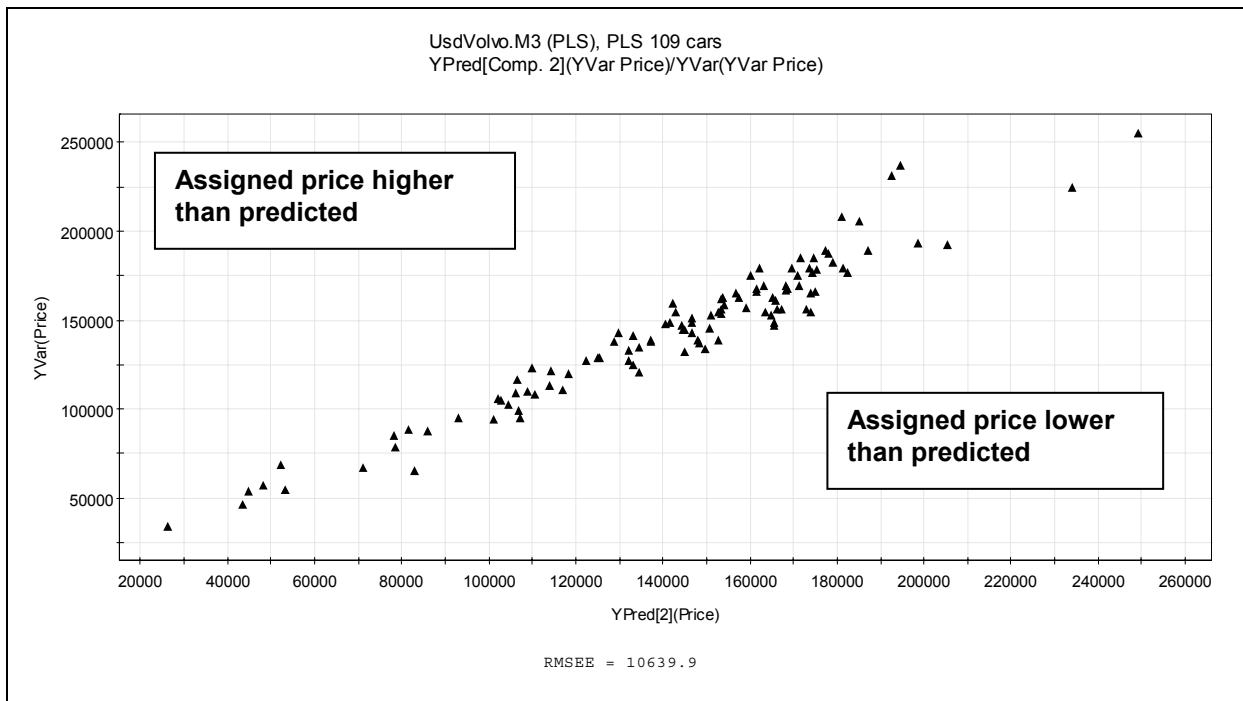


The scaled and centred regression coefficients show the importance of the X variables for the prediction of the price.

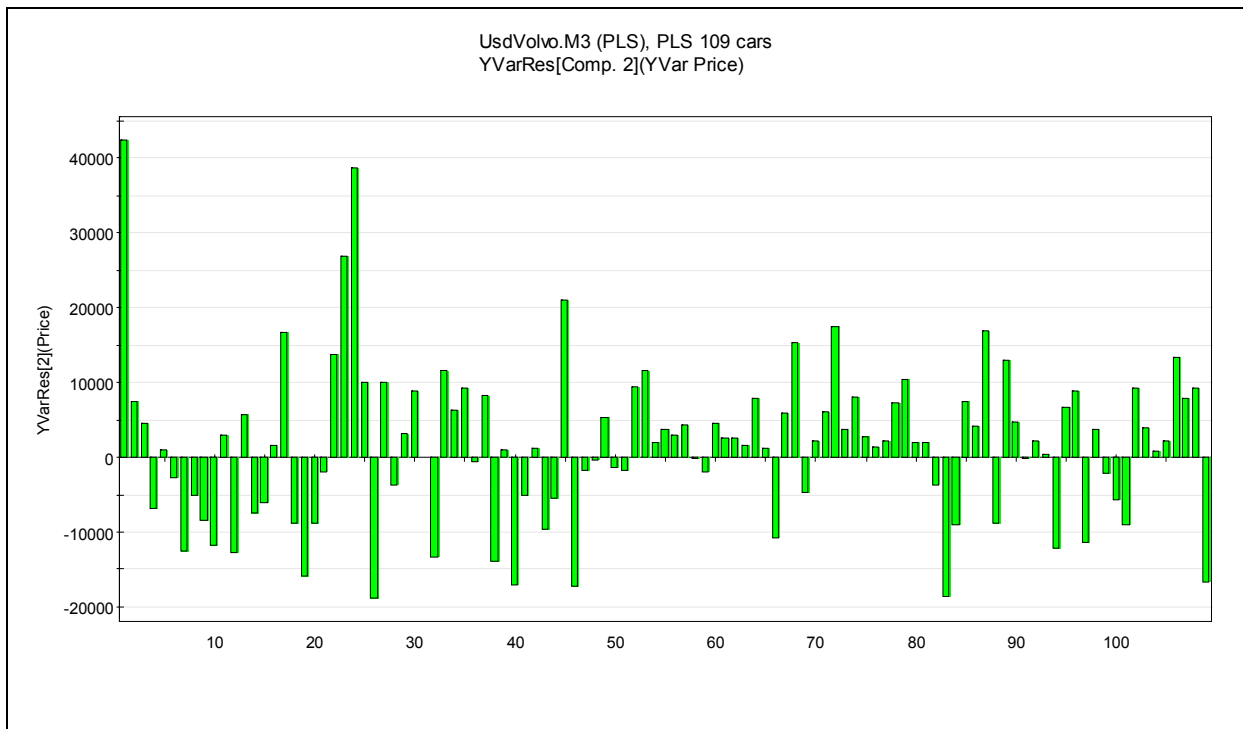


Most important for the price were Engine and Year, followed by Mileage, Combi, 850, and GL. In Sisjön they seemed to have the highest prices.

We may look at the predicted prices and compare them with the assigned prices. Cars of interest to us are of course those where the assigned price is lower than the predicted.



A better way of studying the differences is to plot the Price residuals, i.e., the difference between the assigned and the predicted prices (Hint: *Plot/List | Column Plot | Y VarRes | Price | A=2*).



The most interesting car seems to have a price that is about 18.000 SEK lower than the assigned price. To find the exact differences we used *Predictions* | *YPred* | *List*. Below are parts of this list:

YPredicted				
	1	2	3	4
24	23	Si855GLT96	208000	181131
25	24	Si855GLT97	231000	192361
26	25	Si854Tur94	188000	178054
27	26	Si855__94	155000	173913
28	27	Sä440GL_94	116500	106379
29	28	Sä245GL_90	67500	71216.4
30	29	Sä744GL_87	46500	43305.9
31	30	Sä744GLE87	57000	48060.6

We find that car 26 seems to be the best buy - the assigned price is 155.000, while the predicted price is 174.000. The prediction for the excluded car, number 61, is 163.000, which corresponds well with the assigned price of 171.000.

Conclusions

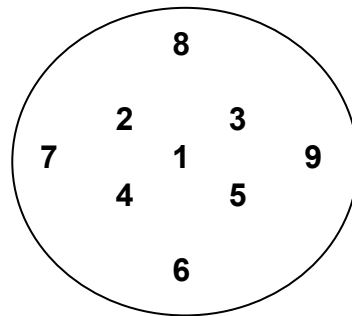
An analysis of the information in the advertisement enabled the best buy, car 26, to be identified. Surprisingly strong relationships were discovered between the car price and the X-variables. The selling site Sisjön tends to have the highest prices.

MVDA-Exercise THICKNESS

Thickness of Polymer Disks

Background

As part of a quality control scheme in a process manufacturing polymer disks, nine thickness measurements were taken on the disks produced. The first five measurements, G1 to G5, were taken near the middle of the disks and the other four, G6 to G9, were made on the periphery (see picture). The objective was to produce disks with uniform thickness within given specification. Problems stemmed from small but expensive increases in the number of disks discarded.



Objective

We would like to answer the following questions:

- How many components are there in the data? Are there outliers?
- After removing outliers, can we interpret the components and detect production failure?
- Are there any groups? Can they be interpreted?
- Do you see any trends?
- Can we find and interpret additional outliers from the DModX plot?

Data

The data set consists of nine thickness measurements made on 184 disks manufactured between October and November 1991. The data are given as deviations from the target.

Tasks

Task 1

Create a new project and import the data, Thicknes.xls. The imported file should contain 184 observations and 9 variables.

Task 2

Make a first PC model for overview using non-scaled data (Hint: *Work Set|New|Scale|Set Scaling|None and press SET*). Compute four PCs. Save the model. Plot t_1/t_2 and t_3/t_4 . Do you see any outliers? Plot p_1/p_2 , and interpret why the outliers in t_1/t_2 differ from the main cluster. Try also the contribution technique to find the reasons for the outliers (Hint: *Contribution|Scores*). Evaluate average against observation 39 for example.

Task 3

Create a new model. Modify the work set by removing the four outliers 39, 40, 111, and 155. Extract four components and save the model. Interpret the model.

Task 4

Plot t_1/t_2 and t_3/t_4 . Do you see any clustering in either plot? Plot the corresponding loading plots and interpret the groupings. Plot t_1 , t_2 , t_3 and t_4 vs Num (run order). Are there trends present? Look in the corresponding loading plots to interpret the trends.

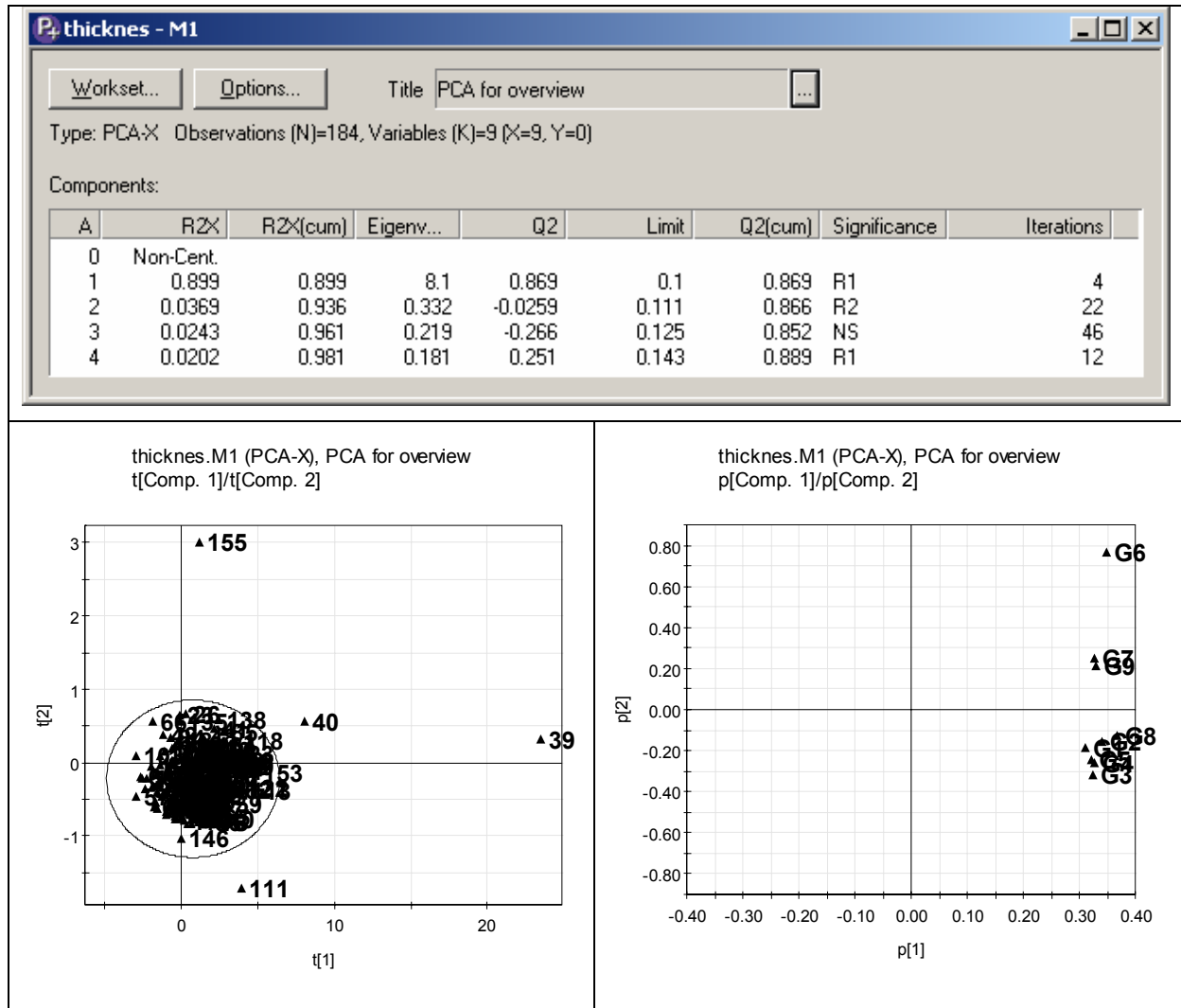
Task 5

Look for moderate outliers and anomalies in the observations (disks) over time by plotting DModX (distance to model), *Analysis|Distance to model*. Do you see any outliers? Looking at the residuals, can you explain why the aberrant samples are outliers? Use *Analysis|Contribution|Distance to model X-block*. Using the list function in SIMCA (*Plot/List|Lists|X Var Res*) it is also possible to list the residuals for specific observations.

Solutions to THICKNESS

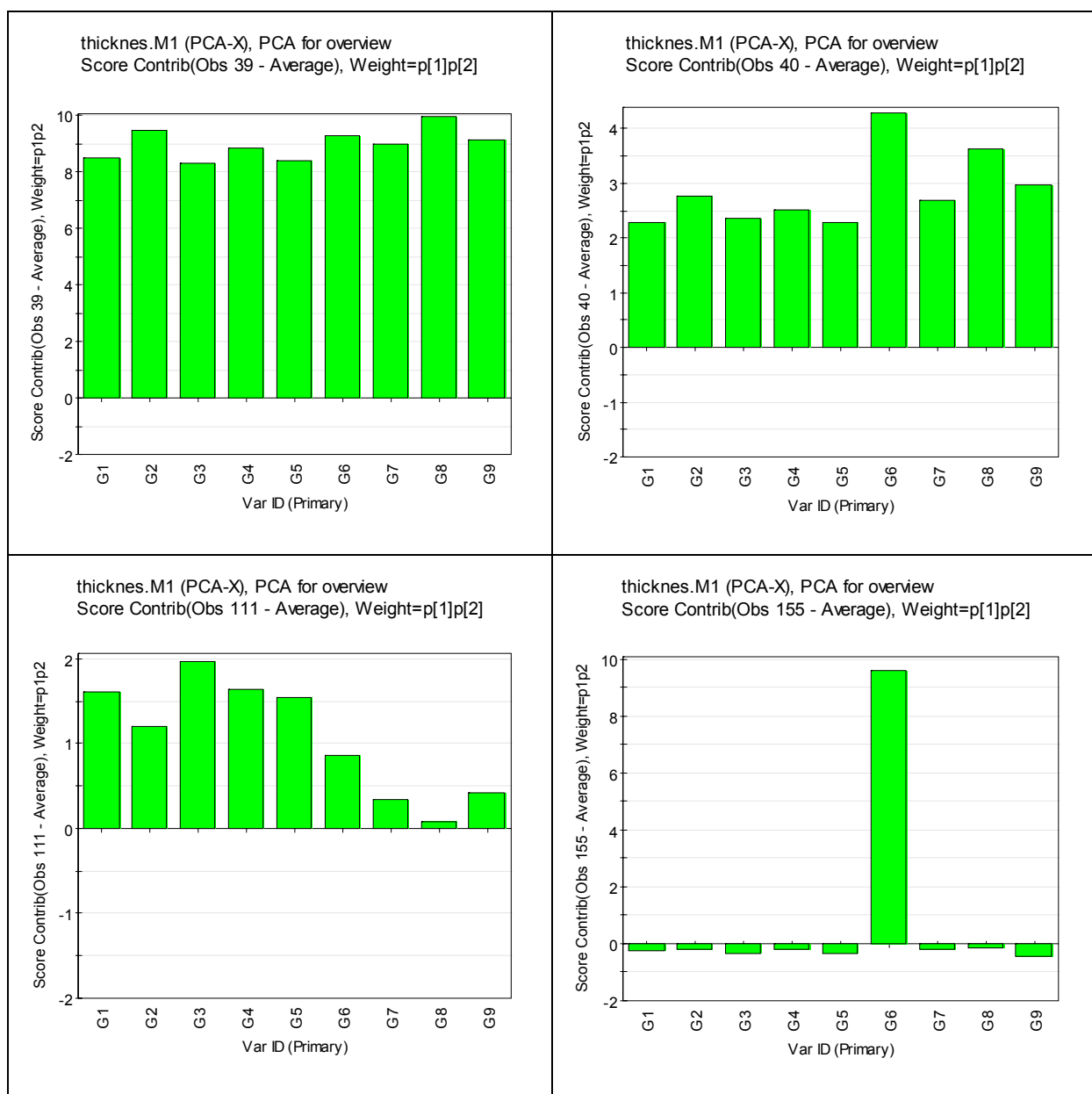
Task 2

A four-component PCA-model was generated to overview the data.



Four clear outliers can be identified in the t_1/t_2 score plot; observations 39, 40, 111 and 155.

Contribution plots reveal which variables are responsible for the outliers. Here we have made contribution plots focussing on the difference from the projection centre to observations 39, 40, 111 and 155.



The interpretation of the four contribution plots indicates that:

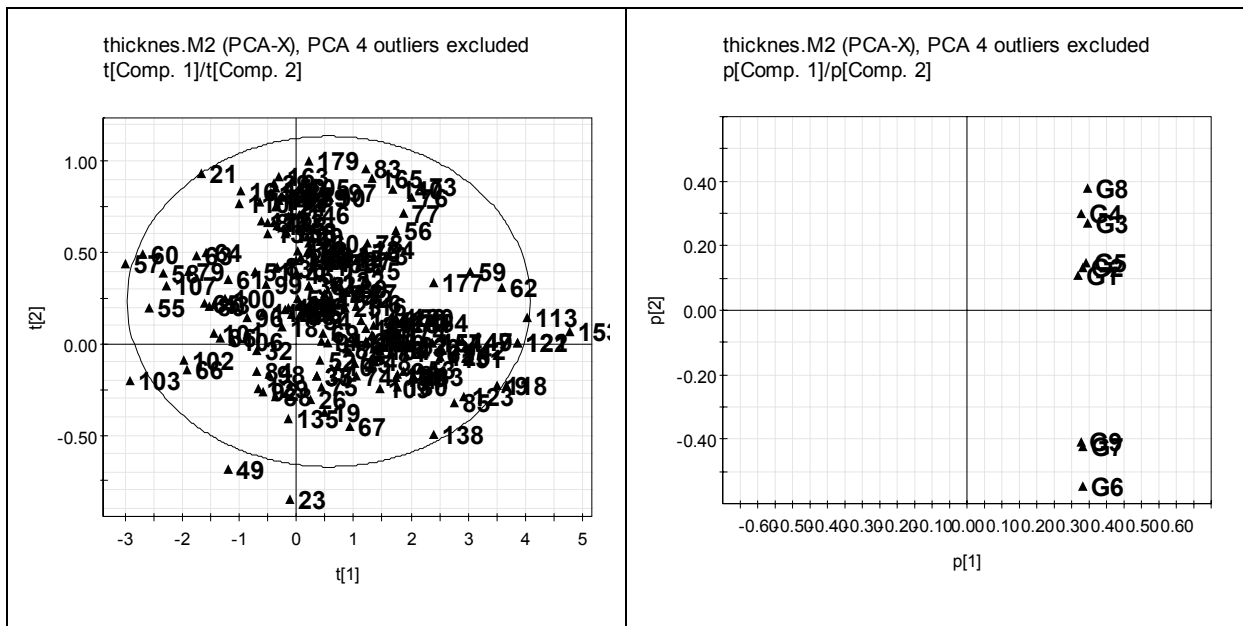
- 39 is thicker in all measurements
- 40 is also thicker all over but not as extreme
- 111 is thicker in the interior part
- 155 is thicker in one of the measurements, G6, on the periphery.

Task 3

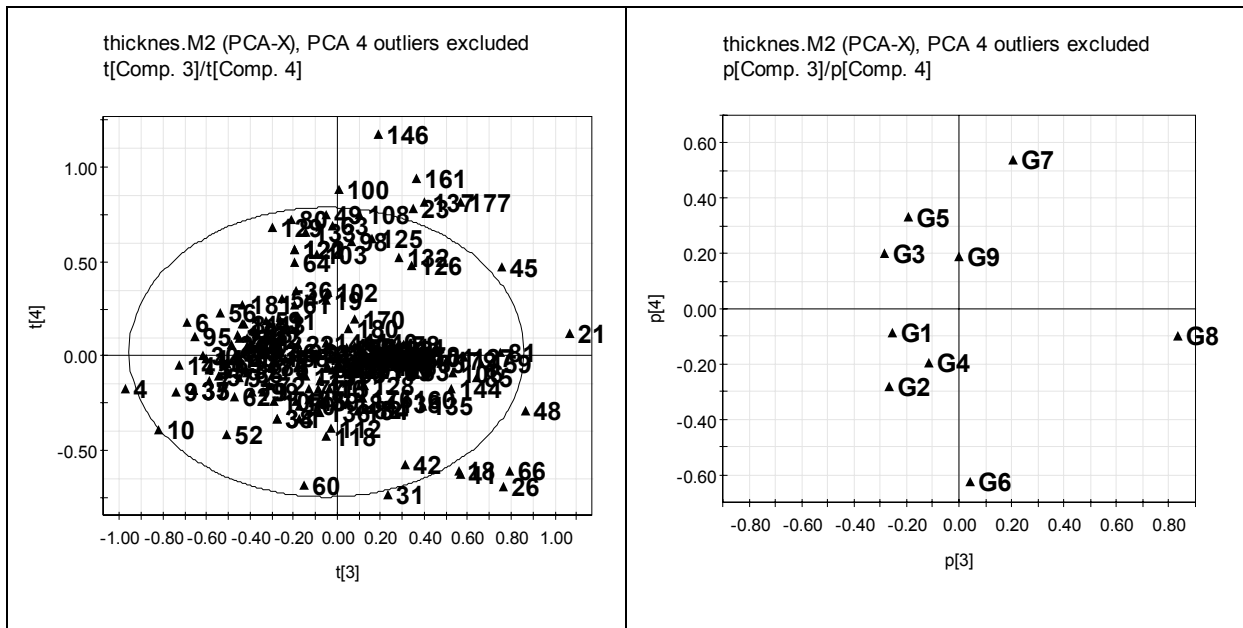
The outliers 39, 40, 111 and 155 were removed. The results after computing four components are shown below.

PCA 4 outliers excluded									
Type: PCA-X Observations (N)=180, Variables (K)=9 (X=9, Y=0)									
Components:									
A	RZX	RZX(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations	
0	Non-Cent.								
1	0.816	0.816	7.34	0.765	0.1	0.765	R1	6	
2	0.0652	0.881	0.587	0.0558	0.111	0.778	R2	36	
3	0.0476	0.929	0.428	-0.176	0.125	0.756	NS	24	
4	0.0334	0.962	0.3	0.22	0.143	0.81	R1	15	

Task 4

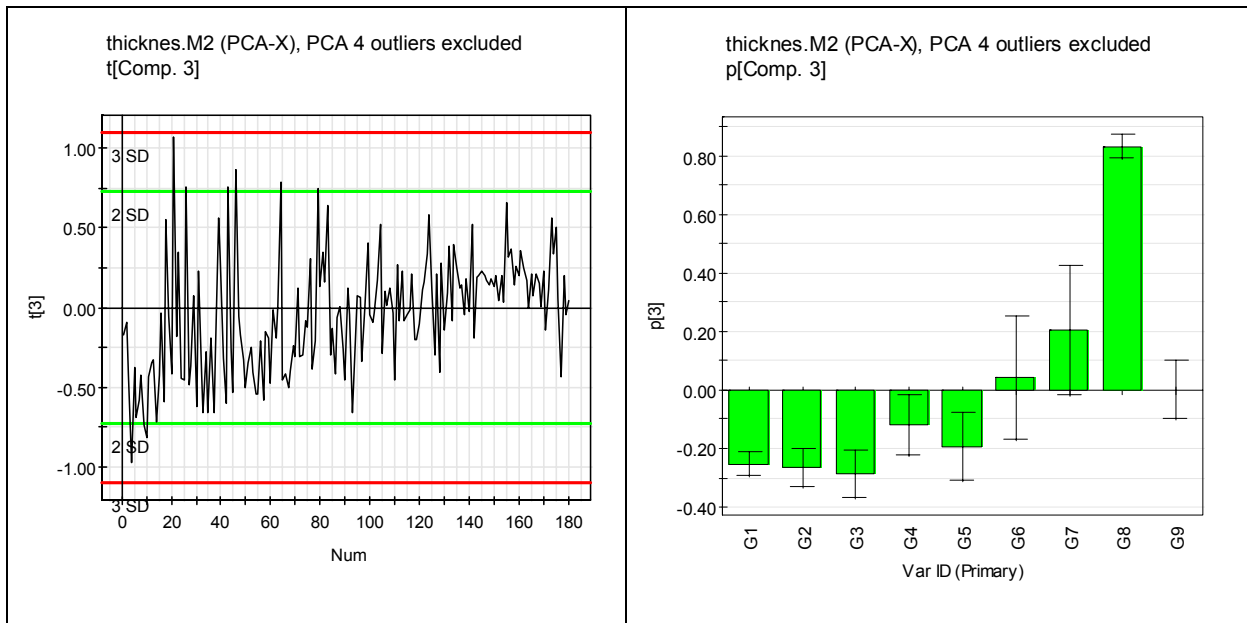


The score plot, t_1/t_2 , shows no groupings, but t_3/t_4 looks a little more suspicious. The first component expresses general disk thickness. The second component reflects shape variations (convexity or concavity). The third component explains a trend in data (discussed below). The fourth component is hard to interpret.

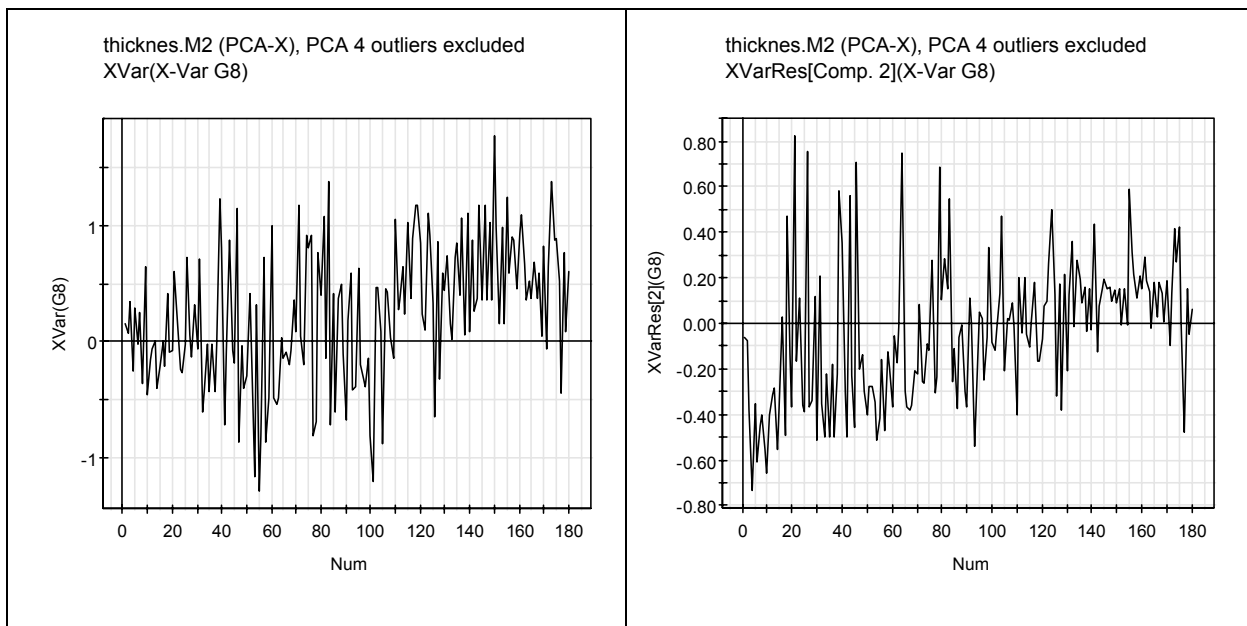


The observations are predominantly located in the middle region of the t_3/t_4 score plot, but there are some observations that are scattered around this main cluster. The periphery variables G6, G7 and G8 are responsible for most of this additional variability.

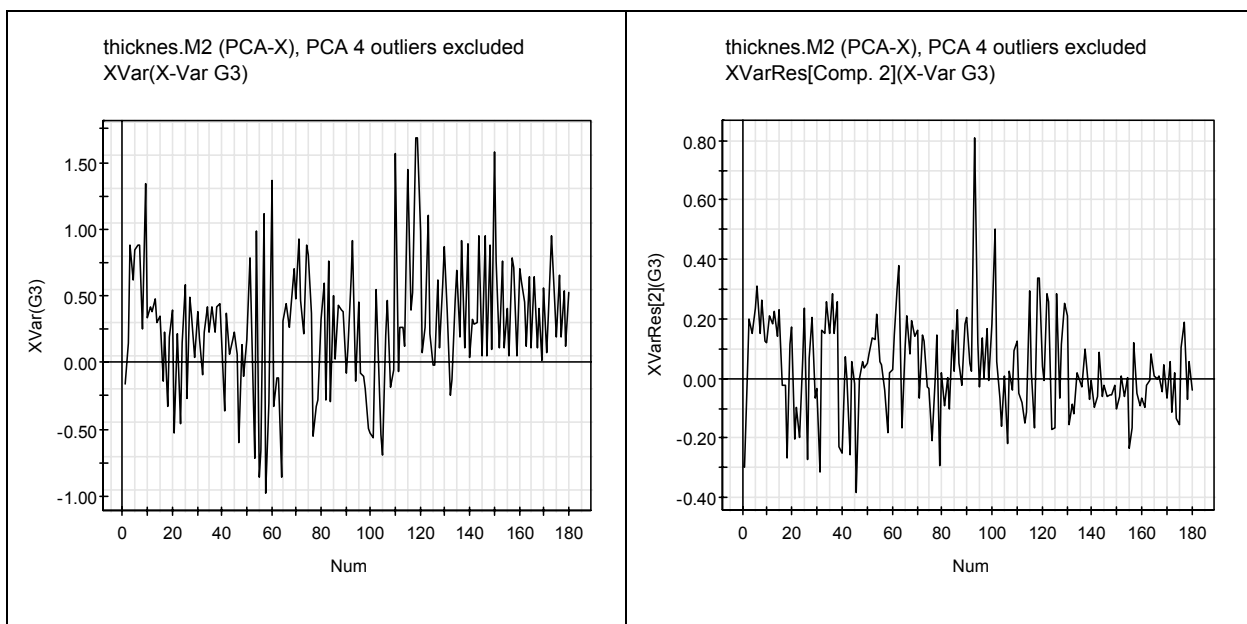
When we look at Num/ t_1 , Num/ t_2 , Num/ t_3 and Num/ t_4 we can see an upward trend in the Num/ t_3 plot.



The loading plot shows that the trend in Num/ t_3 is most likely due to an increase in G8.



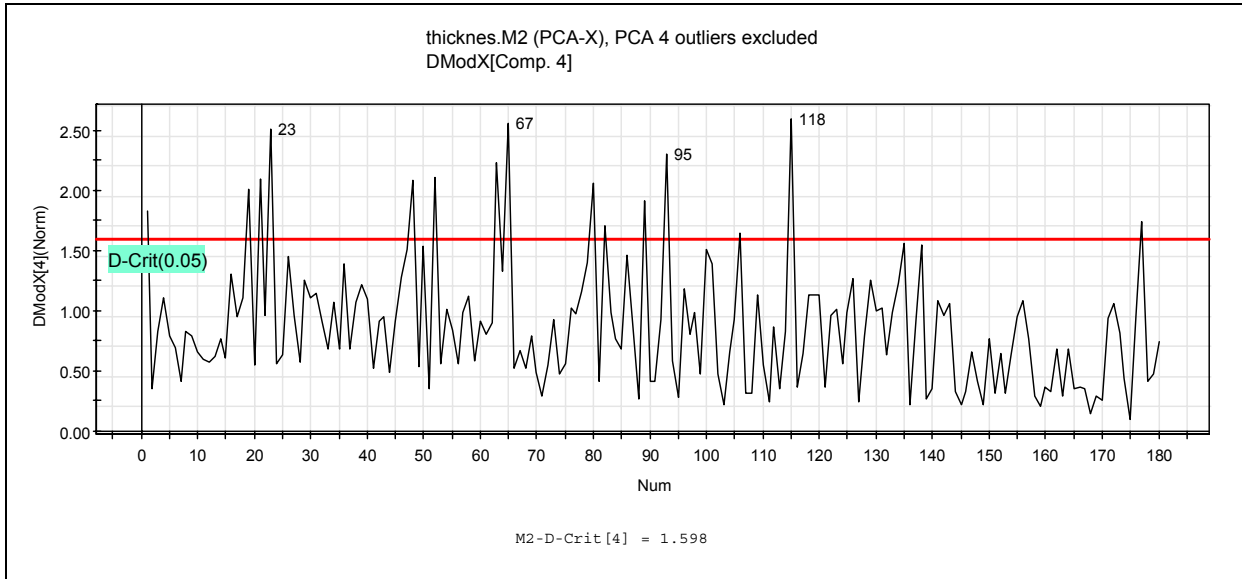
Detecting the change in G8 like this actually hints at the solution to the problem. This shift is not so obvious when we look at the raw data (above left), but if we look at the residual in G8 after two components (above right) we can see a very clear trend. This trend is then modelled by the third component.



If the raw data for G8 are compared with the raw data for G3 (above left), it is hard to detect systematic differences, but if the residuals after two components are compared (above right) it is easy to see a trend in G8 but not in G3.

Task 5

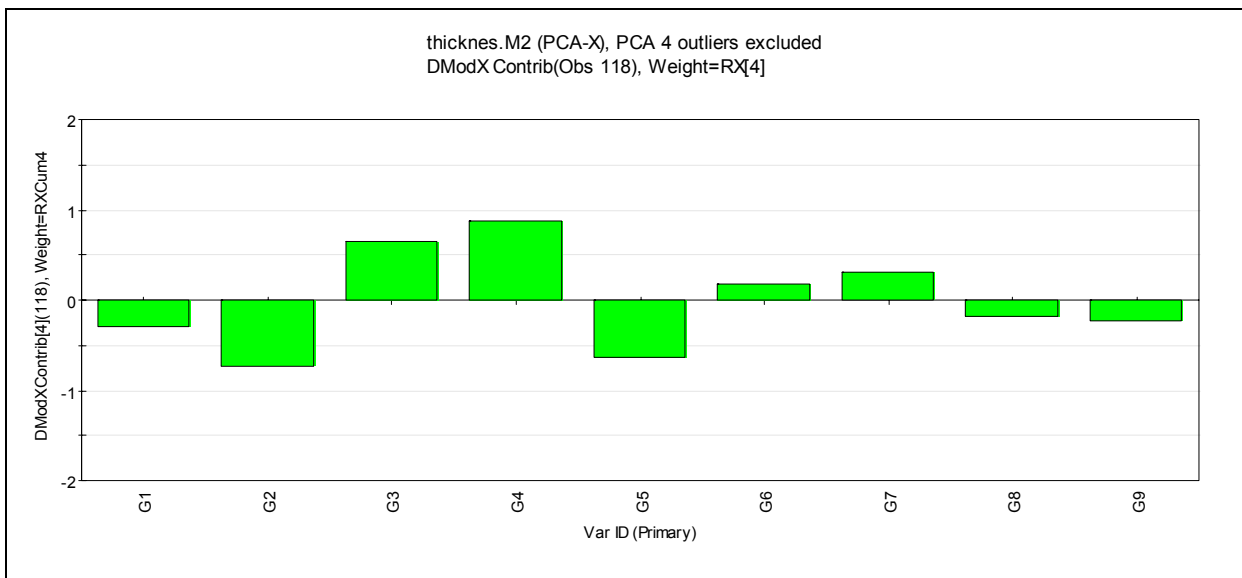
In order to look for moderate outliers, i.e. observations that do not conform to the model and break the general correlation structure, it is recommended to make DModX plots.



There are some moderate outliers in DModX. The causes for moderate outliers can be found in the list of variable residuals (VarRes). Below we see that observation 118 has large residuals, positive and negative, in most variables: a skew disk.

115	08NOV91:08:52	-0.0478348	0.104162	-0.0688835	-0.130647	0.179184	0.0439014	0.00388044	0.0150761	-0.104346
116	08NOV91:08:55	0.0153376	0.0209526	-0.0749038	-0.0108951	0.0557913	-0.0212192	-0.0371368	0.00742719	0.0472173
117	08NOV91:08:57	-0.0472704	0.106502	-0.109305	-0.0905104	0.180579	0.0259939	0.00428896	0.00899075	-0.0822334
118	08NOV91:08:59	-0.145346	-0.39204	0.367836	0.471846	-0.353768	0.10701	0.17638	-0.108204	-0.130022
119	08NOV91:09:12	-0.0244482	0.000894054	0.0371441	-0.0207893	-0.0220188	-0.0160079	-0.0580029	0.0125626	0.087882
120	08NOV91:13:26	0.0415702	0.0230015	-0.115645	0.0141929	0.0333021	-0.0632847	-0.0722373	0.0112777	0.134332
121	08NOV91:13:34	-0.0707664	-0.0402284	0.167343	-0.144632	0.00242643	0.0126807	-0.195748	0.0504382	0.20472

In the contribution plot it can also be seen that observation 118 has large residuals, both positive and negative, see below.



Conclusions

Based on cross-validation, there are two-four significant components for this data set. The first explains the thickness variations, the second explains the shape (if the disk is convex or concave) and the third explains a trend in the data. The fourth component is hard to interpret. There are no groupings in the first two components, but in the third and the fourth there is some clustering. These groupings can be interpreted by looking in the corresponding loading plots. Some moderate outliers can be seen in the DModX chart. These outliers can be interpreted by looking in the contribution/distance to model plot for the respective observation.

MVDA-Exercise CUPRUM

MSPC of an electrolysis process

Background

In an electrolysis process very pure copper (>99.998%) was produced. To monitor the quality of the copper, the levels of eight impurities (Ag, Ni, Pb, Bi, Sb, As, Te, and Se) were determined. These impurities were weighted together to form the total analysis index, TAI, which is used as an overall index to determine the quality of the copper. The problem for the manufacturer was that the univariate TAI index did not provide enough information about the product quality. Hence, the manufacturer wanted to see if multivariate modelling could be used to determine the quality of the copper more accurately.

Objective

The objective with this exercise is to contrast univariate and multivariate control chart approaches, and to demonstrate how the use of the TAI index gives misleading results with regard to outlier detection and copper quality determination. The objective is also to introduce how PCA-results may be displayed graphically in multivariate control charts. Multivariate Shewhart, cumulative sum (CuSum) and exponentially weighted moving average (EWMA) control charts are used and compared.

Data

The dataset has 730 observations (2 samples per day during one year of copper production) and 9 variables (Ag, Ni, Pb, Bi, Sb, As, Te, Se and TAI).

Tasks

Task 1

Create a new project with the Cuprum data in CUPRUM.XLS. First we have to check the data. Make histograms of the variables (*DataSet/Quick Info/Variables* or *Plot/List/Histogram Plots*, Select DS1 and VarDS).

Task 2

Now, we are going to make a univariate quality analysis by looking only at the TAI variable. Create a line plot of TAI as a function of time (*Plot/List/Line Plots*, select *DS1/VarDS/TAI* for the Y-Axis and *DS1/Num* for the X-Axis; Use None as plot label). Investigate samples 111 and 302 (Use the Yellow flag button in the toolbar). What may be said about their quality? Are they similar or dissimilar with regard to copper quality?

Task 3

According to TAI, samples 111 and 302 are comparable. The problem for the manufacturer was that customers did not find samples 111 and 302 similar. We will use PCA to try to understand and explain this. Make sure that variables 1-8 have been log-transformed. (In: *Work set/New*, use the tab marked *Transform*. Select variables 1-8 and log-transformation).

Make sure that the TAI-variable is excluded from the workset before calculating the PC:s. Run PCA and extract the first two components. Create necessary score-, loading- and contribution plots, and interpret these. Why are samples 111 and 302 of dissimilar quality?

Task 4

To overview a PC-model, we may plot Hotelling's T^2 as a function of time. Make a Shewhart control chart of T^2 and DModX (*Plot/List/Control Chart*, select *Shewhart* and *T2 comp 2*). Investigate samples 167, 195, 228, 338, 399, 577 and 611. These samples deviate in different ways from the majority of the copper samples - How? Compare with the time series plot of TAI, which you created in Task 1. Do you find the same information with TAI?

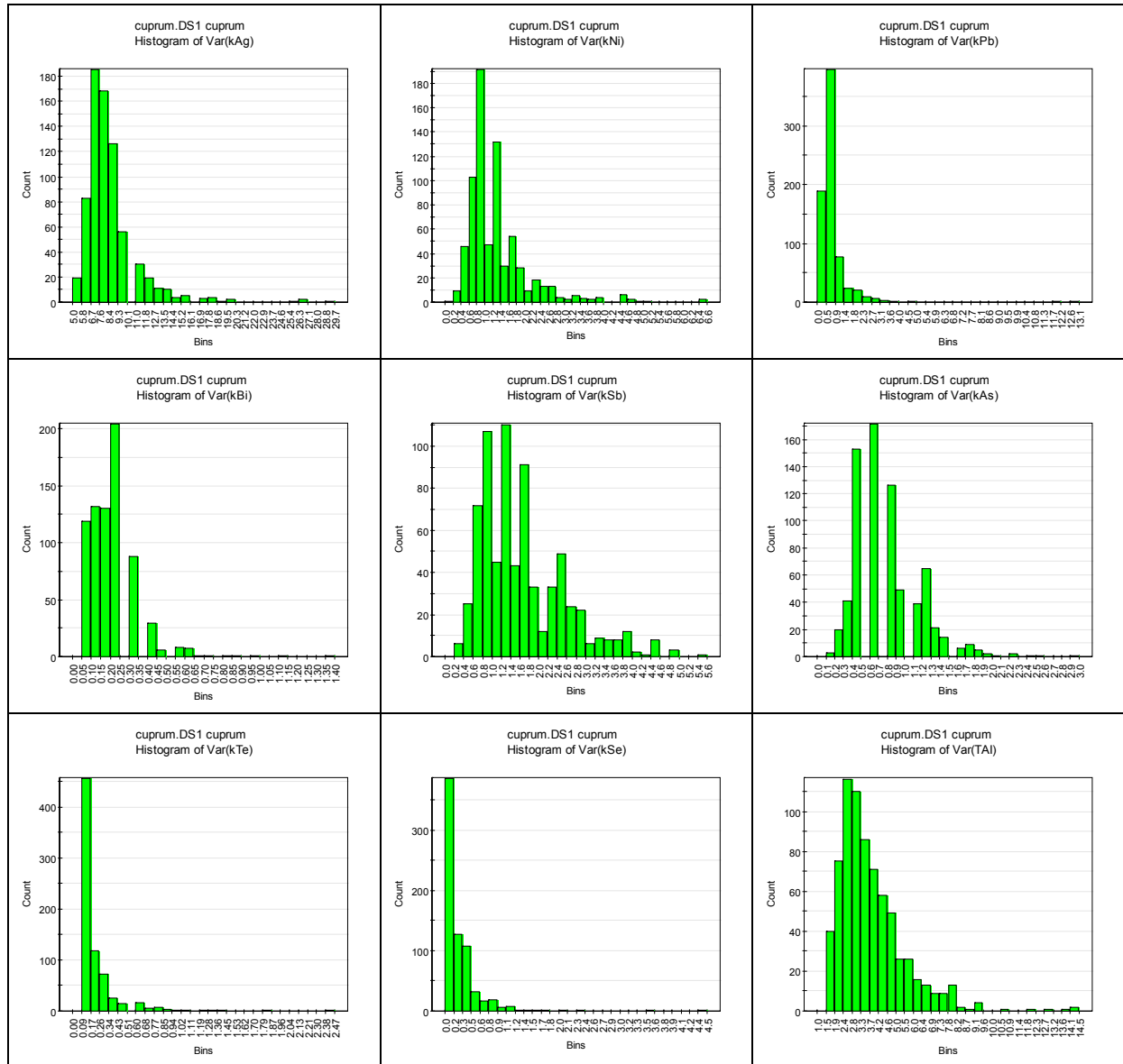
Task 5

It is also possible to monitor the scores themselves and not their summary T^2 . Let us focus on t_1 . Make Shewhart, CUSUM and EWMA control charts of t_1 and compare these. Use subgroup size = 2 for the latter two. Which chart is most useful for the copper electrolysis process?

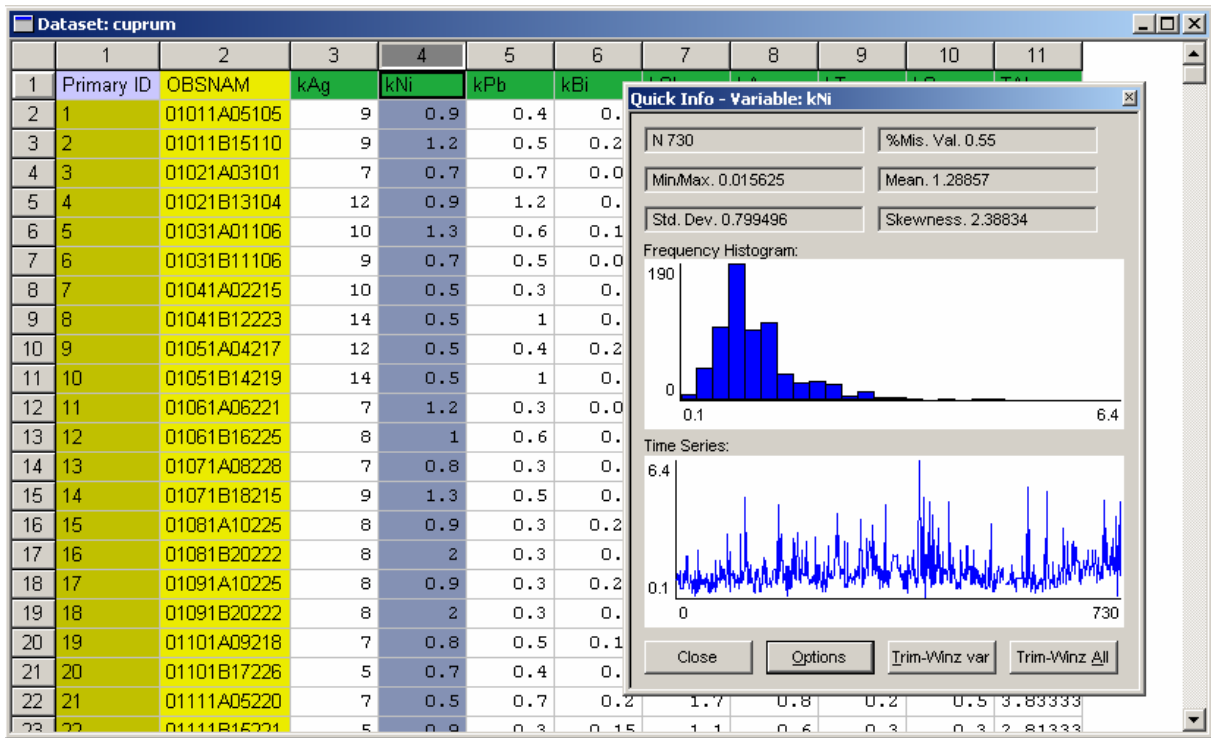
Solutions to CUPRUM

Task 1

We see that all nine variables are skewed to the right. This indicates that the log-transform is appropriate. As a rule of thumb, a skewness test > 2 indicates strongly skewed data. Six variables (Ag, Ni, Pb, Bi, Te and Se) may therefore benefit from log-transformation. The TAI index, being a weighted sum of the other variables, is more normally distributed than most of the variables. To make things easy we log-transform variables 1-8, but leave TAI as it is, since it will not be used in the PCA modelling.

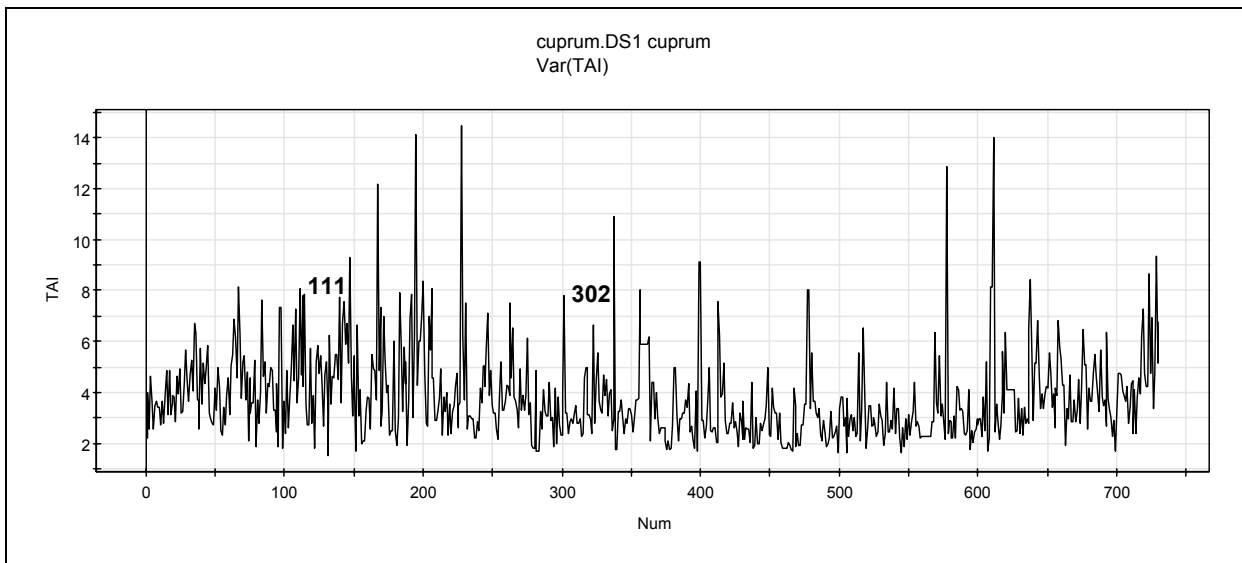


The Quick Info shows, in addition to the histogram, some statistics including the skewness.



Task 2

A line plot of TAI is shown below. A TAI value of 8 is the critical quality limit and copper samples with TAI exceeding 8 will be discarded. Certain samples have been marked for comparative purposes. Sample 111 has TAI = 8.1 and is just outside the critical quality limit and sample 302 has TAI = 7.8 making it just inside this limit. According to TAI, samples 111 and 302 are comparable. However, the problem for the manufacturer is that these samples are perceived by customers to be of very different quality.

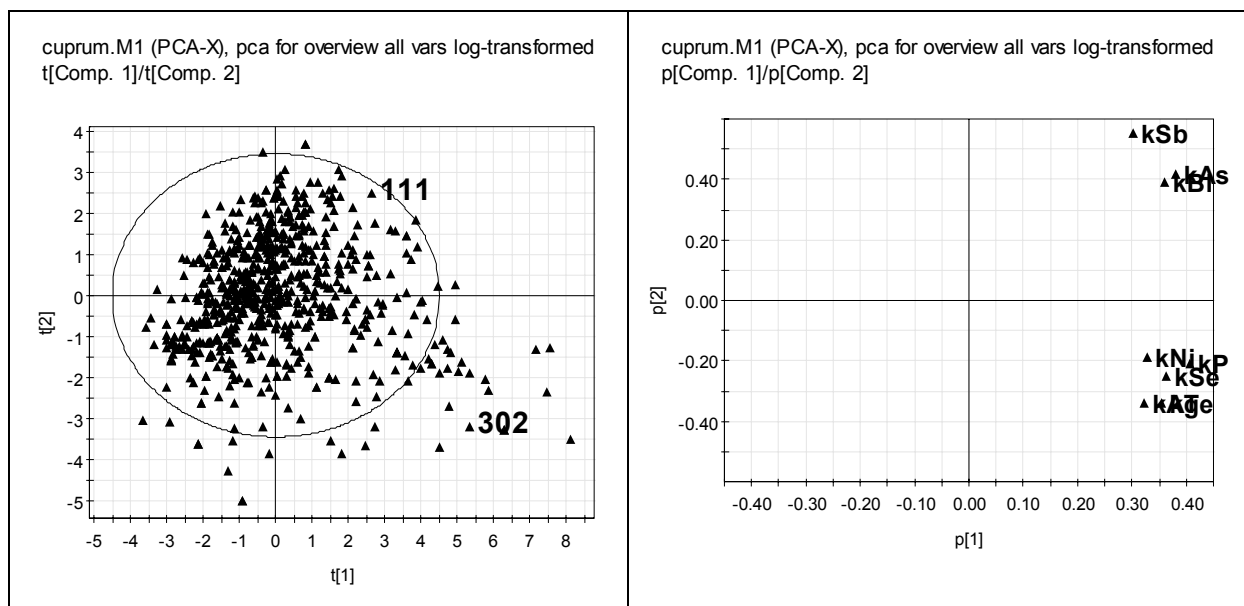


Task 3

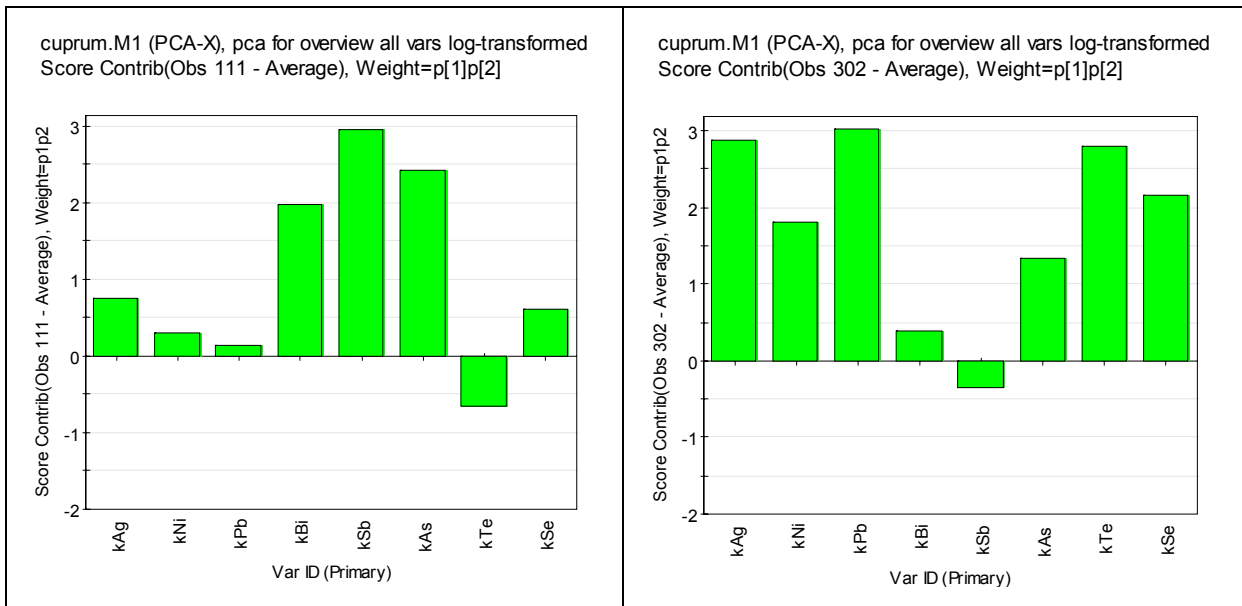
A two-component PCA model gave $R^2X = 67\%$ and $Q^2 = 41\%$, which are reasonable values for this type of process.

A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.							
1	0.425	0.425	3.4	0.17	0.112	0.17	R1	21
2	0.247	0.671	1.97	0.284	0.126	0.406	R1	9

The t_1/t_2 score plot reveals that samples 111 and 302 are not of comparable quality. Using the corresponding loading plot we can interpret the first score as reflecting “average” impurity and the second score is modelling deviations from this “average” impurity. Copper samples to the far left in the score plot have the best quality (lowest levels of impurities) and those on the right-hand side the poorest quality. Samples 111 and 302 have different types of impurities. Sample 111 has comparatively high levels of Bi, Sb and As (see the contribution plot also). Sample 302 has high levels of other impurities, viz. Ag, Ni, Pb, Te, Se and to some extent As. This information was lost when summarising the eight variables as the TAI index.

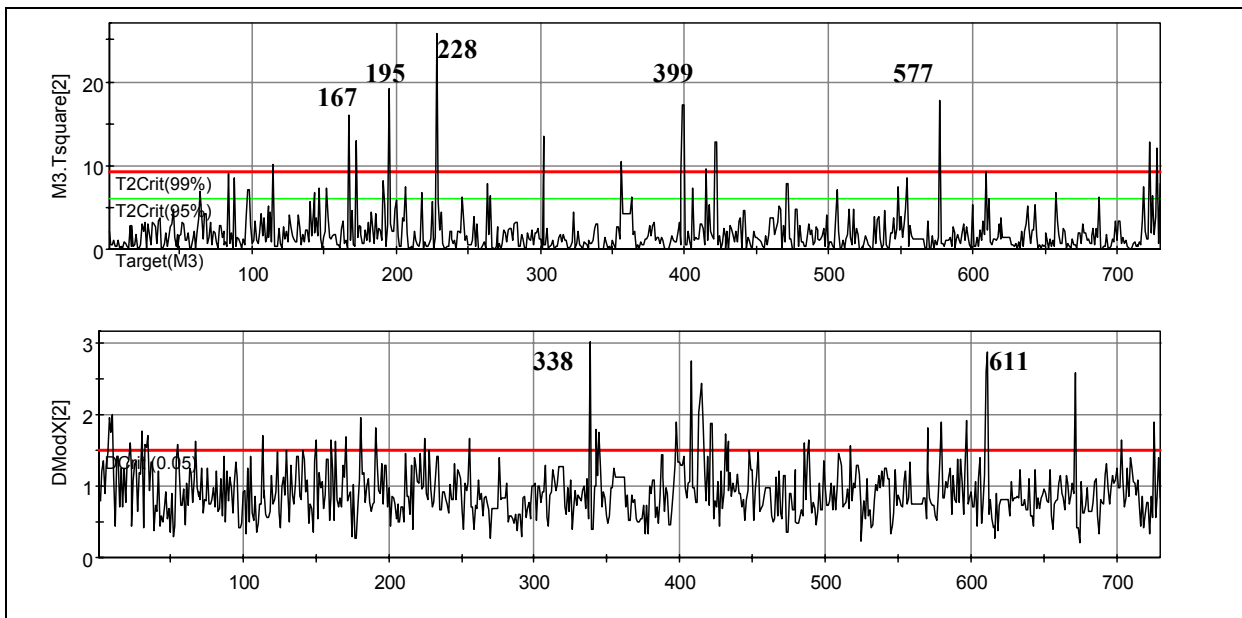


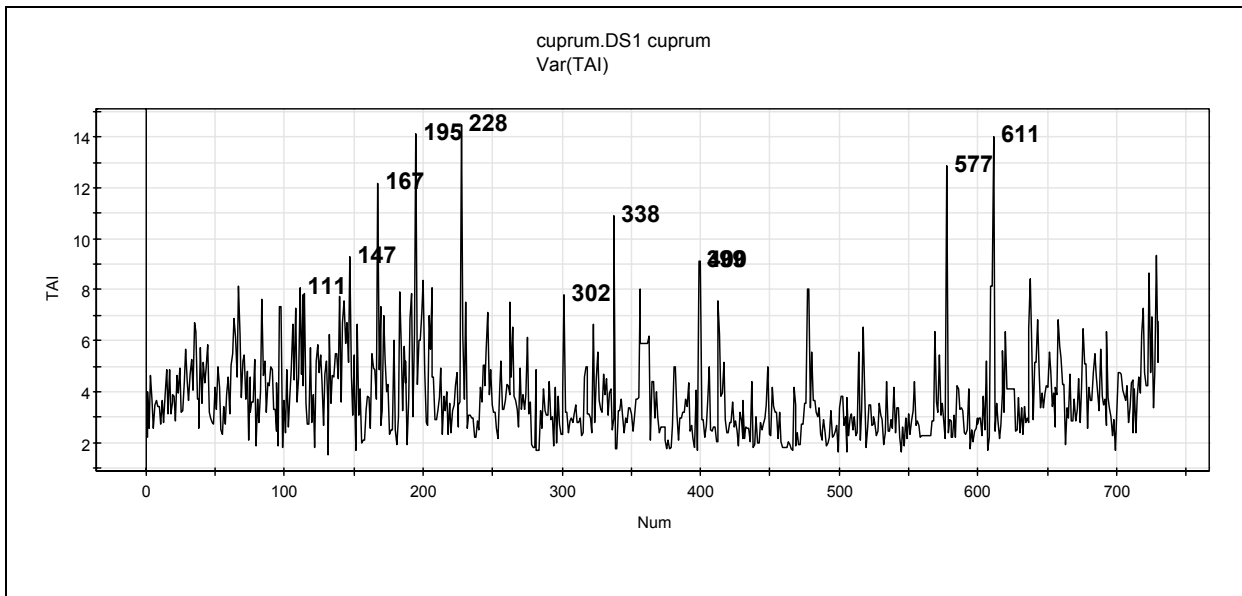
Why was this information lost using TAI? The answer is that the impurity variables are correlated and migrate “up and down” in two clusters. They do not vary independently. Hence, the data analysis must capture these variable correlations. The TAI scale does NOT capture these correlations, because it is a weighted sum of the original variables. The PCA model is based on projections to latent variables, which model these variable correlations, and therefore are better summaries of the original variables.



Task 4

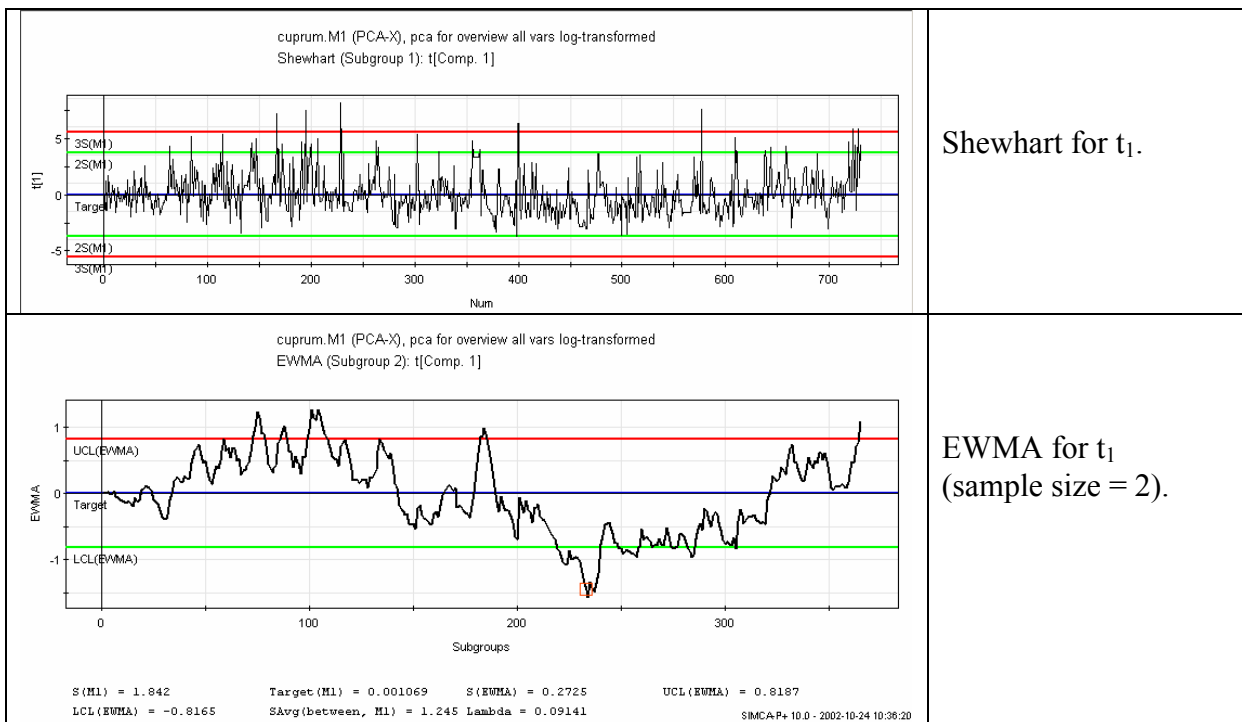
Hotelling's T^2 may be used to find samples which deviate strongly from the model. We can instantly see that samples 167, 195, 228, 399 and 577 are strong outliers. Another tool, DModX may be used to detect moderate outliers such as samples 338 and 611. Moderate outliers *break* the general correlation structure, and are thus the tricky samples to identify. They lie away from the plane of the PC model and hence are not picked up by the Hotelling's T^2 . Having understood these two types of outliers, we now realise the misleading nature of the univariate approach. The TAI scale is unable to distinguish between these two types of outliers and hence vital information goes undetected.

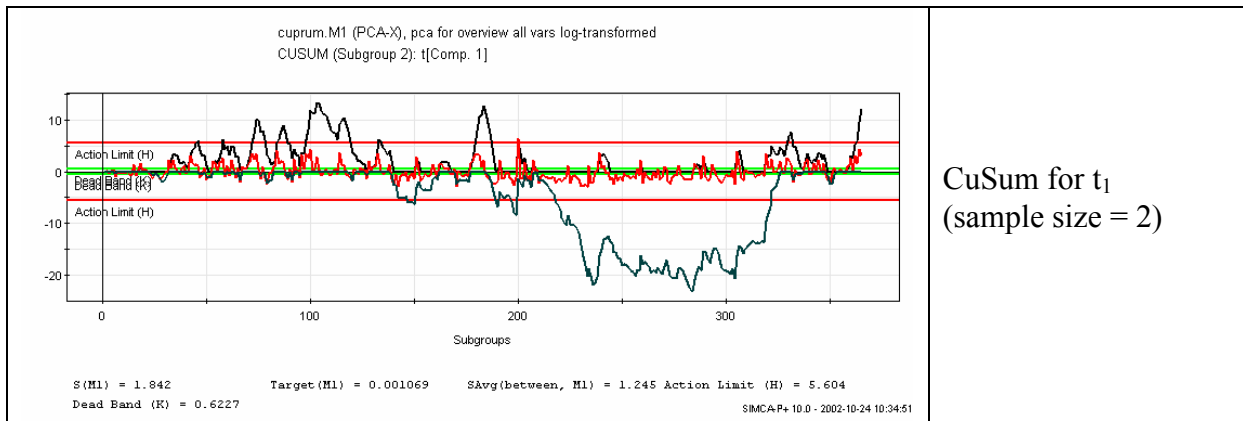




Task 5

The value of λ estimated from the data is 0.09. This corresponds to a long memory and hence the Shewhart chart (which corresponds to $\lambda = 1$) is inadequate for this process. EWMA is more appropriate. The use of subgroup size 2 (sample size) for EWMA results in a smoothing of the data, rendering it easier to detect small process shifts. The small λ -value means that the CUSUM control chart may also be appropriate. In fact, here a relatively long period of systematically negative deviations from the process mean is easier to detect than in the EWMA chart. This is vital information for a process operator or a quality engineer.





Conclusions

In this copper plant, the use of the TAI index was not very successful. The two samples 111 and 302 were claimed to have similar quality, and no indication of their completely different impurity patterns was acquired. However, PCA suggested that these samples were not of comparable quality, because the two samples were situated far apart in the score plot. The loading plot provided clues for the interpretation by indicating two kinds of impurity profiles. This is possible because PCA models the correlation structure among all quality variables. The eight quality variables migrate up and down in clusters and do not vary independently of each other. This information was not shown by the TAI-scale.

MVDA-Exercise SURFACTANT

QSAR modelling of aquatic toxicity and washing performance of non-ionic surfactants

Background

Non-ionic surfactants are important in commercially available detergent mixtures, since they extract the hydrophobic part of the soil from the fabric surface into the washing solution. The non-ionic surfactants used in detergent blends are often ethylene-oxide (EO) based and consist of several kinds of mixtures with varying molecular weight distributions (due to the polymerisation process used in their manufacturing). The surfactant molecule is composed of two parts, a hydrophilic and a hydrophobic moiety. It is the hydrophilic moiety that can be either ionic or non-ionic. Because non-ionic surfactants are increasingly used today, a general awareness exists of their possible deleterious effects on the environment. Quantitative Structure-Activity Relationship (QSAR) modelling provides an interesting instrument for exploring the adverse effects of non-ionic surfactants.

This exercise has two phases. Phase 1 deals exclusively with aquatic toxicity data. Phase 2 also involves washing performance characteristics.

Objective

The objective of this study was to investigate relationships between chemical properties and aquatic toxicity/washing performance of non-ionic surfactants.

Phase 1 Data

Lindgren¹ and Uppgård² carried out a multivariate characterisation of 36 commercial non-ionic surfactants. They were able to compile 19 chemical descriptors. These non-ionic surfactants were tested for their aquatic toxicity towards two species, the fairy shrimp *Thamnocephalus platyurus* (TP) and the rotifer *Brachionus calyciflorus* (BC). The test results were registered as a relative toxicity scale for each species. The chemical and biological variables are explained below. Note that six surfactants, numbers 2, 5, 8, 11, 21 and 31, were not tested in the BC model system.

Phase 1 Tasks

Task 1

Run SIMCA and import SURFACT1.SIM (or *.dif). Give the project a unique name. This file is composed of 36 observations (surfactants) and 21 variables. Define variables 1-19 as X and variables 20 and 21 as Y. Run PCA on the X-block. How many components can you extract using cross-validation? How many of these are really meaningful? Look at score and loading plots. What do you see?

The four first PCs can serve as a basis for multivariate design, in which a representative training set of surfactants is selected. D-optimal design was used to identify a suitable training set consisting of 10 surfactants. The selected ones were 4, 6, 9, 10, 12, 13, 15, 18, 20, and 23, which are encircled in the solutions section. The G-efficiency of this design was 76.2%.

Task 2

Define a new training set consisting of the ten surfactants listed above. Use the 19 X-variables and the 2 Y-variables and run PLS. Interpret the QSAR. Can the chemical properties of the surfactants be used to model the aquatic toxicity? Make predictions for the 26 surfactants in the validation set.

Task 3

Try to identify a 2^{4-1} fractional factorial design (FFD) in the four Principal Properties (PPs) listed in Task 2. Such a design would encode 8 surfactants. Supplement these eight with two centre-points so that you get a training set with 10 members, comparable to the size of the D-optimally selected training set. Run PLS and examine the new QSAR and its predictive ability. Compare with Task 2. Is the FFD strategy purposeful for this data set?

There is no solution given to this task!!!

Phase 2 Data

Lindgren and Uppgård later expanded the multivariate characterisation to embrace 38 non-ionic surfactants, and now their focus was placed on washing performance. For this kind of surfactant there has always been a trade-off between washing performance and biodegradation. Washing performance means the ability to remove soil from dirty fabric. A lot of the intrinsic surfactant properties are regulated by the properties of the surfactant side chain. For instance, branching gives good washing performance but poor biodegradability, whereas straight chains have the reversed pattern. Thus, it appears that a useful compromise might be a straight chain with some degree of branching. This was the working hypothesis of the current study.

The questions asked by the researchers now were: (1) Is it possible to quantitatively model performance of technical blends as a function of chemical properties? (2) Which representative surfactants should be studied?

Phase 2 Tasks

Task 4

Start a new project and import SURFACT2.SIM (*.dif). Give the project a unique name. This data file contains 38 observations and 23 variables. The 19 X-variables are the same as for Phase 1. The four Y-variables are detergency efficiency (Ydet, var 20), concentration of surfactant (Yconc, var 21), and washing temperature (Ytemp, var 22) at optimal washing conditions, and the fairy shrimp toxicity (Ytox, var 23). Define the X- and Y-blocks according to the information above. **The software will now prompt that some observations and variables are to be deleted. Do NOT accept this.** Run PCA on the X-block, make score and loading plots, and interpret the model.

Task 5

In Task 4 it was found that eight surfactants were not chemically interesting. Remove these eight compounds (numbers 3, 10, 13, 14, 17-19, and 29). Remember that this removal can be done interactively, so there is no need to use the WorkSet dialogue. **The software will now prompt that some observations and variables are to be deleted. Do NOT accept this.** Run PCA on the X-block, make score and loading plots, and interpret the refined model.

Task 6

Select surfactants 2, 5, 8, 9, 11, 30, 31, 33, 37, and 38 as the new training set. **The software will now prompt that some observations and variables are to be deleted. Do NOT accept this.** Fit a PLS model between the 19 X-variables and the four Y-variables. Change the number of cross-validation groups (*View/Project Options*) from default = 7 to 10 (because there are ten substances in the training set). Review the fit and interpret the model. Which chemical properties are likely determinants for washing performance and aquatic toxicity? Is it rational to handle all four responses in the same QSAR model, or should more than one QSAR be calculated?

Data table Phase 1

		Mw	C	redC	redC/C	Eow	Griffin	Davis	CPP	redCPP	CP	dCP	Chains	RMChain	F-alcohol	maxEO	w33EO	w66EO	CMC	logP	BATP	BABC
1	B-048	641	13	9	69.23	10	13.75	6.23	0.27	0.37	67	27.5	1	100	4.99	8	11	6	0.12	5.661	0.0469	0.0956
2	B-058	553	13	9	69.23	8	12.75	5.53	0.3	0.41	44	23	1	100	5.28	7	13	6	0.079	5.532	0.07	
3	B-065	684	16	16.1	100	10	12.87	4.75	0.27	0.27	76	4.5	2	95	4	8	16	9	0.001	7.768	-0.8842	-0.9175
4	B-09	669	15	8.5	56.67	10	13.17	5.28	0.27	0.44	54	29	1	100	0	10	9	5.7	0.31	6.109	-0.0788	0.0514
5	B-160	474	14	13.7	100	6	11.14	4.49	0.34	0.34	40	10	2	85	6.9	5	11	7	0.013	6.453	-0.7612	
6	B-267	581	15	8.5	56.67	8	12.13	4.58	0.3	0.5	16		1	100	0	7	8	5.2	0.031	5.98	-0.0611	0.0435
7	B-271	612	15	8.5	56.67	9	12.53	4.82	0.29	0.48	32	10	1	100	0	8	8.3	5.2	0.052	6.044	-0.0036	0.1141
8	B-535	393	11	10.9	98.91	5	11.22	5.43	0.38	0.38	27	10.5	2	85	11.34	5	9.4	5.7	0.11	4.801	-0.2745	
9	BOX-257	515	13	13.2	97.92	7	11.97	4.96	0.32	0.32	47	11.5	7	28	7.3	5	12.2	7.6	0.042	5.988	-0.6891	-0.4368
10	BOX-4511	707	15	14.1	96.23	11	13.7	5.82	0.25	0.26	87	22.5	5	36	0	9	18	10.2	0.23	6.775	-0.7401	-0.5716
11	BOX-915	377	10	9.8	98.69	5	11.68	5.95	0.37	0.38	36	10	6	54	12.99	4	9.2	7	0.3	4.272	-0.0009	
12	BOX-918	508	10	9.6	97.65	8	13.88	7.05	0.3	0.3	77	33	6	42	7.4	6	12	6.8	0.33	4.465	0.291	0.6124
13	GO-100	704	18	17.7	100	10	12.47	4	0.27	0.27	76	9	2	85	3.18	11	16.5	10.6	0.05	8.697	-0.8123	-0.8007
14	GT-110	730	16	16.2	100	11	13.28	5.06	0.25	0.25	90	7.5	2	90	2.77	11	12.5	8.8	0.005	7.833	-0.9397	-0.979
15	GUD-050	393	11	10.5	95.64	5	11.22	5.43	0.38	0.39			2	52	18.85	3	10.8	8.3	0.1	4.801	-0.1271	0.3537
16	GX-060	465	13	9	69.23	6	11.38	4.83	0.34	0.48			1	100	7.45	3	8.9	6.9	0.034	5.404	0.1088	0.0669
17	HDo-11	752	18	17.9	100	11	12.86	4.25	0.25	0.25	76	6	2	95	1.69	10	15.2	7.8	0.1	8.407	-1.094	-0.9471
18	HDo-7	575	18	17.9	100	7	10.68	2.85	0.32	0.32	30		2	95	4.9	6	13.2	8.2	0.005	8.149	-0.8649	-0.8148
19	HDo-9	664	18	17.9	100	9	11.91	3.55	0.28	0.28	58	7	2	95	2.97	8	14	7.8	0.03	8.278	-0.9087	-0.9327
20	Is-11	671	12	8	66.67	11	14.45	7.05	0.25	0.36	72	22	1	100	6.43	10	19	9.1	0.16	5.587	0.5646	0.5254
21	Is-8	539	12	8	66.67	8	13.08	6	0.3	0.43	16		1	100	14	8	15	11.7	0.08	5.393	0.2735	
22	Is-9	583	12	8	66.67	9	13.61	6.35	0.28	0.4	59	16	1	100	10.86	9	15.4	10.9	0.09	5.458	0.4146	0.5396
23	LON-50	379	10	7	69.5	5	11.64	5.9	0.37	0.52	34	9	2	95	13.3	5	11.3	8.9	0.899	3.882	0.8723	1.0786
24	LON-60	423	10	6.9	69	6	12.51	6.25	0.34	0.47	36	13	2	90	14.7	5	9.9	7.4	0.64	3.947	0.8547	0.9235
25	LON-70	647	10	7	69.5	7	13.22	6.6	0.32	0.44	61	18	2	95	12.04	7	11.3	7.6	0.18	4.011	0.9724	1.1068
26	LON-80	511	10	6.9	69	8	13.8	6.95	0.3	0.41	80	23	2	90	5.4	9	15.4	8.4	0.19	4.075	1.0936	1.0963
27	LTO-10	641	13	8.5	65.38	10	13.75	6.23	0.27	0.39	70	23	5	50	6.91	9	15.7	8.6	0.1	5.661	0.0263	-0.2006
28	LTO-8	553	13	8.6	66.15	8	12.75	5.53	0.3	0.43	61	12.5	4	60	7.01	8	15	8	0.07	5.532	-0.1726	0.0583
29	M-1618/10	684	16	16.1	100	10	12.87	4.75	0.27	0.27	69		2	95	3.79	9	15.6	8.8	0.004	7.768	-1.0837	-0.8868
30	M-24/60	477	14	13.9	100	6	11.08	4.4	0.34	0.34	32	3	2	95	10.49	5	10.3	7	0.04	6.453	-0.662	-0.1699
31	MO-11/50	393	11	11	99.55	5	11.22	5.43	0.38	0.38	36	14	2	95	10.71	6	12.5	9.5	0.24	4.801	0.492	
32	MO-13/100	641	13	8.6	66.15	10	13.75	6.23	0.27	0.39	78	16.5	4	60	4.85	10	17.2	8.7	0.07	5.661	0.0772	0.1942
33	MO-13/80	553	13	8.6	66.15	8	12.75	5.53	0.3	0.43	48	17	4	60	23.82	5	10.3	8	0.037	5.532	-0.1131	0.1933
34	So-10	631	12	7.2	58.13	10	13.96	6.56	0.27	0.43	74	29	2	85	0.21	10	12.2	7.2	0.19	5.432	-0.1461	0.2101
35	So-6	455	12	7.2	58.13	6	11.62	5.16	0.34	0.55	20		2	85	0.27	7	9.8	6.3	0.022	5.175	-0.2703	0.1913
36	So-9	587	12	7.2	58.13	9	13.51	6.21	0.28	0.45	59	32	2	85	0.23	9	12	6.9	0.02	5.368	-0.1702	0.2027

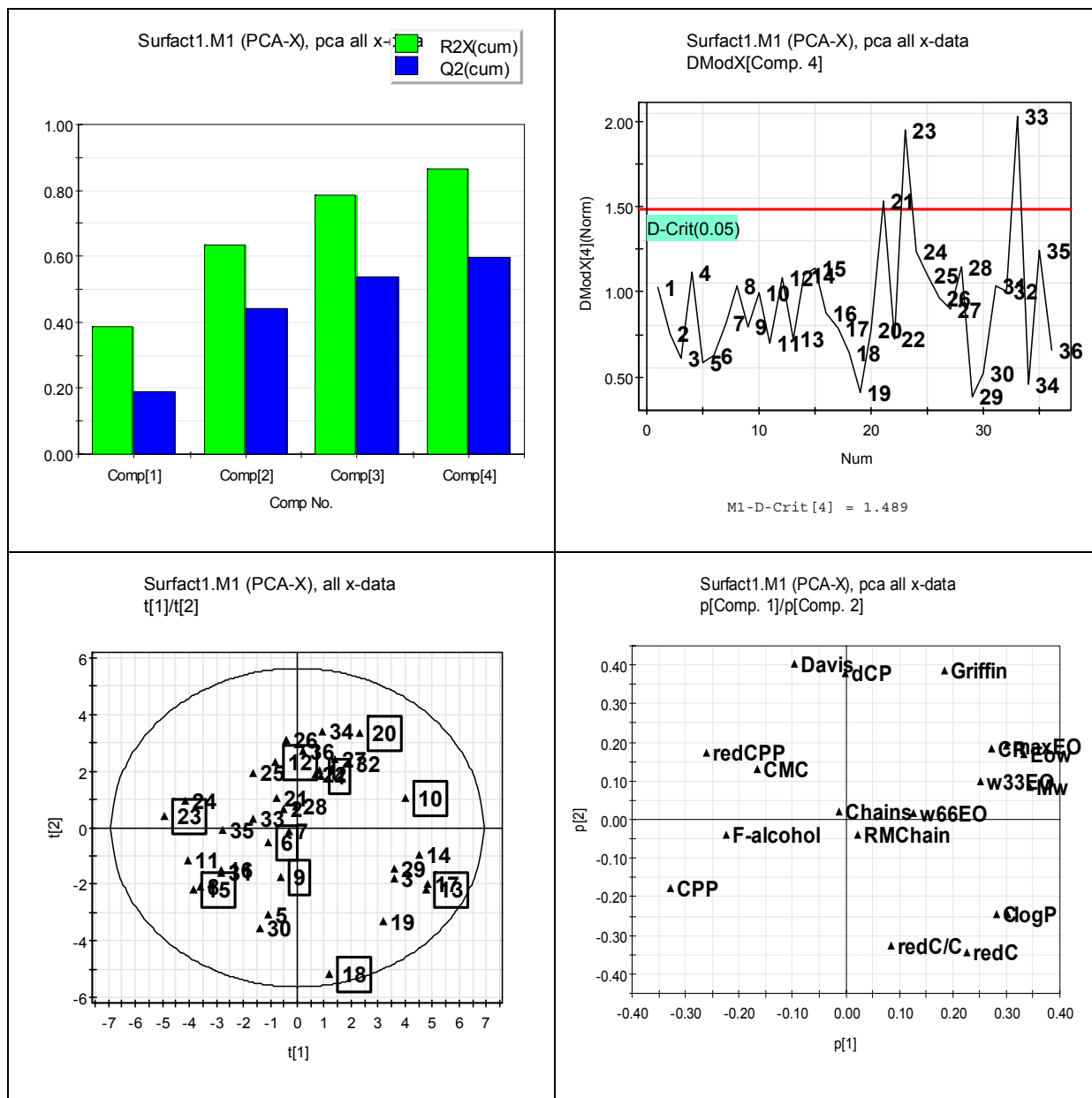
Mw = molecular weight; C = the number of carbon atoms in the hydrophobic part of the surfactant; red C = the number of carbon atoms in the longest chain of the hydrophobic part of the surfactant; redC/C = the ratio between the longest chain and the total number of carbon atoms in the hydrophobic part of the surfactant; Eow = the wanted moles of ethylene oxide per fatty acid alcohol; Griffin = The hydrophilic-lipophilic balance according to Griffin; Davis = the hydrophilic-lipophilic balance according to Davis; CPP = the critical packing parameter according to Israelachvili; redCPP = the critical packing parameter with respect to whether the hydrophobic part is branched or not; CP = the cloud point; dCP = the highest derivative of the transmittance-temperature curve; Chains = the number of different carbon chains in the hydrophobic part; RMChain = the molar ratio between the dominating type of carbon chain and the total carbon chain in the hydrophobic part of the surfactant; F-alcohol = the ratio of non-ethoxylated fatty alcohol in the surfactant product; maxEO = the position of the peak in the ethylene-oxide distribution chromatogram; w33EO = the width of the digitized chromatogram at 33% peak height; w66EO = the width of the digitized chromatogram at 66% peak height; CMC = the critical micellar concentration; log P = the logarithm of the octanol/water partition coefficient; BATP = relative toxicity scale for *Thamnocephalus platyurus* (low values imply high toxicity); BABC = relative toxicity scale for *Brachionus calyciflorus* (low values imply high toxicity).

References: (1) Åsa Lindgren, PhD Thesis, Umeå University, 1995; (2) Lise-Lott Uppgård, Graduate Thesis, Umeå University, 1995.

Phase 1 Solutions

Task 1

PCA gives 4 principal components. The first two PCs account for 64% of the variation and the t_1/t_2 and p_1/p_2 plots are therefore good summaries of the data set. In the score plot a weak grouping can be spotted. The surfactants in the lower right quadrant are the most hydrophobic. The variables C , $\text{red}C$, $\text{red}C/C$ and $\log P$ estimate the surfactants' hydrophobic properties.

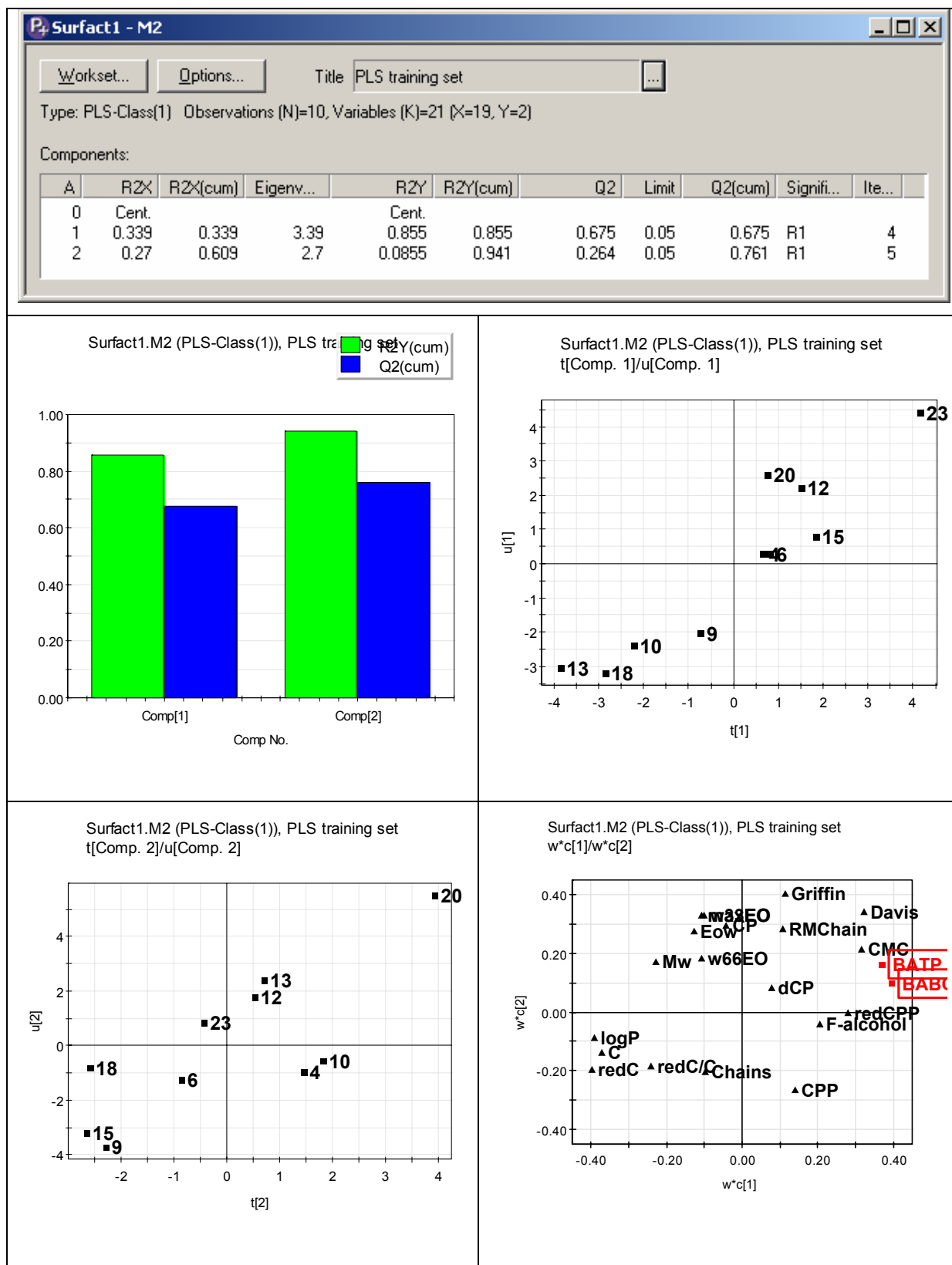


The scores of the first four PCs are given on next page.

ObsNum	ObsName	t[1]	t[2]	t[3]	t[4]
1	B-048	0.83759	2.048	-1.4613	-0.66729
2	B-058	-0.51709	0.67629	-1.5995	-0.31242
3	B-065	3.6044	-1.7779	0.33479	0.85088
4	B-09	0.85234	1.8352	-2.8646	-1.203
5	B-160	-1.0708	-3.0394	-0.07692	-0.40928
6	B-267	-1.1031	-0.52616	-3.611	-1.5443
7	B-271	-0.2828	-0.12617	-3.4053	-1.2217
8	B-535	-3.5852	-2.0931	0.083152	-0.37177
9	BOX-257	-0.59774	-1.7451	3.1552	-2.3
10	BOX-4511	4.0136	1.0717	3.1236	-0.67645
11	BOX-915	-4.0535	-1.1609	2.7084	-1.1022
12	BOX-918	-0.80209	2.3209	3.4349	-2.0985
13	GO-100	4.7551	-2.1895	0.41968	1.1292
14	GT-110	4.5119	-0.93985	0.1025	0.20453
15	GUD-050	-3.8573	-2.16	1.7651	0.53897
16	GX-060	-2.8362	-1.5066	-1.9209	-0.12165
17	HDo-11	4.8436	-2.0024	-0.34121	0.13191
18	HDo-7	1.1991	-5.1752	-0.77536	0.32409
19	HDo-9	3.1694	-3.3127	-0.57714	-0.06867
20	Is-11	2.3224	3.3702	-0.11888	1.8205
21	Is-8	-0.75404	1.0629	-0.34064	3.0001
22	Is-9	0.65145	1.8641	-0.11081	2.4385
23	LON-50	-4.9445	0.40147	0.3202	2.3674
24	LON-60	-4.1427	0.96361	0.10604	1.1818
25	LON-70	-1.6425	1.9531	-0.18382	0.63593
26	LON-80	-0.39122	3.0882	0.4335	0.89223
27	LTO-10	1.3765	2.422	1.6143	-0.832
28	LTO-8	-0.05642	0.76566	0.79037	-0.40352
29	M-1618/10	3.6029	-1.4635	0.13154	0.60842
30	M-24/60	-1.3661	-3.5374	-0.27137	0.14087
31	MO-11/50	-2.8362	-1.5717	0.88482	1.6079
32	MO-13/100	1.8868	2.2799	1.1688	-0.11234
33	MO-13/80	-1.6581	0.3024	1.1508	-0.28345
34	So-10	0.9487	3.3713	-1.0187	-0.69582
35	So-6	-2.7954	-0.07021	-2.2114	-1.1645
36	So-9	0.22218	2.7325	-1.3607	-1.1285

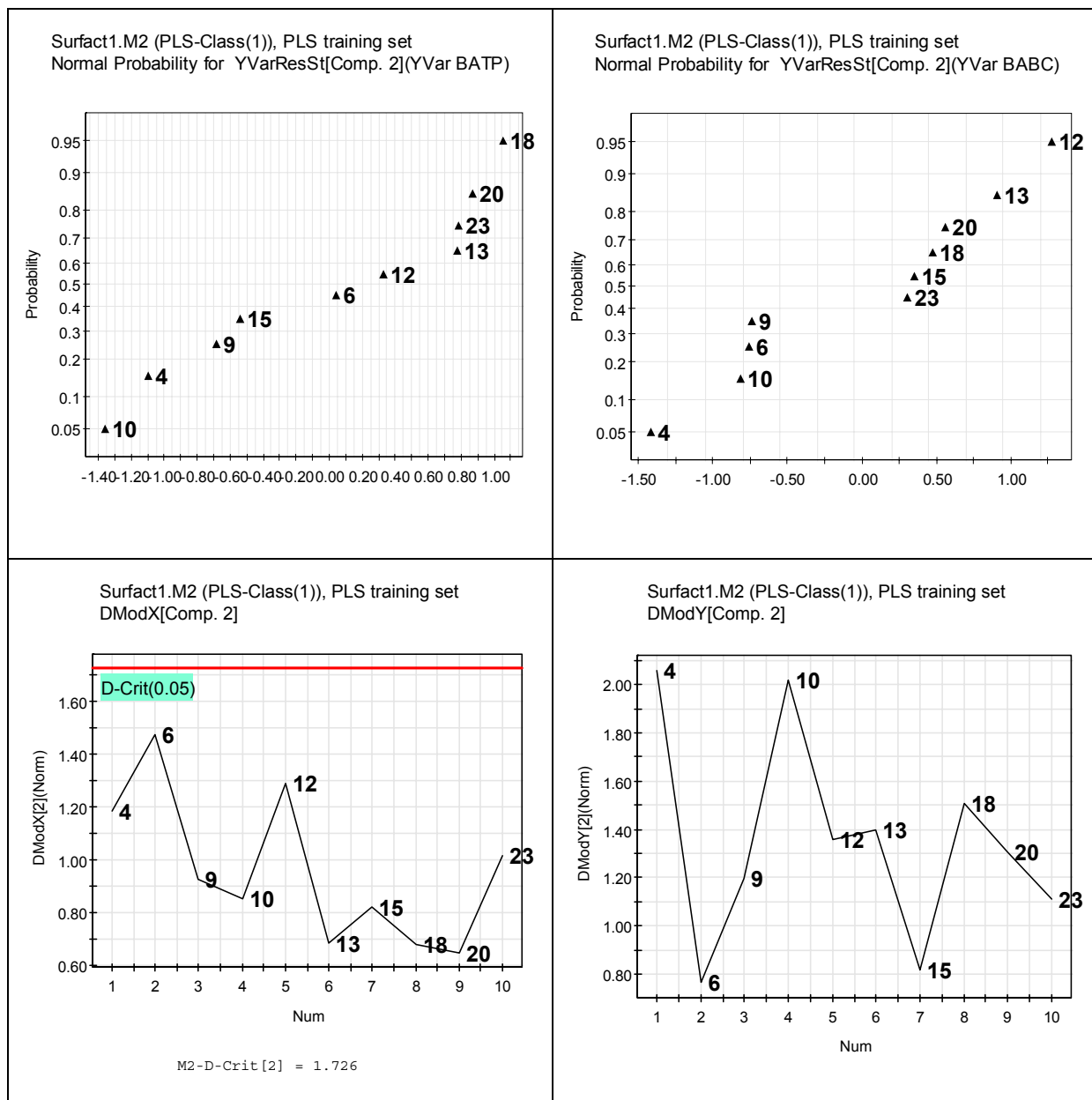
Task 2

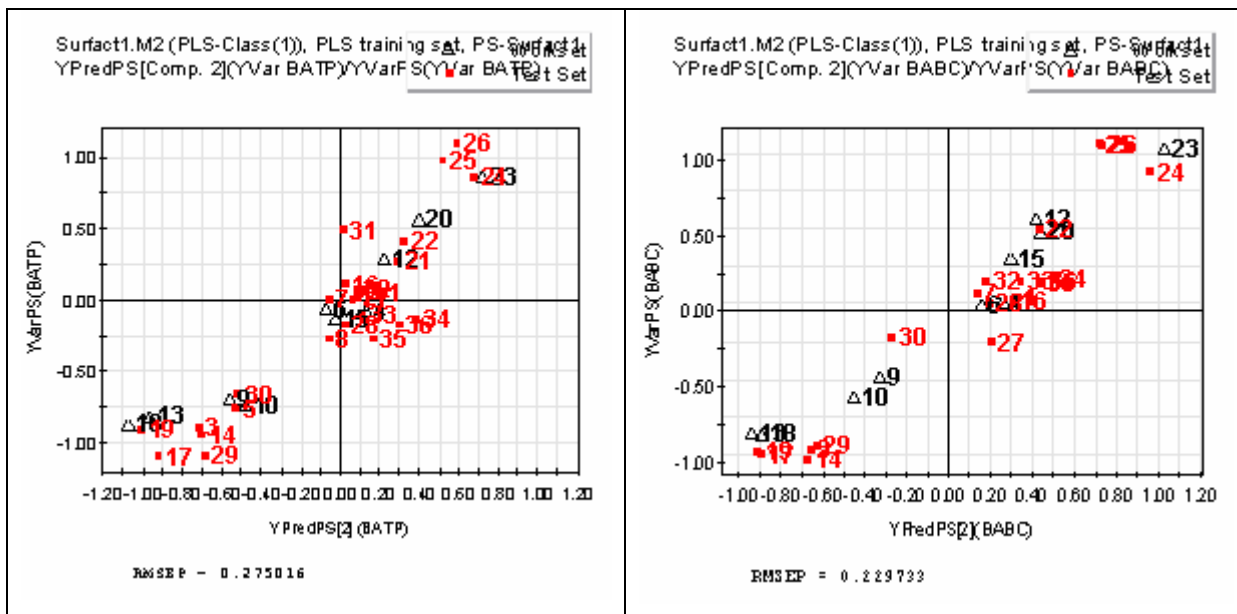
The PLS analysis yielded a two-component model with excellent statistics. The correlation between the chemical properties and the aquatic toxicity endpoints is stable, as seen in the PLS score plots (t_1/u_1 , t_2/u_2). The PLS loading plot, wc_1/wc_2 , suggests that the aquatic toxicity of the surfactants is regulated primarily by the hydrophobic properties. Descriptors like C, redC, log P and CMC are the most influential and relate to the hydrophobicity. In addition, we can see that the two endpoints are strongly correlated.



The residuals are of diagnostic interest. The N-plot of residuals for both responses does not unveil any outliers. Also the DModX and DModY plots indicate that all surfactants in the training set are well modelled. Note that the DModY chart can be viewed as a summary of both N-plots.

The surfactant QSAR is well founded and can be used for predictions of the surfactants in the validation set. Evidently, the model has excellent predictive capability. The training set compounds are marked by open triangles in the last two plots.





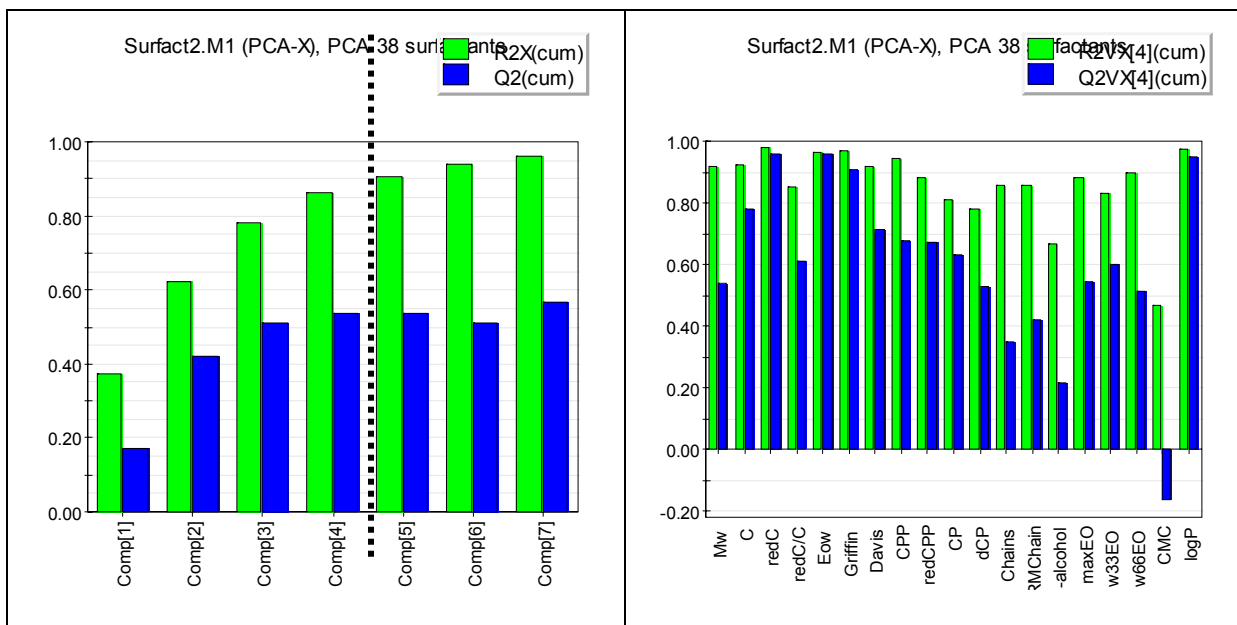
Task 3

There is no solution to this task.

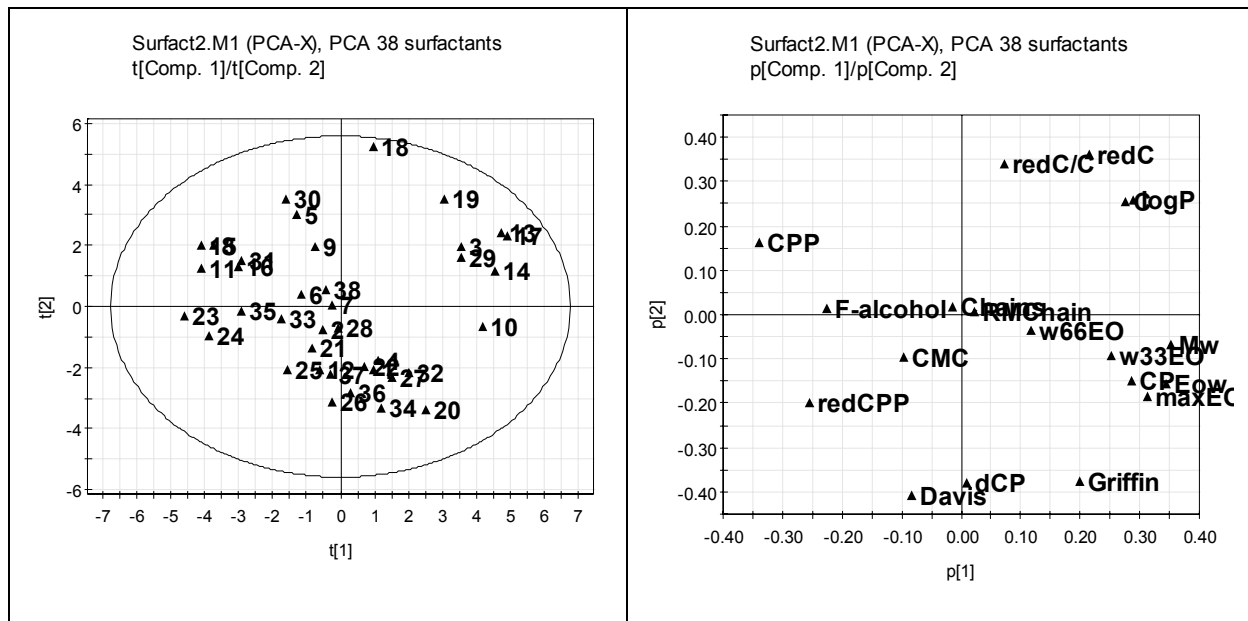
Phase 2 Solutions

Task 4

The PC-analysis gave seven components, but only about four are of appreciable size.



The scores and loadings of the two first components are plotted below. The third and fourth components only account for 15% and 8% of the variation, and are therefore not considered in this brief interpretation.

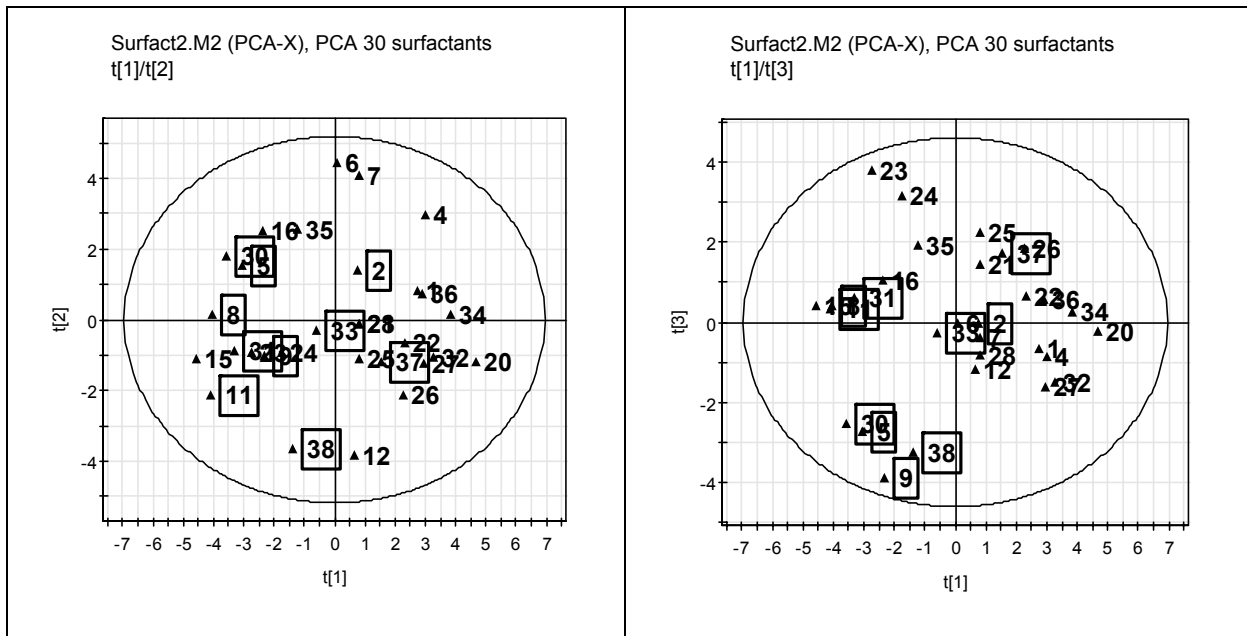


Based on the loadings the authors interpreted the first PP as surfactant lipophilicity, and the second PP as a reflection of the hydrophobic/hydrophilic balance of the surfactants. This because high values for PP1 (high lipophilicity) correspond to high values for log P, ethylene oxide (EO) content, molecular weight (Mw), and cloud point (CP), and low values for the amount of non-ethoxylated fatty alcohol (F-alcohol), the critical packing parameter (CPP) and the reduced packing parameter (redCPP). The interpretation of the second PP was based on the fact that high values of PP2 (low hydrophobic/hydrophilic balance) correspond to low values of the two indices reflecting hydrophobic/hydrophilic balance (Davis and Griffin) and high values of redC/C, i.e., a branched hydrophobic part.

The distribution of surfactants in the score plot made the researchers aware of two strong clusters. Since the smaller cluster, located in the upper right-hand part, comprises surfactants that are too lipophilic for the experimental objective, the eight surfactants belonging to this cluster were excluded. The reason for this action was that if a surfactant is very lipophilic, it might be hard to maintain the desired balance between good washability and biodegradation within reasonable time.

Task 5

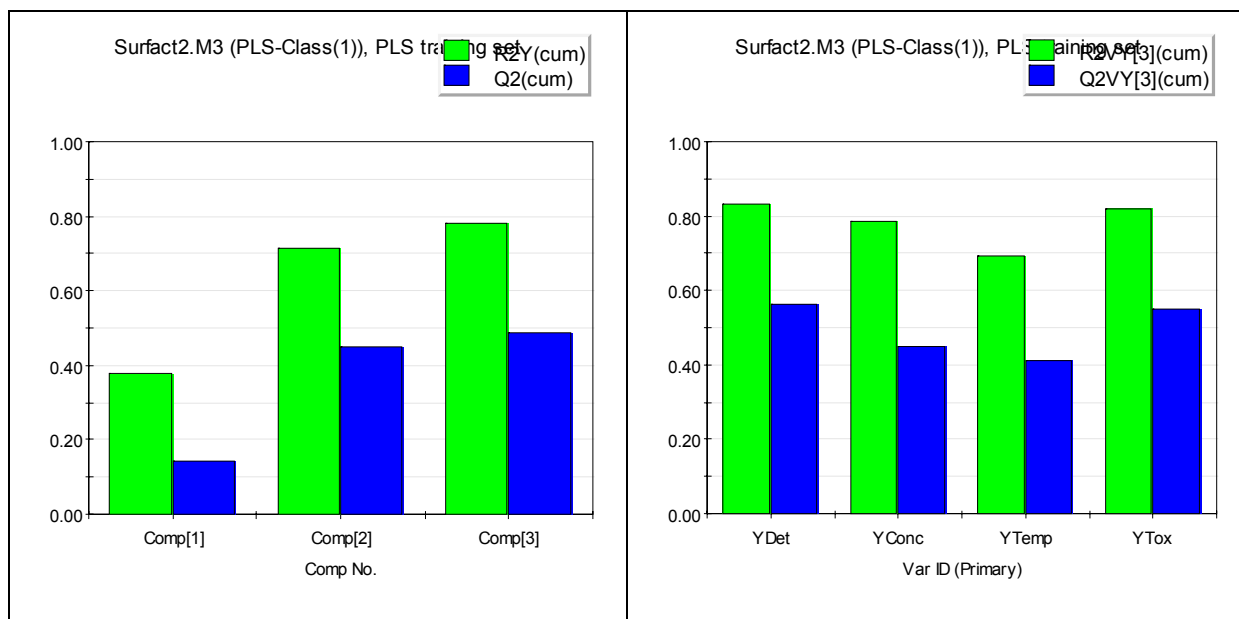
A new PCA was made on the remaining 30 detergents. This model displayed the performance statistics $R^2X = 0.76$ and $Q^2 = 0.52$, after three components. The first PP explains 38%, the second 21%, and the third 17% of the total variation. The updated PPs are plotted below. Their interpretation is similar to the previous PPs.



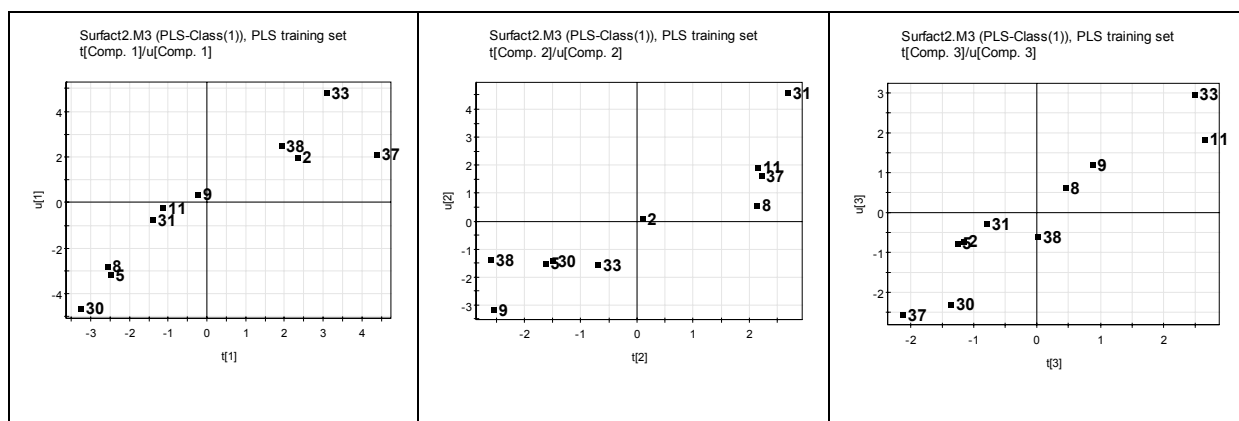
Subsequently, the three updated PPs were used in multivariate design. The authors did not use any statistical experimental design (factorial, fractional factorial, D-optimal,...), but identified ten representative non-ionic surfactants providing a reasonable coverage of a restricted PP-area. The location of the ten surfactants is indicated in the score plots above by open boxes. These surfactants form the training set (surfactants 2, 5, 8, 9, 11, 30, 31, 33, 37, and 38).

Task 6

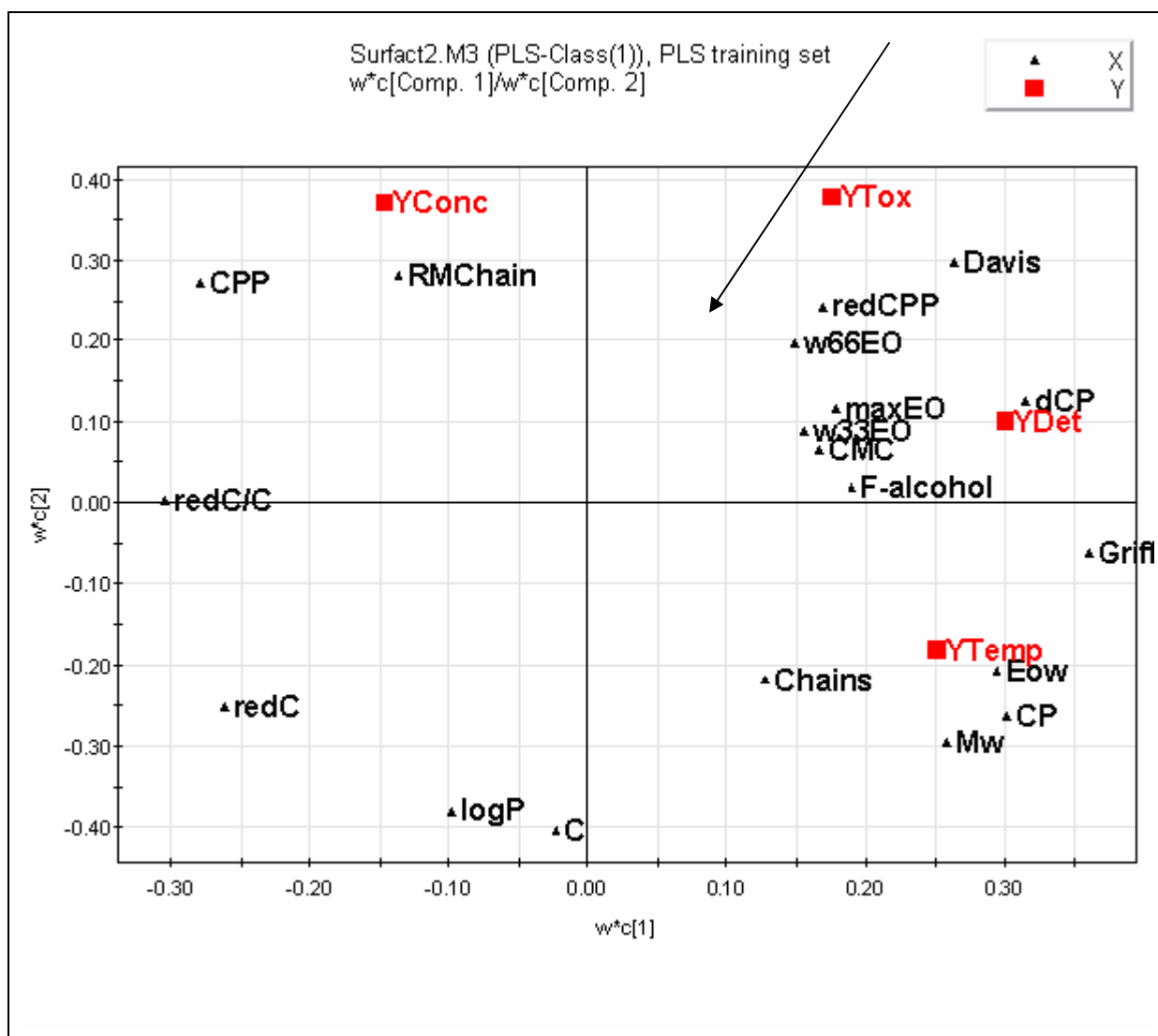
The four responses (YDet, YConc, YTemp & YTox) and the 19 chemical descriptors were modelled simultaneously with PLS. Three PLS components were obtained, and the cumulative R^2Y - and Q^2 -values are plotted below. After three components R^2Y amounts to 0.78 and Q^2Y to 0.49, which are excellent values considering that technical surfactant blends are being investigated, and not pure compounds. The right-hand plot below shows R^2Y and Q^2Y for each response. Evidently, three responses, YDet, YConc, and YTox, are modelled and predicted well, and their Q^2 's range between 0.45 and 0.56. The Q^2 -value of the last response, YTemp, is a little lower (0.41).



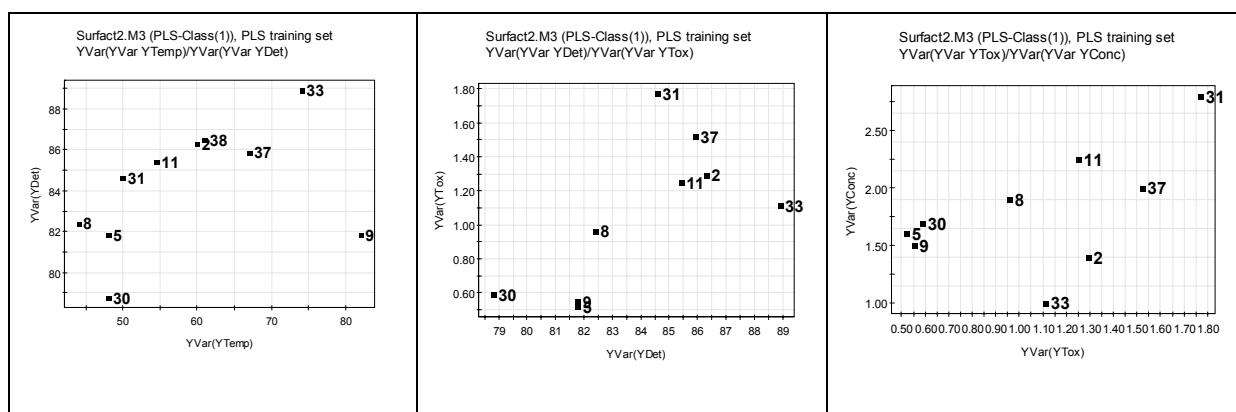
In order to interpret the PLS model we consider the scores and loadings. A close inspection of the t_1/u_1 score plot shows a mild curvature in the relation between X and Y. This suggests that the second model component represents a compensation for non-linearity. However, in the last component the inner relation is linear. The overall conclusion is that there is a strong correlation structure between the 19 chemical descriptors and the four response variables.



The PLS loadings of the first two components are shown below. The third component only accounts for 7% of the response variation and is not plotted here. An interpretation of the third component is integrated in the overall model interpretation presented below.



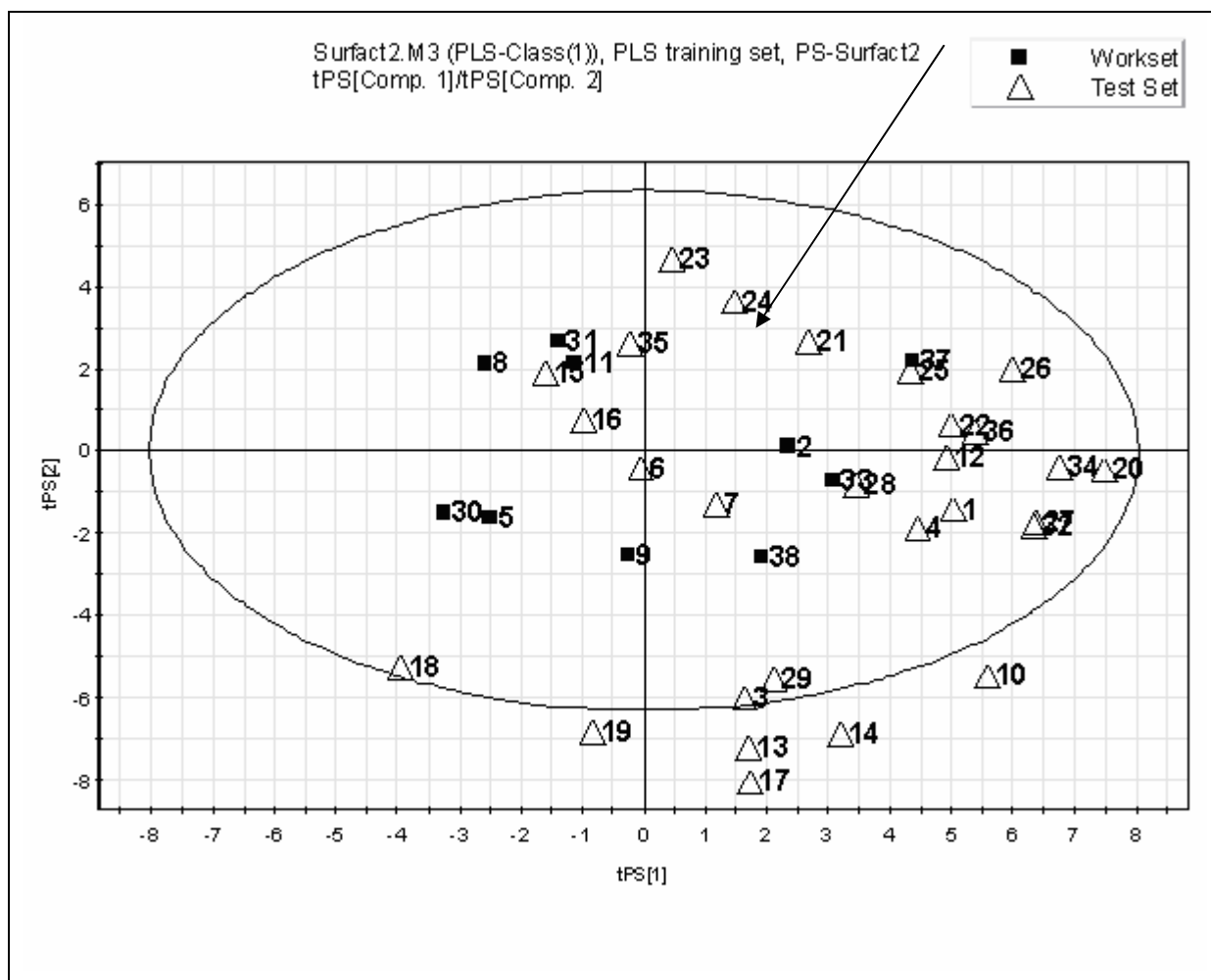
We see that the four responses are not completely correlated. However, YTemp is moderately correlated with YDet, YDet is partly correlated with YTox, and YTox is partially correlated with YConc. This partial linking of the responses motivates their treatment in one single PLS model. Plots of raw data supporting this conclusion are given below.



- **YDet:** In order to achieve high washing efficiency, the surfactant should contain a short carbon chain. This is inferred from the model contribution of the descriptor redC, which denotes the number of carbon atoms in the longest chain of the hydrophobic part.
- **YConc:** A desired low value of surfactant concentration is modelled as obtainable for surfactants with high molecular weight and high molar ratio of dominating chain to total number of carbons.
- **YTemp:** The washing temperature is influenced by descriptors like molecular weight and the number of EO-units. Low molecular weight and few EO-units correspond to low temperature.
- **YTox:** Not unexpectedly, the response acute toxicity is well modelled by log P. This kind of relationship is often found in QSAR applications in aquatic toxicology. To enable low toxicity the surfactants should display low log P (lipophilicity).

Weighting this together indicates that an almost straight side chain, with some degree of branching, should have a good balance between washing performance and toxicity (cf. surfactant no. 37).

One of the most exciting steps of this application arose when the derived PLS model was used to compute predictions of the 28 non-tested surfactants. In this process, the known X-data for these compounds were inserted into the PLS model, and the unknown Y-data were estimated. The score plot below shows the distribution of prediction set surfactants in the t_1/t_2 score plane. Surfactants close to nr 37 are interesting candidates for further studies.



Conclusions

Non-ionic surfactants of the polyethylene-type form a group of detergents undergoing intense research and rapid change. Lindgren and co-workers were able to identify a set of promising surfactants by adhering to multivariate methodology. However, the exact identity of the finally proposed surfactants has never been disclosed due to trade secrecy reasons.

The primary conclusions of this application were:

- All surfactants cannot be tested - multivariate characterisation and design is useful to select representative compounds.
- Strong relationships exist between measured physico-chemical properties of surfactants and their performance profiles.
- Surfactant performance is a multivariate property and must be addressed by measuring a set of response variables.
- Predictions from PLS models are useful to identify interesting surfactants for further performance optimisation.

MVDA-Exercise PULP

Modelling a beating process in Swedish pulp mills and prediction of pulp quality parameters

Background

The normal way of characterising pulp is to use several different physical and chemical parameters. These parameters are highly correlated, which means that pulp can be characterised using only a few informative latent variables. To acquire in-depth knowledge about the effect of beating chemical pulps, a multivariate investigation was carried out by Lars Wallbäcks.

Objective

The main objective of the investigation was to reveal how beating conditions affect pulps of different composition. Another objective was to develop a predictive model for the tear index.

Data

The starting materials were taken from 7 different Swedish pulp mills. A research institute (STFI) performed the physical measurements on the different pulps at 4 times: before beating, after 1 000 revolutions, after 2 000 rev., and after 4 000 rev. Eighteen classical pulp descriptors were recorded in order to carry out the multivariate characterisation. This exercise concentrates on PFI-mill beating.

Tasks

Task 1

Create a new project in SIMCA. Import data from PULP.XLS. Observe that the first two columns are observation ID.s. Give the project a unique name. Make sure that 28 observations and 21 variables have been imported. Note that the observations from the same mill are 7 observations apart. For example, the first mill has numbers 1, 8, 15, and 22 and names (ID.s) 1, 11, 111, 1111.

Task 2

Create a summary of the data. Run PCA on all observations and the first 18 descriptors. Extract the first two components. Construct a plot of the observations (t_1/t_2). Are the observations clustered or do they show some other pattern? Can you identify any trends? Display the corresponding picture of the descriptors (loading plot, p_1/p_2). Can you explain the trend seen in the score plot? Does t_1 separate the observations according to revolution? (Hint: variable 21 is the square root of the number of revolutions/1 000? Use *Plot/List|Scatter Plot* to look at this variable.)

Task 3

Use the same model as that in the previous example and look at tear index (variable 13) vs. t_1 . What is a likely reason for the scatter of the unbeaten samples? How does the pulp react with beating? Looking at score t_1 , what can you say about the spread of the unbeaten observations compared with the fully beaten observations?

Task 4

A common way of using multivariate models is to replace variables that are difficult to measure with variables that are more easily obtainable. To do this we build a multivariate model from the easily measured variables to predict the more complicated variables. Define variables 1-7 as X and tear index as Y. Use observations 11, 33, 55, 77, 111, 333, 555, 777, 1111, 3333, 5555, and 7777 as the training set. Fit a PLS model and make predictions for the remaining observations. Is it possible to compute reasonable predictions for tear index? Repeat this procedure with other variables as Y, e.g., burst index (var no 14) or tensile index (var no 9).

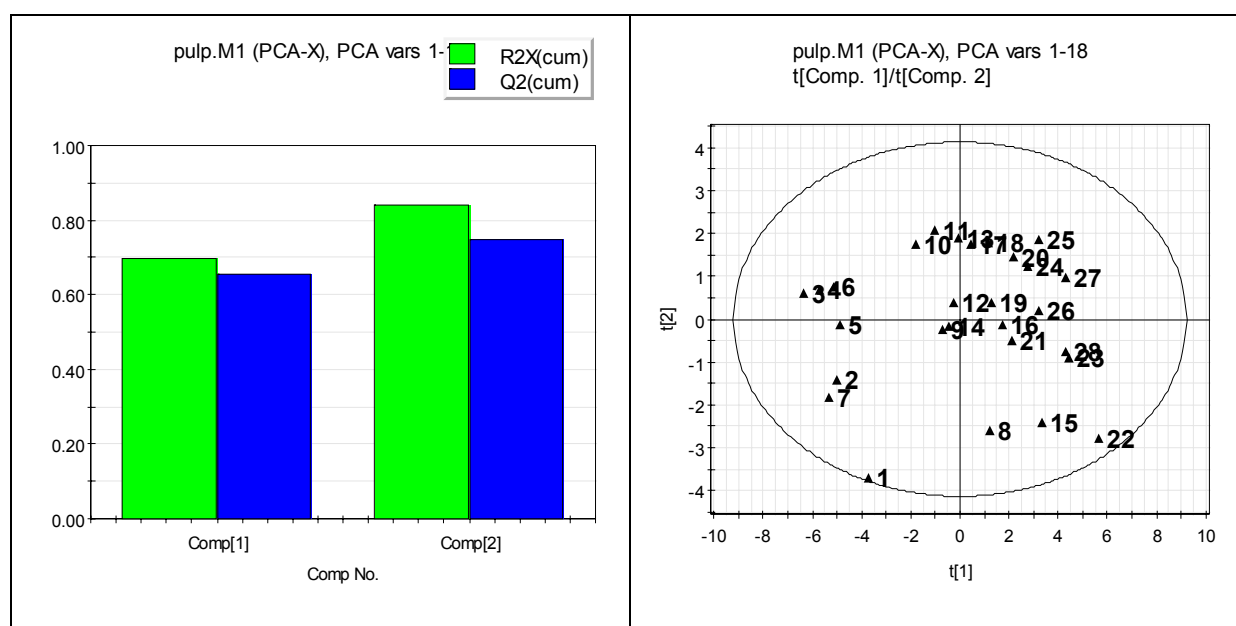
Explanation of data set:

Variable no	Variable name and abbreviation	Swedish abbr'n
1	Shopper Riegler value, SR	SR
2	Water Retention Value, WRV (%)	WRV
3	WRV of the coarse fraction (%)	WRV mesh
4	Fibre length (Kajaani) (mm)	Aritm Fl.
	Arithmetic average	
5	Fibre length (Kajaani) (mm)	L. vikt. Fl.
	Length weighted average	
6	Fibre coarseness (Kajaani) (m/mg)	Fl. vikt
7	Fines fraction, P200 Mesh (%)	Finfrakt.
8	Density (kg/m ³)	Densitet
9	Tensile Index (kNm/g)	Dragi.
10	Stretch to break (%)	Brott/jn.
11	Tensile energy absorption (TEA) (J/kg)	BrArl
12	Tensile stiffness index (kNm/g)	DrStl
13	Tear Index (mNm ² /g)	Rivi.
14	Burst index (kPcm ² /g)	Spr_ngl.
15	Scott Bond (J/m ²)	ScottB
16	Surface roughness (Bendtsen) (ml/min)	Ytr_het
17	Air Permeance (Bendtsen) (ml/min)	Porositet
18	Light Scattering coefficient (m ² /kg)	S
19		K
20		Ljush
21		rmal

Solutions to PULP

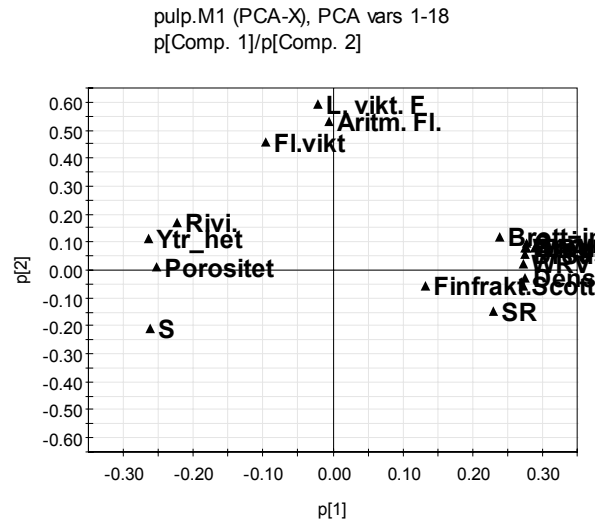
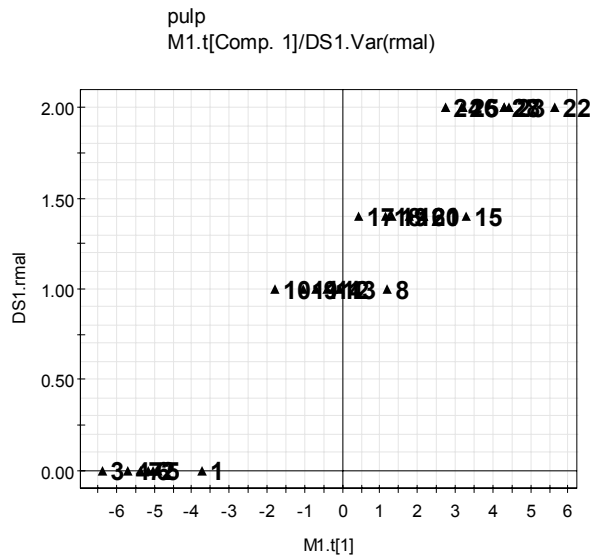
Task 2

The two first components of the PC-model explain 84 % of the variation and predict 75 %. In the score plot below it is clear that the score (numerical value) of the first component increases with increasing beating. The relation is not linear, however. Different pulps appear to be separated both in the first and the second component.



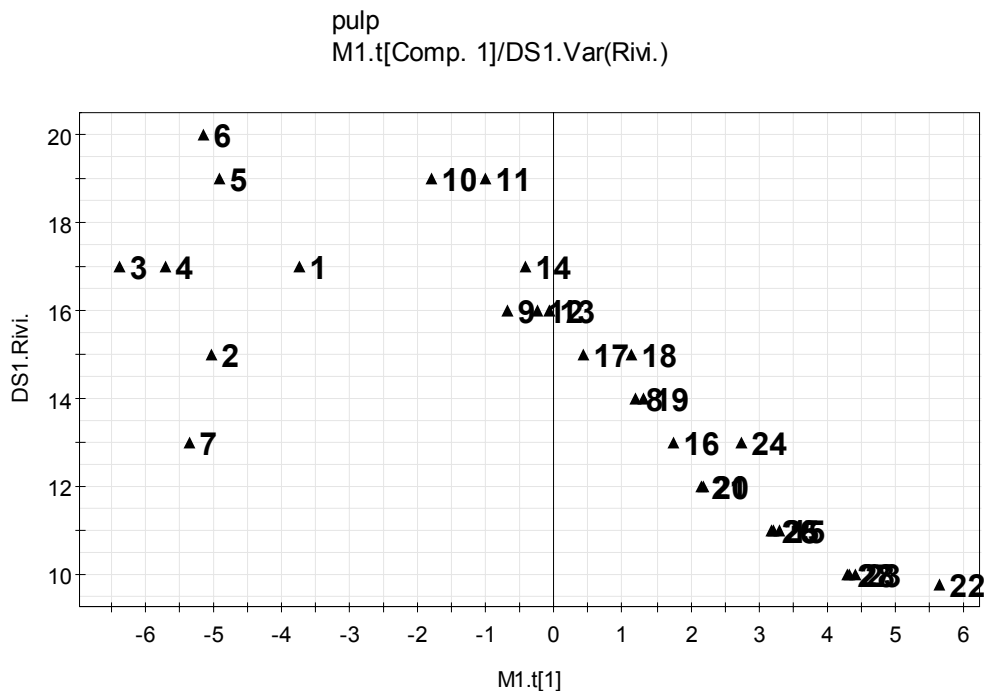
The scatter plot of t_1 against variable 21 (square root of beating/1 000) is shown below. We can see that t_1 separates the observations with respect to beating. Differences in starting material composition are believed to be responsible for the differences in the starting points.

The loading plot indicates an increase in fibre length with increasing beating. This is in agreement with the fibre straightening effect associated with PFI-mill beating.



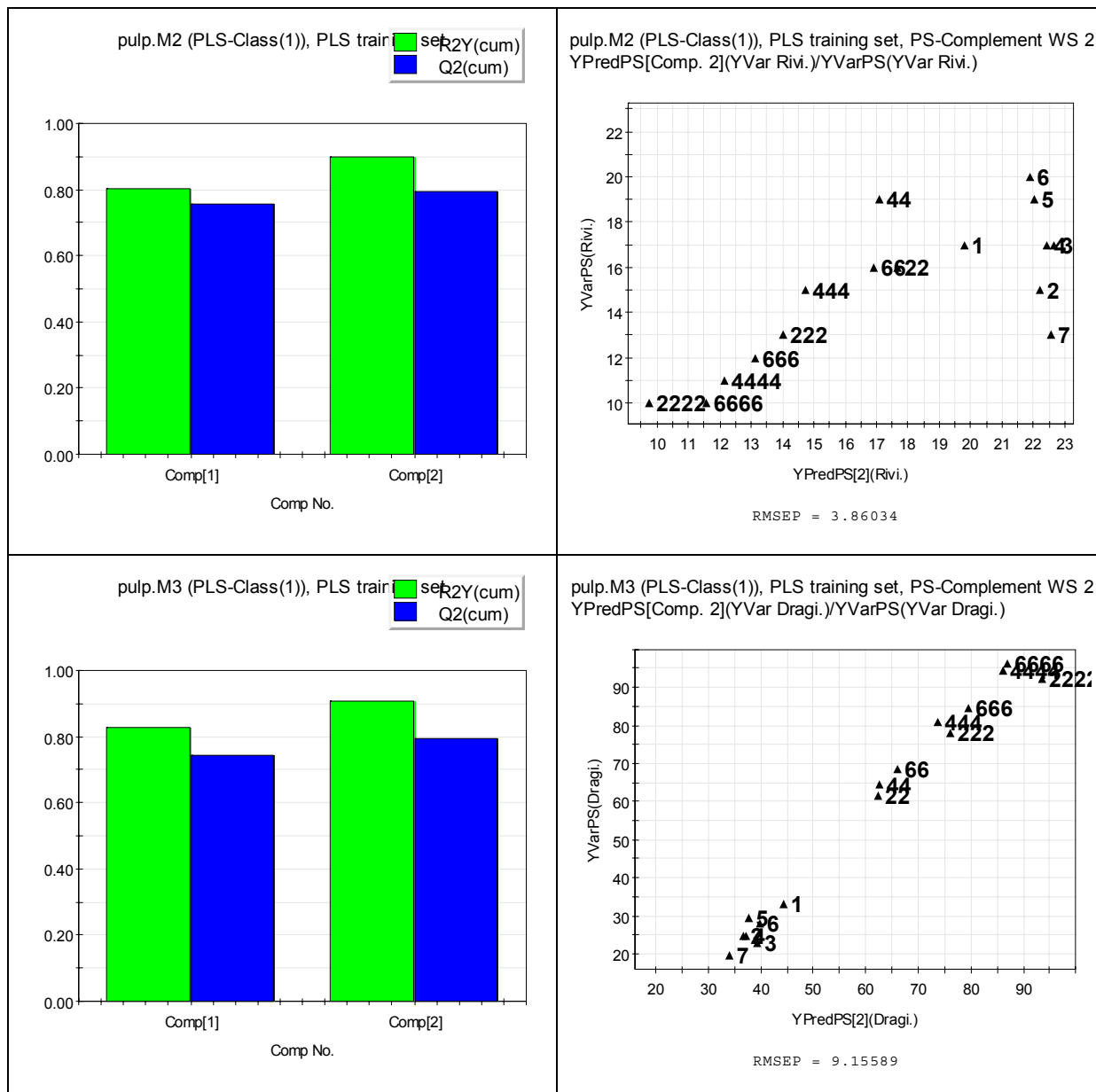
Task 3

The spread of the unbeaten observations indicates a difference in starting material composition. However, after beating, the observations define a linear data structure. This means that the beating process has made the different pulps more alike. Hence we can develop a model describing tear index as a function of beating, but only on beaten samples.



Task 4

PLS analysis yielded a two-component model with $R^2Y = 0.90$ and $Q^2 = 0.79$. According to the prediction plot, reliable predictions of the tear index are possible except for the unbeaten samples. The next set of plots show modelling and prediction results when using tensile index as the y-variable. Interestingly, tensile index is well forecast also for the unbeaten samples.



Conclusions

Although there seems to be fundamental compositional differences between the unbeaten pulps, the beating process is able to unify the properties of the various pulps. Important pulp quality characteristics, like tear index or tensile index, are well modelled and predicted from the numerous easily measured physical and chemical descriptors.

MVDA-Exercise SUGAR

Multivariate calibration

Background

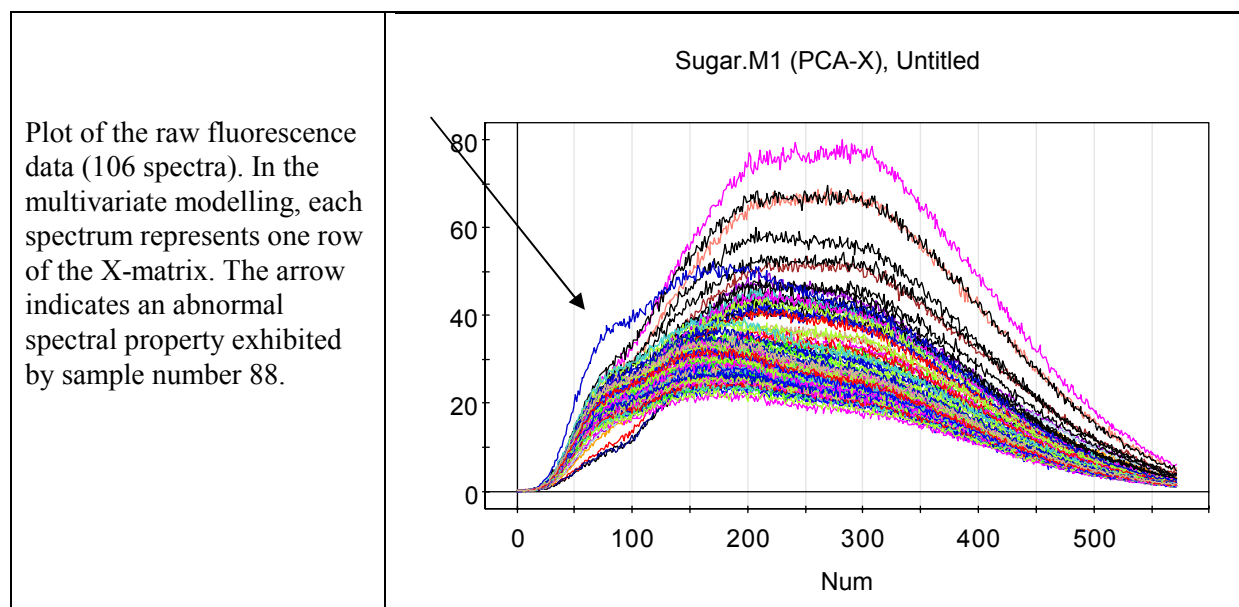
Multivariate calibration is an important application of multivariate data analysis. After spectroscopic measurements of the samples, with known characteristics (e.g. concentrations/levels/qualities/etc.), have been taken a regression model is built using the spectral data as X and the sample characteristics as Y. The intention with this exercise is to demonstrate multivariate calibration based on fluorescence data.

Objective

At a sugar plant in Scandinavia an important quality parameter was the ash content of the sugar, but the conventional measurement technique for this response variable was judged to be arduous and time-consuming. As a consequence, the goal was to replace traditional wet-chemistry measurements with rapid on-line fluorescence techniques. To determine the feasibility of this new technology, the applicability of multivariate calibration was explored. More details can be found in the original literature source [Bro, 1996].

Data

Process data were measured during a campaign of 2.5 days, right from the start-up of the new factory. A total of 106 samples (time points, observations) were compiled and by digitising the fluorescence spectra 571 X-variables were created.



Tasks

Task 1

It is instructive to precede the regression analysis with a PCA of the X-data. This tells us about the practical rank of X and if there are spectral outliers. Open the file SUGAR.SIM (or *.DIF). It contains 106 observations and 573 variables. Variables 572 (V1_2, impurity) and 573 (V2_2, colour) are the two responses. Give the project a unique name. Select all observations and the 571 X-variables and run PCA. Compute the first two principal components. Create the necessary score and loading plots. Do you see any spectral outliers in the score plot? What can you say about the loadings? Do they resemble the real spectra?

Task 2

Change the scaling of the X-variables so that they are mean-centred but not scaled. This is accomplished by the commands (*WorkSet /New (as model) /Scale*), and selecting Base “Ctr”. Then run PCA and extract two components. Create the same score and loading plots as in the previous task. Are there any spectral outliers? What about the loadings? Examine the model residuals as well (DModX); do they reveal any moderate outliers outside the critical model distance?

Task 3

Next, we want to examine whether the fluorescence data carry any information that is useful for modelling and predicting the quality of the final sugar product. When we do this analysis, the unrepresentative samples 1-15 and 88 are excluded. Open the file SUG_SORT_ID.DIF, which contains 90 observations sorted according to ash content. The first three columns represent Primary_ID, Class_ID, and Real_ID. The X-data starts in the fourth column. Define variables 1-571 as X and 572 as a single Y-variable. We will not use variable 573. Remember to set the scaling of the X-variables to base weight “Ctr”.

Divide the data in two classes, one containing the 45 odd-numbered observations (class 1), and another containing the 45 even-numbered observations (class 2). Use unscaled and mean-centred X and run PLS. Compute a PLS model on one class and verify predictive ability with the other. Then change the role of the two sub-sets and repeat the analysis procedure. Can we make a model that is able to predict ash content?

Task 4

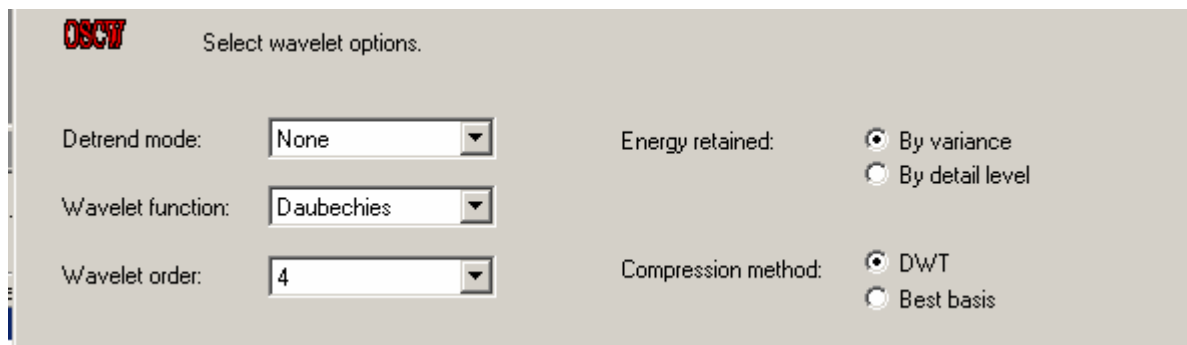
It might be possible to enhance the predictive power of the calibration models. In order to test whether this is possible we will use orthogonal signal correction (OSC). Go to *DataSet/Spectral Filters/OSC*. Select variables 1-571 as X and variable 572 as Y. Select the 45 odd-numbered observations. Use default options for transformations and scaling (don't change the settings). Press *Next* to extract the first OSC component. Save after the first component. Make sure that the additional observations are stored in the secondary data set.

After switching to the new project, prepare for PLS modelling. Select X and Y variables, and only Centre the X-data. You may use the *AutoFit* option in the modelling, but save only one component. Evaluate the model, and make external predictions for the secondary data set, which should contain the even-numbered samples. Is it possible to improve the predictive power of the model?

Repeat the above procedure, but select the even-numbered observations for “OSCing”, and the odd-numbered observations as secondary data set.

Task 5

We will now test the applicability of wavelet analysis for signal compression. Use the same two subsets of data as in the previous task. Go to *DataSet/Spectral Filters/Combination OSC/Wavelet*. Select variables 1-571 as X and variable 572 as Y. Select the 45 odd-numbered observations. Use default options for transformations and scaling (don't change the settings). Press *Next* to extract the first OSC component. Press *Next* to select wavelet function, wavelet order, and compression method. Select the following:



The screenshot shows a dialog box titled "OSCWT" with the subtitle "Select wavelet options." The dialog contains the following settings:

- Detrend mode: None (dropdown menu)
- Wavelet function: Daubechies (dropdown menu)
- Wavelet order: 4 (dropdown menu)
- Energy retained: By variance, By detail level
- Compression method: DWT, Best basis

Press *Next* and select the number of wavelet coefficients representing 99.5% of the total variance. Press *Next* and *Save* odd-numbered samples in the new primary data set and the even-numbered observations in the new secondary data set.

After switching to the new project, prepare for PLS modelling. Select X and Y variables, and only Centre the X-data. You may use the Autofit option in the modelling, but save only the first component. Evaluate the model, and make external predictions for the secondary data set, which should contain the even-numbered samples. Is it possible to compress data and maintain the same predictive power?

Repeat the above procedure, but select the even-numbered observations for “OSCing” and wavelet analysis, and the odd-numbered observations as secondary data set.

You may also re-apply the outlined procedure where wavelet coefficients describe only 95% of the total variance.

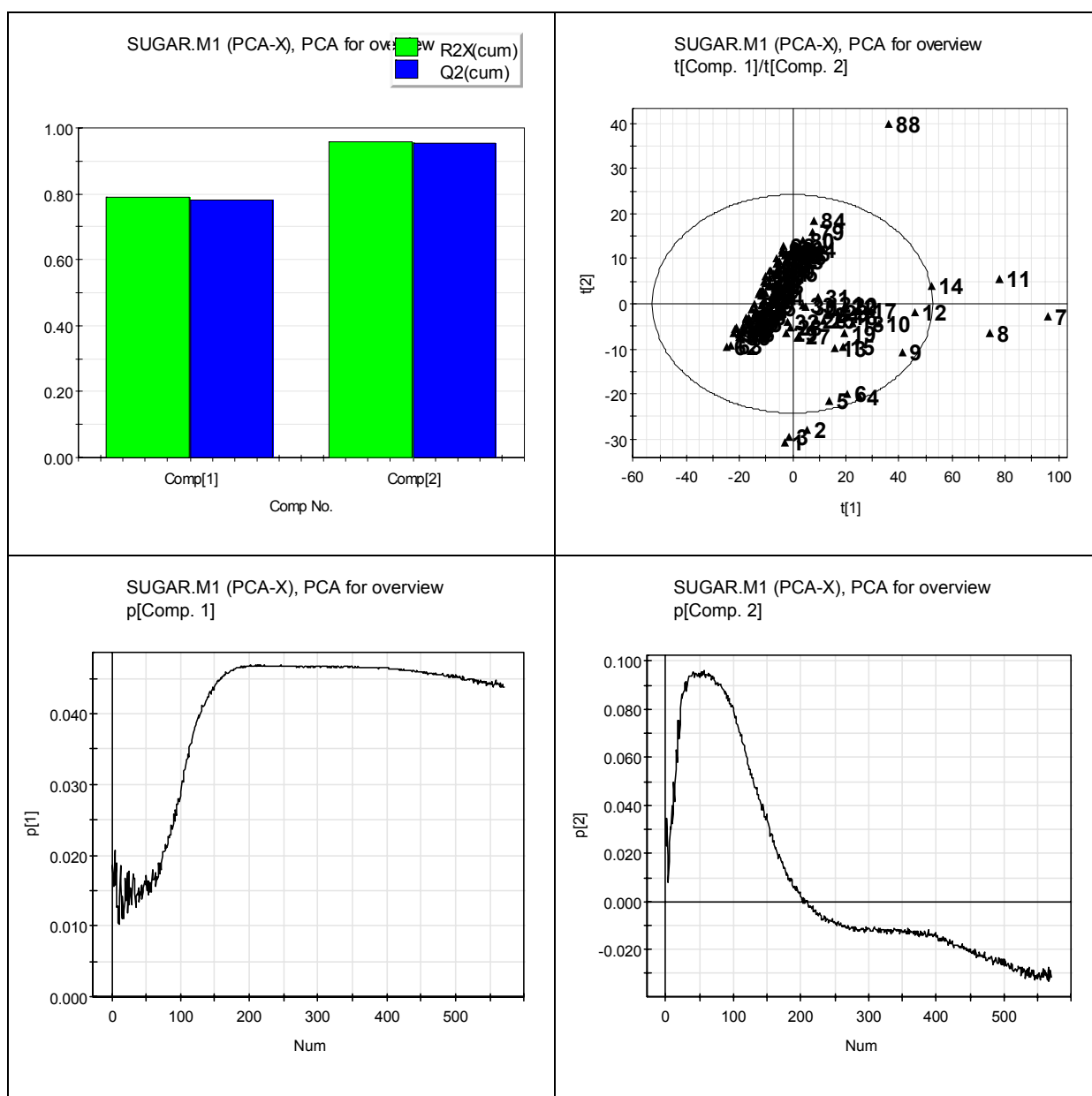
Task 6

Repeat Tasks 3-5, but for the second response, variable 573, the sugar colour. There are NO solutions provided to this task.

Solutions to SUGAR

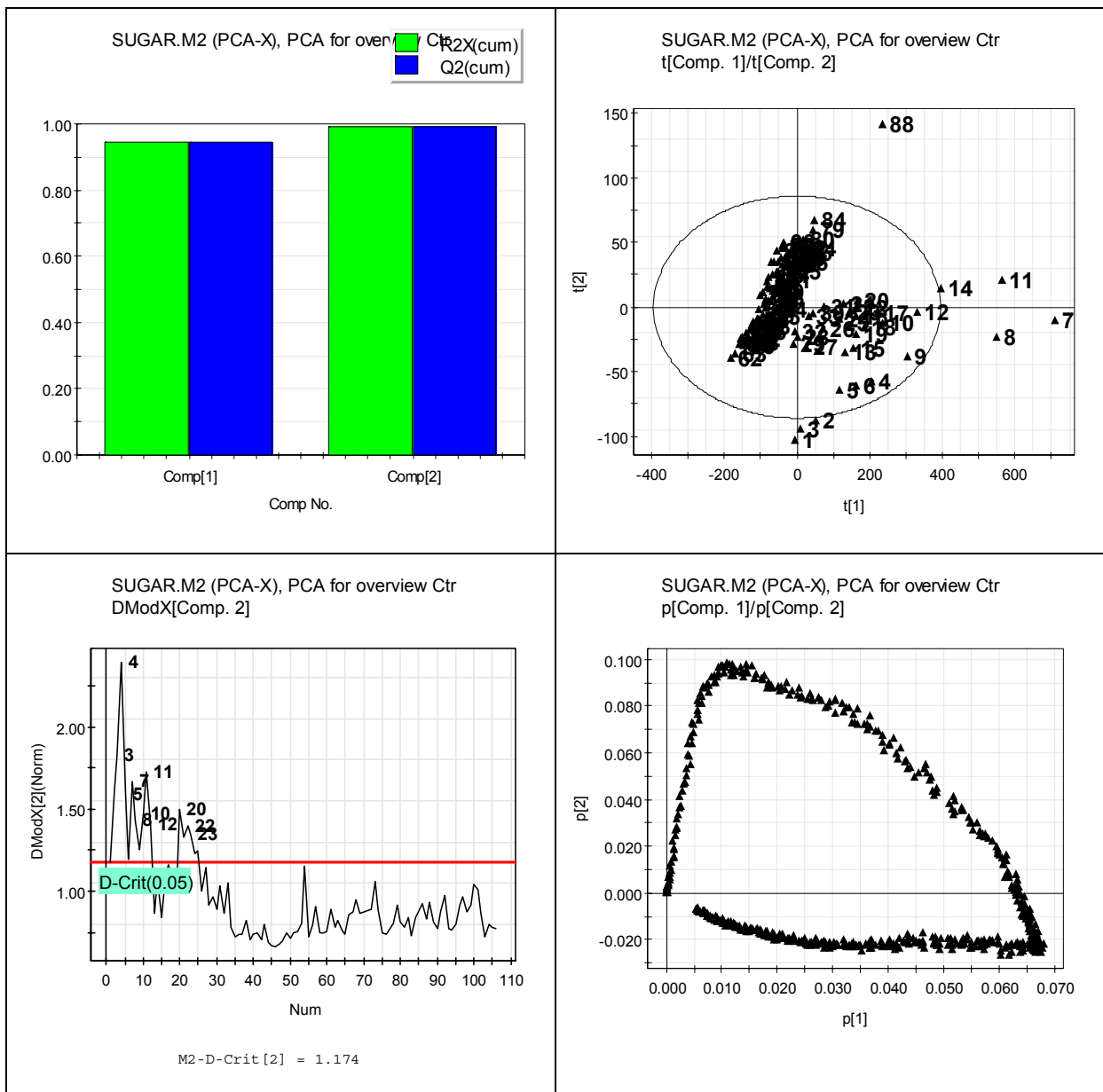
Task 1

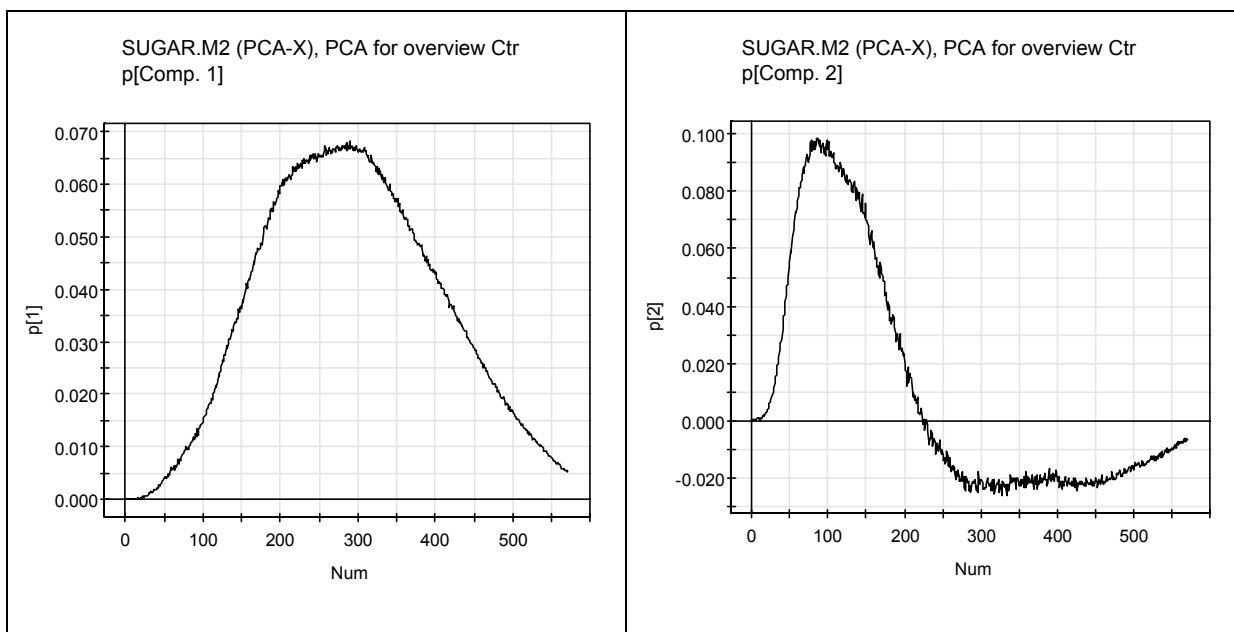
The first two principal components gave $R^2 = 0.95$. In the t_1/t_2 score plot we can see how the manufacturing process has drifted over time. This is data from a new plant taken from the very start of sugar production. We can see that it took a while for the process to reach stability and eventually found a region of relatively constant process conditions (thick cluster in the left-hand part). Moreover, sample #88 appears to be extreme, which is discernible also in the plot of the raw data (see previous pages). Here, we have plotted the loadings versus “Num”, because there is a natural spectral order among the sugar variables. With spectral data, the first loading usually closely resembles the average raw data spectrum, but in this case it does not. Why? The explanation is that we have run the PCA with UV-scaled variables. This means that we have scaled up those parts of the spectra with low signal amplitude variation, at the expense of the contribution from those parts that have large signal amplitude variation. The solution is to re-run the PCA, but without scaling the X-variables.



Task 2

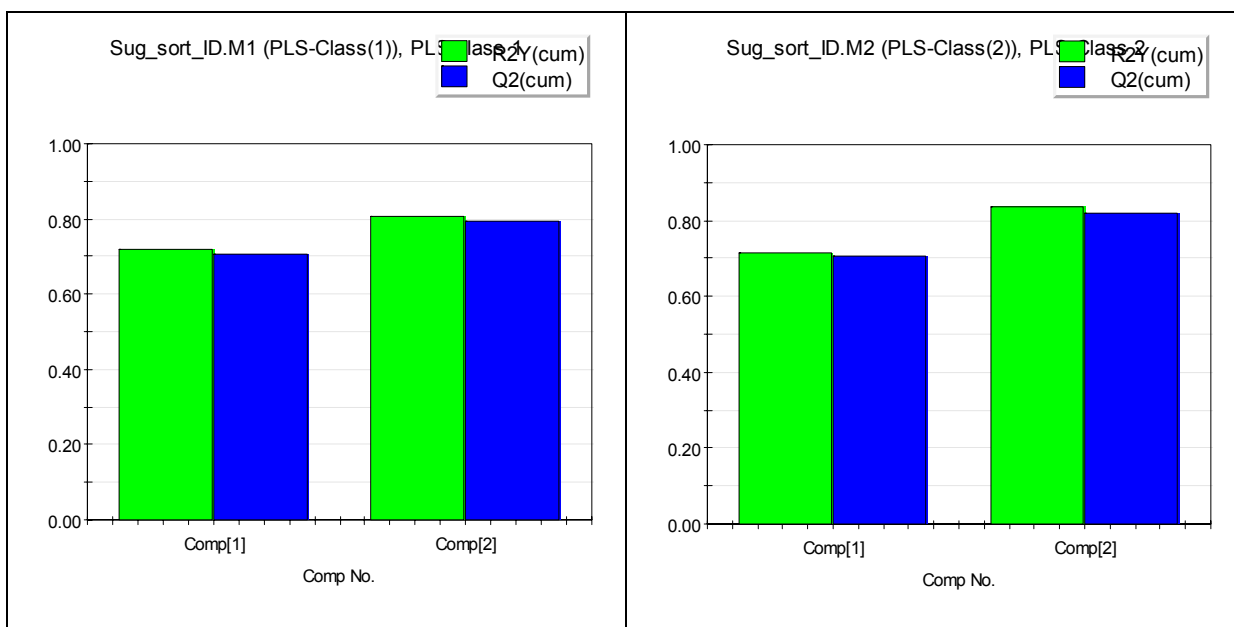
The first two components yielded an R^2 of 0.99. Evidently, there is a comparatively large drift going on until around sample 15. From sample 16 onwards, with the exception of sample 88, the process has reached a state of comparatively stable operating conditions. The DModX chart also suggests that the largest process variation is found at the beginning of the sampling period. The three loading plots below indicate that the data contain two independent spectral contributions. However, the first loading plot, the scatter plot, is not very informative. This is because loadings from spectral data often get a “worm-like” appearance. Instead of the scatter plot, loadings against spectral wavelength is more appropriate. The first loading spectrum has a systematic structure and resembles the average spectrum. The second loading spectrum captures a small peak in the lower wavelength region of the spectra. Evidently, the use of unscaled but mean-centred data seems appropriate in order to enhance interpretation.



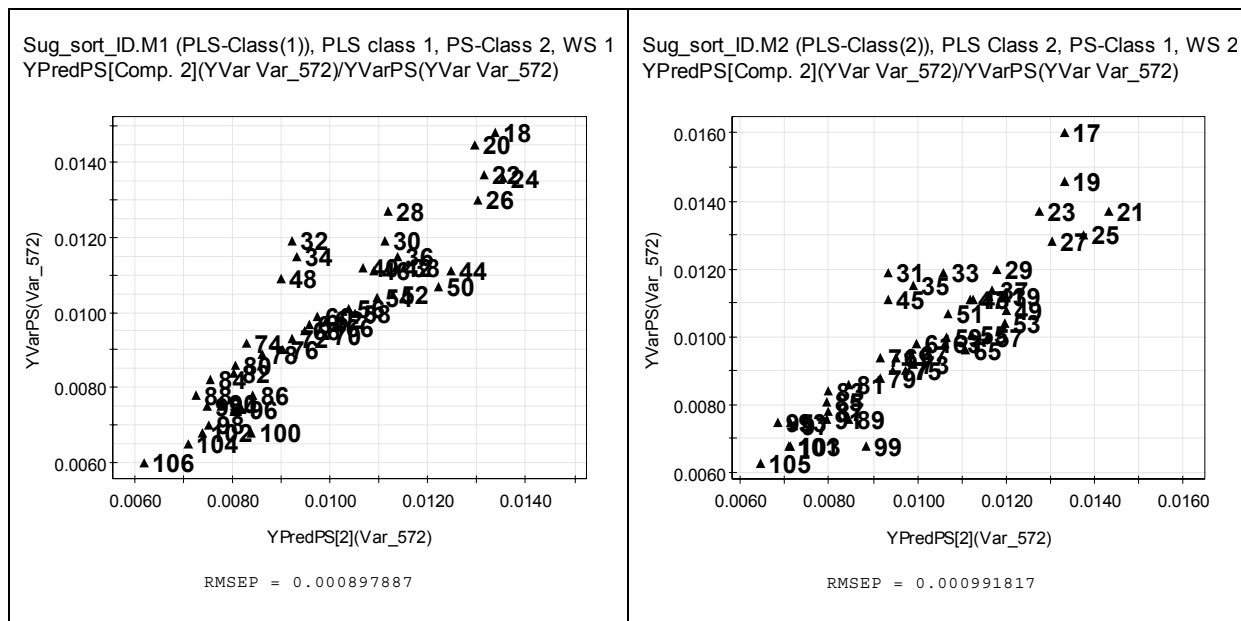


Task 3

The application of PLS to these two sub-sets produced two-dimensional models of sound explanatory and predictive powers. For the odd-numbered training set, the produced model had a goodness-of-fit of 81% ($R^2Y = 0.81$) and a goodness-of-prediction of 79% ($Q^2_{int} = 0.79$). The corresponding values for the model based on the even-numbered training set were 84% ($R^2Y = 0.84$) and 82% ($Q^2_{int} = 0.82$). These values are plotted below. The subscript “int” of the Q^2 's shows that these estimates were derived using internal cross-validation. The similarity of these R^2 - and Q^2 -values indicates that the two sub-sets are rather well balanced and have similar characteristics. Hence, the sub-division of the observations by even/odd appears valid.



By using the two external validation sets it is possible to further test the predictive power of the two models. In doing so, we obtained the results that are summarised by the two figures below. The external Q^2 -values, Q^2_{ext} , amount to 0.83 (RMSEP = 0.00090) and 0.80 (RMSEP = 0.00099) for the two models, which are similar to the internal Q^2 -values of 0.79 and 0.82. Based on these model performance statistics, we conclude that both cross-validation and validation by external prediction, yield similar estimates of an important quantity: the ability of the model to predict the ash content of new sugar samples.



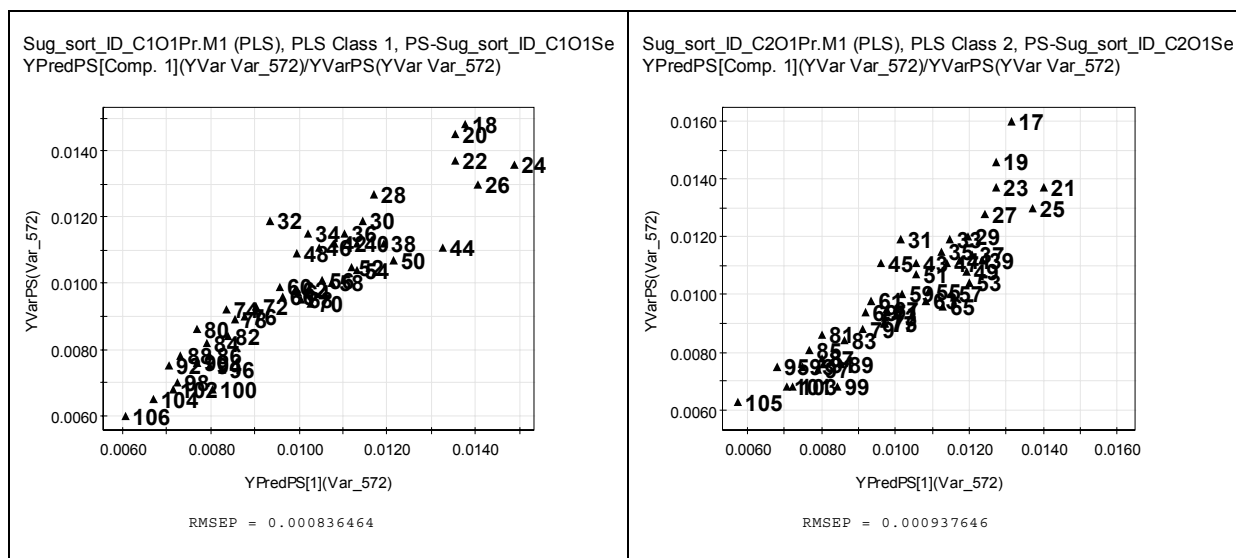
Task 4

By using OSC for signal correction and removing one component of unrelated X-information, we can see that it is possible to increase the predictive power for each model by 2%. This is shown in the second row of the table below. Interestingly, only one PLS component is necessary in both cases. For comparison, the corresponding results for MSC and SNV pre-processed calibration models are given. We can see that using MSC and SNV for signal correction gave no improvements in predictive power for this application.

Model	RMSEP(o)	$Q^2_{ext}(o)$	A(o)	RMSEP(e)	$Q^2_{ext}(e)$	A(e)
PLS	0.00090	0.83	2	0.00099	0.80	2
OSC1	0.00084	0.85	1	0.00094	0.82	1
MSC	0.00174	0.36	5	0.00173	0.33	5
SNV	0.00163	0.43	5	0.00160	0.43	5
1 st Der.	0.00111	0.81	2	0.00122	0.73	2
2 nd Der.	0.00223	0.00	2	0.00258	0.00	2

Caption to table: (o) training set odd-numbered observations and prediction set even-numbered observations; (e) training set even-numbered observations and prediction set odd-numbered observations; RMSEP = root mean square error of prediction for prediction set; Q^2_{ext} = goodness of prediction for external prediction set; A = number of PLS components.

The two figures below show the relationships between predicted and observed ash contents for the two external prediction sets (OSC pre-treated data).



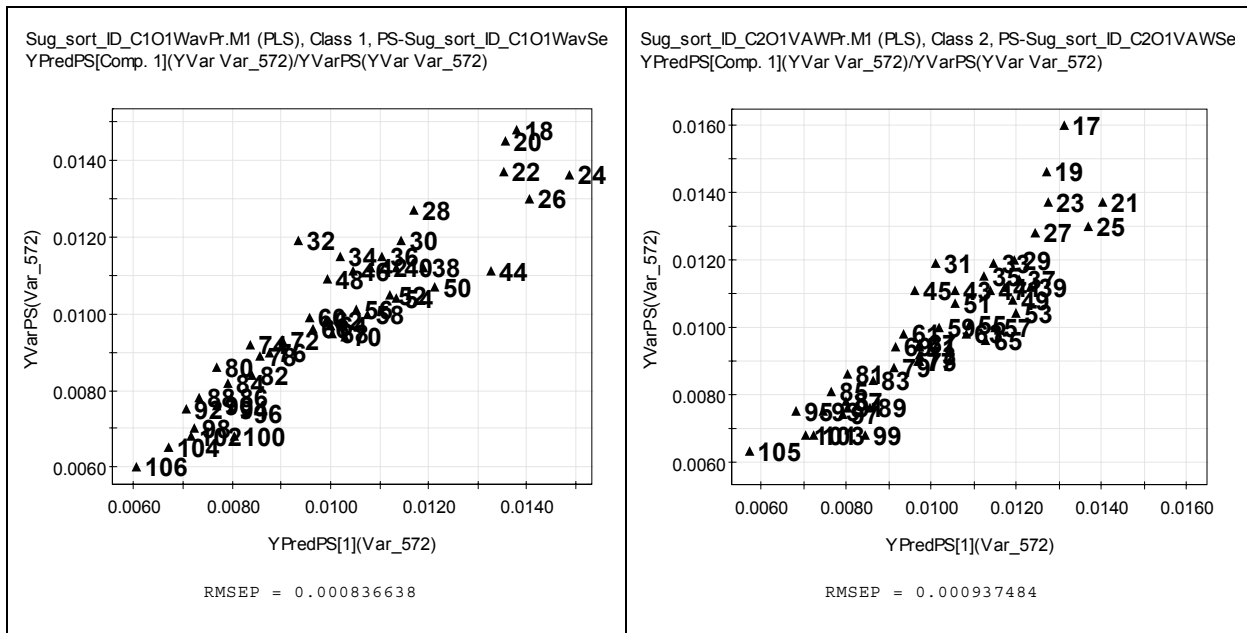
There is a slight improvement in predictive power (2%) when applying OSC to the SUGAR data. At first glance, such a small improvement might seem negligible, but ultimately it may be a decisive factor in upholding a competitive edge in production. OSC also works better than MSC and SNV in the current application. Since MSC and SNV have other underlying rationales than OSC (baseline removal and amplitude adjustments), this result is not completely surprising. Also, the MSC and SNV models are less parsimonious than the corresponding PLS and OSC1-PLS calibration models.

Task 5

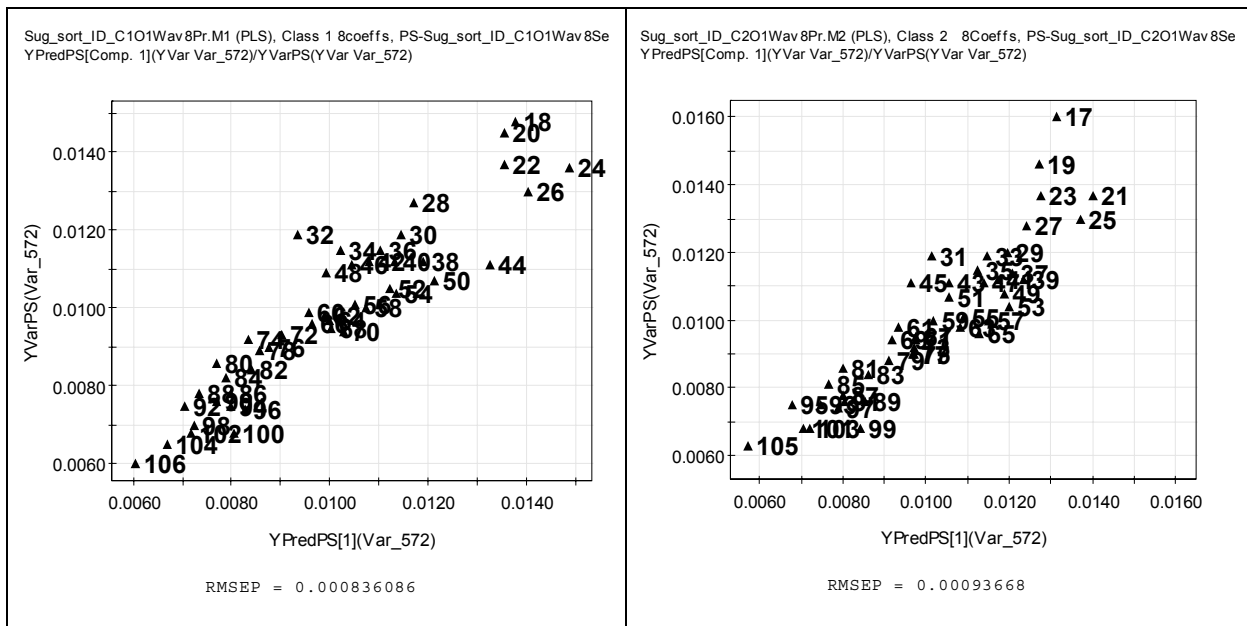
By using OSC for signal correction and wavelet analysis for signal compression, we can see that we obtained results very similar to the case where only OSC was used ($K = 571$). We tested to save wavelet coefficients representing 95% ($K = 8$) and 99.5% (K approx. 330) of the total variance explained. We conclude that the addition of wavelet analysis did not add any extra predictive power, however, it enabled significant signal compression.

Model	RMSEP (o)	Q2ext	RMSEP (o)	Q2ext
PLS	0.00090	0.83	0.00099	0.80
OSC1-PLS	0.00084	0.85	0.00094	0.82
OSC1-Wav99.5%-PLS	0.00084	0.85	0.00094	0.82
OSC1-Wav95%-PLS	0.00084	0.85	0.00094	0.82

The two plots below show the results of using 322 (odd-numbered training set) and 343 (even-numbered training set) wavelet coefficients representing 99.5% of the total spectral variance. Note the similarity with corresponding plots in Task 4.



The two plots below show the results of using 8 (odd-numbered training set) and 8 (even-numbered training set) wavelet coefficients representing 95% of the total spectral variance. Note the similarity with corresponding plots in Task 4.



Conclusions

It is possible to use fluorescence measurements to model and predict the quality of produced sugar. A predictive capacity (Q^2) in the range 0.80-0.85 is considered to be very satisfactory for on-line situations. Such results have also been seen in additional studies done at other sugar plants across Scandinavia, where models of similar or even higher predictive abilities have been established.

Furthermore, this example demonstrates the benefit of first conducting PCA followed by PLS. PCA often provides a good understanding of what is going on in a data set. In this way, the analyst is better prepared for subsequent PLS modelling which can be done faster and more accurately.

There is a slight improvement in predictive power (2%) after applying OSC to the SUGAR data. At first glance, such a small improvement might seem negligible, but over time it may be a decisive factor in upholding a competitive edge in production. OSC is more appropriate than MSC and SNV in the current application because the latter have other underlying rationales than OSC (baseline removal and amplitude adjustments). Also, the MSC and SNV models are less parsimonious than the corresponding PLS and OSC1-PLS calibration models.

MVDA-Exercise NIR_Chip

Multivariate characterisation and classification of wood chips

Background

In the particleboard industry it is important to understand the properties of the chips that are fed into the plant. Spectroscopic techniques, like NIR, are being increasingly used in the particleboard industry for on-line assessment of the quality of the starting material. This information is vital for the production of particleboard with the required properties.

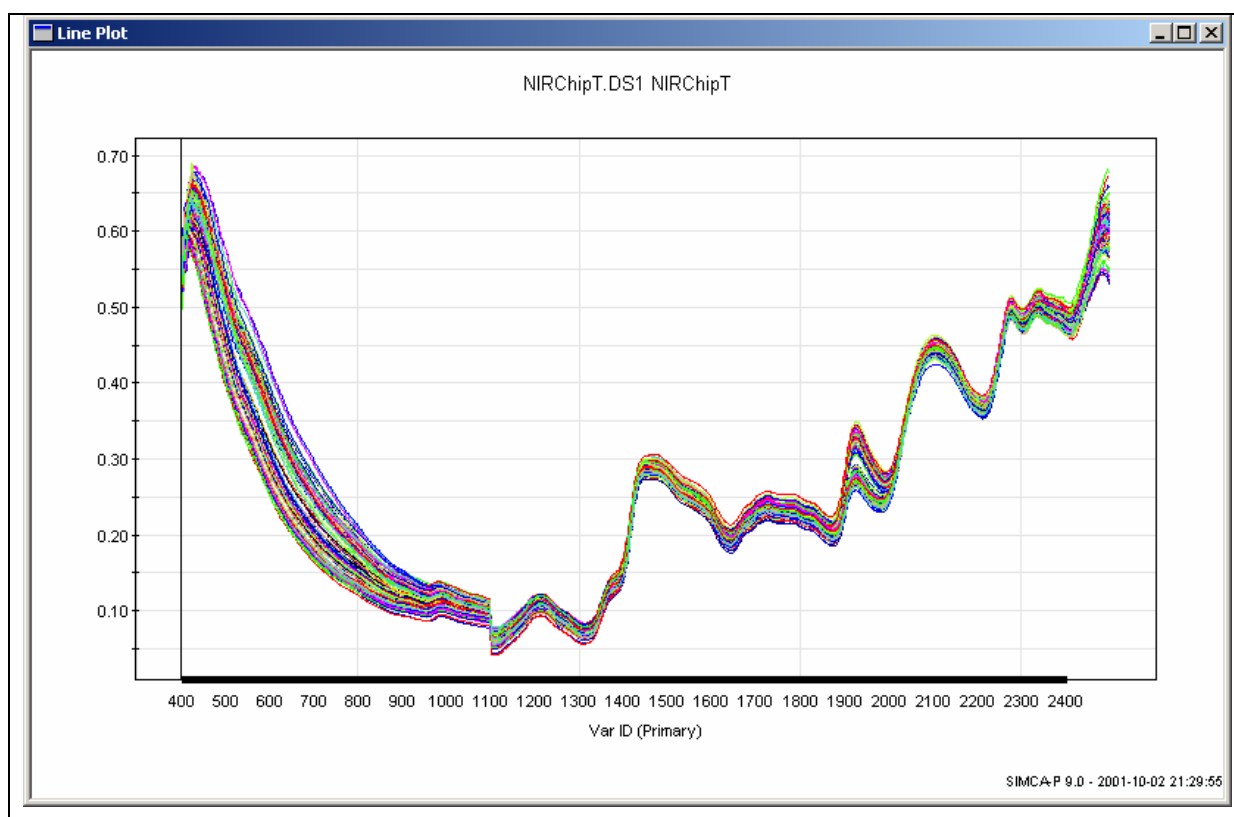
Objective

The objective of this investigation was to examine whether four types of wood chip could be distinguished from each other using NIR characterisation and multivariate data analysis.

Data

This data set contains wood chips of four different categories (varying particle size, varying moisture content), compiled at a Northern particleboard factory. The primary data set has 140 observations (chip samples) and 1050 spectral variables. In this set, class assignment is known for each observation, i.e., there are four classes consisting of 35 samples each. In the secondary prediction data set, there are 78 additional observations of unknown class.

Plot of spectral data for the training set:



Tasks

Task 1

Initiate a new project in SIMCA and import the primary data set NIRChipT.Sim and give the project a unique name. This data set has 140 observations and 1050 variables. In *WorkSet|New|Scale* select Ctr as base weight to centre all variables. Press *Set*. Go to *Observations* and define the four classes, thus: Class 1: 1-35; Class 2: 36-70; Class 3: 71-105; Class 4: 106-140.

Task 2

Compute an overview PCA-model. Cross-validation will suggest 12 components, but for overview purposes it is sufficient to consider PC1 and PC2. Create score and loading plots and interpret the model.

Task 3

Compute class-specific models for each class (*Analysis|Autofit Class Models*). Cross-validation will indicate rather too many components, but you should store three components for each class model. Then import the prediction data from the file NIRChipS.Sim and give this data set a unique name. Go to *Predictions|Specify Predictionset|Dataset* and select the secondary data set for predictions. Create DModX plots (*Predictions|Distance to Model|X-block*) for each class and evaluate the classification results. Try to ascribe a likely class membership for each observation in the prediction data set.

Note: Here, we are interested in classification; therefore, in this exercise we should use DModX+. When the objective is to use DModX as a guidance for contribution plotting, we use the alternative measure DModX.

Task 4

Another method used to discriminate observations is called PLS-DA (Discriminant Analysis). Go to *Analysis/Change Model Type* and select *PLS-DA*. Run autofit and investigate the resulting t_1/t_2 score plot. Interpret the model. Is it possible to discriminate between the four classes of observations? Apply the PLS-DA model to the prediction data. Compare classification results with preceding task.

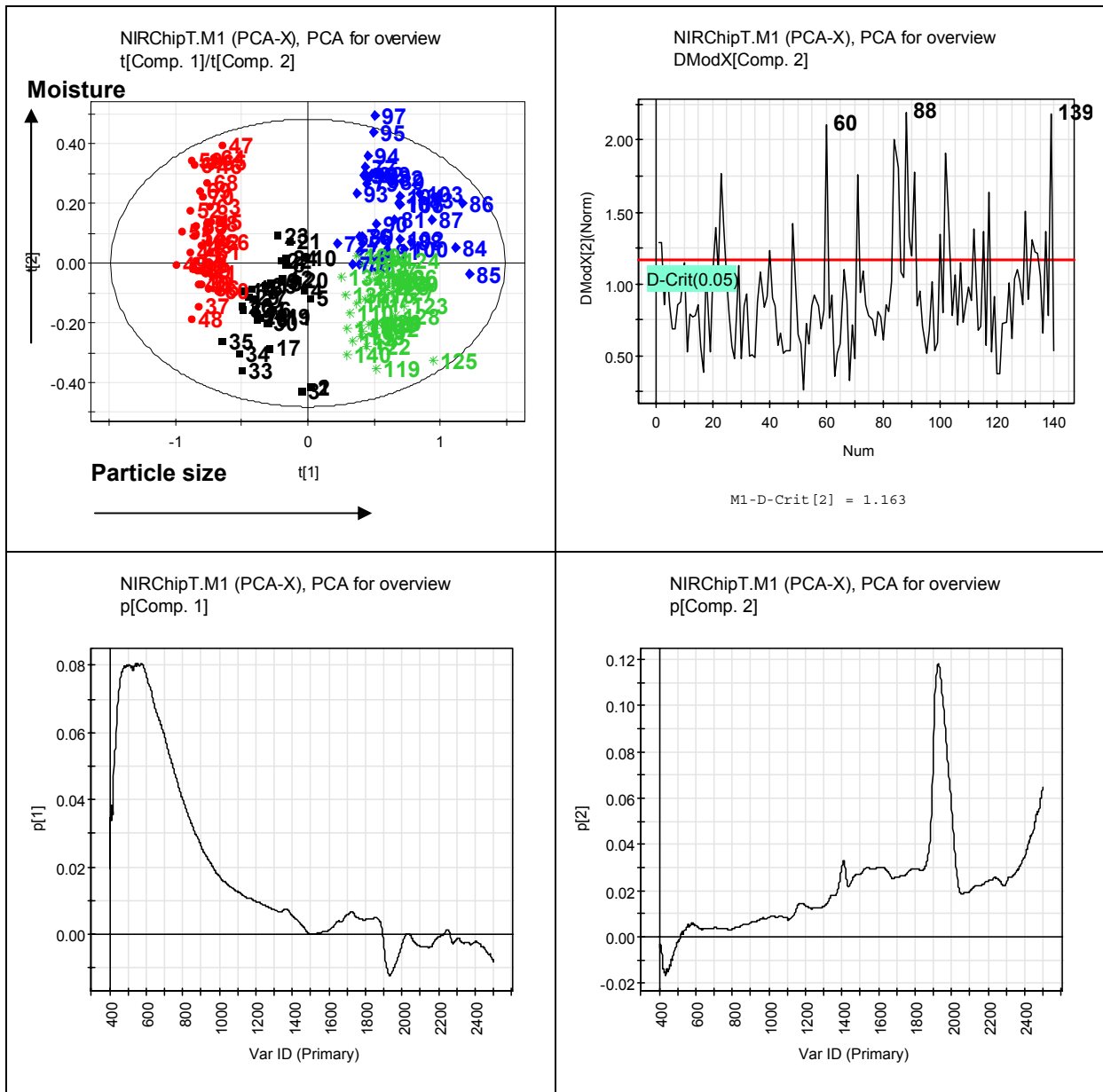
Task 5

Repeat the classification step of Task 3, but use more components than three for each class model. Is it possible to sharpen classification accuracy? ***There is no solution given to this task!***

Solutions to NIR_Chip

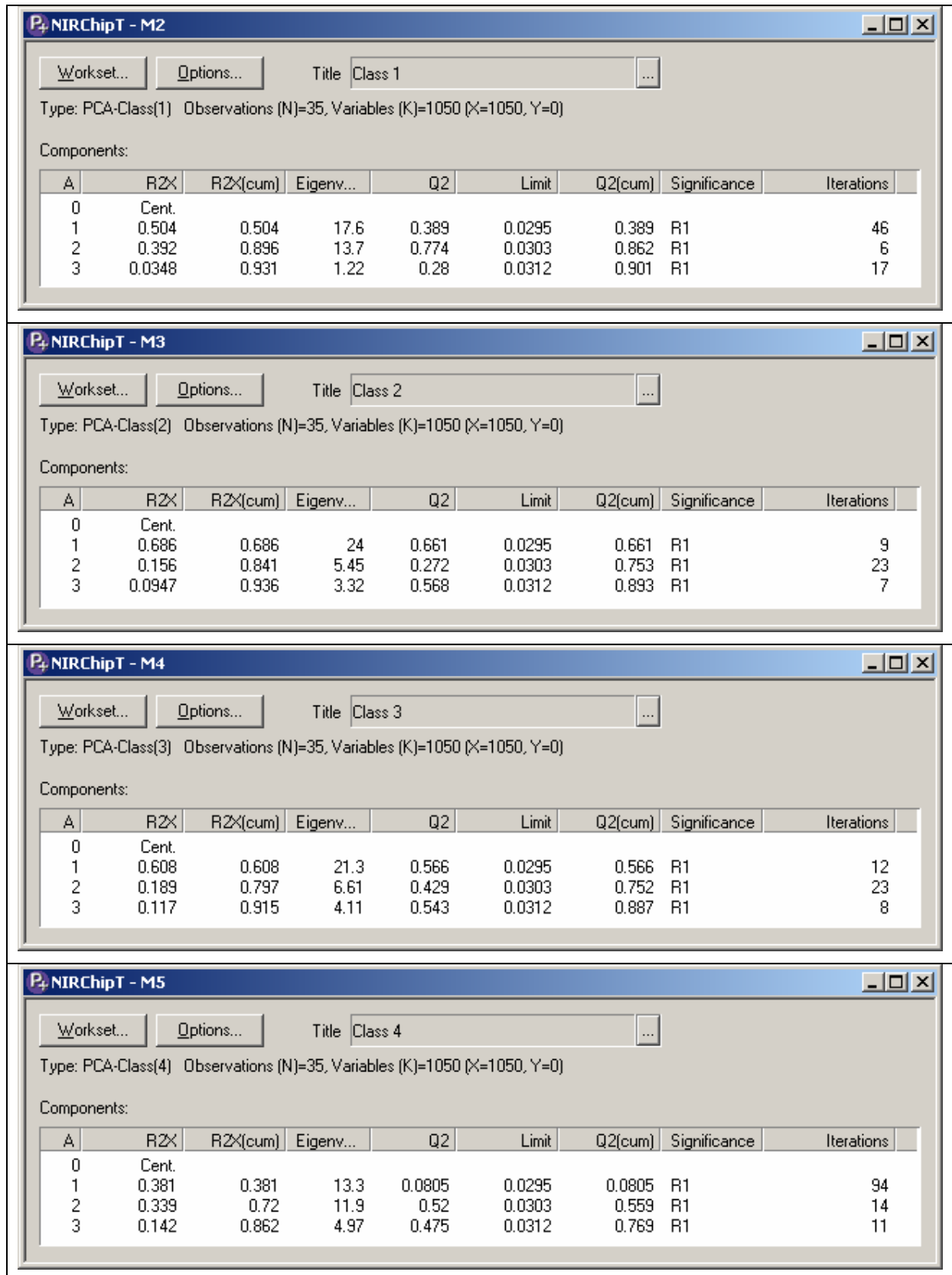
Task 2

After two components $R^2X = 0.94$ and $Q^2 = 0.94$. The score plot shows some grouping according to class assignment. However, classes 3 and 4 seem to overlap. Particle size and moisture govern the separation among the chip samples. Samples 60, 88, and 139 are moderate outliers according to the DModX plot. The first loading spectrum resembles the average NIR spectrum in the low wavelength region, and the second loading spectrum captures a peak at higher wavelengths.



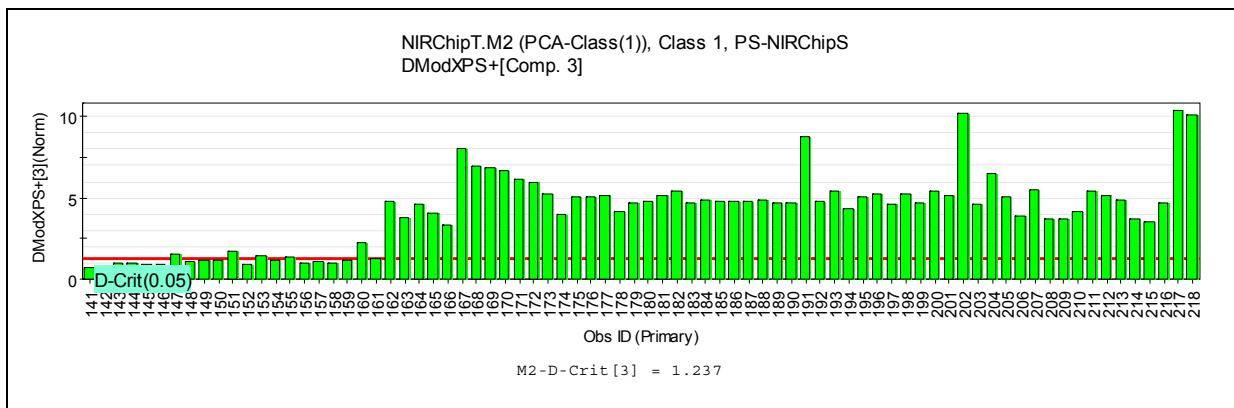
Task 3

Four class-specific PCA models were computed and are summarised below:

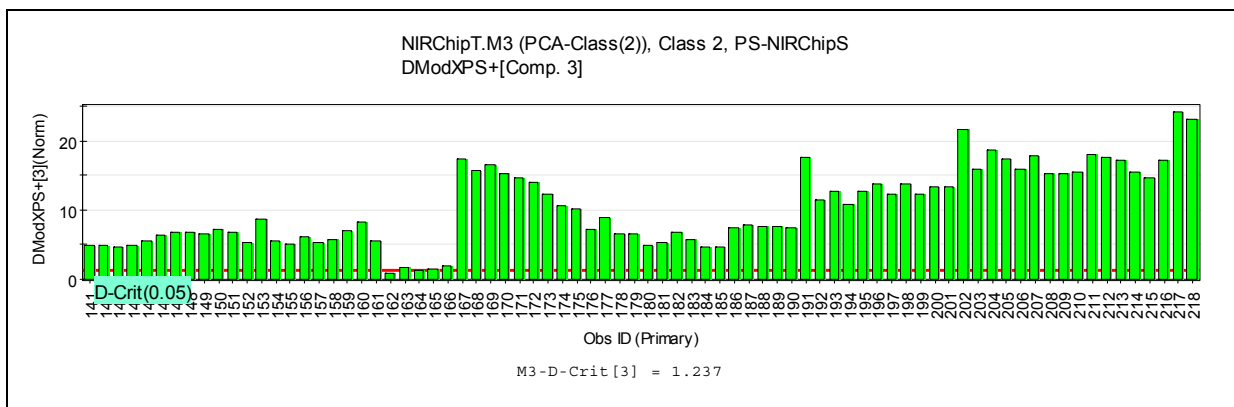


These models were executed on the 78 samples in the prediction set, the results of which are summarised below in four DModX-plots.

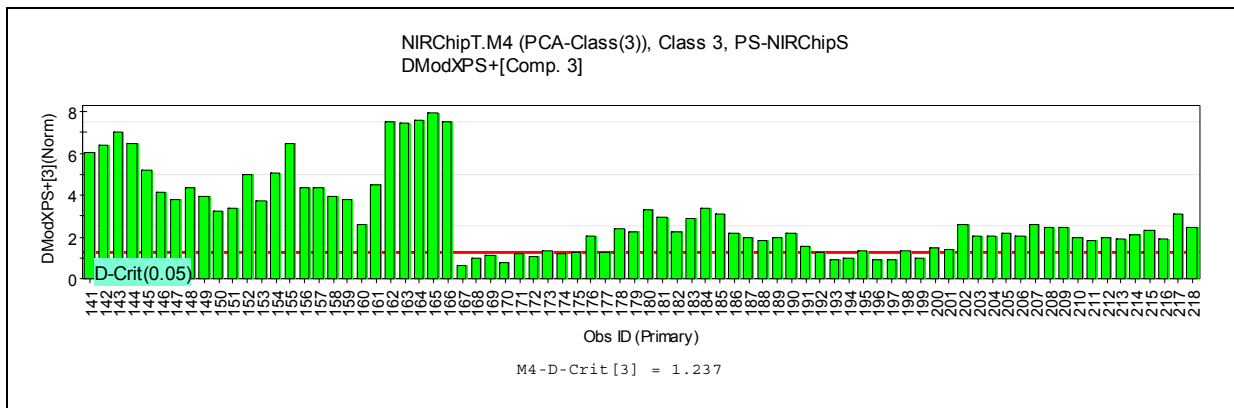
Observations 141-161 are classified as close to Class 1.



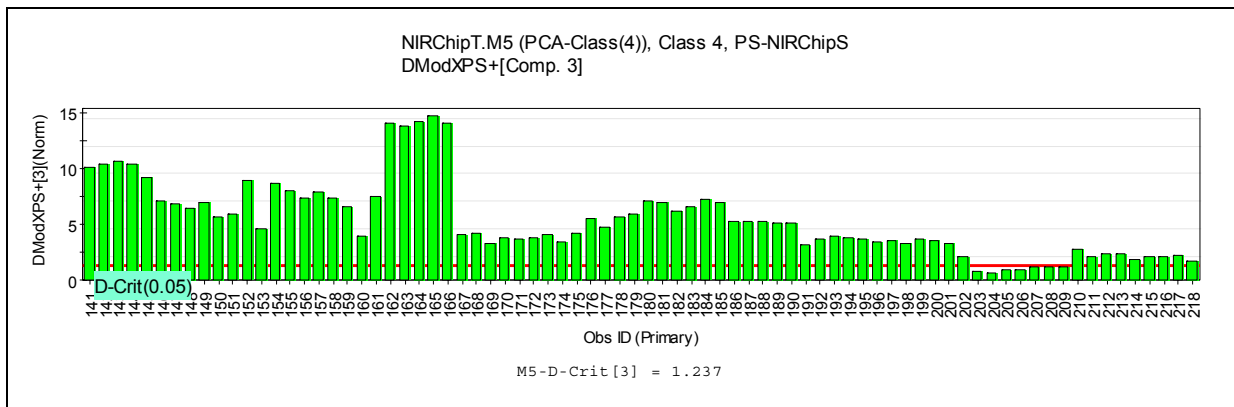
Observations 162-166 are classified as close to Class 2.



Observations 167-175, 177, and 192-201 are classified as close to Class 3. Samples 176, 178-191, and 202-218 are also classified as comparatively similar to class 3.



Observations 203-209 are classified as very close to class 4. Samples 202 and 210-218 are classified as close to class 4.



Task 4

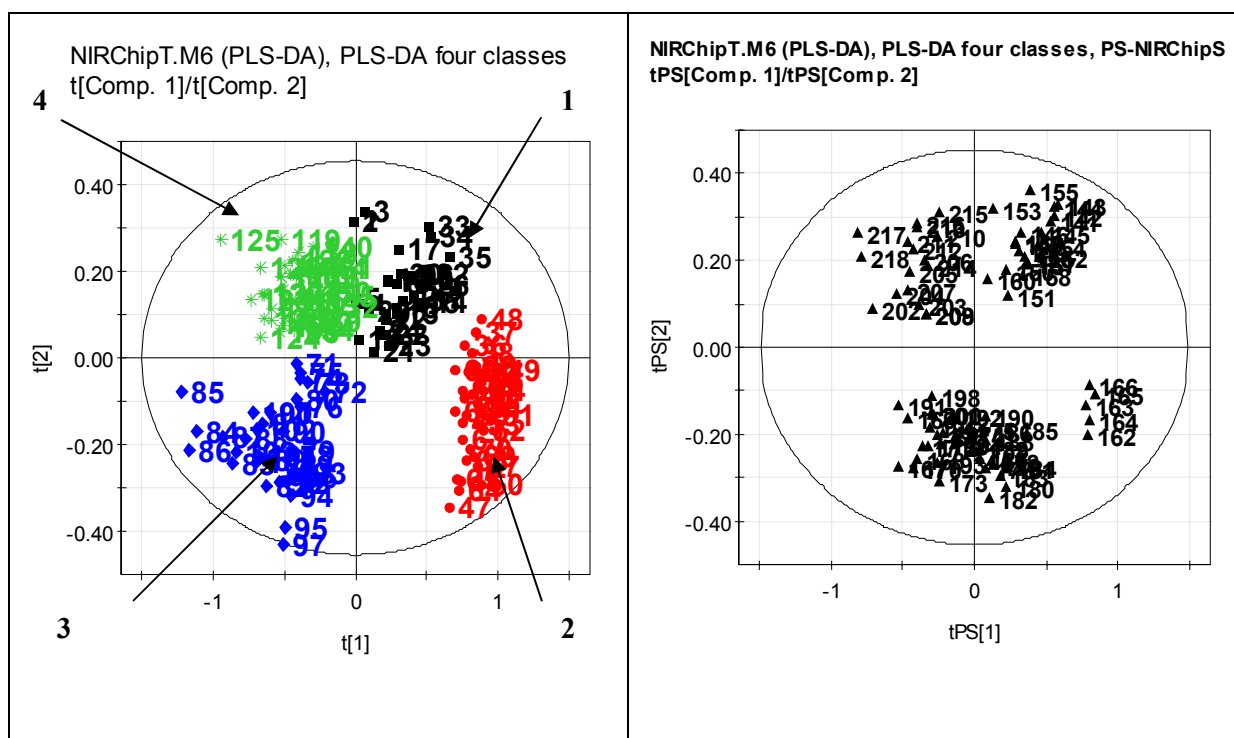
The PLS-DA modelling yielded a highly significant three-component model, with $R^2X = 0.97$, $R^2Y = 0.61$ and $Q^2 = 0.61$.

NIRChipT - M6											
Workset...		Options...		Title PLS-DA four classes							
Type: PLS-DA Observations (N)=140, Variables (K)=1054 (X=1050, Y=4)											
Components:											
A	R2X	R2X(cum)	Eigenv...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signifi...	Ite...	
0	Cent.	0.85	119	Cent.	0.307	0.307	0	0.307	R1	6	
1	0.85	0.935	11.8	0.254	0.561	0.363	0	0.559	R1	4	
2	0.0845	0.97	4.95	0.0524	0.614	0.115	0	0.609	R1	6	

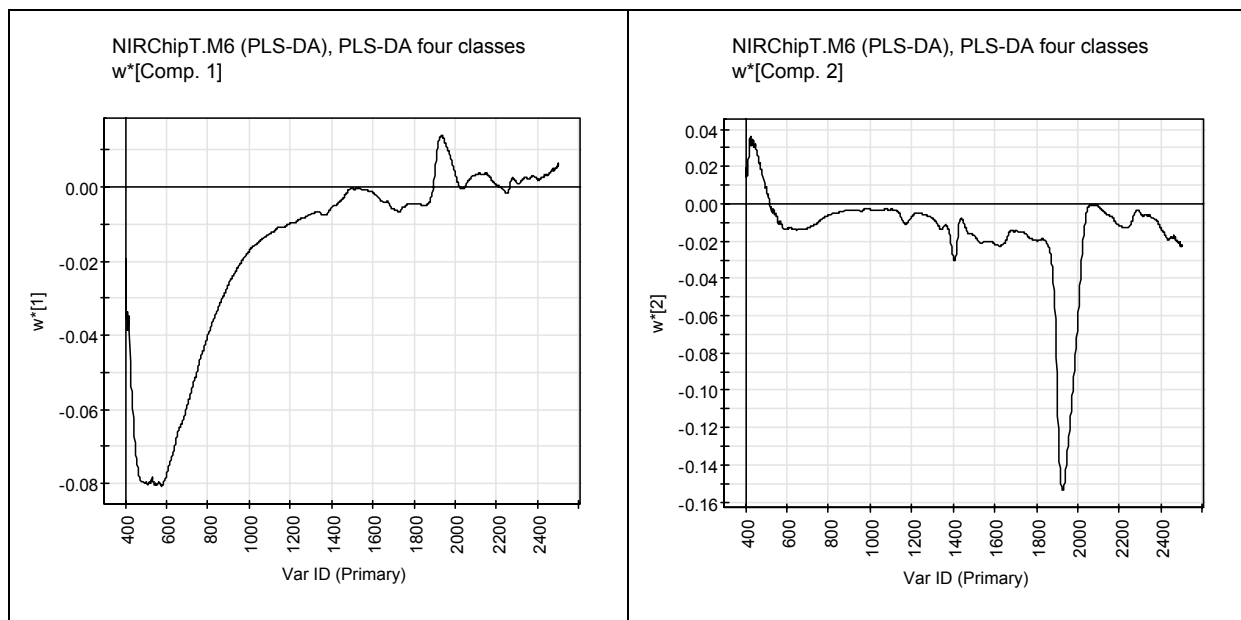
Usually in PLS there is an interest in plotting t/u score plots. However, PLS-DA is a special case where the plotting of t/t score plots is more relevant. This is because we want to see if the X-data carry class discriminating information. Indeed, the t_1/t_2 plot shown below indicates the existence of class discriminating information. The right-hand score plot below shows prediction results. Four clusters of observations are seen. A summary of the prediction (classification) results indicates:

- Observations 141-161 \Leftrightarrow class 1
- Observations 162-166 \Leftrightarrow class 2
- Observations 167-201 \Leftrightarrow class 3
- Observations 202-218 \Leftrightarrow class 4.

However, note that the classification of samples 167-201 is not clear-cut. In the prediction score plot (right) parts of these samples are situated close to the area where we find class 2 observations in the training set score plot (left). This classification ambiguity with regard to class 3 was also observed in the previous task.



In order to see which variables carry class discriminating information, we look at the PLS weight spectra.



Apparently the left-hand part of the wavelength region reflects differences in the t_1 (horizontal) direction. This component separates the observations based on particle size. The second component (t_2 , vertical direction) is heavily influenced by variables 725-811. This part of the spectral region (1848 – 2020 nm) contains discriminating information about chip moisture content.

Conclusions

The main conclusion drawn from this investigation is that NIR characterisation coupled with multivariate data analysis is useful for on-line discrimination of four types of starting material in the particleboard industry. Classification results ranged from good to excellent for the prediction data. The worst classification was for class 3 samples.

The correct class memberships are as follows:

- Observations 141-161 \Leftrightarrow class 1
- Observations 162-166 \Leftrightarrow class 2
- Observations 167-201 \Leftrightarrow class 3
- Observations 202-218 \Leftrightarrow class 4.

This study hints at how multivariate characterisation for classification of raw materials can be carried out when apparently the same starting material is delivered in different batches, or supplied by different manufacturers. This approach is common practice in the pharmaceutical and particleboard industries.

MVDA-Exercise CELLULOSE

Modelling Viscosity of Cellulose Powder

Background

This example illustrates the use of spectral filtering and wavelet compression with multivariate calibration.

The data set of this example was collected at Akzo Nobel, Örnsköldsvik, in Sweden. The raw material for their cellulose derivative process is delivered to the factory in form of cellulose sheets. Before entering the process the cellulose sheets are controlled by a viscosity measurement, which functions as a steering parameter for that particular batch.

In this data set NIR spectra for 180 cellulose sheets were collected after the sheets had been sent through a grinding process. Hence the NIR spectra were measured on the cellulose raw material in powder form. For calculation of the calibration model 90 spectra were used. The remaining 90 spectra were used for model validation.

Objective

The objective of this study is to develop a good calibration model with the calibration set of 90 samples and validate this model with the test set of 90 samples.

We will use signal filtering to possibly improve the calibration model, and we will compress the X matrix, with wavelets, for efficiency and fast computation.

The results of the model after filtering and compression will be compared with the results of the model with the original data.

Data

The data-set consists of:

- X: 1201 wavelengths in the VIS-NIR region (400-2500) nm
- Y: Viscosity of cellulose powder.

Tasks

Task 1

Create a new project in SIMCA by importing the data from *CELLULOSE.DIF* (*File/New*). Mark the first column in the data set as primary observation ID. Mark the second column (called Class ID) as secondary observation ID. Mark the third column (Viscosity) as the Y-variable. Mark the first row as primary variable ID. Finish the import

With the dataset open and active, right click and select Plot Xobs to plot the spectra. What do you see?

Task 2

Divide the 180 observations into a calibration set (ClassID = 1) and a prediction set (Class ID = 2). (Hint: Go to *Workset/New/Observations* and apply the search function to the secondary observation ID.)

When working with spectral data it is often appropriate to work with Pareto-scaled data. To Pareto-scale and mean-center the spectral data, follow these steps: Select the *Scale* tab, and mark all the X-variables. Under *Set Scaling/Base* select “Par”. Press *Set*. Now you have scaled the data appropriately.

Fit a PLS-model to the calibration set (Class 1). Use this model to predict viscosity values for the predictions set (Class 2). This model will be our reference model. Review the fit and interpret the model.

Task 3

NIR data often contain systematic variation that is not related to the response Y. We will apply signal filtering to the X block (the NIR data) to remove variation that might not be of relevance for Y.

SIMCA supports a number of spectral filters, including MSC, SNV, OSC, wavelets, and first and second derivatives. Explore the impact of these and see if the predictive power may be improved. Maintain the previous division of the observations in two classes, i.e., Class 1 = calibration set and Class 2 = prediction set.

Create line plots of filtered (“corrected”) spectral data and evaluate the impact of the various filtering approaches. Compare external predictive ability.

Task 4

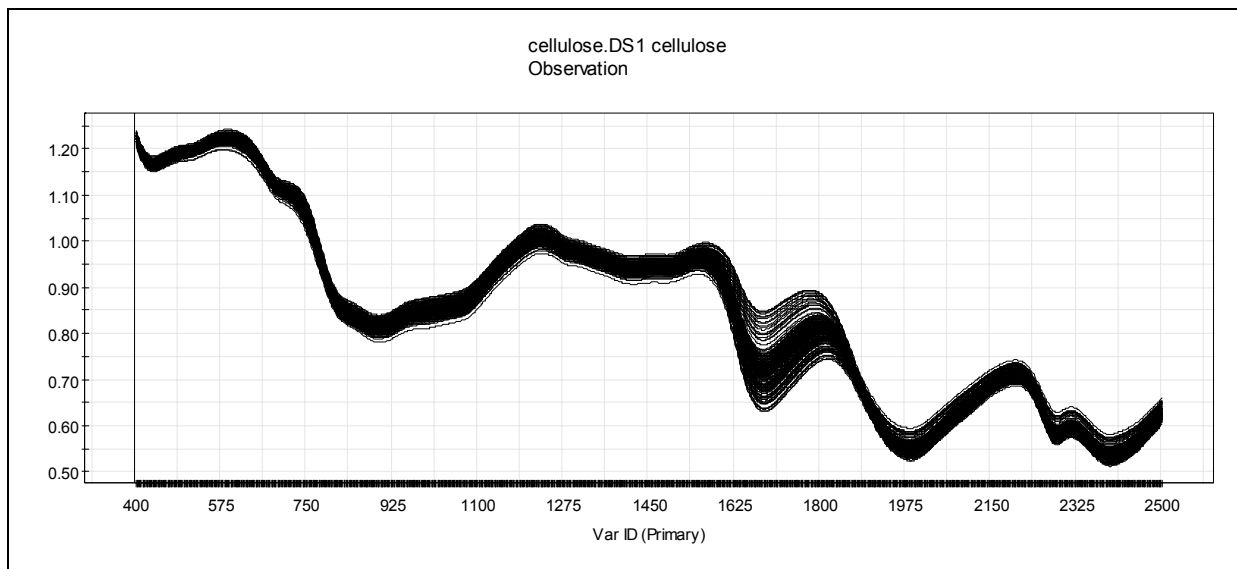
Swap the roles of Classes 1 and 2, and redo Task 3.

There is no solution provided to this Task.

Solutions to CELLULOSE

Task 1

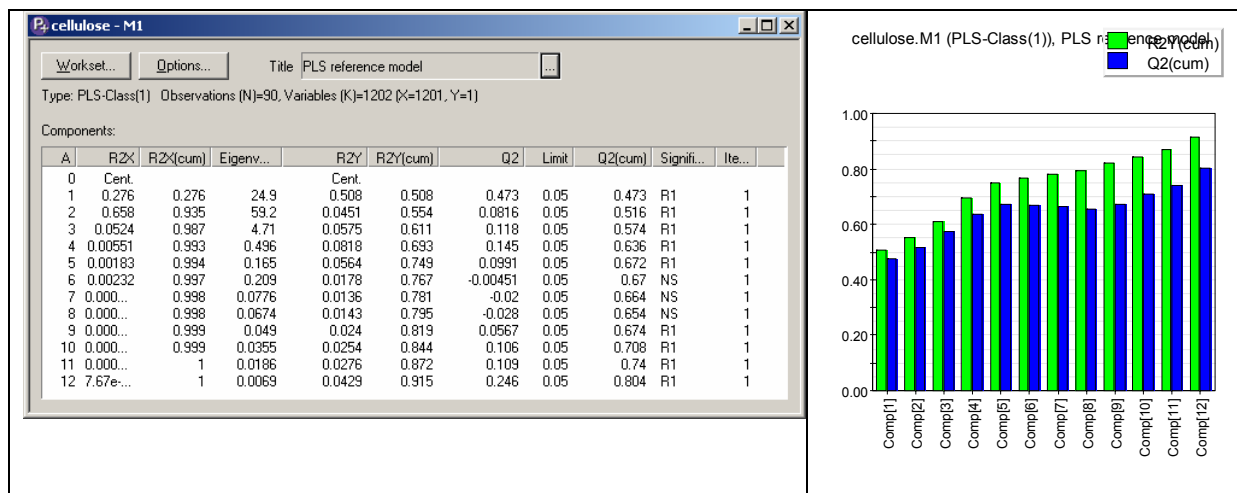
A plot of spectral data, prior to pre-processing, for all 180 observations is given below. We can see that the most extensive spectral variation occurs in the 1600-1800 nm region.



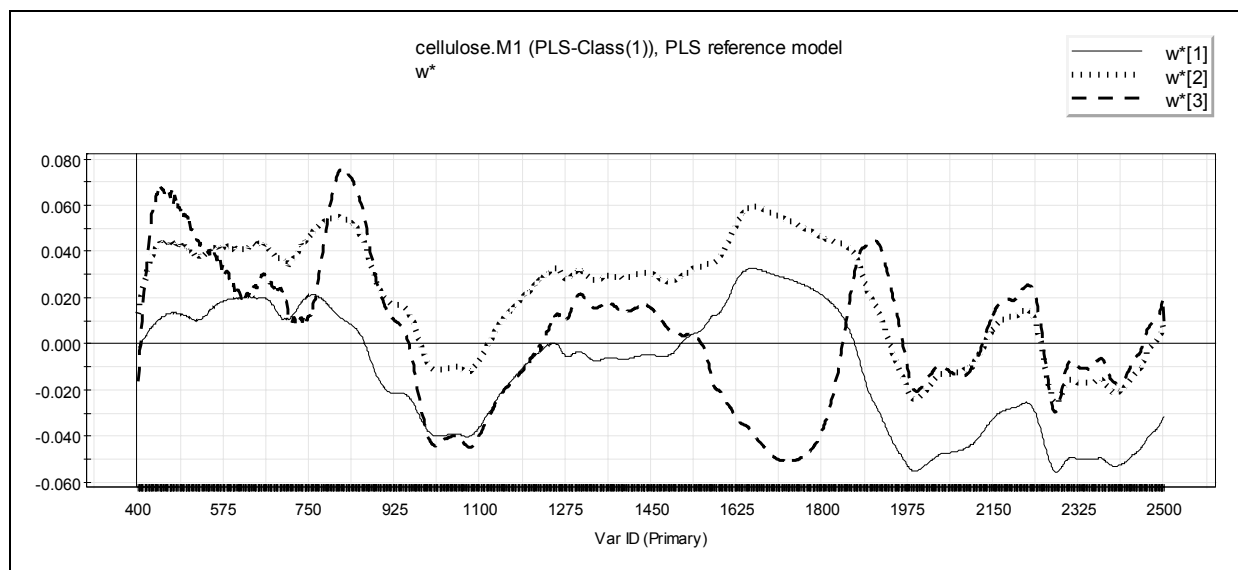
Task 2

We fitted a PLS-model with 12 components, although cross-validation suggested only five. The plot below shows that cross-validation is in this case trapped by a first local peak in Q^2 . Further augmentation of the model leads forward to a second peak. The best predictions for the prediction set are obtained with 12 components. Experience also shows that rather many components are often needed with NIR-data. Hence, we decided to use 12 components.

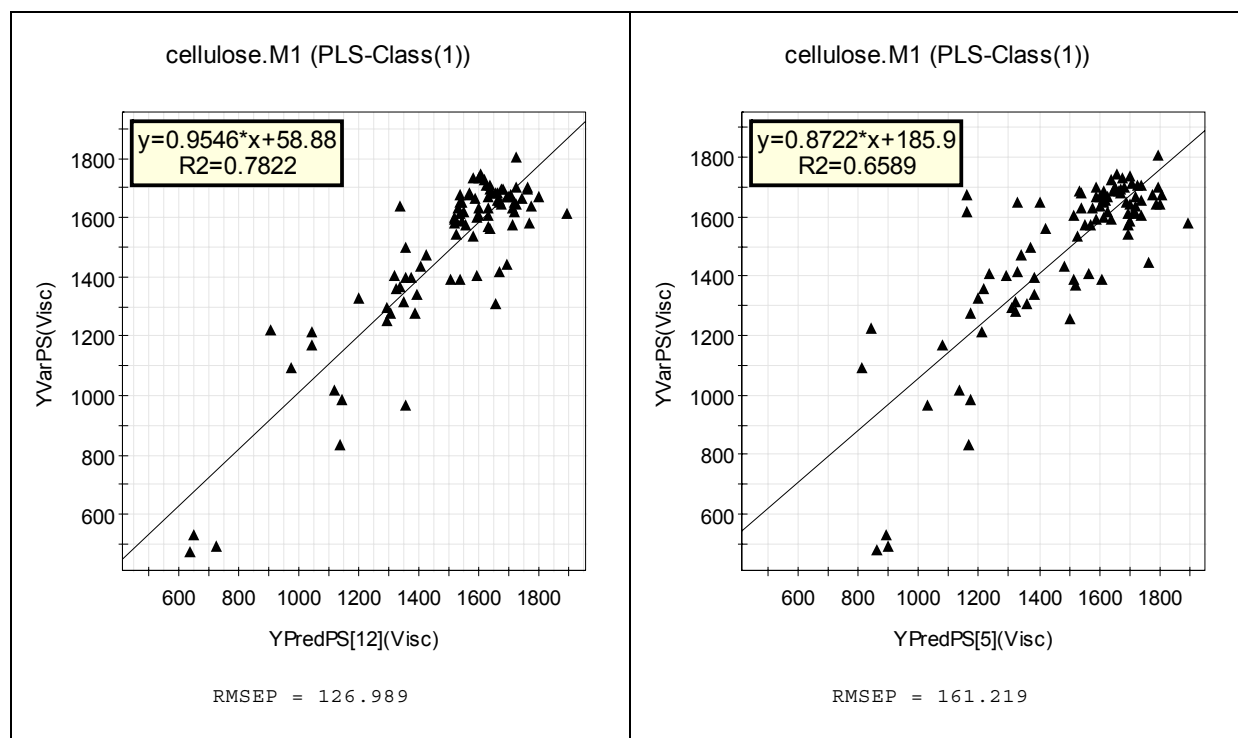
As shown by the summary table, the second component captures a lot of spectral variation that is not particularly related to the viscosity measurements.



The line plots of $w^*_1 - w^*_3$ show how the various spectral regions contribute to the first, second, and third components. It is reasonable to focus on these three as they capture almost 97% of the variation of Y. The other components are small corrections. The wavelength regions 950-1100, 1600-1825 and 1950-2100 carry a lot of information.



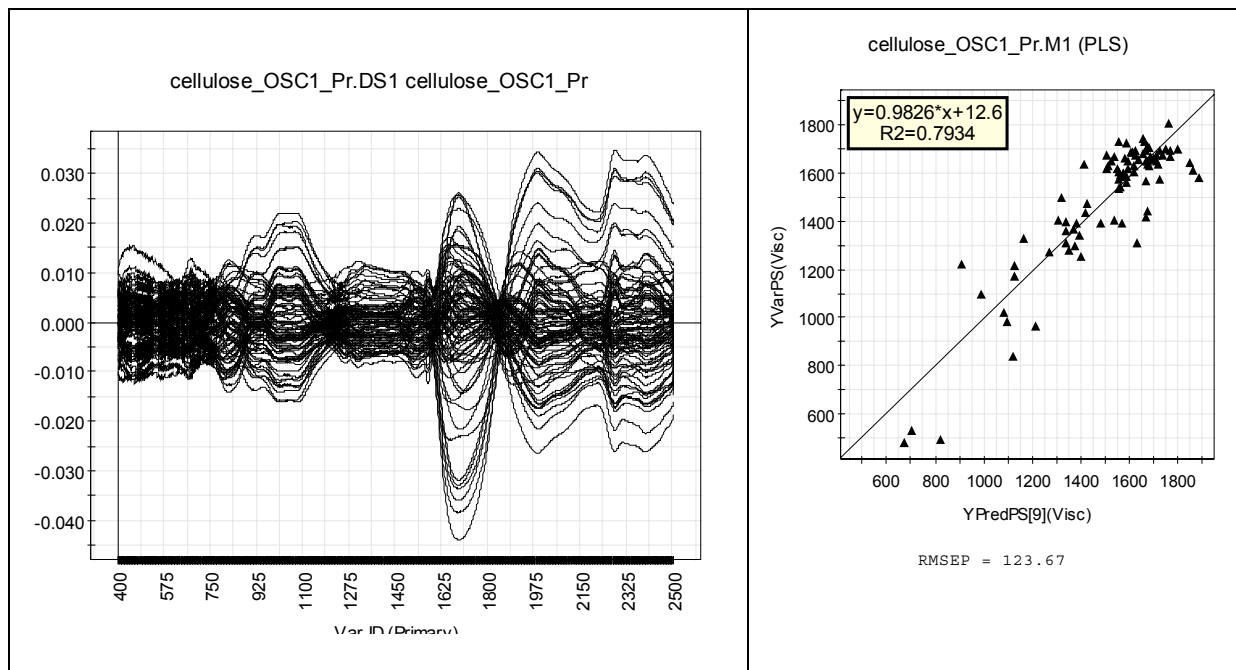
The predictive power is documented in the two plots below, after 12 and 5 components. External RMSEP is lower with 12 components. One way to get hold of the external Q^2 is to use the *Show/Hide regression line*-button in SIMCA. The R^2 displayed is equivalent to Q^2_{ext} , and a value of 0.78 must be considered very good in the light of the cross-validated Q^2_{int} of 0.80.



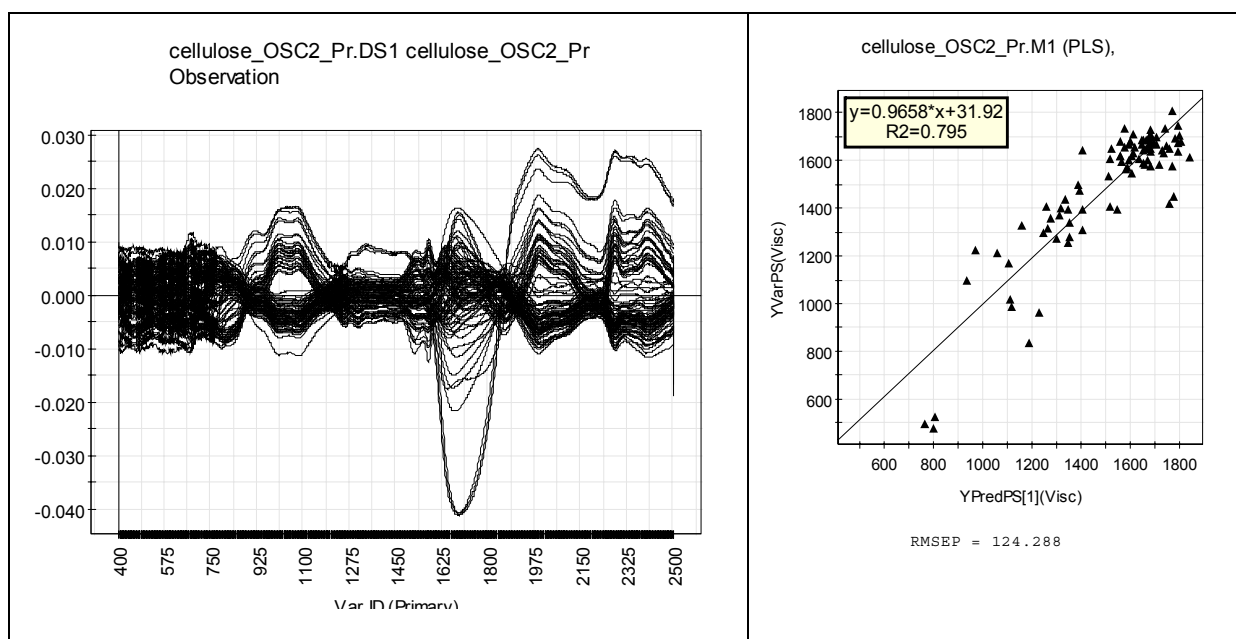
Task 3

In all the models presented below, we did not change default scaling in the spectral filtering routines. However, when doing PLS analysis on the filtered data, default scaling (Ctr) was changed (to Par) for all X-variables.

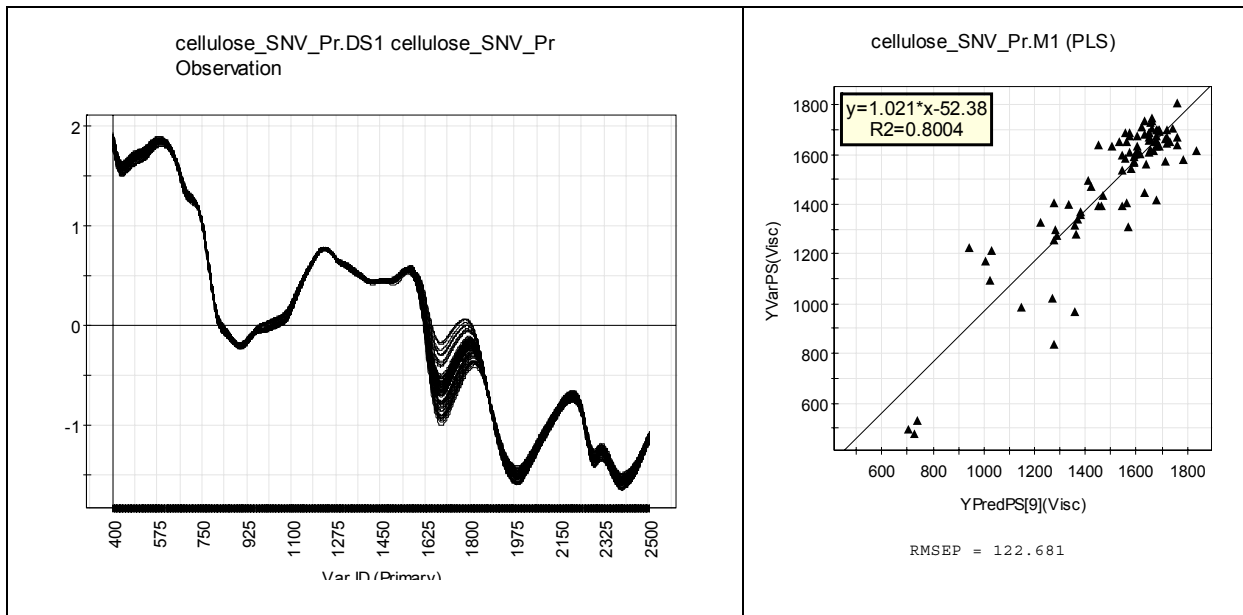
Orthogonal signal correction - 1: One OSC-component was computed removing 69% of X. The resulting PLS-model based on the filtered data has 9 components. External predictive power, Q^2_{ext} , is = 0.793.



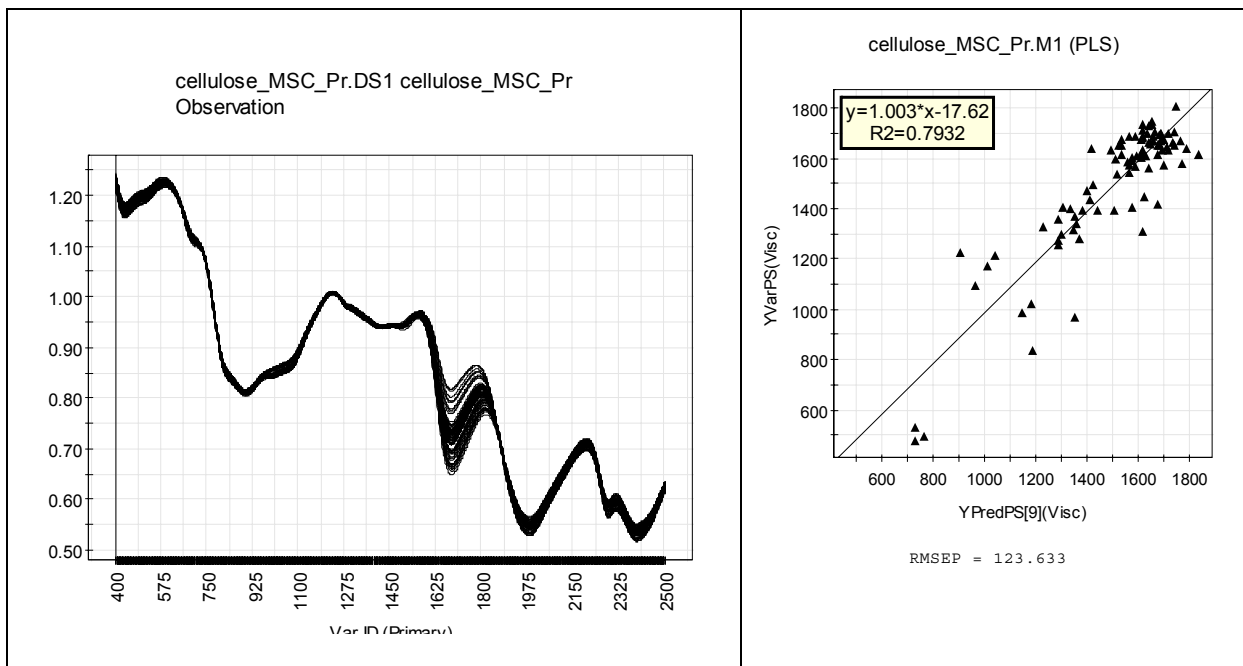
Orthogonal signal correction - 2: Two OSC-components were computed removing 82% of X. The resulting PLS-model based on the filtered data had 1 component. External predictive power, Q^2_{ext} , is = 0.795.



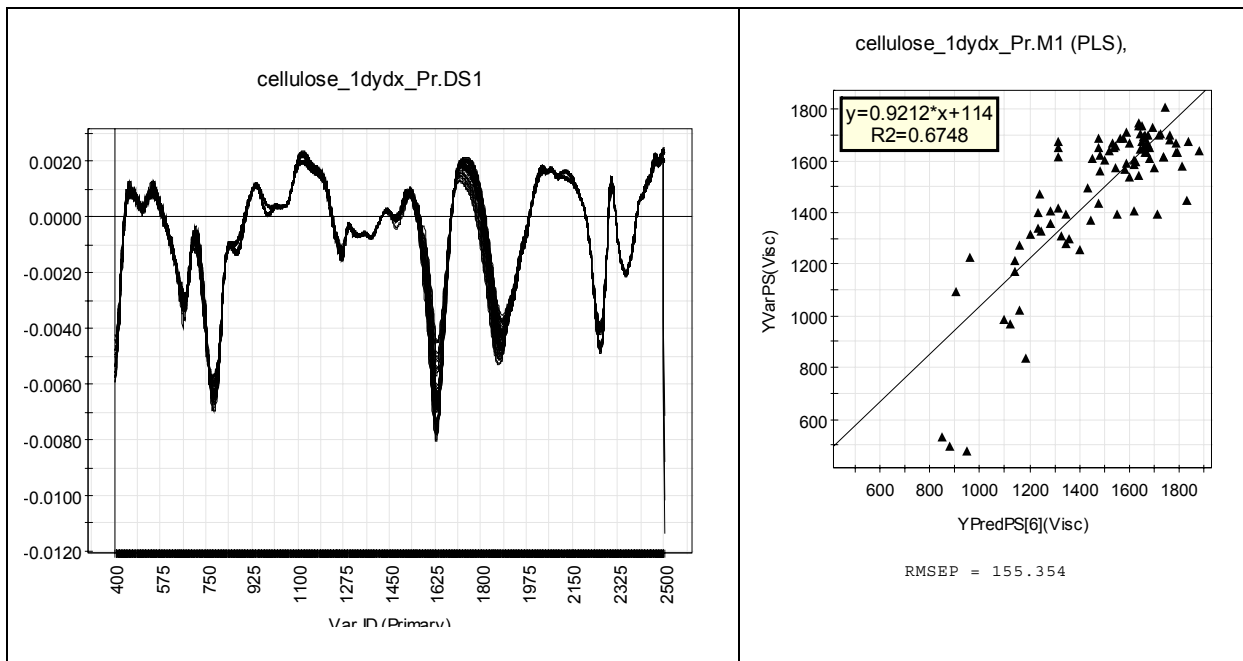
Standard normal variate: The resulting PLS-model based on the filtered data has 9 components. External predictive power, Q^2_{ext} , is = 0.800.



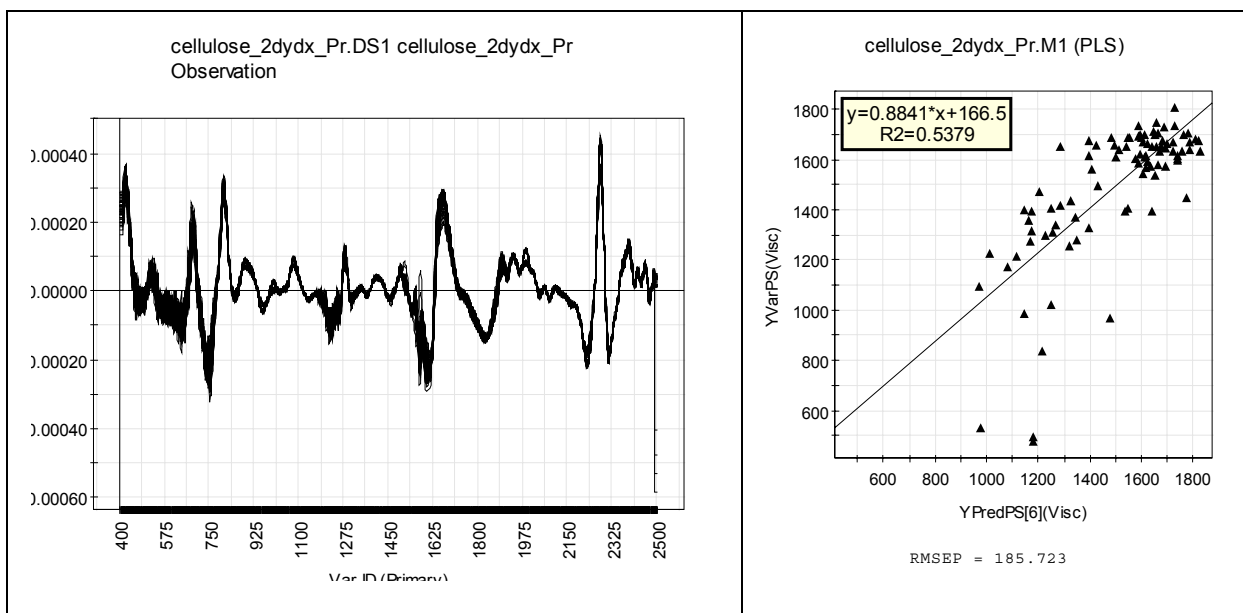
Multiplicative signal correction: The resulting PLS-model based on the filtered data has 9 components. External predictive power, Q^2_{ext} , is = 0.793.



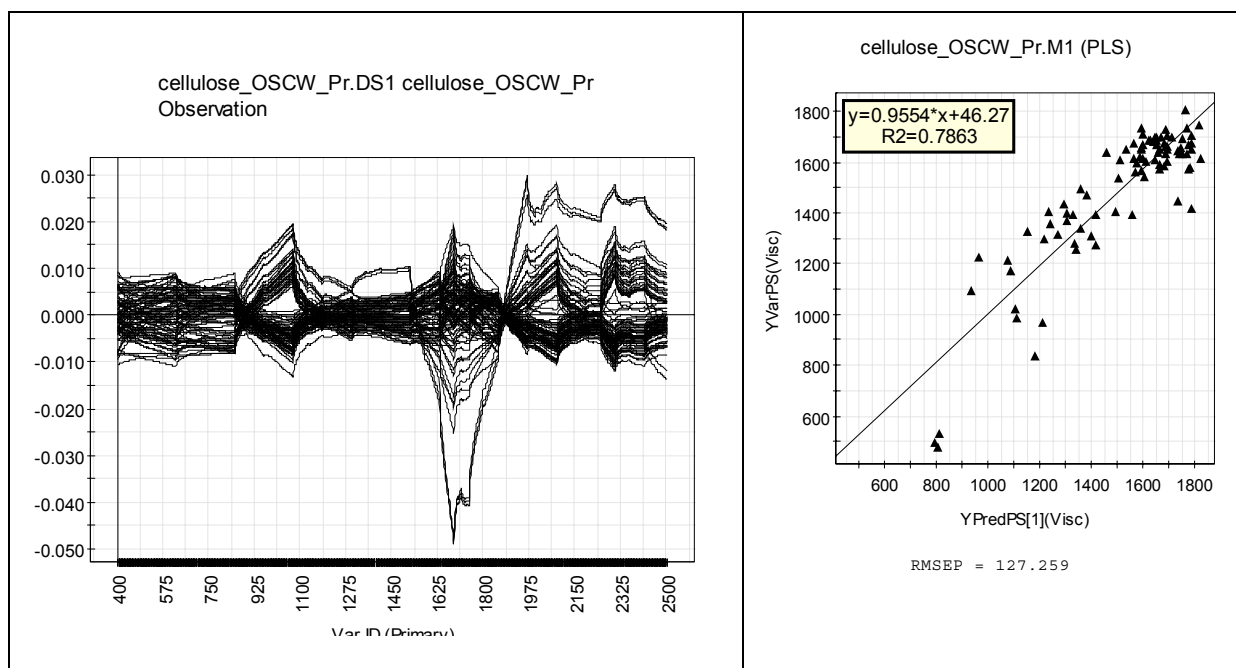
1st derivative: The resulting PLS-model based on the filtered data has 6 components. External predictive power, Q^2_{ext} , is = 0.675. The SG-smoothing was accomplished using a window size of five points and a quadratic polynomial.



2nd derivative: The resulting PLS-model based on the filtered data has 6 components. External predictive power, Q^2_{ext} , is = 0.538. The SG-smoothing was accomplished using a window size of eleven points and a quadratic polynomial.

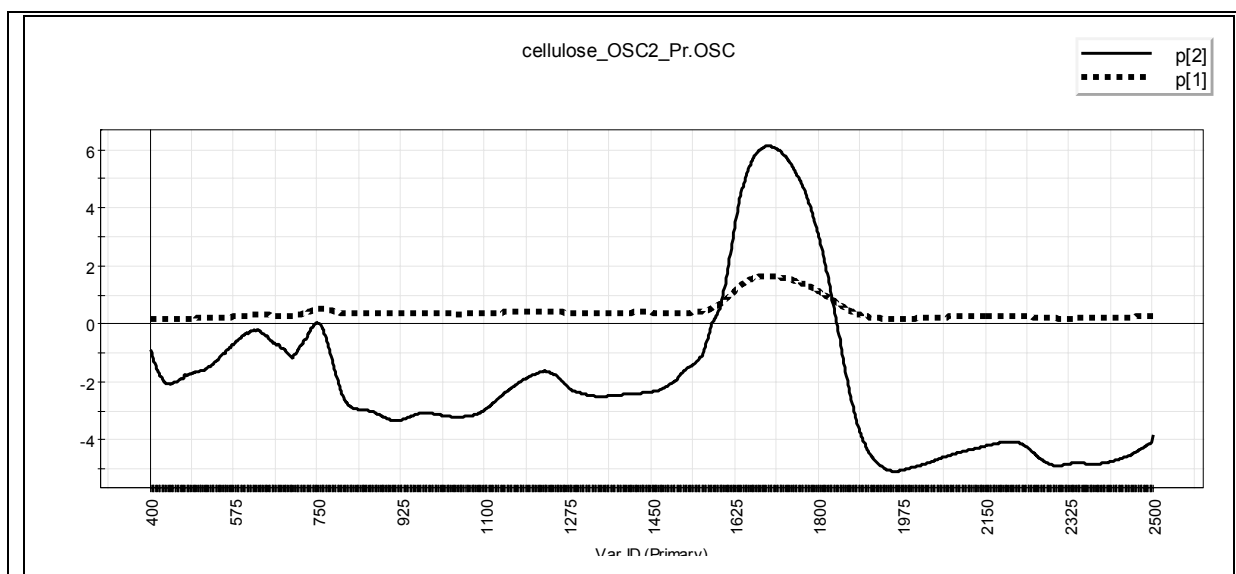


OSC2 and wavelet compression: In addition to the OSC2 filtering, compression of spectral data was accomplished using Dau-4. Energy retained was by variance, and compression method was DWT. By storing 16 wavelet coefficients 95% of the variance in X is explained. The resulting PLS-model based on the filtered and compressed data has 1 component. External predictive power, Q^2_{ext} , is = 0.786. This implies a compression efficiency of 98.7% (16/1201) with no loss of predictive power.



Conclusions

This example illustrates how signal filtering and compression can be used in multivariate calibration. NIR data often contain large systematic variation not related to Y, such as baseline shifts etc. Hence, signal filtering may improve model transparency. In this example OSC, SNV, and MSC performed well, a far better than derivation. The coupling of OSC and wavelet compression showed that the signal could be efficiently compressed 1201 to 50 data points without any loss of information. An attractive property of OSC is that it actually makes a model of what was peeled-off from the X-matrix. The OSC loadings p_1 and p_2 plotted below indicates which spectral regions were modified by OSC, i.e., predominantly between 1600 – 1800 and 1850 – 2500 nm.



MVDA-Exercise SOVRING

Process monitoring and optimisation of a mineral sorting plant

Background

Multivariate methods are becoming increasingly common in the mining industry. The current data set originates from LKAB in Malmberget, Sweden, where multivariate data analysis has been very successful and is currently used as a natural part of process development. In this case the main goal was to find a way to control the amount and quality of the two products PAR and FAR. The process set points were varied according to an experimental RSM design in the variables (factors) Ton_in, HS_1, and HS_2.

Objective

We want to answer the following questions:

- Can the data be used to model the process?
- Is it possible to monitor and identify process upsets?
- Are there trends, groups, or different states of the process?
- Can we understand and interpret the relationship between input and output variables?
- Can we make predictions?
- What are the best conditions to maximise amount and quality of product?

Data

X-variables (Predictors)

Total feed	Ton_in	Design variable
Load crusher 30	KR30_IN	
Load crusher 40	KR40_IN	
PARmull	PARM	
Speed separator 1	HS_1	Design variable
Speed separator 2	HS_2	Design variable
Power crusher 30	PKR_30	
Power crusher 40	PKR_40	
Waste rock	GBA	
Load separator 3	TON_S3	
Tailing crusher	KRAV_F	
Tailing total	TOTAVF	

Y-variables (Responses)

Amount concentrate type 1	PAR
Amount concentrate type 2	FAR
Relative amount	r-FAR
Iron in FAR (quality)	%Fe_FAR
Phosphorous in FAR (quality)	%P_FAR
Iron content in crude ore (quality)	%Fe_malm

Tasks

Task 1

Import the data file SOVRING.XLS and create a SIMCA project with a unique name. The data set has 572 observations and 18 variables. During the import, set the column labeled “ONUM” as the primary observation ID, and the two columns labeled “Date/Time” and “Select” as secondary observation ID.s. Three of the variables have more than 50% missing data because laboratory measurements were not performed for all observations. SIMCA checks for missing data at the import and warns if the limit 50% is exceeded. In this case it is OK to include all variables since the missing data are well distributed in the data table. The first step is to make an overview of the data, a PC model with 2 components. Plot the scores and loadings of this model. To make the observation names more informative (time), use positions 7-10 as label (right mouse button-click, *Properties|Label Types*, Start 7, Length 4). Interpret what you see and check your conclusions with additional tools. Hint: make contribution plots (scores mode), time series plots (*Analysis|Scores|Line plot*), and inspect your data table. Save the model and give it an informative name (for example “all data”). To give all models concise names is good practice.

Exclude the extreme observations and compute a new PC model. In this case it is justified to remove the deviating observations because there is a logical reason for doing so (no material feed). The resulting score plot will now be more interesting. Check the correlation structure of the variables.

Task 2

Now, we are going to create a PLS model incorporating all X- and Y-variables. Before any calculations can be made we have to select observations in which the design was made and therefore have response data. The observations that have missing data in the responses are marked with an “O” in the third column. The observations that have measured data in the responses are marked with an “S” in the third column. Make a selection with the “FIND” function (*Workset|New|Observations|Find*) searching for “O”-marked observations. Exclude the selected observations and use the remaining 85 ones.

- a) Define variables 1-12 as X and variables 13-18 as Y. Expand the design variables (1, 5, 6) with cross and square terms: (*Expand|Square and Cross*). Fit the PLS model. Check the cross-validation value (Q^2). Study the distribution of observations (*Analysis|Score|Scatter plots*). Which type of diagnostics are available and appropriate? Examine the correlation between the variables.
- b) What is the difference between loading plots and coefficient plots? Make sure that you understand the meaning of VIP. Check the correlations between the design variables and other measured variables.
- c) How should we optimise the system if we want to maximise the amount of iron and simultaneously minimise the amount of phosphorous? Illustrate this with contour plots and discuss the problems with this presentation technique.
- d) There is a clear indication that the amount of product and the quality of product are dependent on different X-variables. Do we have something to gain by making different models for the amount of product and quality of product, that is, one model for responses 13-15 and another for responses 16 and 17? Do not forget to expand the three designed X-variables. Compare the results with those from the model containing all response variables. Which conclusions can you draw?
- e) Response variable no 18 is very difficult to measure. Use the first PLS model and make predictions for no 18 for all observations. Which model parameter is important to check when making predictions of response variables? Illustrate the variation in Y18 with time. Hint: Use *Predictions|Time series*, select YPredPS and YVarPS as items for the variables.

Task 3

Use the same selection of observations as in the previous task. Choose the response PAR. Make a PLS model using only the design variables (var 1, 5 & 6) with expansions as X and compare with a model with the twelve original variables as X (design variables expanded). Compare R^2Y . Also, check what happens when you make contour plots with different models and discuss the outcome and problems with correlated predictors.

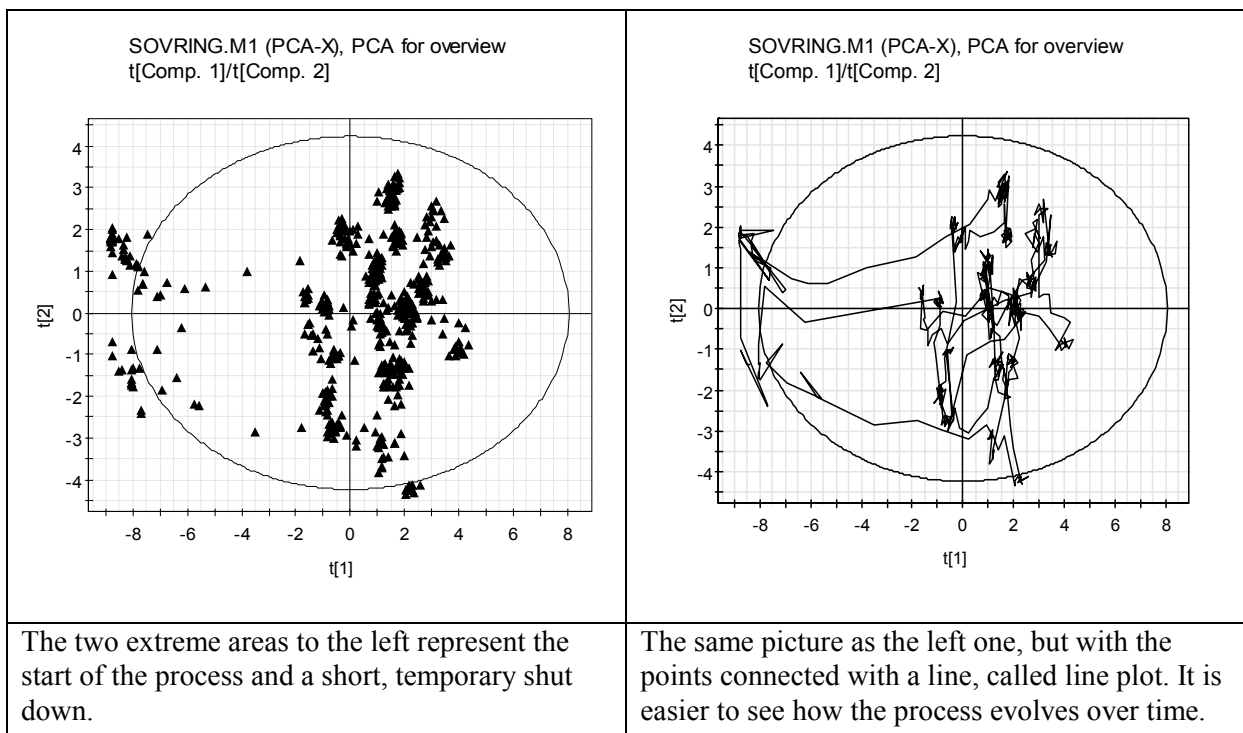
Solutions to SOVRING

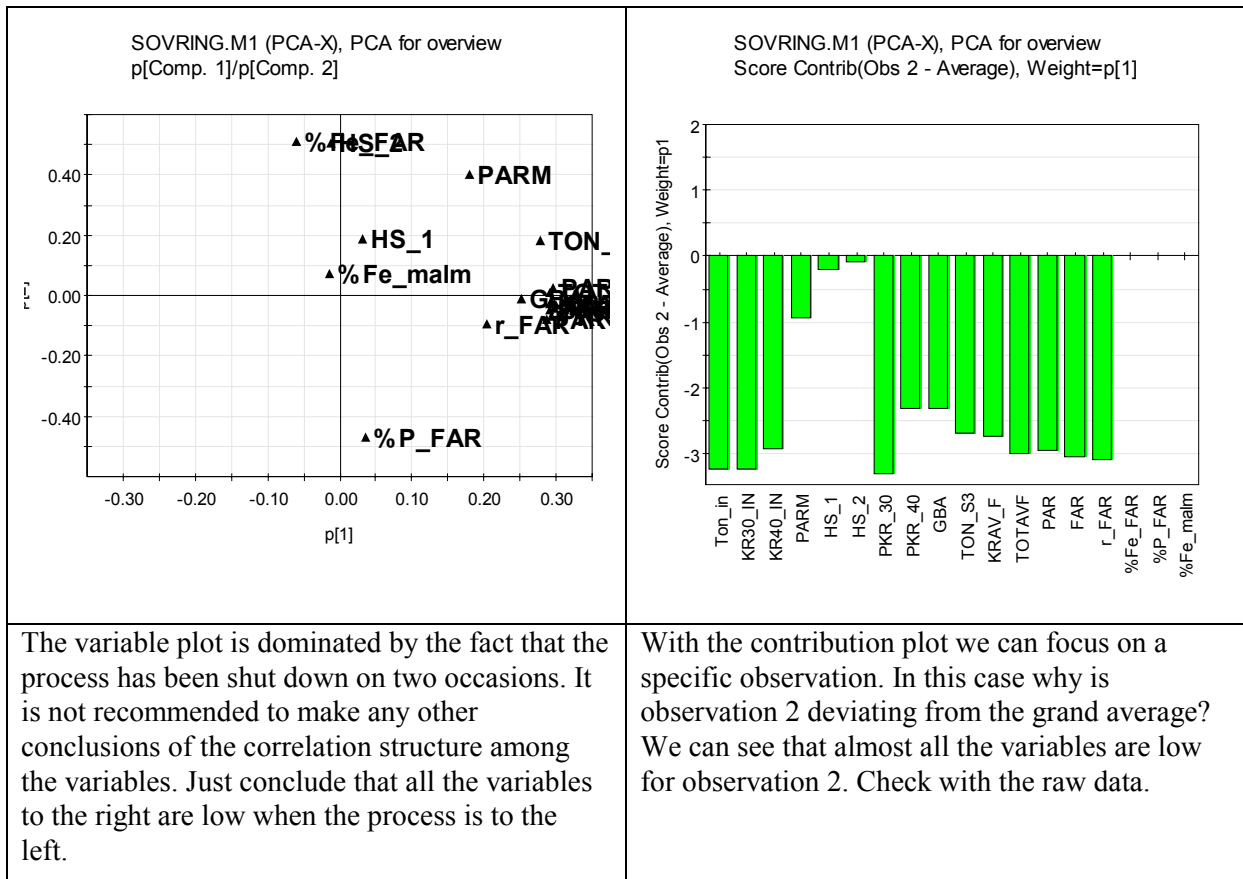
Task 1

A PC model with all variables as X was computed. The two first components explained 79% of the variability in the data.

PCA for overview								
Type: PCA-X Observations (N)=572, Variables (K)=18 (X=18, Y=0)								
Components:								
A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.							
1	0.663	0.663	11.9	0.642	0.0543	0.642	R1	7
2	0.129	0.791	2.32	0.168	0.0572	0.702	R1	22

In the score plot we see that the process has behaved differently on two occasions. We can also see that the process is moving between clusters due to different settings in the process parameters. The variable plot and the contribution plot tell us that the process feed rate is very low -- and that all parameters correlated to this phenomenon are low -- when we are positioned in the left-hand part of the score plot.





Task 2a

With variables 1-12 as X and variables 13-18 as Y, and variables 1, 5, and 6 expanded with cross and square terms, you will produce a PLS model with 6 components.

SOVRING - M2

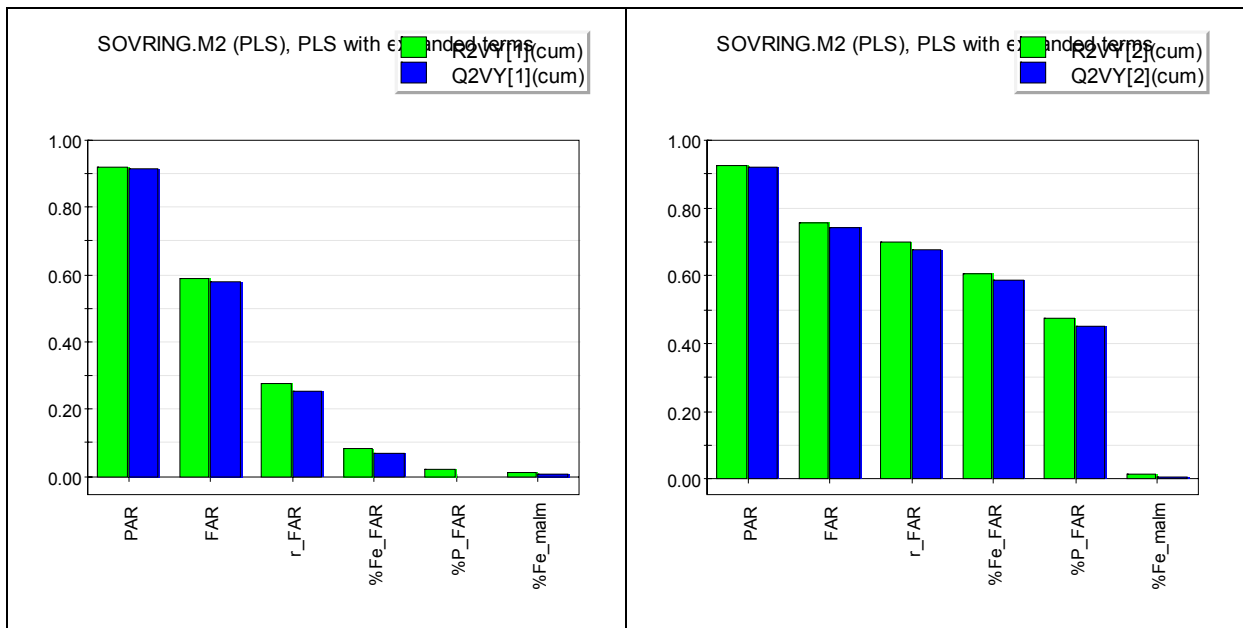
Workset... Options... Title: PLS with expanded terms

Type: PLS Observations (N)=85, Variables (K)=24 (X=18, Y=6), Expanded=6

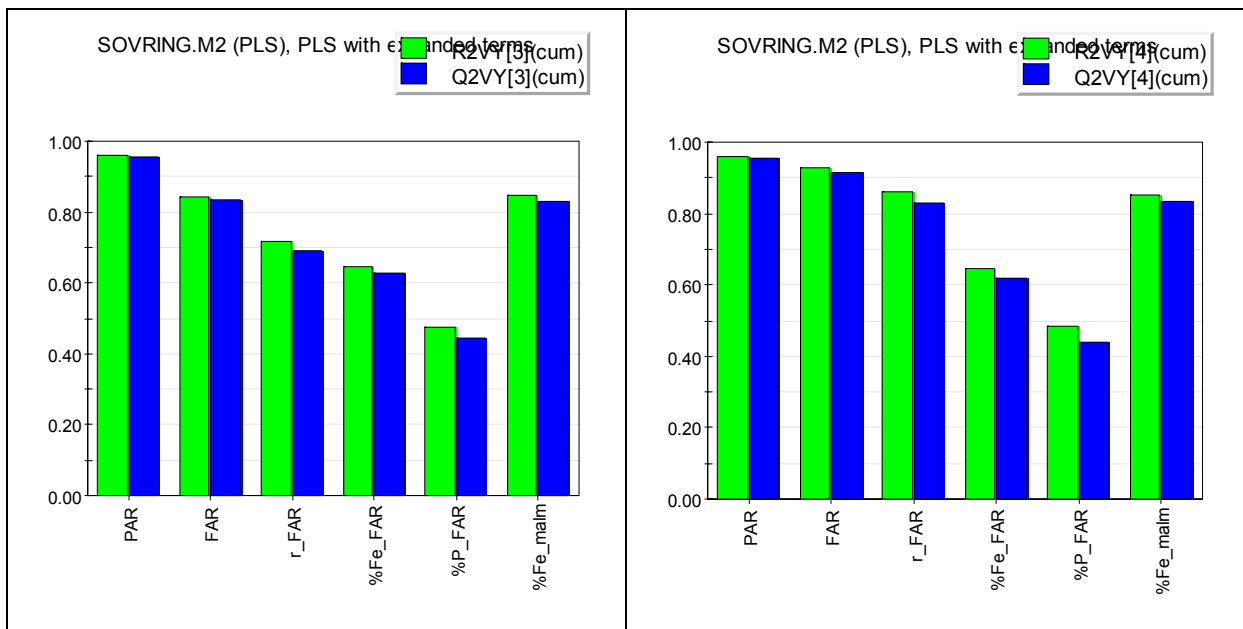
Components:

A	R2X	R2X(cum)	Eigenv...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signifi...	Ite...
0	Cent.			Cent.						
1	0.393	0.393	7.08	0.317	0.317	0.303	0.05	0.303	R1	9
2	0.131	0.525	2.36	0.263	0.58	0.373	0.05	0.563	R1	19
3	0.0914	0.616	1.64	0.168	0.748	0.387	0.05	0.732	R1	7
4	0.0641	0.68	1.15	0.0409	0.789	0.132	0.05	0.767	R1	29
5	0.0558	0.736	1.01	0.0331	0.822	0.114	0.05	0.794	R1	10
6	0.0395	0.776	0.711	0.0167	0.839	0.0578	0.05	0.806	R1	15

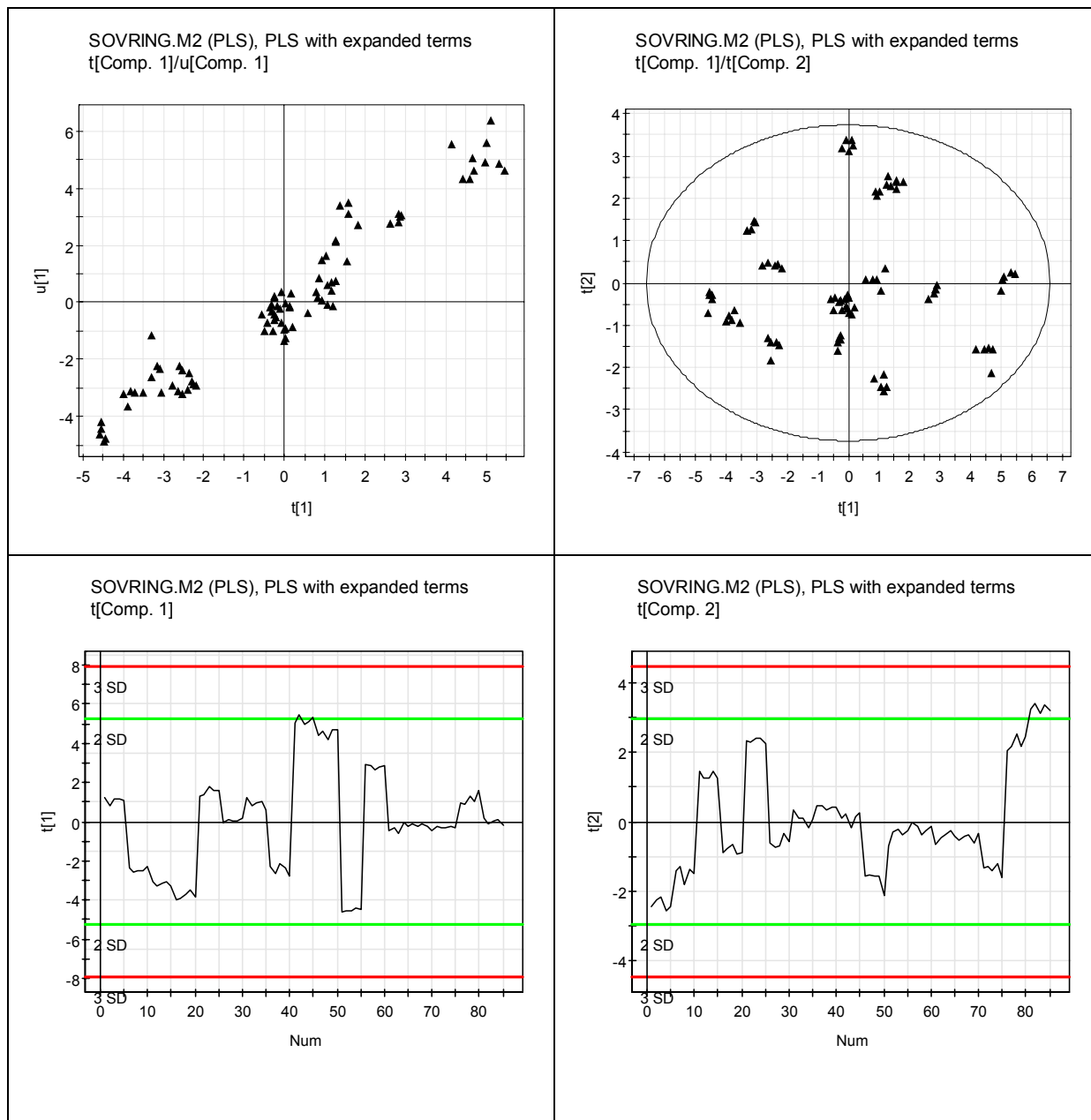
The two first components are clearly the most important. The graphs shown below are created under *Analysis|Summary|X/Y overview|Plot*. To change the component number, right click in the plot and select *Properties* in the menu.



According to the component contribution plot, the variability in PAR is explained by the first component, whereas FAR and r_FAR need one extra component. The chemical information is modelled by components 2-6. We may conclude (see $w \cdot c_1 / w \cdot c_2$ plot) that the total feed and the quality are affected by different factors in the process. One important and difficult parameter to measure is %Fe in to the process. It is modelled well by the 3rd component.

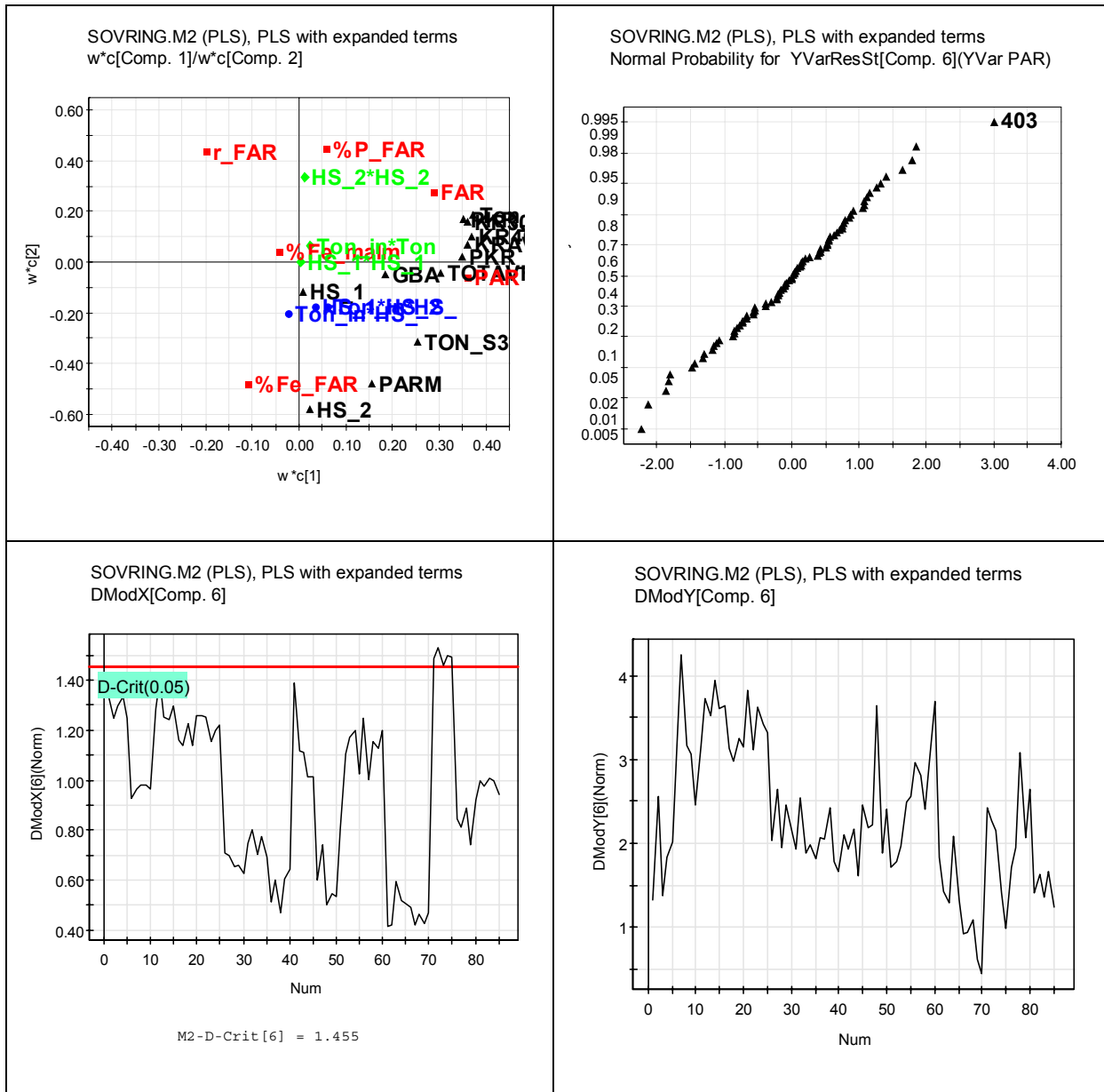


The t_1/u_1 score plot shows an obvious strong correlation between X and Y. In the t_1 vs t_2 plot we can see that the process is stationary at the different design points. The stability for different settings of the process parameters is also clearly shown in the Num vs t-plots. There are no deviating observations.



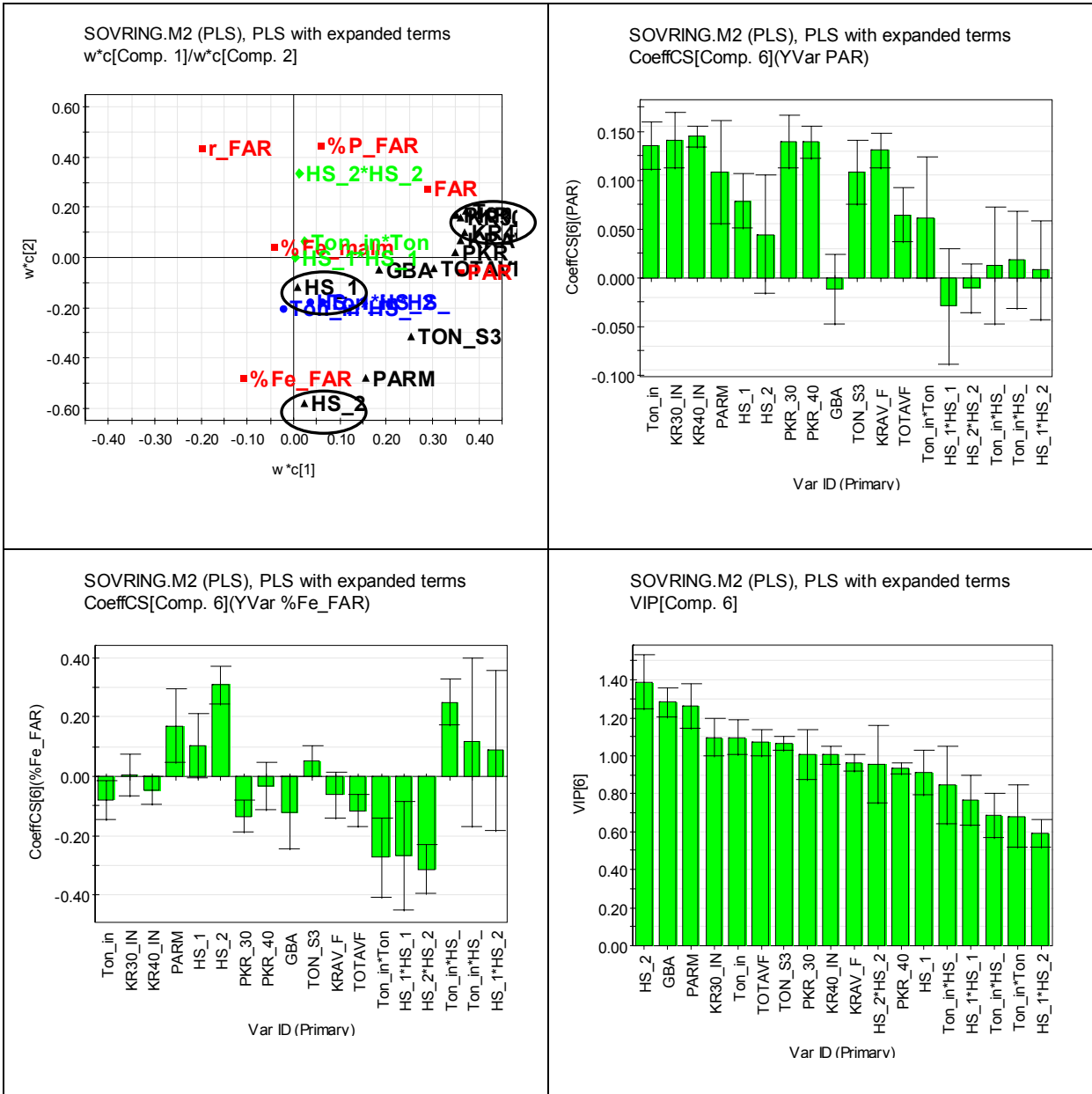
The controlled variable Ton_in (total feed) correlates with many of the measured variables. The interesting point is that the important quality responses are correlated to other variables than the feed variables. For example %P_Far is orthogonal to total feed and is mainly connected to HS_2, which is another controlled variable.

Diagnostics of the residuals is an important issue. In the N-Plot you can check if the residuals are normally distributed. The X-axis represents the standard deviation of the residuals. All observations are within $\pm 3SD$ and lie on a straight line for the response PAR. It might be interesting to check observation 403. Do not forget the tools Distance to Model for X and Y-blocks. What is the difference between all these plots?



Task 2b

Even when there are many components in the model, the first loading plot is usually the most interesting. It will show the main correlation structure between the variables and how all responses are connected (the design variables are ringed). The coefficient plot is very useful, especially in models with many components. This example is a typical illustration of when we really need the coefficients to understand the influence on all the responses. The loading plot together with the coefficients is a good variable map. VIP gives a normalised order of importance of all X variables in the total model.



Task 2c

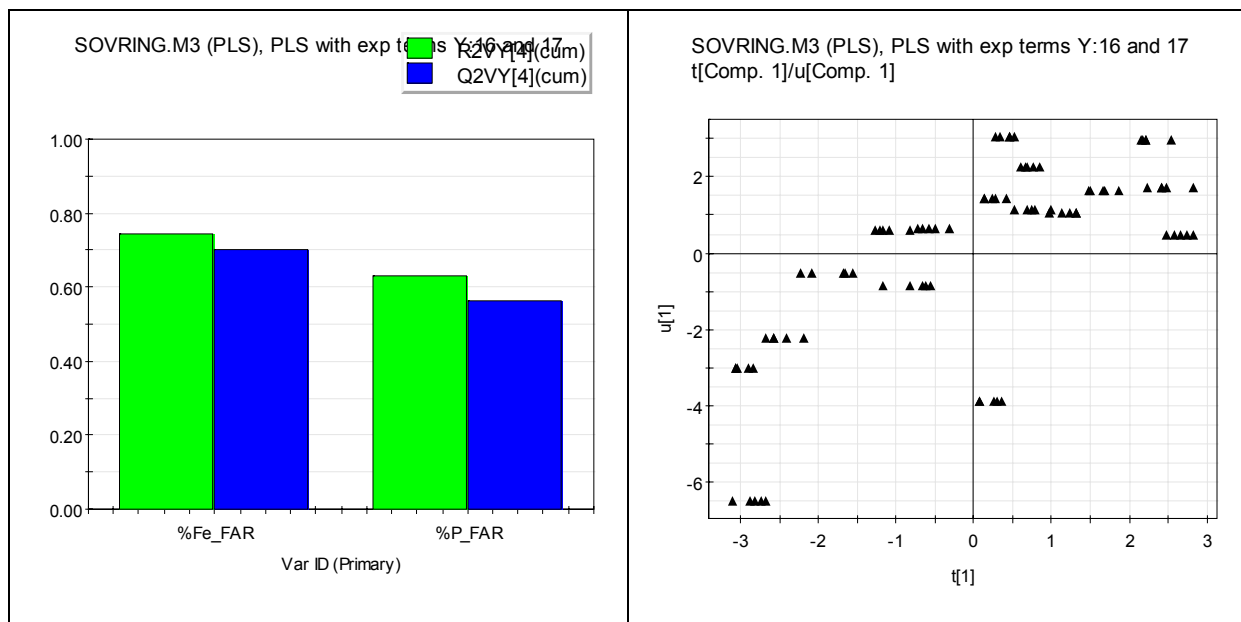
So far we expect that high production and a minimum of P with maximum of Fe in the product, are achievable goals. How well we are able to fulfil these goals is not so simple to understand from only the model. A convenient way of illustrating this is with contour plots of the responses. The problem in the multivariate case is that a contour plot is dependent on two orthogonal axes and will have to set the other X variables at a constant level. There is a risk that this could be incorrect due to correlations between the X variables. When one X variable is varied others will go together with it and cannot therefore be considered as constant. To illustrate this we will make different models later to check the influence of this problem.

Task 2d

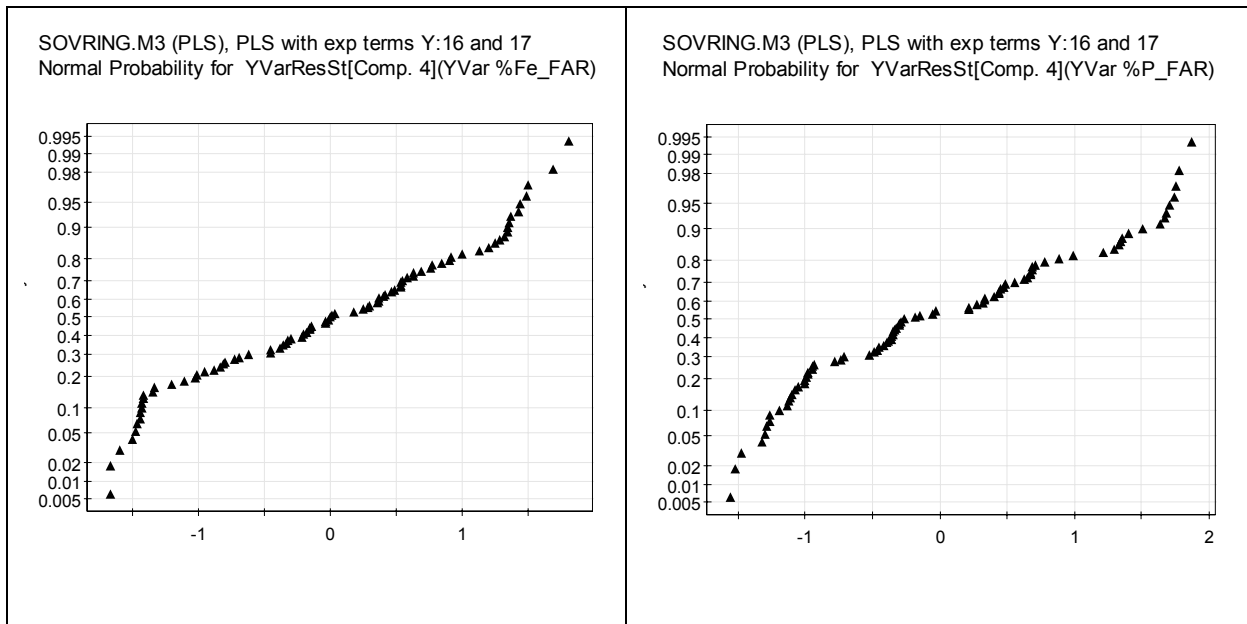
A PLS model was computed with X-block 1-12 and Y-block 16-17 (product quality). This resulted in a model with only four components.

SOVRING - M3											
Workset...		Options...		Title PLS with exp terms Y:16 and 17							
Type: PLS Observations (N)=85, Variables (K)=20 (X=18, Y=2), Expanded=6											
Components:											
A	R2X	R2X(cum)	Eigenv...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signifi...	Ite...	
0	Cent.			Cent.							
1	0.242	0.242	4.35	0.464	0.464	0.437	0.05	0.437	R1	5	
2	0.279	0.52	5.02	0.129	0.593	0.225	0.05	0.564	R1	5	
3	0.0624	0.583	1.12	0.0658	0.659	0.0986	0.05	0.607	R1	13	
4	0.0688	0.652	1.24	0.0285	0.687	0.0646	0.05	0.632	R1	12	

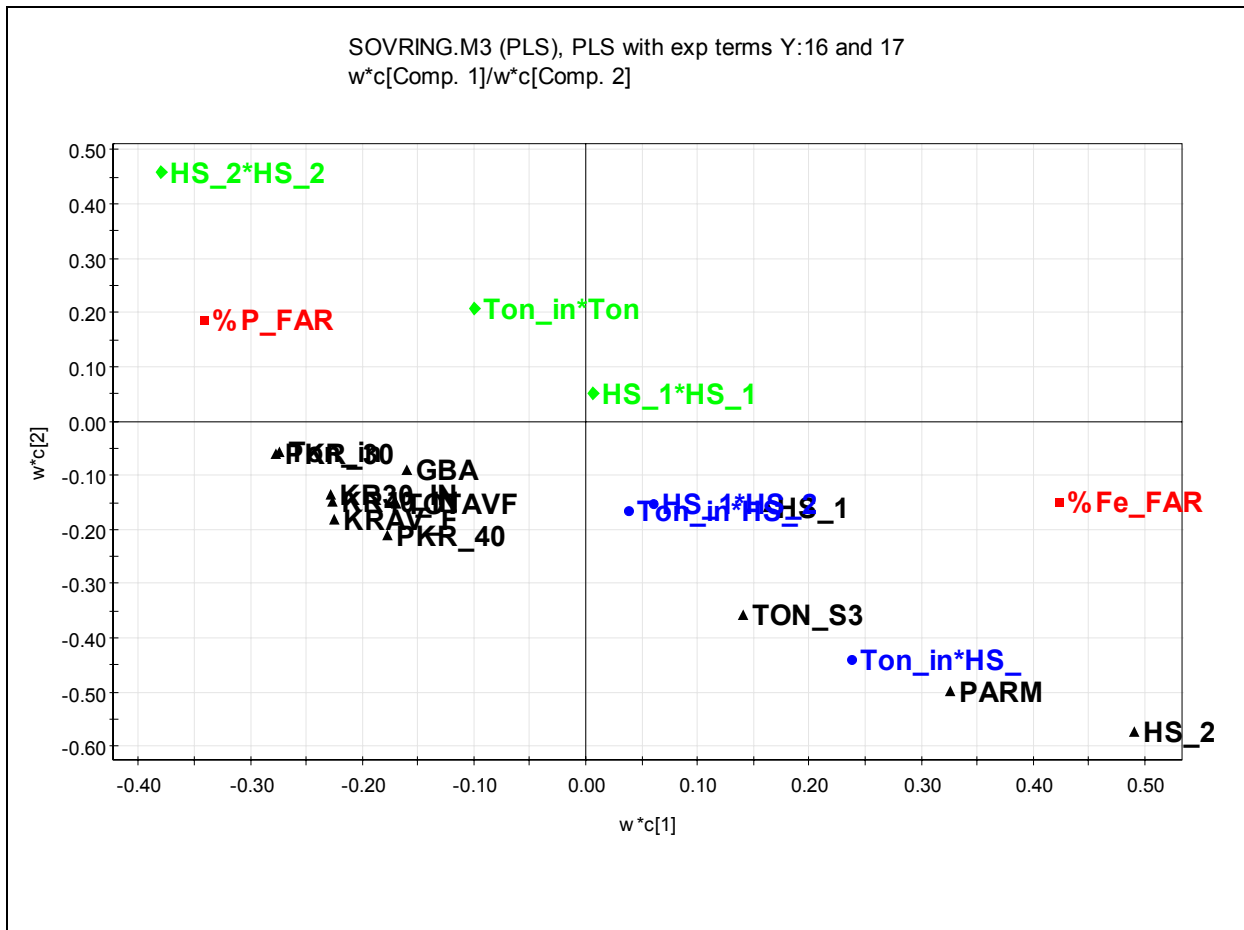
This model is similar to the overall model with all Y-data.



In the t_1/u_1 score plot all observations with the same settings for X are close to each other. Below we see that there are no residual distribution problems for this model.

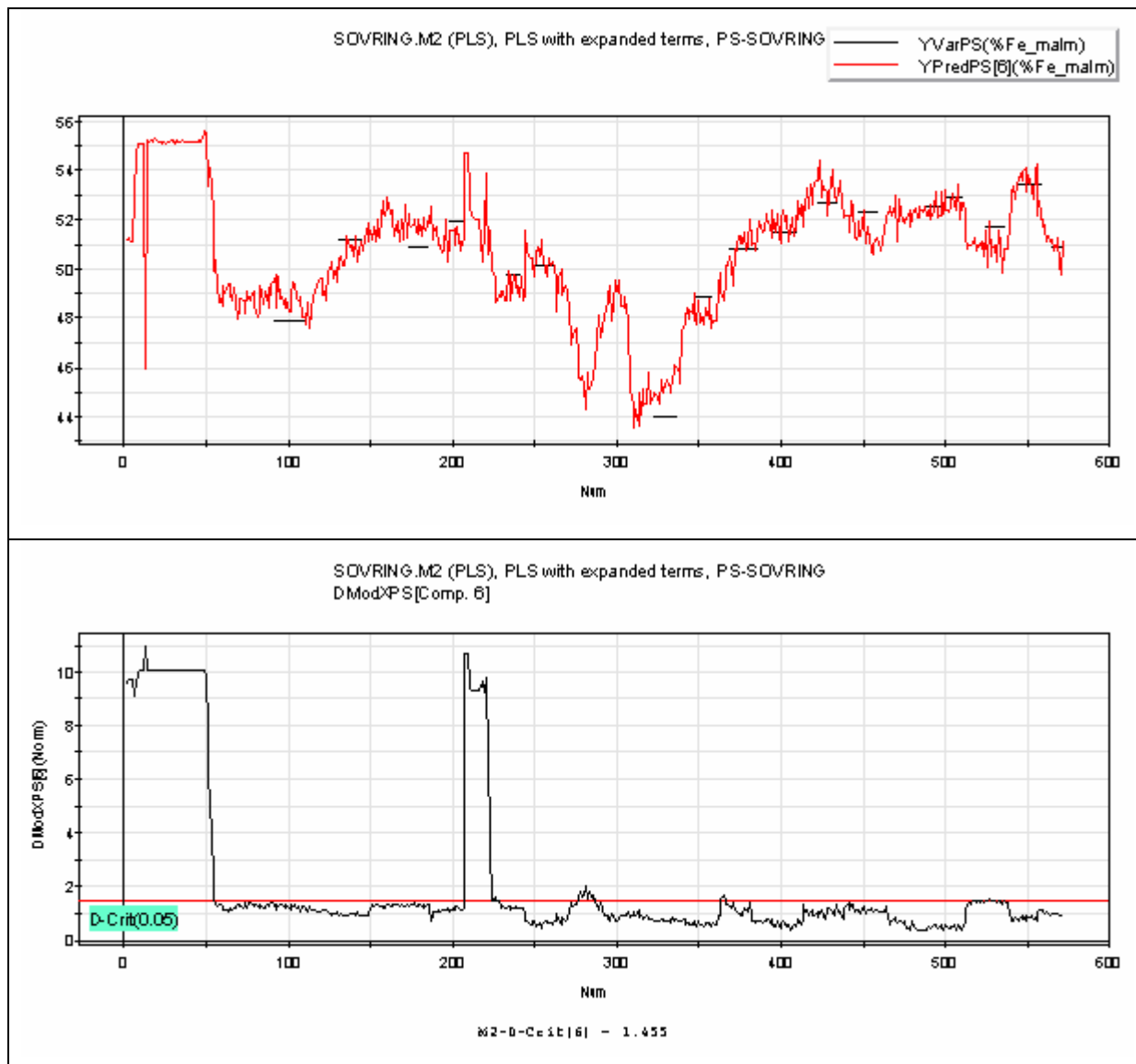


In the loading plot we can see that %Fe_FAR and %P_FAR are negatively correlated, as expected. Before you study the loading plots and the coefficient plots examine the t_1/t_2 score plot and the distance to model plots to check for unexpected groupings and outliers.



Task 2e

In the prediction graph below we can see the small errors in the predictions. Moreover, the DModX plot shows where the model is unreliable.



Task 3

The first model uses only the designed X variables; Ton_in, HS_1 & HS_2, with cross and square term expansions.

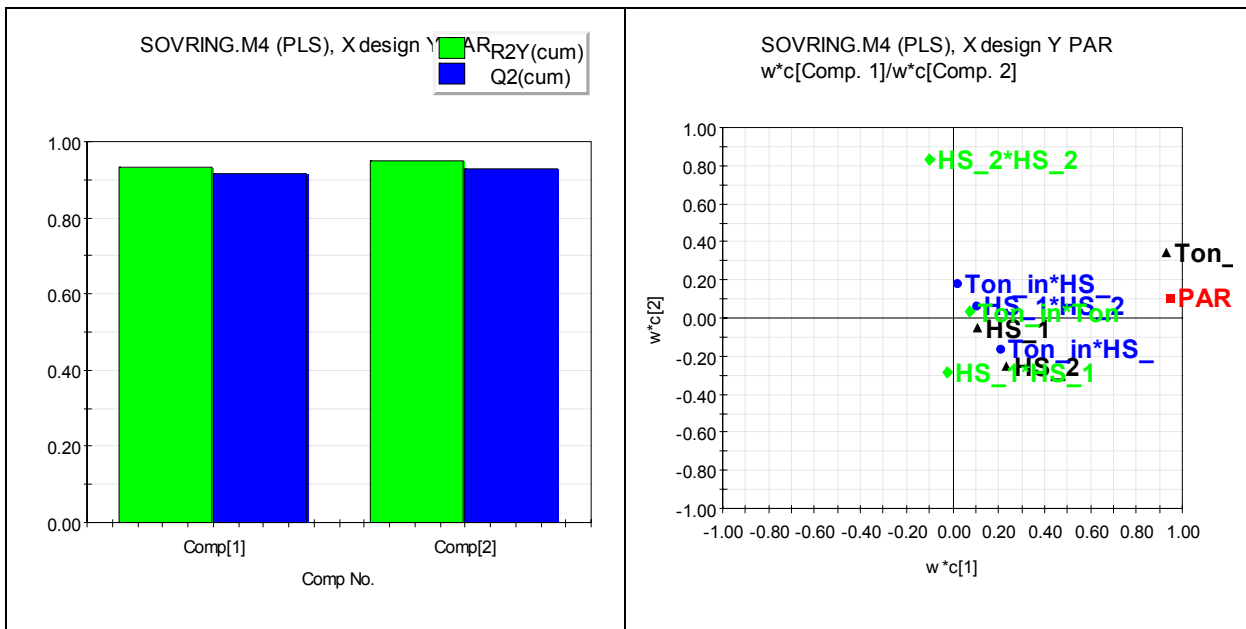
SOVRING - M4

Workset... Options... Title X design Y PAR

Type: PLS Observations (N)=85, Variables (K)=10 (X=9, Y=1), Expanded=6

Components:

A	R2X	R2X(cum)	Eigenv...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signifi...	Ite...
0	Cent.			Cent.						
1	0.119	0.119	1.07	0.932	0.932	0.916	0.05	0.916	R1	1
2	0.168	0.287	1.51	0.0176	0.95	0.171	0.05	0.93	R1	1



The second model contains all X variables with expansion in the design variables.

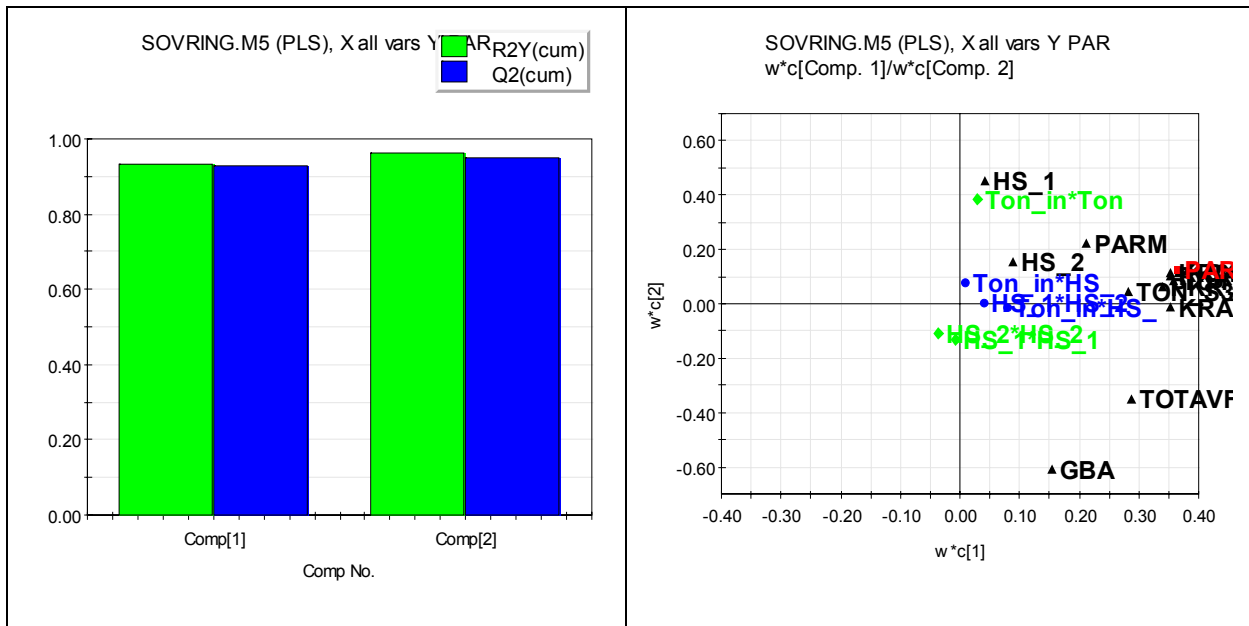
SOVRING - M5

Workset... Options... Title X all vars Y PAR

Type: PLS Observations (N)=85, Variables (K)=19 (X=18, Y=1), Expanded=6

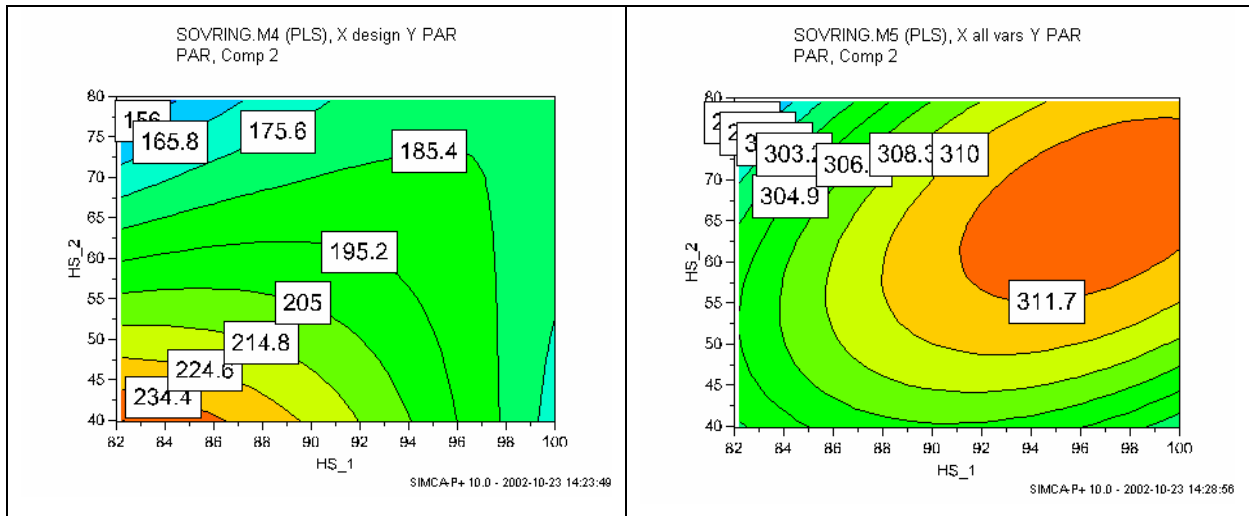
Components:

A	R2X	R2X(cum)	Eigenv...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signifi...	Ite...
0	Cent.			Cent.						
1	0.393	0.393	7.08	0.934	0.934	0.93	0.05	0.93	R1	1
2	0.0971	0.491	1.75	0.0267	0.961	0.306	0.05	0.951	R1	1

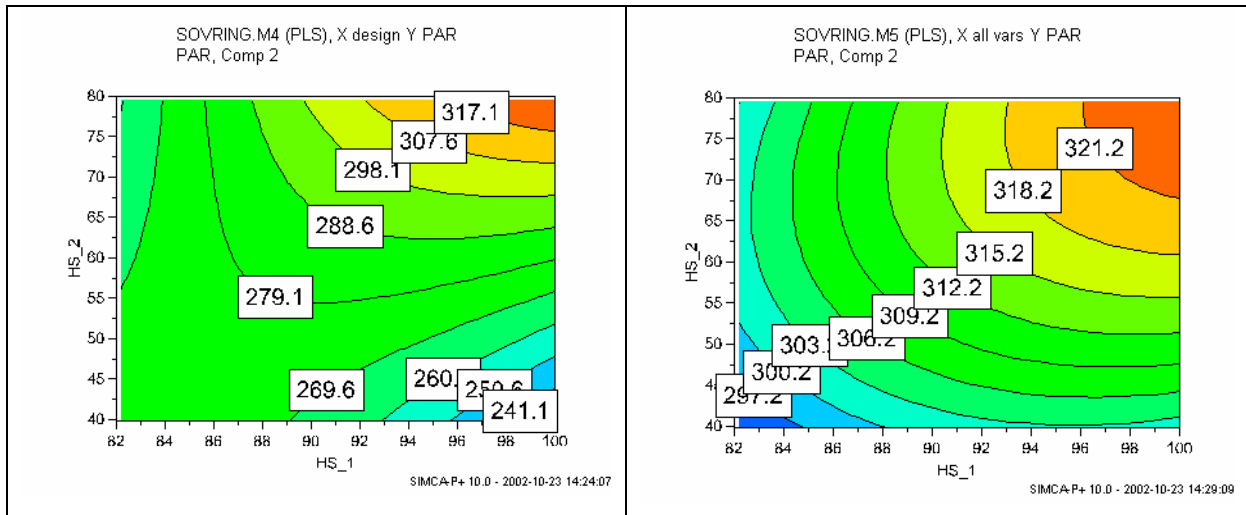


Both models are very similar in the explanation of Y. The contour plots are, however, very different with regard to both shape and level. The correct plots are the ones on the left. These are constructed from only the design variables. The other contour plots are misleading because all variables (except the two in the plot) have to be kept constant, which is unrealistic due to the strong correlations among the X-variables.

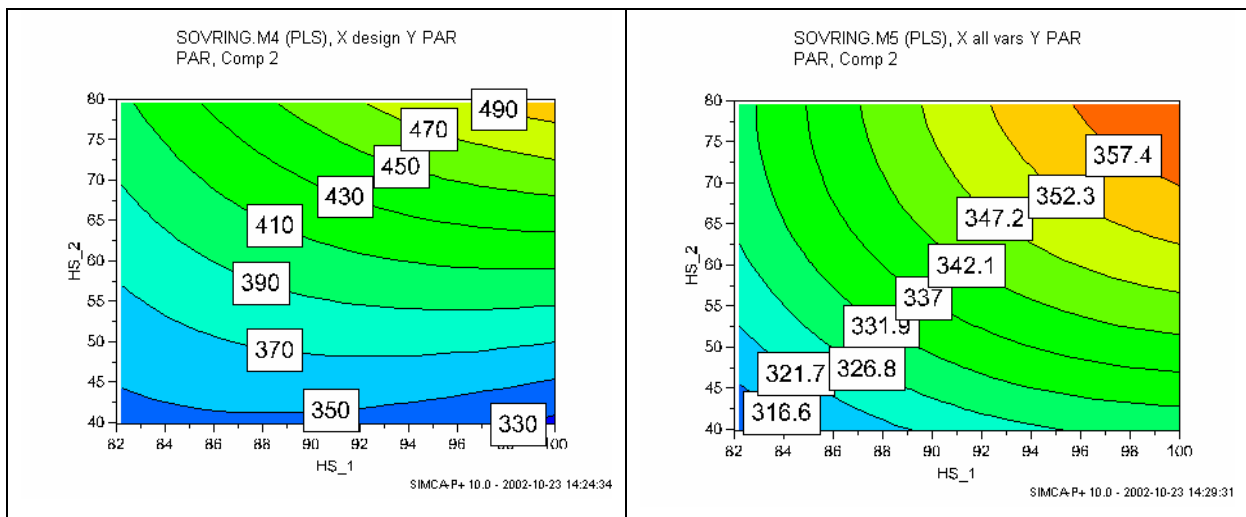
Ton_in set to its low value (845 t/h):



Ton_in set to its centre value (1252 t/h)



Ton_in set to its high value (1658 t/h)



Conclusions

This example shows that statistical experimental design in the dominating process variables gives data with high quality that can be used to develop good predictive process models. By initiating the data analysis with PCA, it was possible to discover two periods corresponding to no or low feed of starting material. These were excluded in the PLS analysis. The PLS model founded on the set of 85 representative samples, enabled excellent predictions of PAR and FAR. The prediction quality was a little lower for the other responses. However, prediction of iron content in incoming ore can be accomplished with reasonable confidence.

MVDA-Exercise PROC1A

Fault Detection using Control Charts

Background

This example deals with a chemical production plant manufacturing a polymer. It is a continuous process which went out of control at around time point 80 after a fairly successful campaign to decrease the side product (y6).

Objective

The manufacturing objective was to minimise the yield of side product (y6) and maximise product strength (y8). The data analysis objective of this exercise is to investigate whether MSPC could have detected the process upset earlier and thus prevented the shutdown.

Data

The data set contains 33 variables and 92 hourly observations. The measured variables are distributed as seven controlled process variables (x1in – x7in), 18 intermediate process variables (x8md – xpen), and eight output variables (y1 – y8). All the data are coded so as to not reveal any proprietary information.

Tasks

Task 1

Create a new project by reading in the spreadsheet PROC1A.xls. Use *Plot/Lists|Time Series Plot* to look at the raw data in blocks of 8-9 variables. Can you spot anything unusual about the latter observations?

Task 2

Make a new WorkSet. Exclude observations 70-92 and compute a PC model to overview the data.

Task 3

Use *Predictions|T Predicted* or *Plot/List/Line Plots* to generate a T Predicted scatter plot of all the observations. Confirm that observations 80-92 fall outside the tolerance region.

Task 4

Use *Predictions|Contribution|Scores/T2* to find how observation 80 differs from the average observation (use pp 1,2 weighting). Which are the most deviating variables at this time point? Use *Predictions|Contribution|Distance to Model X-block* to investigate why observation 33 is a moderate outlier.

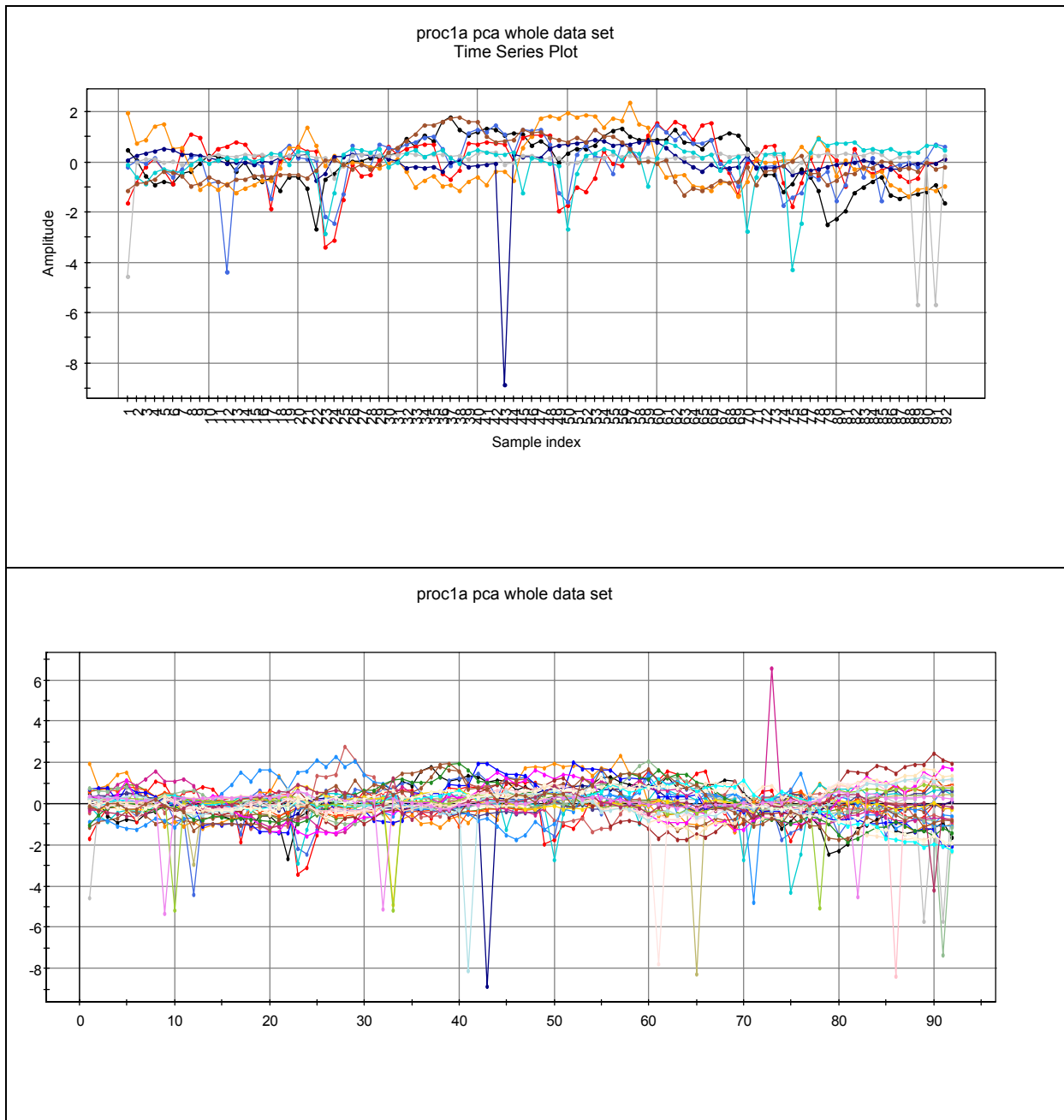
Task 5

Make a Shewhart control chart (*Plot/List/Control Charts*) of T^2 (the combined scores). Experiment with other control charts as well. Make CuSum plots for t_1 and t_2 and try to interpret them. Make combined Shewhart/EWMA plots with low lambda (long memory), and high lambda (short memory) and observe the effect of changing lambda.

Solutions to PROC1A

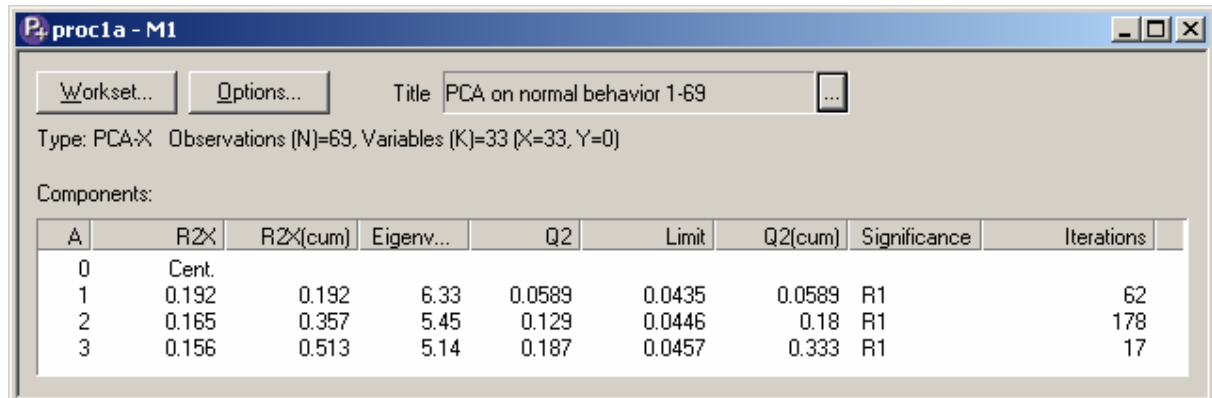
Task 1

Below are time series plots of the first 8 variables and the complete set of 33 signals. It is not easy to spot any deviations apart from the spikes although it is known that the process went out of control at time point 80 and eventually had to be shut down.



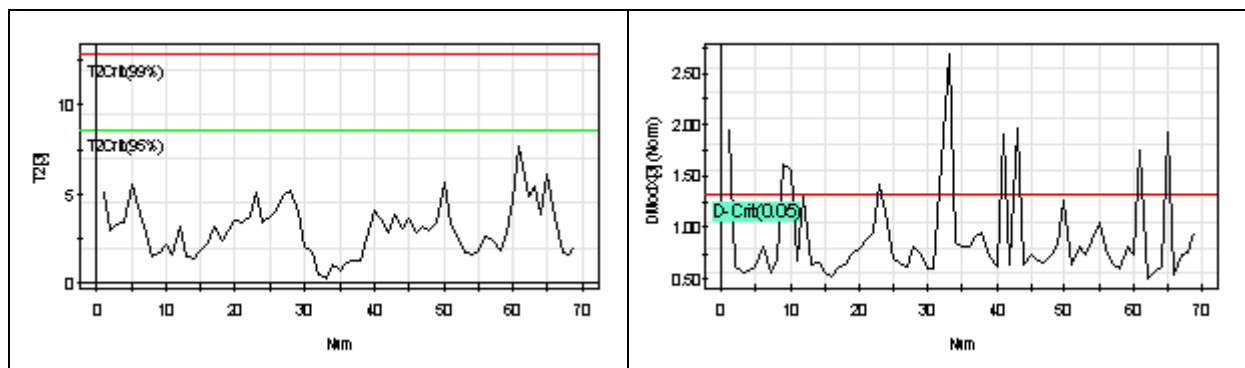
Task 2

A PC-model was fitted to the first 69 samples. Eight components were obtained using cross-validation, but we will use only the first three which explain 51% of the variation.



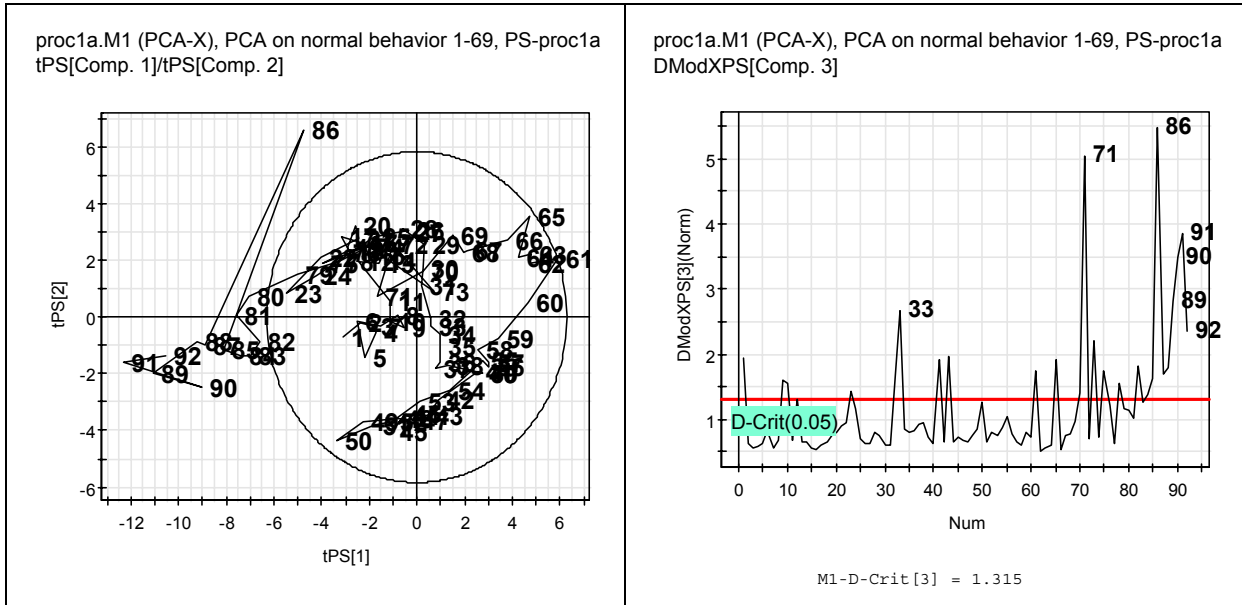
A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.							
1	0.192	0.192	6.33	0.0589	0.0435	0.0589	R1	62
2	0.165	0.357	5.45	0.129	0.0446	0.18	R1	178
3	0.156	0.513	5.14	0.187	0.0457	0.333	R1	17

Now, we are going to use the historical data to define control limits within which the process is in control. One type of control limit is provided by Hotelling's T^2 which models the deviation of the process within the PC-model hyperplane. The first 69 observations fall inside the 95% tolerance limit (below, left). A second type of control chart is provided by DModX (below, right) which monitors the deviation of the process from the PC-model hyperplane. There are several samples which exceed the critical limit of DModX. Usually, deviating observations in DModX indicate minor process upsets not captured by the model. However, if many observations in a sequence show consistently high DModX-values, then that is normally an indication of a new process event.



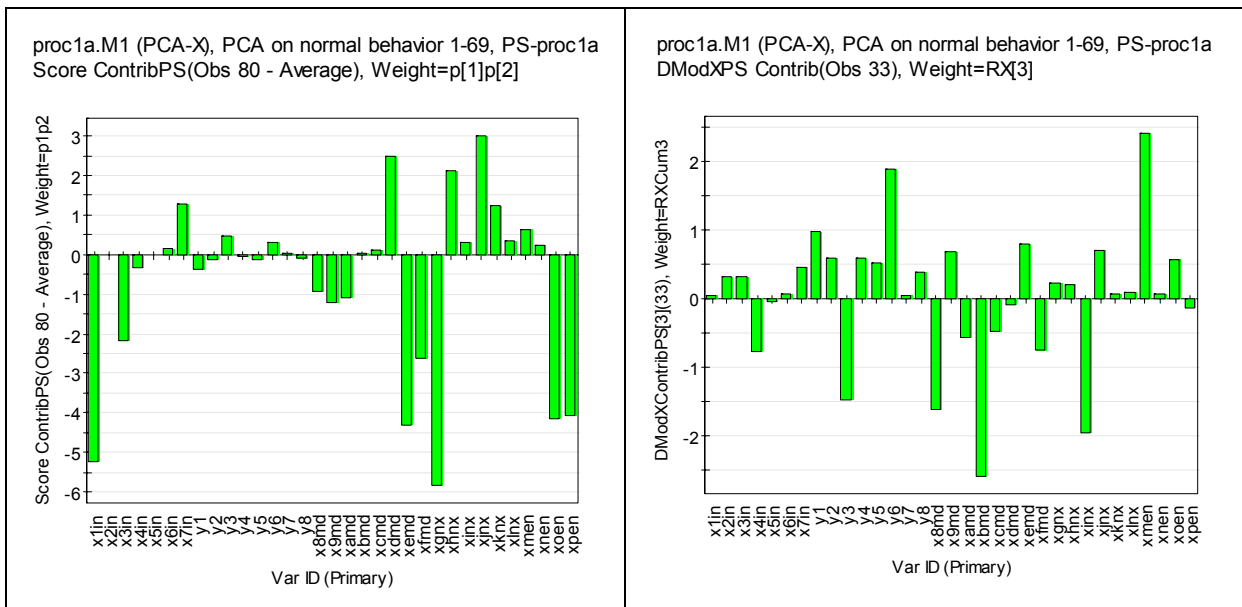
Task 3

The model obtained was applied to the remaining 23 observations (70-92). The score plot indicates that the last 13 observations fall outside the tolerance region, suggesting that a new process behaviour has occurred. Actually, the shift in process operating conditions is seen already from time point 70. It might also be worth inspecting observation 33 more closely, as it displays the highest DModX among the training set observations.



Task 4

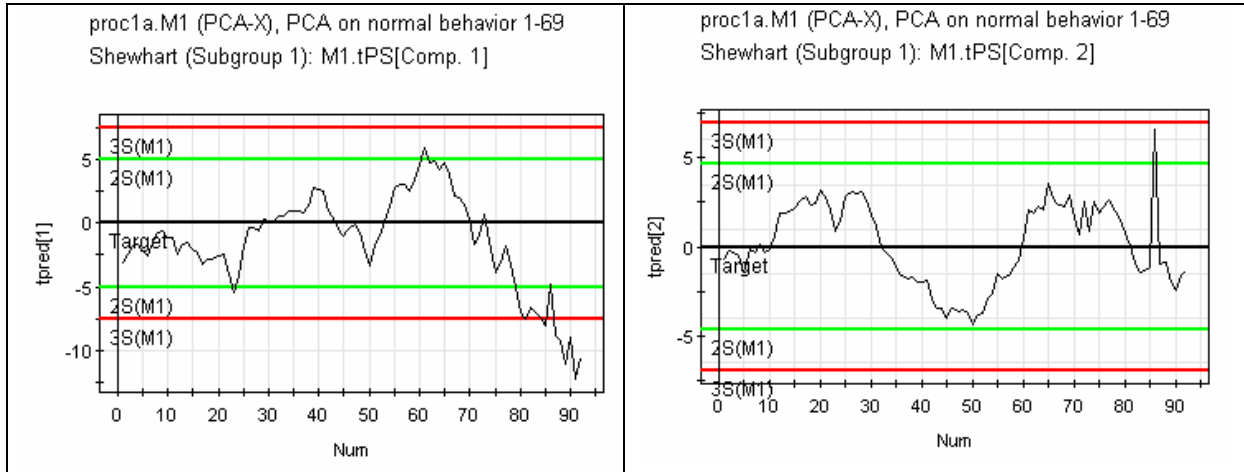
The contribution plot for observation 80 is shown below and highlights which variables are contributing to the difference between observation 80 and the average process point. Evidently, compared with the average process point, observation 80 has significantly lower values for variables x1in, xemd, xgnx, xoex, and xpen.



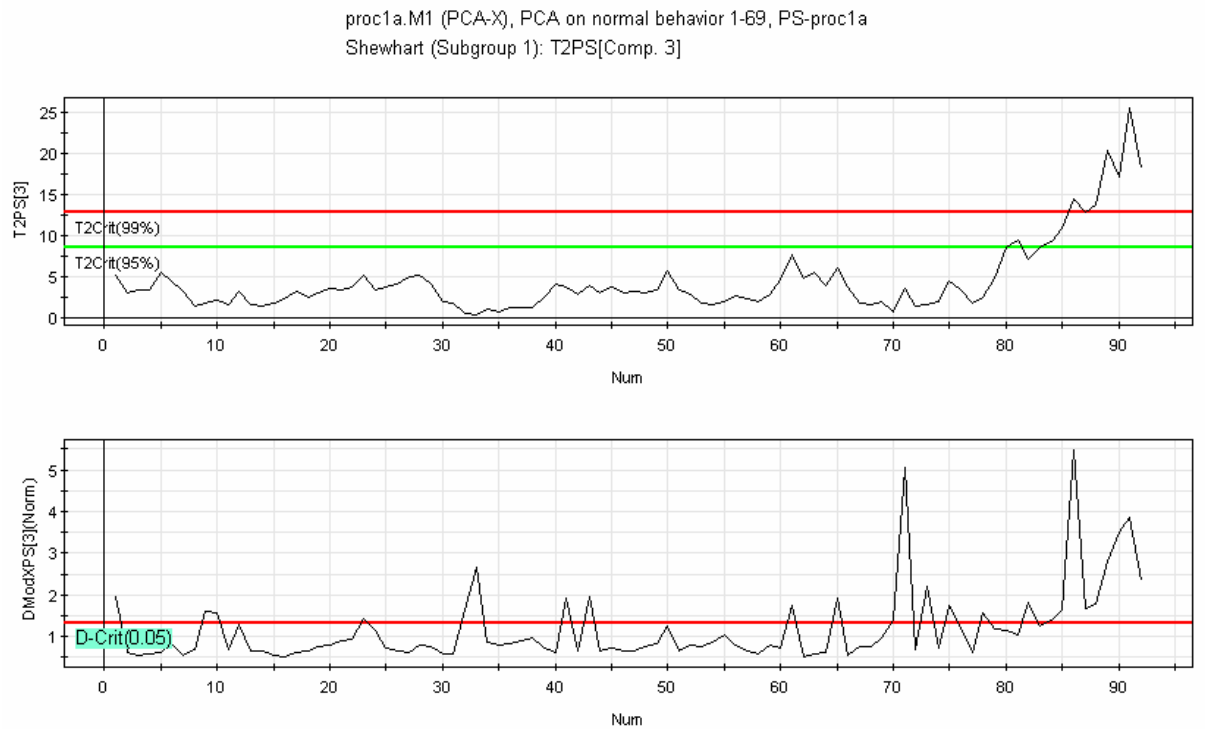
Furthermore, moderate outliers can be inspected in a contribution plot of DModX. As an example, consider the contribution plot of observation 33. It is mainly the two variables xemd and xmen which cause the large residual for this observation.

Task 5

The graphs below represent time series plots of the first two predicted scores (t_1 and t_2). The process drifts significantly in both scores until the end of the campaign when mainly negative values for both scores are apparent.



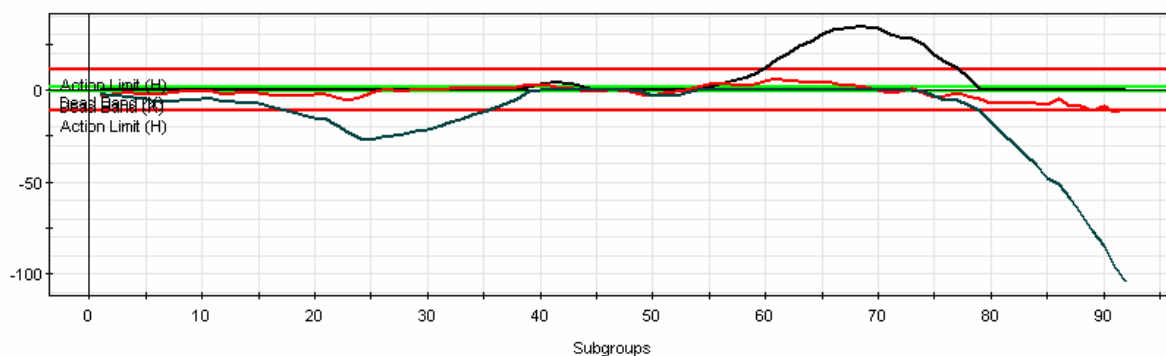
The T2-control chart below shows the problem at the end of the sampling campaign. We can also see how a number of spikes in the residual data are detected using DModX.



S(M1) = 1.465 Target(M1) = 3 M1-D-Crit[3] = 1.315 SIMCAP+ 10.0 - 2002-10-24 08:49:53

The two CuSum charts below give early warning of problems. The CuSum chart is designed to detect shifts in the mean and that is what happens as the process target values are changed during the optimisation campaign.

proc1a.M1 (PCA-X), PCA on normal behavior 1-69, PS-proc1a
CUSUM (Subgroup 1): tPS[Comp. 1]



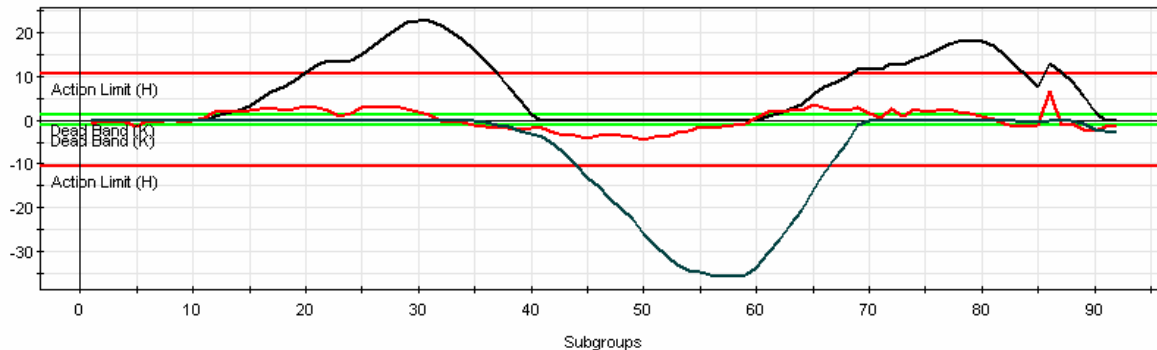
S (M1) = 2.516

Target (M1) = 0

Action Limit (H) = 11.32 Dead Band (K) = 1.258

SIMCAP+ 10.0 - 2002-10-24 08:51:42

proc1a.M1 (PCA-X), PCA on normal behavior 1-69, PS-proc1a
CUSUM (Subgroup 1): tPS[Comp. 2]



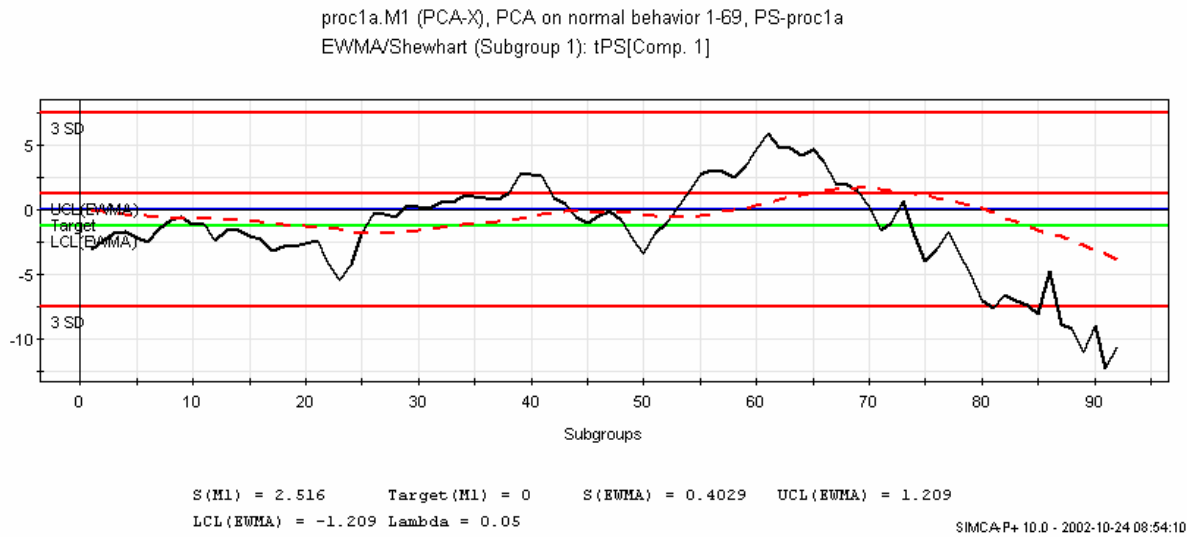
S (M1) = 2.334

Target (M1) = 0

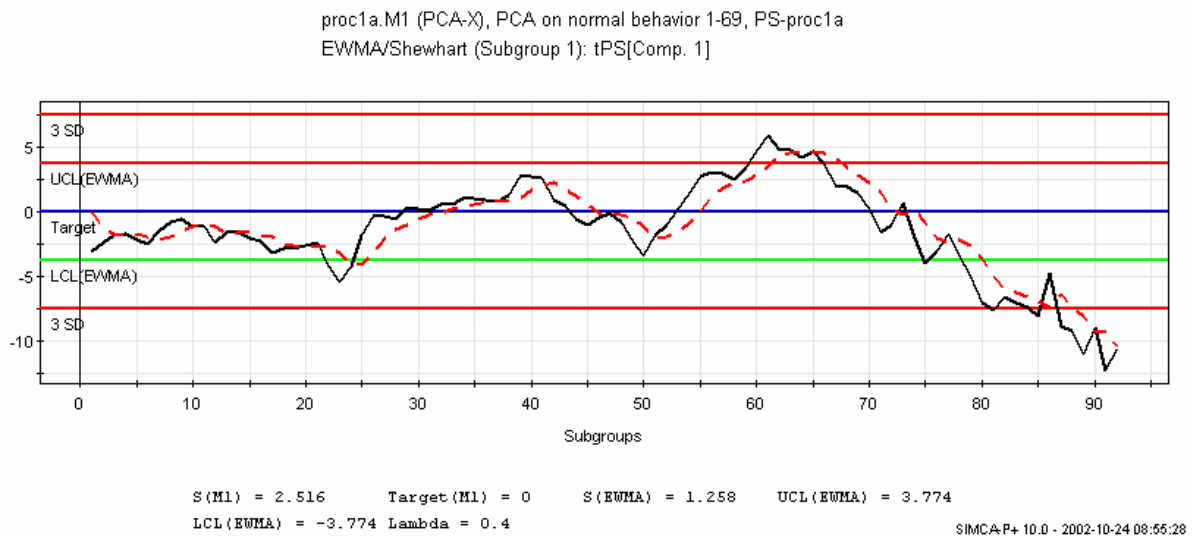
Action Limit (H) = 10.5 Dead Band (K) = 1.167

SIMCAP+ 10.0 - 2002-10-24 08:53:00

With low lambda (0.05) the EWMA graph is only slowly affected by the changes in t_1 and a smooth curve is generated.



As lambda increases (in plot below = 0.4) the EWMA graph more closely resembles the original Shewhart plot.



A value of $lambda = 0.2$ is often used when EWMA is used for forecasting, providing a good balance between memory and adaptability.

Conclusions

Multivariate modelling provides the means to monitor and supervise this process. The score plot of t_1/t_2 elegantly shows how the process moves around in space before finally going out of control. In practice, inspection of control charts of T^2 and $DModX$ would have provided early indications of the process problems.

MVDA-Exercise Baker's Yeast

Batch modelling (BSPC) of a Baker's yeast process

Background

The production of Baker's yeast takes about 14 hours for the final production stage. In this data set, each batch showed variability due to changes in the type of molasses used, temperature, pH, etc. Multivariate data analysis was the preferred choice to overview the data. We thank Jästbolaget AB for the data set.

Objective

The objective of this study was to determine if the yeast manufacturing process could be monitored more efficiently than by examining each individual variable. It was also interesting to establish predictive models for the final quality of the yeast.

Data

There are 33 batches, of which 20 were selected as reference batches. All batches have the same length, 83 observations. Altogether there are $33 \cdot 83 = 2739$ observations.

Observation names ("IDs"):

- Pos1 "r" for reference observations, "t" for test observations
- Pos2 Batch identifier
- Pos3,4 Observation order number in the batch (1-83)

Variable names (for observation level):

- Ethanol
- Temp Temperature
- Molasses Feed of molasses
- NH₃ NH₃ feed
- Air Air flow
- Level Level in tank
- pH

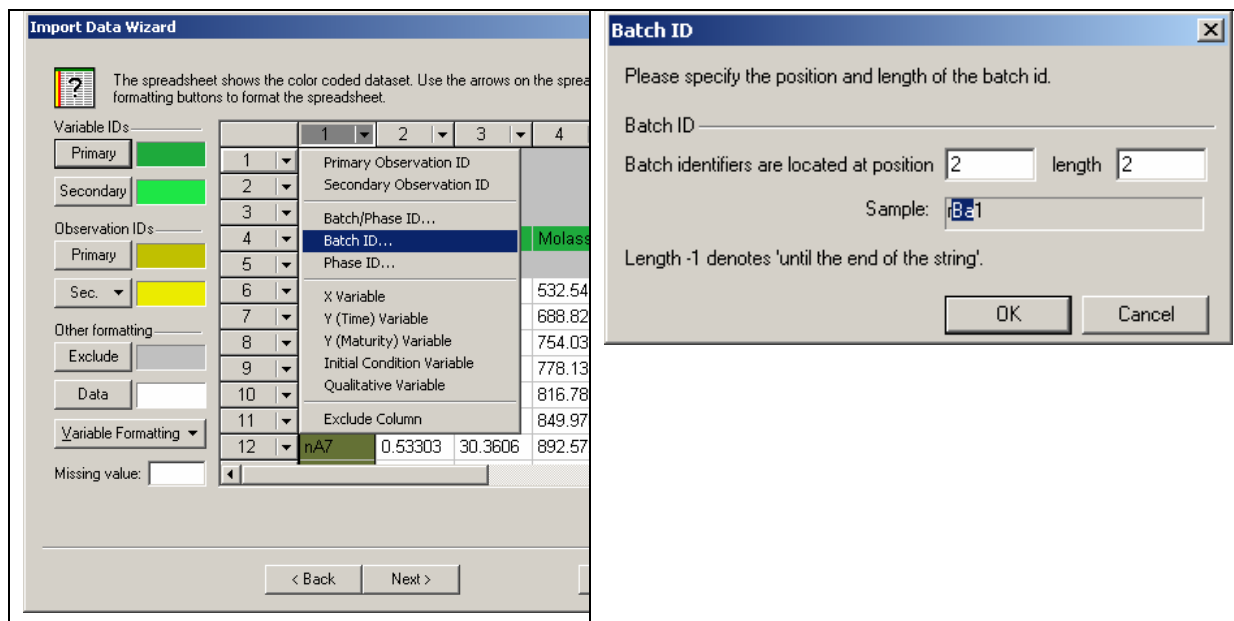
Variable names (for batch level):

- X: Innoc Total amount of dry substance added at start (in kg)
- Y: QP1 Quality Parameter 1 (to be high)
- Y: QP2 Quality Parameter 2 (to be high)
- Y: Amount Total amount of product (yeast) at batch termination (in kg)
- Y: Yield Amount of yeast (corrected for amount of molasses used)

Tasks

Task 1 (Observation level)

Make a batch project in SIMCA. Find the file *Bakers Yeast Primary.XLS*. In the *Import Data Wizard* dialog check the box “*SIMCA-P Batch Project*”. Press Next. Then mark the first column as *Batch ID*. The batch identifier is in position 2 and has length 2. Press OK. This will create a new column in the data-set. There are no phase identifiers in the batches.



Now mark the second column (which used to be the first column) and set it as *Secondary Observation ID*. Also mark the five right-most columns (Innoc, QP1, QP2, Amount & Yield) as *Initial Condition Variables* (they will not be used on the observation level). Press *Next*, *Next* and *Finish*. SIMCA will now autogenerate batch time.

Select the 20 reference batches (Ba, Ca, Ia, Ma, Na, Qa, Ra, Ta, Va, Xa, Za, ab, bb, cb, db, eb, fb, gb, hb & ib).

Make a PLS-model vs time. Look at the score plot t_1/t_2 and the control charts (*Analysis/Batch Control Charts*). Interpret the model.

Task 2 (Observation level)

Import the secondary data set (*File/Import Secondary Dataset* and select the file *Bakers Yeast Secondary.XLS*). Note: You have to specify this data set identically to the primary data set. This data set includes data for ten batches.

Use *Predictions/ Specify Predictionset* and select the newly imported data. Monitor the unmodelled batches in terms of the various batch control charts (*Predictions/Batch Control Charts*).

The thirteen unmodelled batches are Aa, Da, Ga, Ha, Ja, La, Oa, Pa, Ua, Ya, jb, kb & lb. Classify the batches. How do they comply with the control limits? Only four test batches have developmental trajectories very similar to those of the 20 reference batches – which? You may also want open the XLS-file (for the secondary data set) to watch the Y-variables.

Task 3 (Observation level)

In order to highlight how contribution plotting can help us to interpret deviations (i.e., to find “assignable causes”) we are going to focus on batches Ja, and Pa. Batch Ja deviated at the beginning and in the middle, but was OK at the end. Batch Pa deviated towards the end where it produced ethanol instead of yeast.

Make a score contribution plot for these batches to understand how they deviate.

Hint: Use the *Contribution Tool* from the *Marking Toolbar* and double-click in any interesting score or DModX-plot.

Task 4 (Batch level)

Create a batch level project in *File | Create Batch Level Project* and switch to this project. Bring the scores from the observation level. Do not forget to click *Bring secondary dataset to the batch level project*.

Now we are going to use the initial condition (Innoc) and response (QP1, QP2, Amount, Yield) variables together with the scores from the lower level project. Set the four response variables as Y.

Change model type to PCA-X. Autofit this model. Use this model and make classify the batches in the secondary data set. Which four batches evolve as the 20 reference batches?

Task 5 (Batch level)

Your next step is to make the batch level PLS model. However, because the four response variables are correlated in groups of two (QP1/QP2 and Amount/Yield) we will develop two PLS models

The first PLS model will have Amount and Yield as Y, and all other variables as X. Keep the batches Ba, Ca, and Va in the data set although they have missing Y-data, since their X-data stabilise the X matrix.

Compute 2 components. Use this model to predict Amount and Yield for the test batches. What can you say about the predictive qualities of the model?

Study the coefficient plot to establish the influence of the scores on Amount and Yield. What can you say about the predictive power?

Task 6 (Batch level)

Repeat Task 5 but select the two responses QP1 and QP2. Is it possible to predict these two quality parameters for the test set batches?

There is no solution provided to this task!!!!

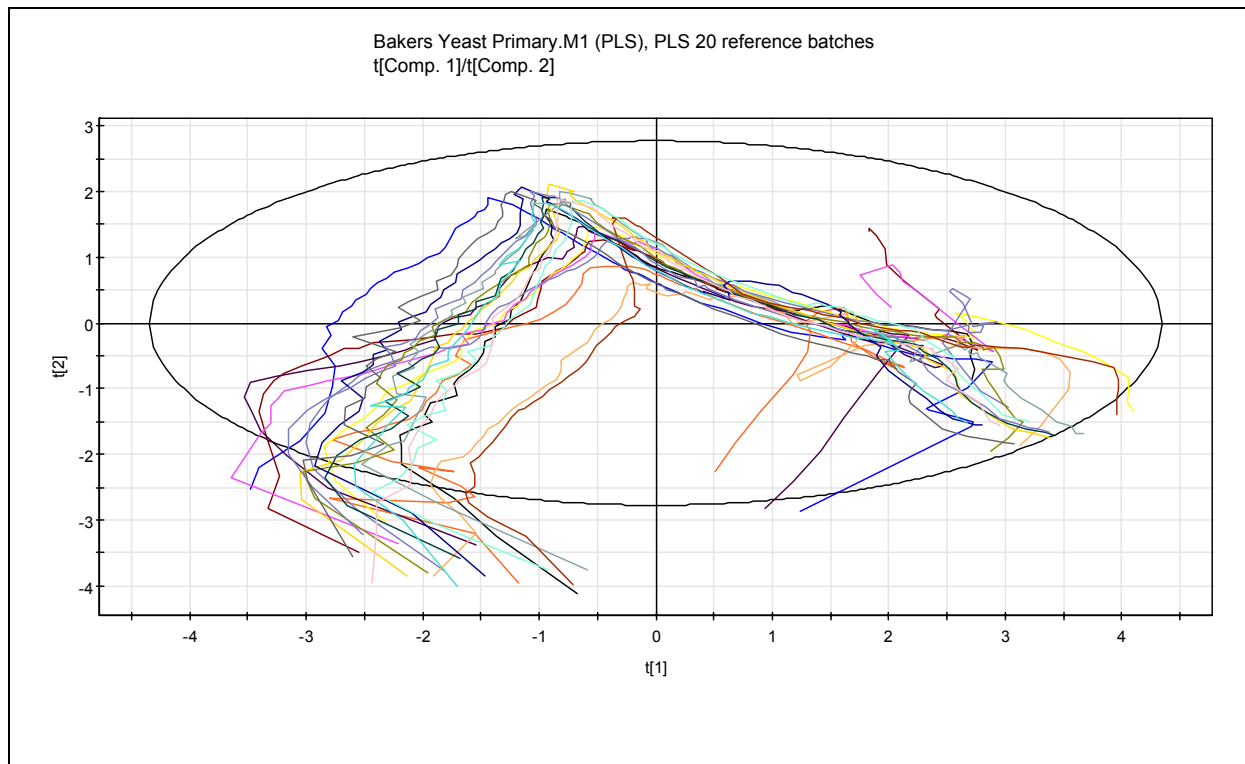
Solutions to Baker's Yeast

Task 1

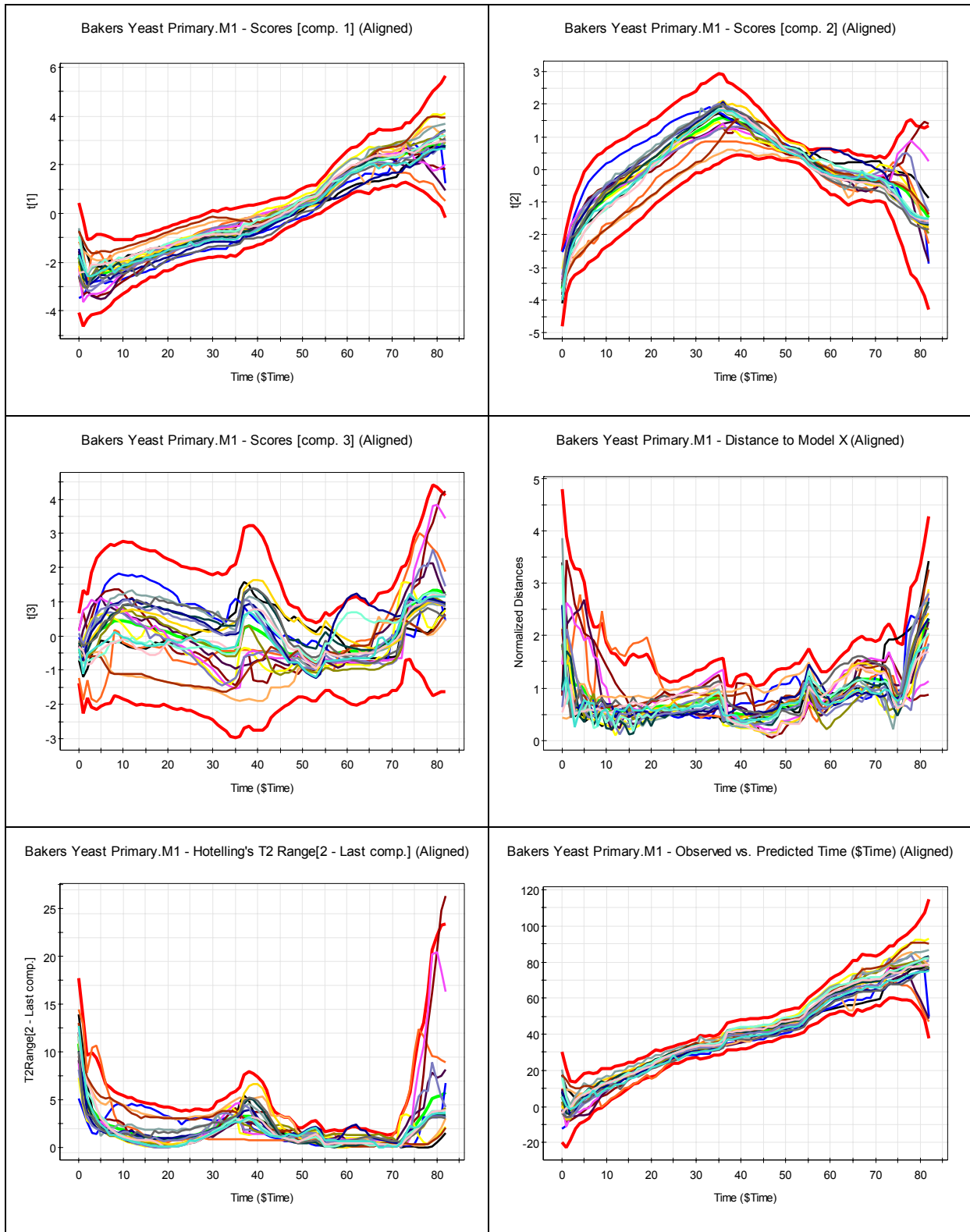
A three-component model was obtained.

Bakers Yeast Primary - M1											
Workset...		Options...		Title PLS 20 reference batches							
Type: PLS Observations (N)=1660, Variables (K)=8 (X=7, Y=1)											
Components: Included Batches:20											
A	R2X	R2X(cum)	Eigenv...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signifi...	Ite...	
0	Cent.			Cent.							
1	0.452	0.452	3.16	0.931	0.931	0.931	0	0.931	RB1	1	
2	0.245	0.697	1.72	0.0197	0.951	0.284	0	0.95	RB1	1	
3	0.133	0.83	0.928	0.0118	0.962	0.238	0	0.962	RB1	1	

In the t_1/t_2 score plot we can see a similar performance profile for all bathes.



Then we evaluate the batch control charts for t_1 - t_3 , DModX, Hot T^2 , and TimePred.

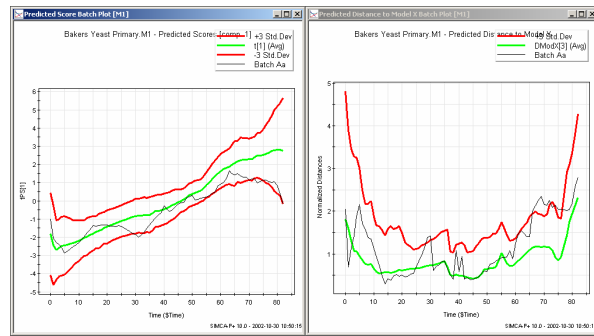


A good new batch should evolve similar to the reference batches and its trace should stay within the control limits.

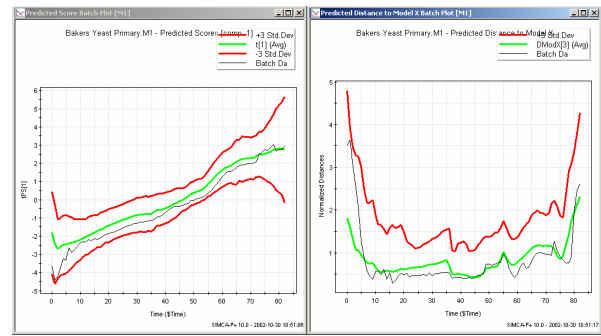
Task 2

For reasons of brevity only two control charts (t_1 & DModX) are plotted for each test batch. This is reasonable since the first score distance vector explains most of the variance.

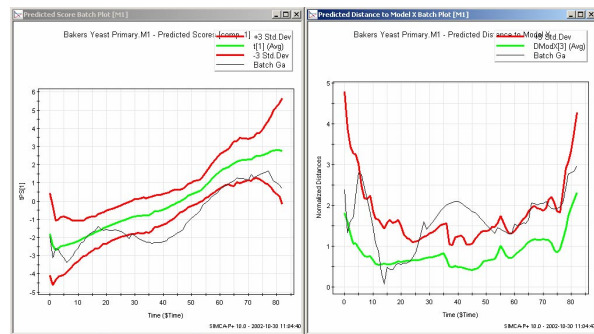
Aa – Minor problems & low yield



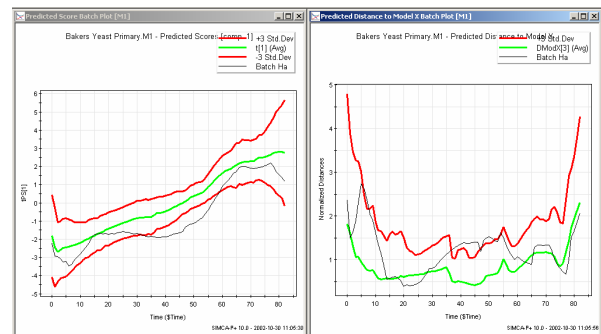
Da – OK all the way



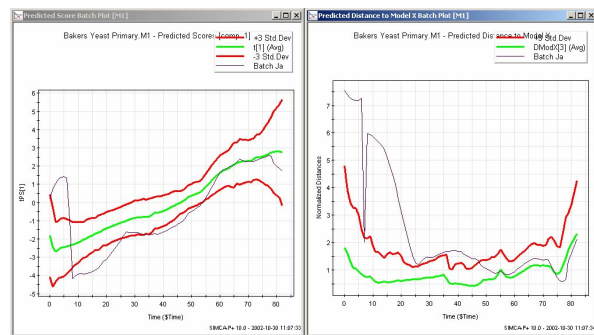
Ga – Problems in the middle, OK at the end



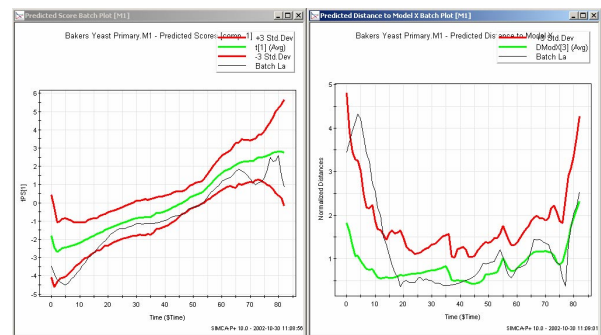
Ha – Problems in the middle, OK at end



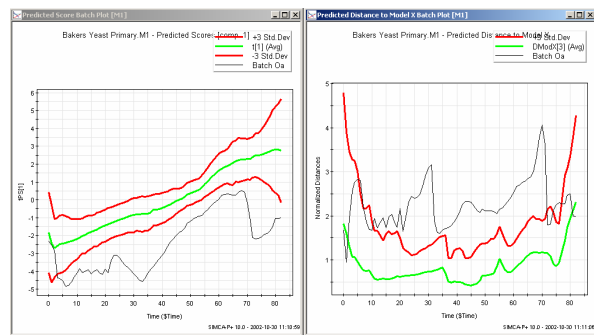
Ja – Problems at start, OK at the end



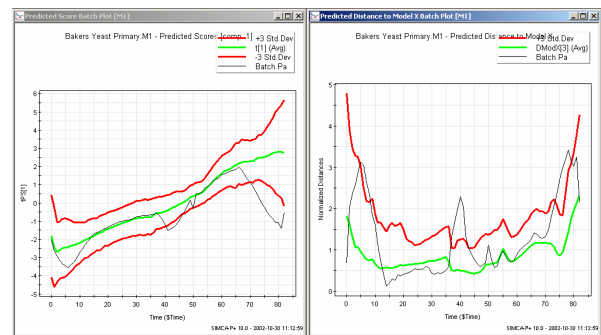
La – Early high DModX, but finally OK



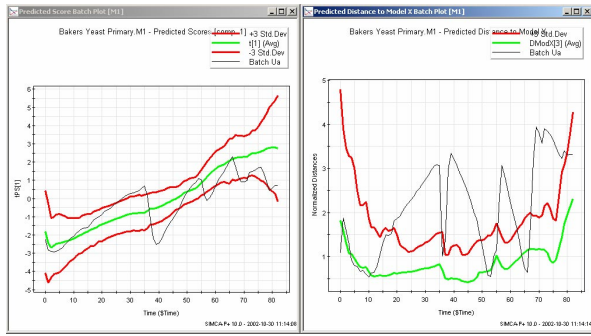
Oa – Problematic batch with low yield



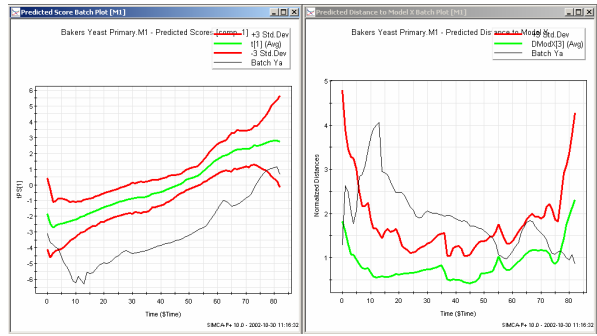
Pa – Failing batch which produced ethanol



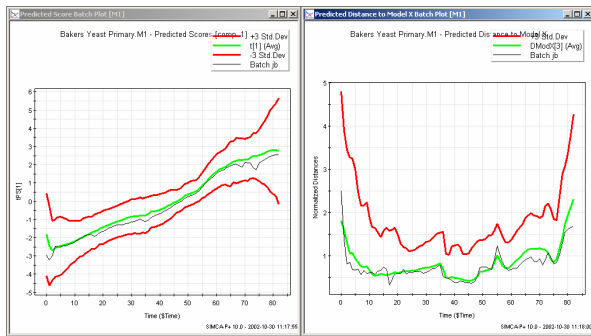
Ua – Problematic batch with low yield



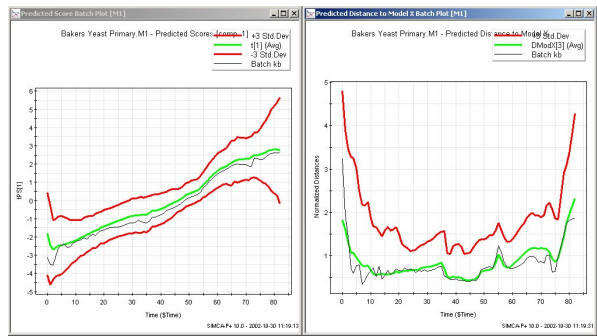
Ya – Problematic batch with low yield



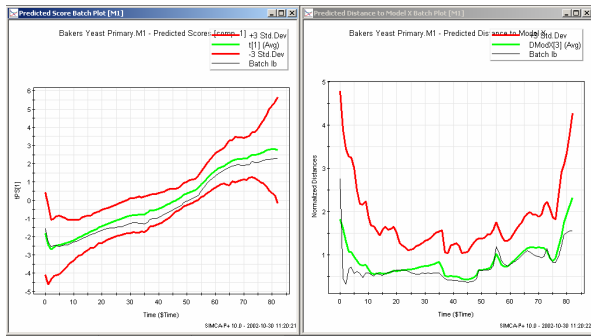
jb – OK all the way



kb – OK all the way

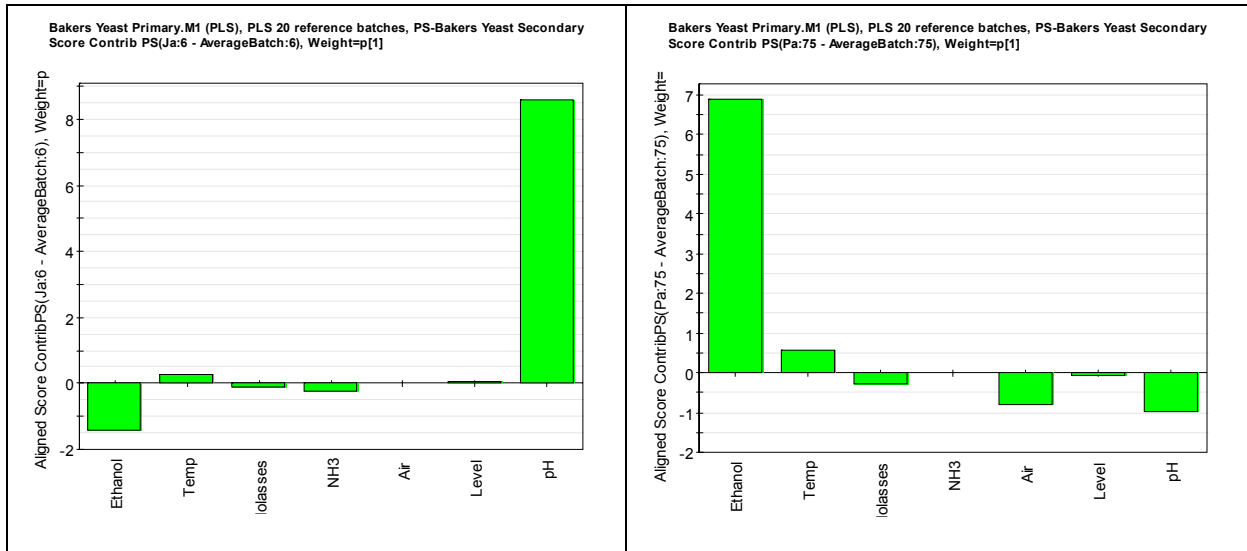


lb – OK all the way



Task 3

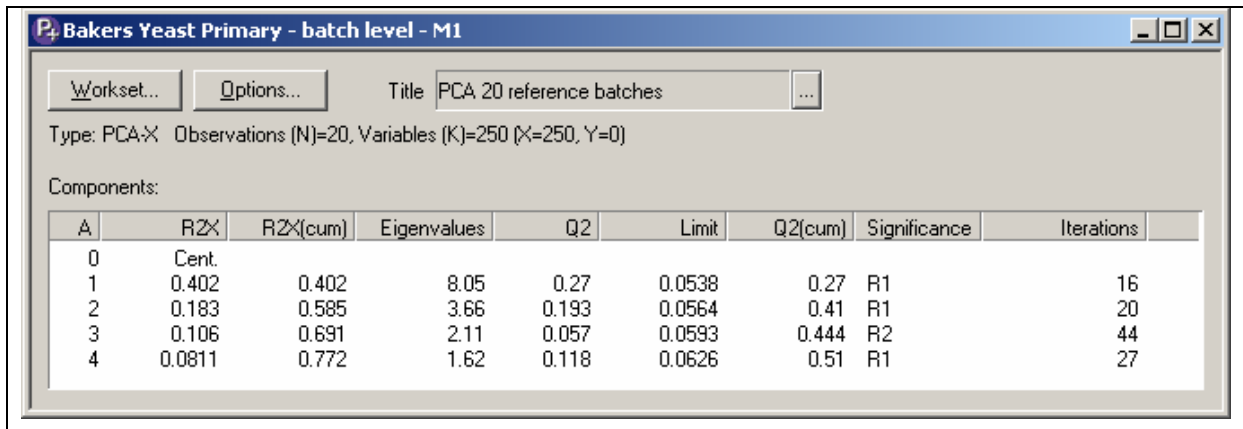
The left-hand score contribution plot below contrasts Ja at time point 6 with the average batch. The variable that is responsible for the deviation of Ja is pH, which is more than 8 standard deviations higher than for the average batch. Similarly, the right-hand plot, developed for time point 75, suggests that the increase of ethanol is almost 7 standard deviations for batch Pa.



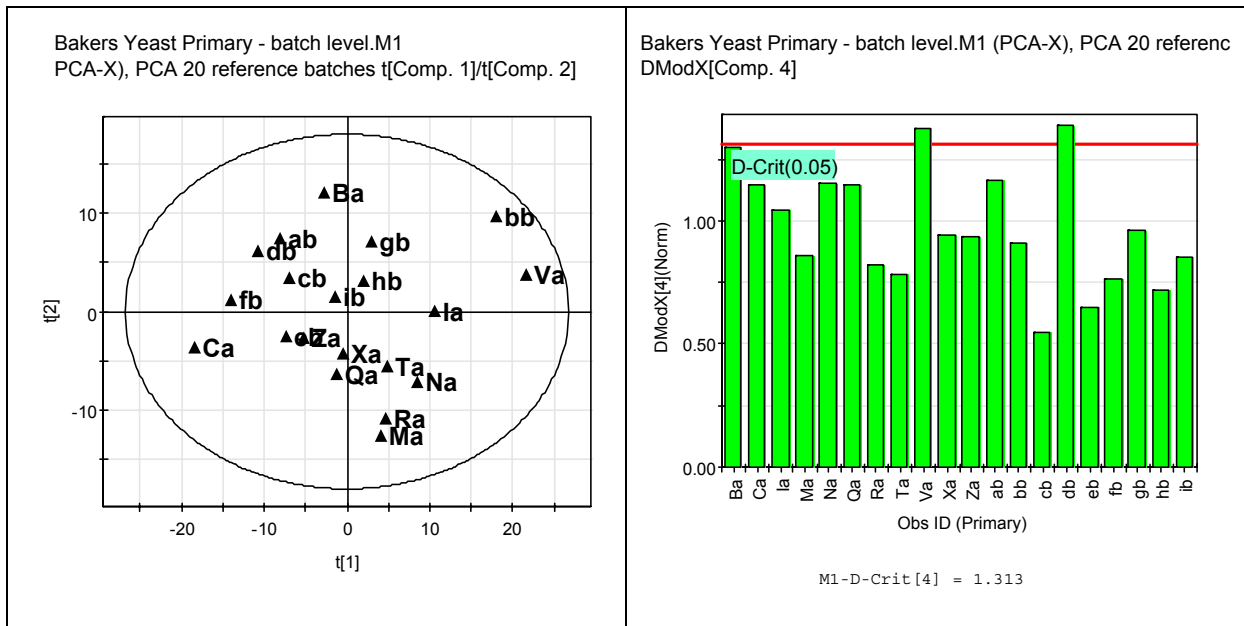
Many other contribution plots may be developed for the rest of the batches, but such an exercise is not pursued here.

Task 4

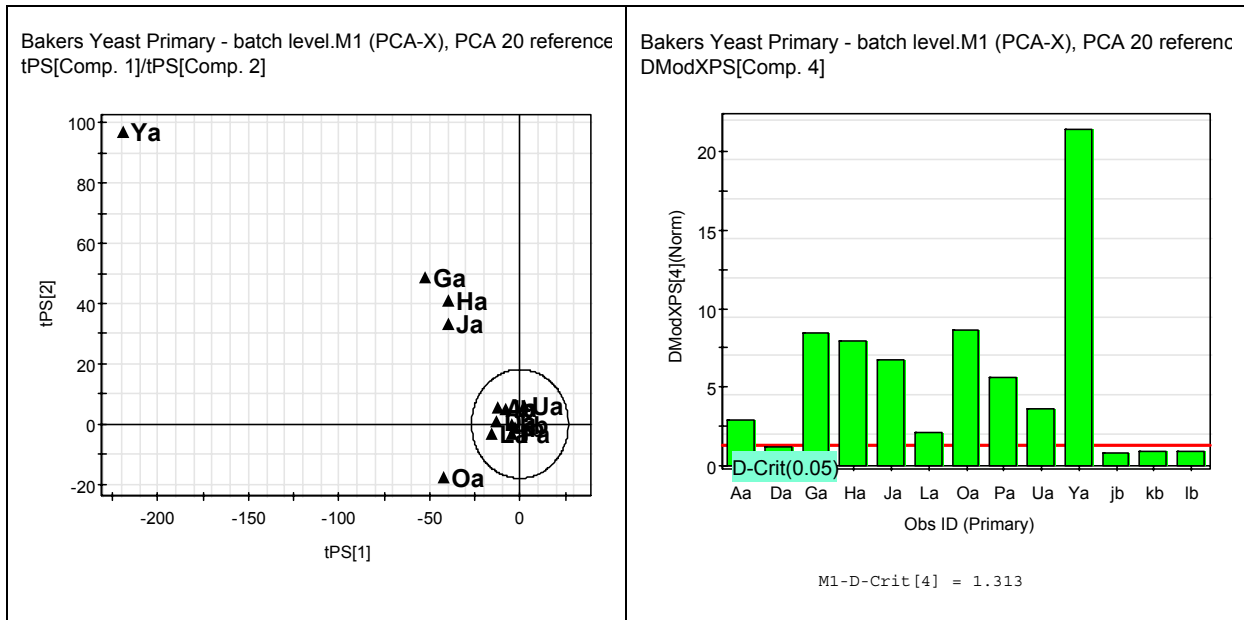
PCA with cross-validation indicates four large principal components with $R^2X = 0.77$ and $Q^2X = 0.51$.



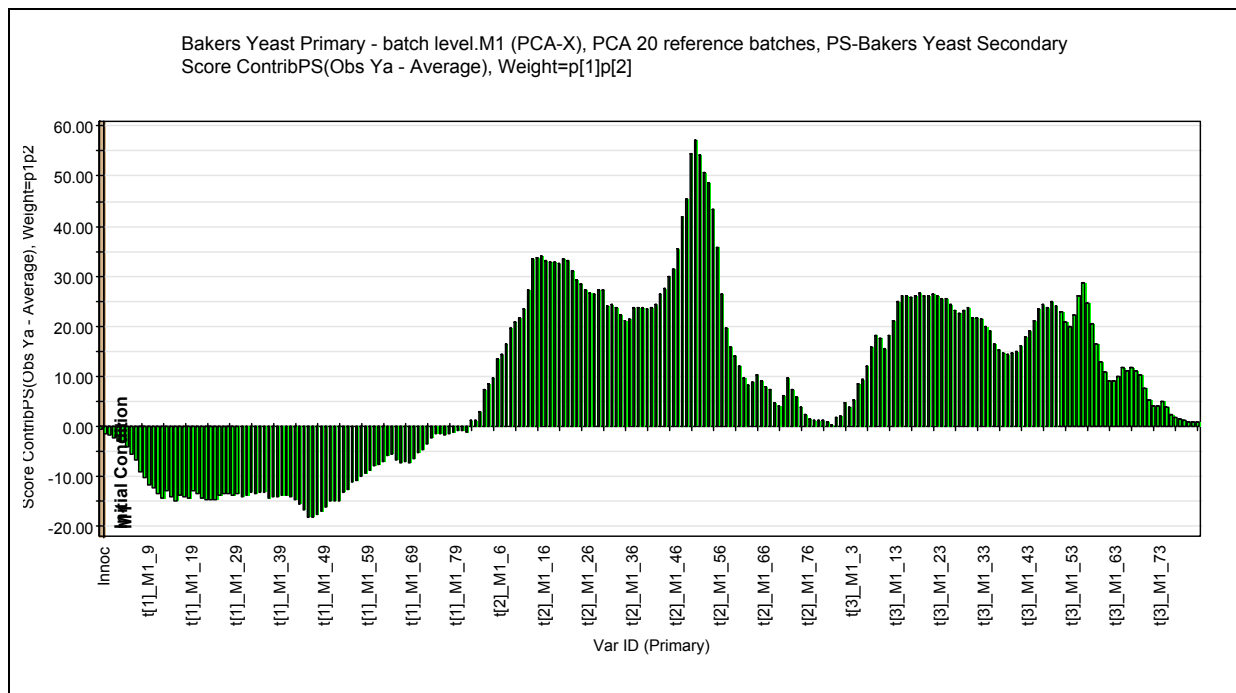
The score plot below shows a homogenous distribution of the batches. Only two out of the 20 batches are positioned outside the tolerance volume of the model, which is OK.



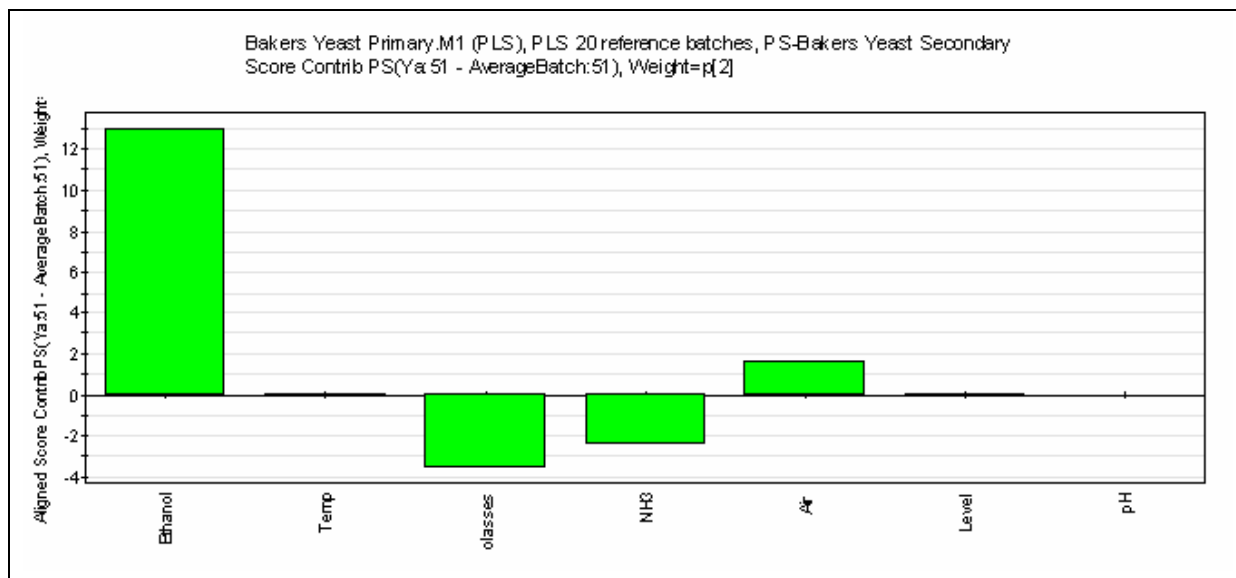
The results of the classification phase are seen below. Based on their trajectorial properties five batches Ya, Ga, Ha, Ja, and Oa are found very different from the 20 reference batches (left-hand score plot below). The right-hand DModX plot identifies the four batches Da, jb, kb, and lb to best conform with the 20 reference batches. Being close to Dcrit, batches like Aa, La, and Ua are also classified as rather similar to the reference batches.



We may use a contribution plot to investigate why Ya is so remotely positioned in the score plane. The contribution plot (scores mode) -- use the contribution tool and double-click on Ya in the score plot -- below suggests that Ya behaves differently in score t_2 in the time interval 47-56. The top peak occurs at time point 51.



A simple double-click on e.g. the coefficient at time point 51 will reveal how the seven original variables are behaving at this occasion. As shown by the plot below, it is mainly ethanol that is problematic and it is 12 standard deviations higher for Ya than for the average reference batch.



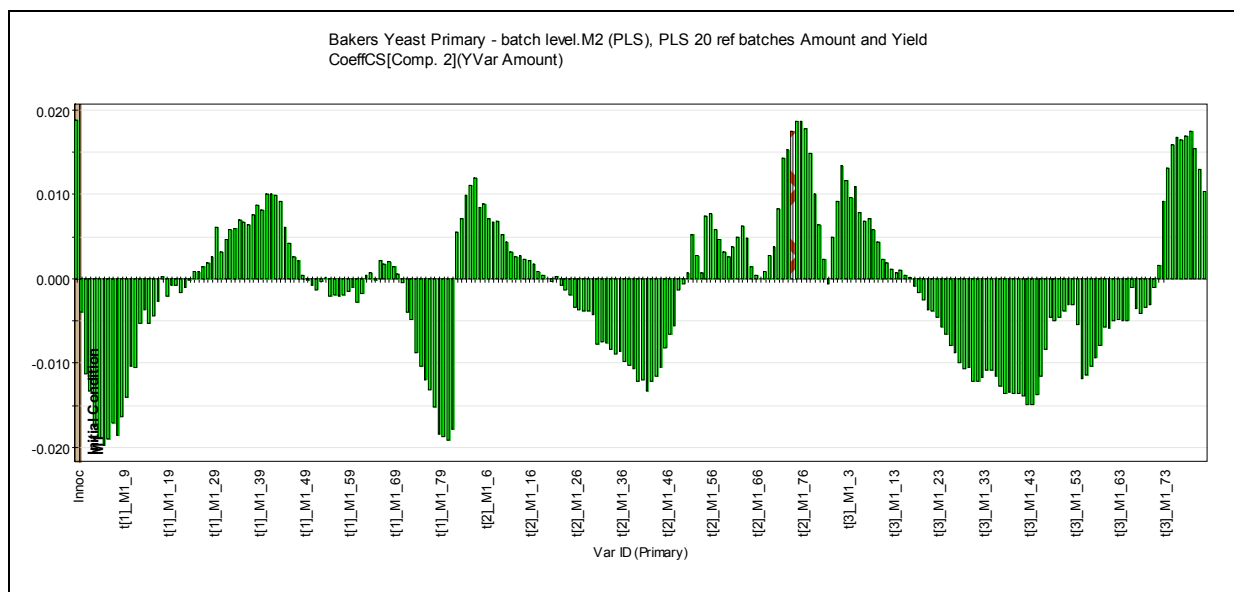
Task 5

A two-component PLS model was computed.

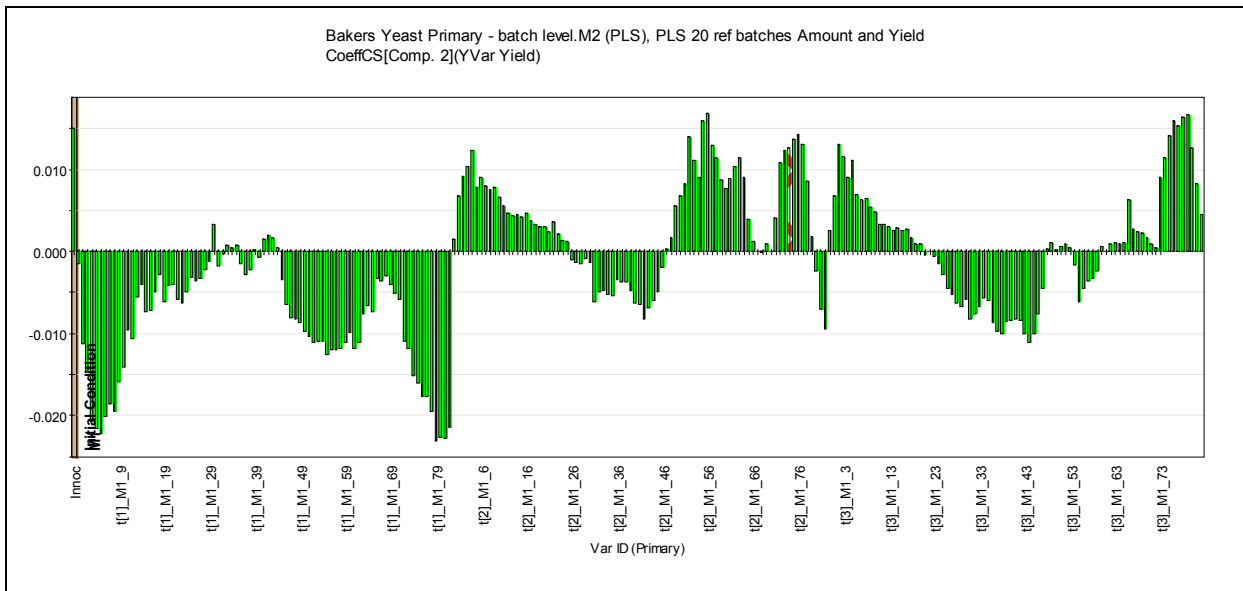
Bakers Yeast Primary - batch level - M2												
Workset...		Options...		Title PLS 20 ref batches Amount and Yield							...	
Type: PLS Observations (N)=20, Variables (K)=252 (X=250, Y=2)												
Components:												
A	R2X	R2X(cum)	Eigenv...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signifi...	Ite...		
0	Cent.			Cent.								
1	0.242	0.242	4.84	0.427	0.427	0.181	0.05	0.181	R1	8		
2	0.232	0.474	4.65	0.182	0.609	-0.0345	0.05	0.153	NS	7		

To interpret the model we may look at the loadings or the coefficients. The first coefficient plot relates to Amount. Here we see the influence of the scores as function of batch maturity. First comes Innoc, then t_1 : 0-82, then t_2 : 0-82, then t_3 : 0-82. Use the plot magnifier, *Scale X*, to look closer at this plot.

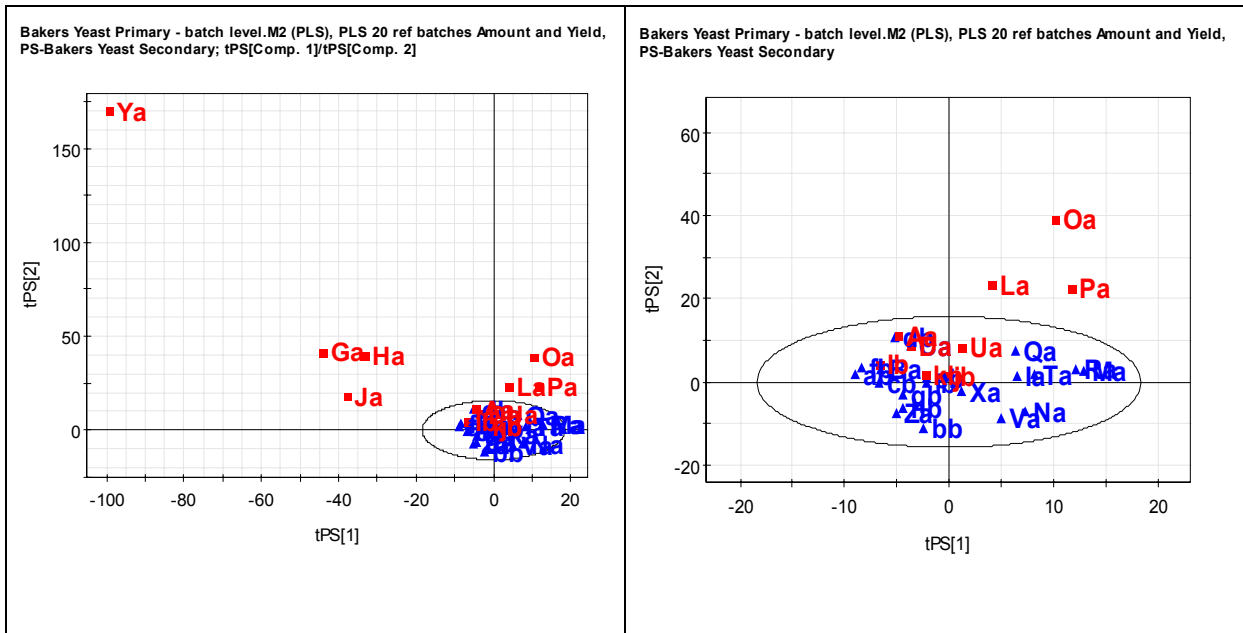
This response depends to a great extent on the initial condition variable Innoc, but it can be seen that contributions from many of the score variables are not negligible. Particularly, t_1 -score variables in the time spans 3 – 9 and 78 – 82 correlate significantly with Amount. The latter time segment is important also for the t_2 - and t_3 -based score variables. Hence, knowing Innoc alone is inferior compared with using also the score trajectories when modelling Amount.



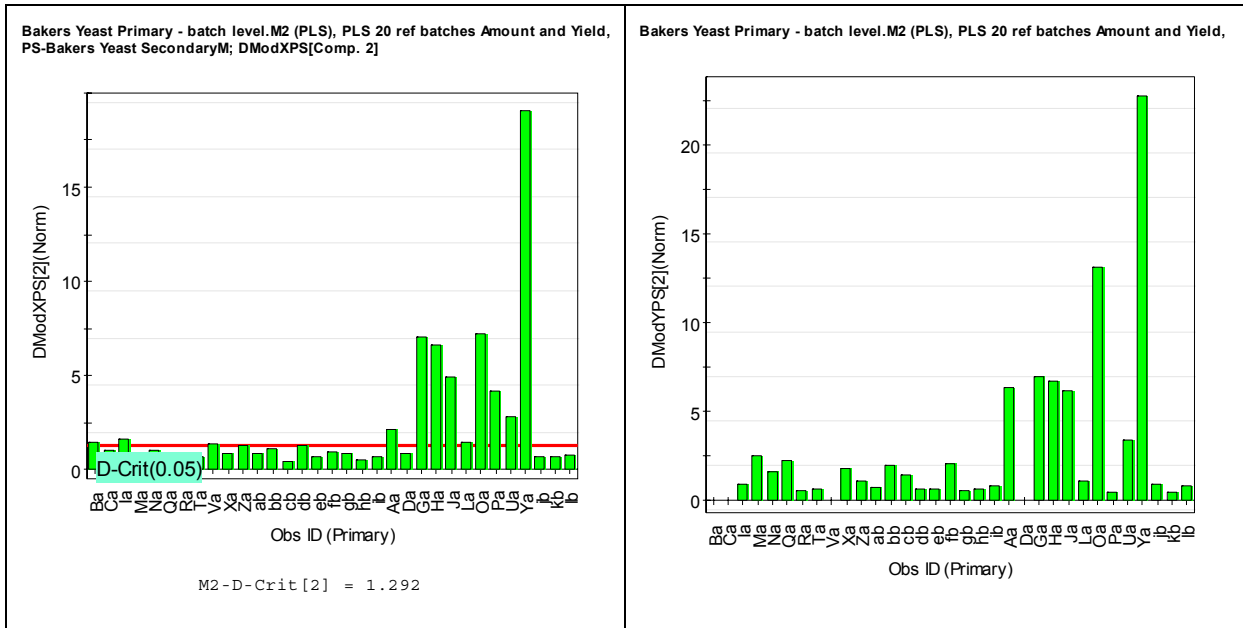
The second coefficient plot reveals which variables are influential for Yield. Essentially, the first ten and the last ten time points of t_1 dominate the model. The extent to which the second and third set of score variables participate in the model formation is much lower. Knowing the amount of explained variance by the first component of the lower level model, this fact is perhaps not so surprising.



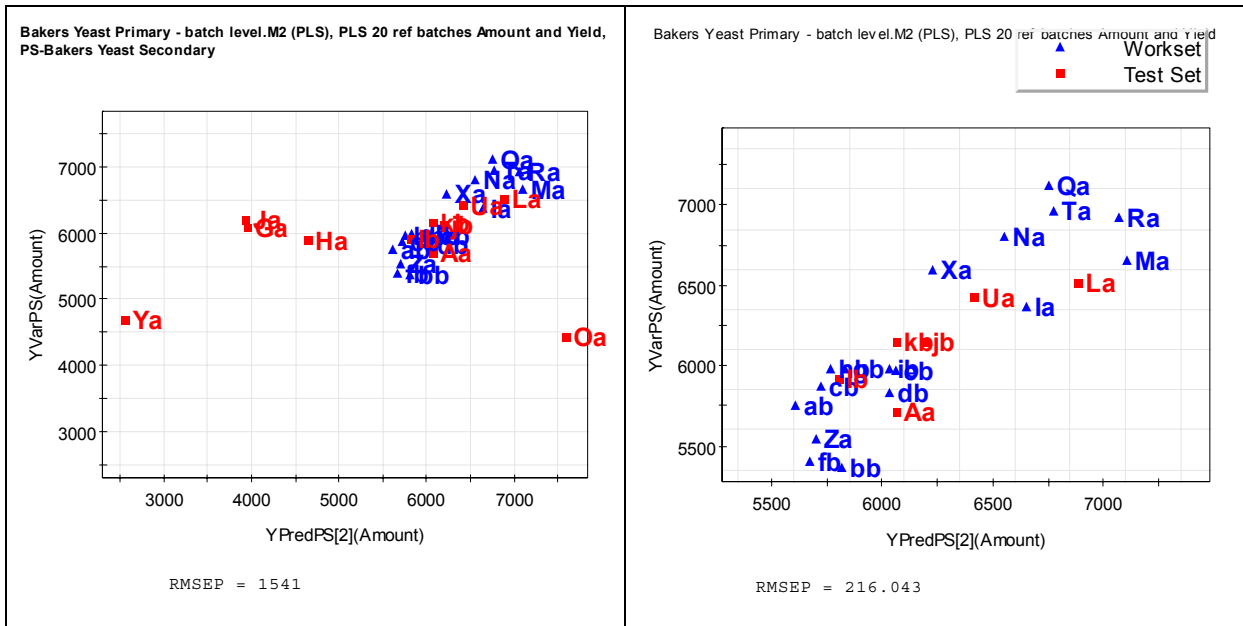
We now move on to the prediction phase. In order to see how the 13 test batches were accommodated by the model, we created the predicted score plot seen below (left-hand plot). Test batches are marked by squares and reference batches by triangles. There are five test batches positioned quite far outside the Hotelling's T2 tolerance region, i.e., test batches Ga, Ha, Ja, Oa, and Ya. The right-hand plot below is an enlargement of the area around the tolerance region. Here we can spot that also test batches La and Pa are outside, but not by much. The remaining six test batches (Aa, Da, Ua, jb, kb, and lb) fit the model much better.



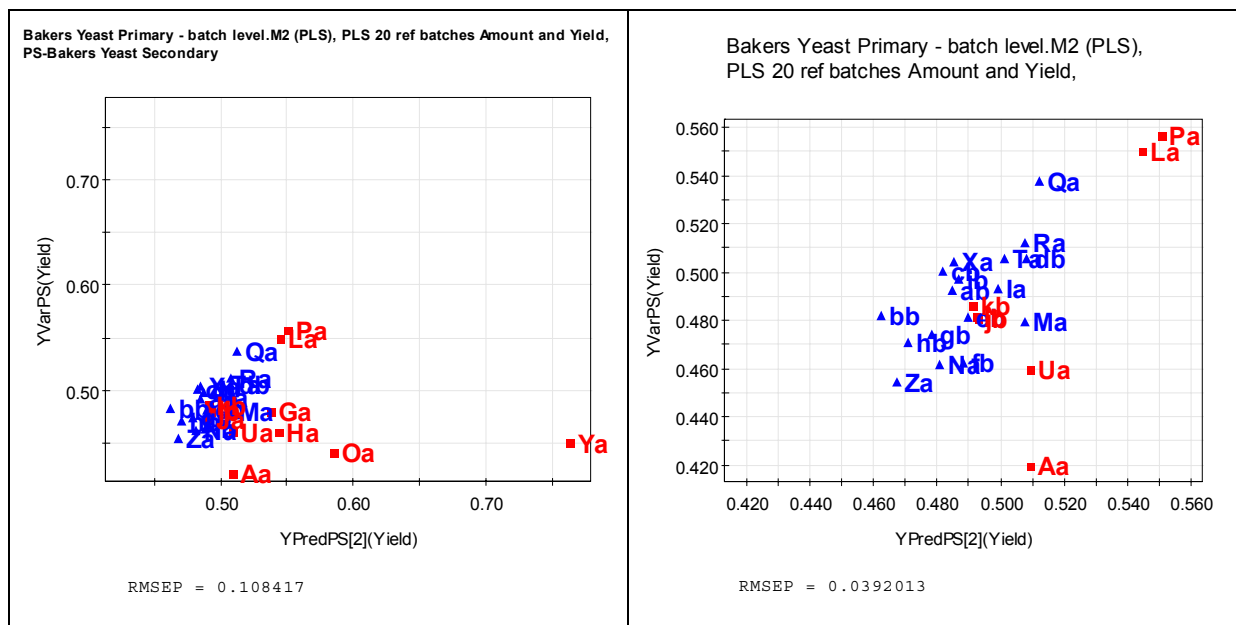
The DModX and DModY residual plots point in the same direction, i.e., that five batches (Ga, Ha, Ja, Oa, and Ya) are very different from the reference batches, whereas La and Pa are more similar. Strictly speaking, however, only four test batches comply perfectly with the model, Da, jb, kb and lb. These were the batches detected already on the observation level.



We will now test the predictive ability of the model. The two plots below display the relationship between measured (“observed”) Amount and the corresponding values computed by the model. It is clearly evident that five batches (Ga, Ha, Ja, Oa, and Ya) are very poorly predicted. This exactly the quintet of batches identified above as not fitting the model. Any prediction computed for anyone of these five tricky batches would correspond to a large extrapolation outside the model validity range and hence be very uncertain. Consequently, these five tricky batches were removed (see prediction results in right-hand plot below). The external RMSEP computed only on true test batches amounts to 216. This corresponds to a Q^2_{ext} of 0.67, which is an excellent result.



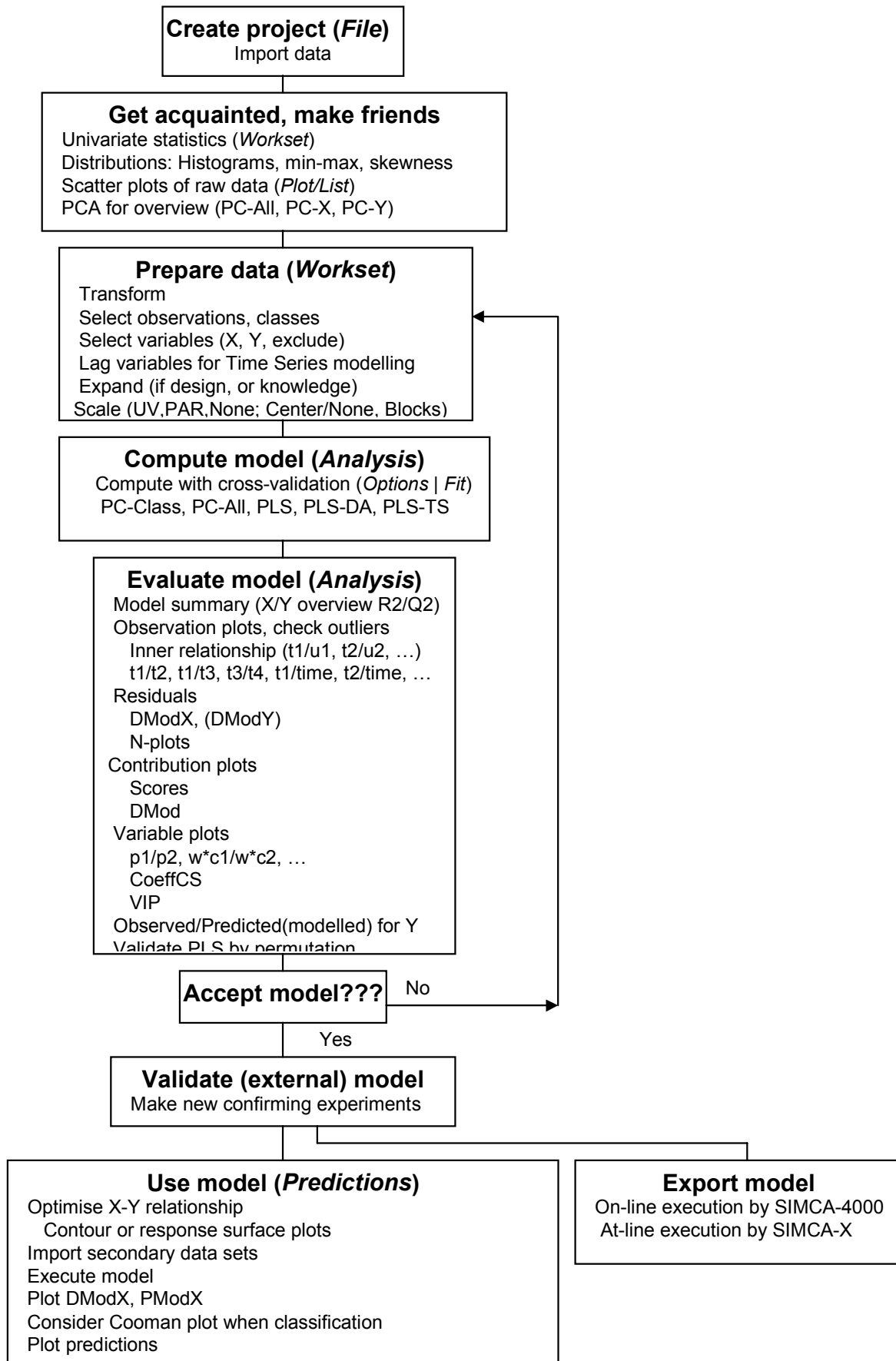
The next set of plots reflect predictions of Yield. If the five non-conformant test batches are omitted, RMSEP becomes 0.039 and $Q^2_{ext} = 0.52$ for the remaining test batches. These results are also very encouraging.




Conclusions

Twenty reference batches were used to train a model to recognize “good” operating conditions. This model was able to categorize between good and problematic batches. Predictive power of the upper level PLS model was very good for test batches, except for five batches that were clearly unrepresentative cases.


Flow-chart of MVDA in SIMCA



				1(9)
Uppgjord - Prepared		Datum - Date	Rev A	Dokumentnr - Document no
Godkänd - Approved		File/reference		

References for MVDA book

- Albano, C., Dunn, III, W.J., Edlund, U., Johansson, E., Nordén, B., Sjöström, M., and Wold, S., (1978), *Four Levels of Pattern Recognition*, *Analytica Chimica Acta*, 103, 429-443.
- Alsberg, B.K., Woodward, A.M, and Kell, D.B., (1997), *An Introduction to Wavelet Transforms for Chemometricians: A Time-Frequency Approach*, *Chemometrics and Intelligent Laboratory Systems*, 37, 215-239.
- Andersson, G., Kaufmann, P., and Renberg, L, (1996), *Non-Linear Modelling with a Coupled Neural Network – PLS Regression System*, *Journal of Chemometrics*, 10, 605-614.
- Andersson, P., Haglund, P., Tysklind, M., (1997a), *The Internal Barriers of Rotation for the 209 Polychlorinated Biphenyls*, *Environmental Science and Pollution Research*, 4, 75-81.
- Andersson, P., Haglund, P., and Tysklind, M., (1997b), *Ultraviolet Absorption Spectra of All 209 Polychlorinated Biphenyls Evaluated by Principal Component Analysis*, *Fresenius Journal of Analytical Chemistry*, 357, 1088-1092.
- Andersson, G., (1998), *Novel Nonlinear Multivariate Calibration Methods*, Ph.D. Thesis, The Royal Institute of Technology, Stockholm, Sweden.
- Andersson, P.M., Sjöström, M., and Lundstedt, T., (1998), *Preprocessing Peptide Sequences for Multivariate Sequence-Property Analysis*, 42, 41-50.
- Andersson, P.M., Sjöström, M., Wold, S., and Lundstedt, T., (2001), *Strategies for Subset Selection of Parts of an In-house Chemical Library*, *Journal of Chemometrics*, 15, 353-369.
- Anonymous, SIMCA-P manual, Umetrics AB.
- Austel, V., (1995), *Experimental Design*, in: Mannhold, R., Krogsgaard-Larsen, P., and Timmerman, H., (Eds.), *Methods and Principles in Medicinal Chemistry*, Vol 2, VCH, Weinheim, Germany, pp. 49-62.
- Barnes, R.J., Dhanoa, M.S., and Lister, S.J., (1989), *Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra*, *Applied Spectroscopy*, 43, 772-777.
- Baroni, M., Clementi, S., Cruciani, G., Kettaneh-Wold, S., and Wold, S., (1993a), *D-Optimal Designs in QSAR*, *Quantitative Structure-Activity Relationships*, 12, 225-231.
- Baroni, M., Constatino, G., Cruciani, G., Riganelli, D., Valigi, R., and Clementi, S., (1993b), *Generating Optimal PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems*, *Quantitative Structure-Activity Relationships*, 12, 9-20.
- Barnett, V., and Lewis, T., (1994), *Outliers in Statistical Data*, John Wiley & Sons, Chichester, England.
- Berglund, A., and Wold, S., (1997a), *INLR, Implicit Non-Linear Latent Variable Regression*, *Journal of Chemometrics*, 11, 141-156.
- Berglund, A., DeRosa, M.C., and Wold, S., (1997b), *Alignment of Flexible Molecules at Their Receptor Site Using 3D Descriptors and Hi-PCA*, *Journal of Computer-Aided Molecular Design*, 11, 601-612.
- Berglund, A., and Wold, S., (1999), *A Serial Extension of Multi Block PLS*, *Journal of Chemometrics*, 13, 461-471.
- Berglund, A., Kettaneh, N., Uppgård, L.L., Wold, S., Bandwell, N., and Cameron, D.R., (2001), *The GIFI Approach to Non-Linear PLS Modelling*, *Journal of Chemometrics*, 15, 321-336.
- Blanco, M., Coello, J., Iturriaga, H., Maspocho, S., and Pagès, J., (2000), *NIR Calibration in Non-linear Systems: Different PLS Approaches and Artificial Neural Networks*, *Chemometrics and Intelligent Laboratory Systems*, 50, 75-82.
- Blum, D.J.W., and Speece, R.E., (1990), *Determining Chemical Toxicity to Aquatic Species*, *Environmental Science and Technology*, 24, 284-293.
- Box, G.E.P., Hunter, W.G., and Hunter, J.S., (1978), *Statistics for Experimenters*, John Wiley & Sons, Inc., New York.
- Box, G.E.P., Jenkins, G.M., and Reinsel, G.C., (1994), *Time Series Analysis - Forecasting and Control*, 3rd edition, Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Bro, R., (1996), *Håndbog i Multivariabel Kalibrering*, KVL, Copenhagen, Denmark.

				2(9)
Uppgjord - Prepared		Datum - Date	Rev	Dokumentnr - Document no
Godkänd - Approved		A		
		File/reference		

Bro, R., (1999), *Exploratory Study of Sugar Production Using Fluorescence Spectroscopy and Multi-way analysis*, Chemometrics and Intelligent Laboratory System, 46, 133-147.

Burnham, A.J., Viveros, R., and MacGregor, J.F., (1996), *Frameworks for Latent Variable Multivariate Regression*, Journal of Chemometrics, 10, 31-45.

Burnham, A.J., MacGregor, J.F., and Viveros, R., (1999), *A Statistical Framework for Multivariate Latent Variable Regression Methods Based on Maximum Likelihood*, Journal of Chemometrics, 13, 49-65.

Burnham, A.J., MacGregor, J., and Viveros, R., (2001), *Interpretation of Regression Coefficients Under a Latent Variable Regression Model*, Journal of Chemometrics, 15, 265-284.

Börås, L., Sjöström, J., and Gatenholm, P., (1997), *Characterization of Surfaces of CMTP Fibers Using Inverse Gas Chromatography Combined with Multivariate Data Analysis*, Nordic Pulp & Paper Research Journal, 12, 220-224.

Carlson, R., Lundstedt, T., and Albano, C., (1985), *Screening of Suitable Solvents in Organic Synthesis. Strategies for Solvent Selection*, Acta Chemica Scandinavica, B39, 79-91.

Cocchi, M., and Johansson, E., (1993), *Amino Acids Characterization by GRID and Multivariate Data Analysis*, Quantitative Structure-Activity Relationships, 12, 1-8.

Connor, K., Safe, S., Jefcoate, C.R., and Larsen, M., (1995), *Structure-Dependent Induction of CYP2B by Polychlorinated Biphenyl Congeners in Female Sprague-Dawley Rats*, Biochemical Pharmacology, 50, 1913-1917.

Coomans, D., Broeckaert, I., Derde, M.P., Tassin, A., Massart, D.L., and Wold, S., (1984), *Use of a Microcomputer for the Definition of Multivariate Confidence Regions in Medical Diagnosis Based on Clinical Laboratory Profiles*, Comp. Biomed. Res., 17, 1-14.

Cramer, R.D., Patterson, D.E., and Bunce, J.D., (1988), *Comparative Molecular Field Analysis (CoMFA). I. Effects of Shape on Binding of Steroids to Carried Proteins*, Journal of American Chemical Society, 110, 5959.

Dahl, K.S., Piovoso, M.J., and Kosanovich, K.A., (1999), *Translating Third-order Data Analysis Methods to Chemical Batch Processes*, Chemometrics and Intelligent Laboratory Systems, 46, 161-180.

Davis, O.L., and Goldsmith, P.L., (1986), *Statistical Methods in Research and Production*, Longman, New York.

Dayal, B., MacGregor, J.F., Taylor, P.A., Kildaw, R., and Marcikic, S., (1994), *Application of Feedforward Neural Networks and Partial Least Squares Regression for Modelling Kappa Number in a Continuous Kamyra Digester*, Pulp and Paper Canada, 95, 26-32.

DeAguiar, P.F., Bourguignon, B., Khots, M.S., Massart, D.L., and Phan-Thuan-Luu, R., (1995), *D-Optimal Designs*, Chemometrics and Intelligent Laboratory Systems, 30, 199-210.

Dearden, J., (1985), *Partitioning and Lipophilicity in Quantitative Structure-Activity Relationships*, Environmental Health Perspectives, 61, 203-228.

DeJong, S., (1993), *PLS Fits Closer Than PCR*, Journal of Chemometrics, 7, 551-557.

De Jong, S., Wise, B.M., and Ricker, N.L., (2001), *Canonical Partial Least Squares and Continuum Power Regression*, Journal of Chemometrics, 15, 85-100.


Deneer, J.W., Sinnige, T.L., Seinen, W., and Hermens, J.L.M., (1987), *Quantitative Structure-Activity Relationships for the Toxicity and Bioconcentration Factor of Nitrobenzene Derivatives Towards the Guppy (Poecilia reticulata)*, Aquatic Toxicology, 10, 115-129.

Deneer, J.W., van Leeuwen, C.J., Seinen, W., Maas-Diepeveen, J.L., and Hermens, J.L.M., (1989), *QSAR Study of the Toxicity of Nitrobenzene Derivatives Towards Daphnia magna, Chlorella pyrenoidosa and Photobacterium phosphoreum*, Aquatic Toxicology, 15, 83-98.

Denham, M.C., (1997), *Prediction Intervals in Partial Least Squares*, Journal of Chemometrics, 11, 39-52.

Drewry, D.H., and Young, S.S., (1999), *Approaches to the Design of Combinatorial Libraries*, Chemometrics and Intelligent Laboratory Systems, 48, 1-20.

Dunn, III, W.J., (1989), *Quantitative Structure-Activity Relationships (QSAR)*, Chemometrics and Intelligent Laboratory Systems, 6, 181-190.

				3(9)
Uppgjord - Prepared		Datum - Date	Rev	Dokumentnr - Document no
Godkänd - Approved		File/reference		
		A		

Efron, B., and Gong, G., (1983), *A Leisurely Look at the Bootstrap, the Jack-knife, and Cross-validation*, American Statistician, 37, 36-48.

Ergon, R., (1998), *Dynamic System Multivariate Calibration by System Identification Methods*, Modelling, Identification and Control, 19, 77-97.

Eriksson, L., Jonsson, J., Sjöström, M., Wikström, C., and Wold, S., (1988), *Multivariate Derivation of Descriptive Scales for Monosaccharides*, Acta Chemica Scandinavica, 42, 504-514.

Eriksson, L., Jonsson, J., Hellberg, S., Lindgren, F., Sjöström, M., Wold, S., Sandström, B., and Svensson, I., (1991), *A Strategy for Ranking Environmentally Occurring Chemicals. Part V: The Development of two Genotoxicity QSARs for Halogenated Aliphatics*. Environmental Toxicology and Chemistry, 10, 585-596.

Eriksson, L., Sandström, B.E., Tysklind, M., and Wold, S., (1993), *Modelling the Cytotoxicity of Halogenated Aliphatic Hydrocarbons. Quantitative Structure-Activity Relationships for the IC50 to Human HeLa Cells*, Quantitative Structure-Activity Relationships, 12, 124-131.

Eriksson, L., and Hermens, J.L.M., (1995a), *A Multivariate Approach to Quantitative Structure-Activity and Structure-Property Relationships*, in: The Handbook of Environmental Chemistry, (ed.) J. Einax, Vol2H, Chemometrics in Environmental Chemistry, Springer Verlag, Berlin.

Eriksson, L., Hermens, J.L.M., Johansson, E., Verhaar, H.J.M. and Wold, S., (1995b), *Multivariate Analysis of Aquatic Toxicity Data with PLS*, Aquatic Sciences, 57, 217-241.

Eriksson, L., and Johansson, E., (1996), *Multivariate Design and Modelling in QSAR*, Chemometrics and Intelligent Laboratory Systems, 34, 1-19.

Eriksson, L., Johansson, E., Müller, M., and Wold, S., (1997), *Cluster-based Design in Environmental QSAR*, Quantitative Structure-Activity Relationships, 16, 383-390.

Eriksson, L., Johansson, E., Tysklind, M., and Wold, S., (1998), *Pre-processing of QSAR Data by Means of Orthogonal Signal Correction*, The QSAR and Modelling Society, Newsletter 1998, www/pharma-ethz.ch./qsar.

Eriksson, L., Andersson, P., Johansson, E., Tysklind, M., Sandberg, M., and Wold, S., (1999a), *The Constrained Principal Property Space in QSAR – Directional and Non-Directional Modelling Approaches*, Proceedings 12th European Symposium on QSAR, August 1998, Copenhagen, Denmark.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wikström, C., and Wold, S., (1999b), *Design of Experiments – Principles and Applications*, Umetrics AB.

Eriksson, L., Trygg, J., Johansson, E., Bro, R., and Wold, S., (2000a), *Orthogonal Signal Correction, Wavelet Analysis, and Multivariate Calibration of Complicated Process Fluorescence Data*, Analytica Chimica Acta Analytica, 420, 181-195.

Eriksson, L., Johansson, E., Lindgren, F., and Wold, S., (2000b), *GIFI-PLS: Modeling of Non-Linearities and Discontinuities in QSAR*, Quantitative Structure-Activity Relationships, 19, 345-355.

Eriksson, L., Johansson, E., Müller, M., and Wold, S., (2000c), *On the selection of training set in environmental QSAR when compounds are clustered*, Journal of Chemometrics, 14, 599-616.

Eriksson, L., Hagberg, P., Johansson, E., Rännar, S., Whelehan, O., Åström, A., and Lindgren, T., (2001), *Multivariate Process Monitoring of a Newsprint Mill. Application to Modelling and Predicting COD Load Resulting from Deinking of Recycled Paper*. Journal of Chemometrics, 15, 337-352.

Fisher, R.A., (1936), *The use of Multiple Measurements in Taxonomic Problems*, Ann. Eugenics, 7, 179-188.


Frank, I.E., and Friedman, J.H., (1993), *A Statistical View of Some Chemometrics Regression Tools*, Technometrics, 35, 109-135.

Frank, I., (1995), *Modern Nonlinear Regression Methods*, Chemometrics and Intelligent Laboratory Systems, 27, 1-19.

Gallop, M.A., Barrett, R.W., Dower, W.J., Fodor, S.P.A., and Gordon, E.M., (1994), *Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries*, Journal of Medicinal Chemistry, 37, 1233-1251.

Geladi, P., MacDougall, D., Martens, H., (1985), *Linearization and Scatter-correction for Near-infrared Reflectance Spectra of Meat*, Applied Spectroscopy, 3, 491-500.

Goodford, P., (1996), *Multivariate Characterization of Molecules for QSAR Analysis*, Journal of Chemometrics, 10, 107-117.

		4(9)	
Uppgjord - Prepared	Datum - Date	Rev A	Dokumentnr - Document no
Godkänd - Approved	File/reference		

Gordon, E.M., Barret, R.W., Dower, W.J., Fodor, S.P.A., and Gallop, M.A., (1994), *Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions*, Journal of Medicinal Chemistry, 37, 1385-1401.

Gottfries, J., Blennow, K., Wallin, A., and Gottfries, C.G., (1995), *Diagnosis of Dementias Using Partial Least Squares Discriminant Analysis*, Dementia, 6, 83-88.

Granberg, R., (1998), *Solubility and Crystal Growth of Paracetamol in Various Solvents*, Ph.D. Thesis, Royal Institute of Technology, Stockholm, Sweden.

Gustafsson, A., (1993), *QFD and Conjoint Analysis – The Key to Customer Oriented Products*, Ph.D. Thesis, Linköping University, Linköping, Sweden.

Hammett, L.P., (1970), *Physical Organic Chemistry*, 2nd edn., McGraw-Hill, New York.

Hansch, C., and Leo, A.J., (1970), *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley, New York.

Hartnett, M.K., Lightbody, G., and Irwin, G.W., (1999), *Identification of State Models Using Principal Components Analysis*, Chemometrics and Intelligent Laboratory Systems, 46, 181-196.

Hellberg, S., (1986), *A Multivariate Approach to QSAR*, Ph.D. Thesis, Umeå University, Umeå, Sweden.

Hellberg, S., Sjöström, M., and Wold, S., (1986), *The Prediction of Bradykinin Potentiating Potency of Pentapeptides. An Example of a Peptide Quantitative Structure-Activity Relationship*, Acta Chemica Scandinavica, B40, 135-140.

Hellberg, S., Eriksson, L., Jonsson, J., Lindgren, F., Sjöström, M., Skagerberg, B., Wold, S., and Andrews, P., (1991), *Minimum Analogue Peptide Sets (MAPS) for Quantitative Structure-Activity Relationships*, International Journal of Peptide and Protein Research, 37, 414-424.

Hermens, J.L.M., (1989), *Quantitative Structure-Activity Relationships of Environmental Pollutants*. in: Hutzinger, O., (Ed.), Handbook of Environmental Chemistry, Vol 2E, Reactions and Processes. Springer-Verlag, Berlin, 1989, pp. 111-162.

Hunter, J.S., (1986), *The Exponentially Weighted Moving Average*, Journal of Quality Technology, 18, 203-210.

Höskuldsson, A., (1996), *Prediction Methods in Science and Technology*, Thor Publishing, Copenhagen, Denmark.

Höskuldsson, A., (1998), *The Heisenberg Modelling Procedure and Application to Nonlinear Modelling*, Chemometrics and Intelligent Laboratory Systems, 44, 15-30.

Jackson, J.E. (1991), *A User's Guide to Principal Components*, John Wiley, New York. (ISBN 0-471-62267-2).

Janné, K., Pettersen, J., Lindberg, N.O., and Lundstedt, T., (2001), *Hierarchical Principal Component Analysis (PCA) and Projection to Latent Structure Technique (PLS) on Spectroscopic Data as a Data Pretreatment for Calibration*, Journal of Chemometrics, 15, 203-213.

Jonsson, J., Eriksson, L., Hellberg, S., Sjöström, M., and Wold, S., (1989a), *Multivariate Parametrization of 55 Coded and Non-Coded Amino Acids*, Quantitative Structure-Activity Relationships, 8, 204-209.

Jonsson, J., Eriksson, L., Hellberg, S., Sjöström, M., and Wold, S., (1989b), *A Multivariate Approach to Saccharide Quantitative Structure-Activity Relationships Exemplified by Two Series of 9-Hydroxyellipticine Glycosides*, Acta Chemica Scandinavica, 43, 286-289.


Jonsson, J., Eriksson, L., Hellberg, S., Lindgren, F., Sjöström, M., and Wold, S., (1991), *A Multivariate Representation and Analysis of DNA Sequence Data*, Acta Chemica Scandinavica, 45, 186-192.

Jonsson, J., (1992), *Quantitative sequence-activity modelling (QSAM)*, Ph. D. Thesis, Umeå University, Umeå, Sweden.

Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C., and Wold, S., (1993), *Quantitative Sequence-Activity Models (QSAM) – Tools for Sequence Design*, Nucleic Acids Research, 21, 733-739.

Kassidas, A., MacGregor, J.F., and Taylor, P.A., (1998), *Synchronization of Batch Trajectories Using Dynamic Time Warping*, AIChE Journal, 44, 864-875.

Kettaneh-Wold, N., MacGregor, J.F., Dayal, B., and Wold, S., (1994), *Multivariate Design of Process Experiments (M-DOPE)*, Chemometrics and Intelligent Laboratory Systems, 23, 39-50.

				5(9)
Uppgjord - Prepared		Datum - Date	Rev A	Dokumentnr - Document no
Godkänd - Approved		File/reference		

Kim, K.H., (1993), *Non-linear Dependence in Comparative Molecular Field Analysis*, Journal of Computer-Aided Molecular Design, 7, 71-82.

Kimura, T., Miyashita, Y., Funatsu, K., and Sasaki, S., (1996), *Quantitative Structure-Activity Relationships of the Synthetic Substrates for Elastase Enzyme Using Nonlinear Partial Least Squares Regression*, Journal of Chemical Information and Computer Science, 36, 185-189.

Kubinyi, H., (1990), *Quantitative Structure-Activity Relationships and Molecular Modelling in Cancer Research*, Journal Cancer Research & Clinical Oncology, 116, 529-537, 1990.

Könemann, H., (1981), *Quantitative Structure-Activity Relationships in Fish Studies. Part 1: Relationship for 50 Industrial Pollutants*, Toxicology, 19, 209-221.

Larsson, U., Carlson, R., and Leroy, J., (1993), *Synthesis of Amino Acids with Modified Principal Properties 1. Amino Acids with Fluorinated Side Chains*, Acta Chemica Scandinavica, 47, 380-390.

Lewi, P.J., (1995), Spectral Mapping of Drug-Test Specificities, In: H. van de Waterbeemd (Ed.), *Chemometric Methods in Molecular Design*, VCH, Weinheim.

Linderholm, J., and Lundberg, E., (1994), *Chemical Characterization of Various Archaeological Soil Samples using Main and Trace Elements Determined by Inductively Coupled Plasma Atomic Emission Spectrometry*, Journal of Archaeological Science, 21, 303-314.

Lindgren, Å., Sjöström, M., and Wold, S., (1996), *Quantitative Structure-Effect Relationships for Some Technical Nonionic Surfactants*, JAOCS, 7, 863-875.

Lindgren, Å., (2000), *Use of Multivariate Methods and DOE to Improve Industrial-Scale Production Quality of a Cellulose Derivative*. Journal of Chemometrics, 14, 657-665.

Linusson, A., Gottfries, J., Lindgren, F., and Wold, S., (2000), *Statistical Molecular Design of Building Blocks for Combinatorial Chemistry*, Journal of Medicinal Chemistry, 43, 1320-1328.

Linusson, A., (2000), *Efficient Library Selection in Combinatorial Chemistry*, Ph.D. Thesis, Umeå University, Umeå, Sweden.

Louwerse, D.J., Bates, A.A., Smilde, A.K., Koot, G.L.M., and Berndt, H., (1999), *PLS Discriminant Analysis with Contribution Plots to Determine Differences Between Parallel Batch Reactors in the Process Industry*, Chemometrics and Intelligent Laboratory Systems, 46, 197-206.

Lowe, G., (1995), *Combinatorial Chemistry*, Chemical Society Reviews, 24, 309-317.

Lundstedt, T., (1991), *A QSAR Strategy for Screening of Drugs and Predicting Their Clinical Activity*, Drug News & Perspectives, 4, 468-475.

Lundstedt, T., Carlson, R., and Shabana, R., (1987), *Optimum Conditions, for the Willgerodt-Kindler Reaction. 3. Amine Variation*, Acta Chemica Scandinavica, B41, 157-163.

Lundstedt, T., Clementi, S., Cruciani, G., Pastor, M., Kettaneh, N., Andersson, P.M., Linusson, A., Sjöström, M., Wold, S., and Nordén, B., (1997), *Intelligent Combinatorial Libraries*, in: H. van de Waterbeemd, B. Testa and G. Folkers, eds., *Computer-Assisted Lead Finding and Optimization, Current Tools for Medicinal Chemistry*, Wiley-VCH, Weinheim.

MacFie, H., Moore, P.B., and Wakeling, I., (1999), *Changes in the Sensory Properties and Consumer Preferences for Dessert Apples*, Apples and Pears Research Council, UK.

MacGregor, J.F., and Nomikos, P., (1992), *Monitoring Batch Processes*, Proceedings NATO Advanced Study Institute for Batch Processing Systems Eng., May 29 – June 7, 1992, Antalya, Turkey.


MacGregor, J.F., and Kourti, T., (1995), *Statistical Process Control of Multivariate Processes*, Control Eng. Practice, 3, 403-414.

MacGregor, J.F., (1996), *Using On-Line Process Data to Improve Quality*, ASQC Statistics Division Newsletter, 16, 6-13.

Martens, H., and Naes, T., (1989), *Multivariate Calibration*, John Wiley, New York.

Martens, H., and Martens, M., (2000), *Modified Jack-Knife Estimation of Parameter Uncertainty in Bilinear Modeling (PLSR)*, Food Quality and Preference, 11, 5-16.

Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K., and Moos, W.H., (1995), *Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery*, Journal of Medicinal Chemistry, 38, 1431-1436.

				6(9)
Uppgjord - Prepared		Datum - Date	Rev A	Dokumentnr - Document no
Godkänd - Approved		File/reference		

Marvanova, S., Nagata, Y., Wimmerova, M., Sykorova, J., Hynkova, K., and Damborsky, J., (2001), *Biochemical Characterization of Broad-specificity Enzymes Using Multivariate Experimental Design and a Colorimetric Microplate Assay: Characterization of the Haloalkane Dehalogenase Mutants*, Journal of Microbiological Methods, 44, 149-157.

Massart, B., (1997), *Environmental Monitoring and Forecasting by Means of Multivariate Methods*, Ph.D. Thesis, University of Bergen, Bergen, Norway.

Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y., and Kaufman, L., (1988), *Chemometrics: A Textbook*, Elsevier.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., Lewi, P.J., and Smeyers-Verbeke, J., (1998), *Handbook of Chemometrics and Qualimetrics*, Elsevier.

McCloskey, J.T., Newman, M.C., and Clark, S.B., (1996), *Predicting the Relative Toxicity of Metal Ions Using Ion Characteristics: Microtox Bioluminescence Assay*, Environmental Toxicology and Chemistry, 15, 1730-1737.

McEwan, J.A., Earchy, P.J., and Ducher, C., (1998), *Preference Mapping – A Review*, Review No. 6, Campden & Chorleywood Food Research Association, Gloucestershire, UK.

McEwan, J.A., and Ducher, C., (1998), *Preference Mapping – Case Studies*, Review No. 7, Campden & Chorleywood Food Research Association, Gloucestershire, UK.

Michailidis, G., and de Leeuw, J., (1998), *The GIFI System of Descriptive Multivariate Analysis*, Statistical Science, 13, 307-336.

Mullet, G.M., (1976), *Why Regression Coefficients have the Wrong Sign*, Journal of Quality Technology, 8, 121-126.

Munck, L., Nørgaard, L., Engelsen, S.B., Bro, R., and Andersen, C., (1998), *Chemometrics in Food Science – A Demonstration of the Feasibility of a Highly Exploratory, Inductive Evaluation Strategy of Fundamental Scientific Significance*, Chemometrics and Intelligent Laboratory, 44, 31-60.

Naes, T., and Indahl, U., (1998), *A Unified Description of Classical Classification Methods for Multicollinear Data*, Journal of Chemometrics, 12, 205-220.

Nijhuis, A., de Jong, S., and Vandeginste, B.G.M., (1997), *Multivariate Statistical Process Control in Chromatography*, Chemometrics and Intelligent Laboratory Systems, 38, 51-62.

Nomikos, P., and MacGregor, J.F., (1995a), *Multivariate SPC Charts for Monitoring Batch Processes*, Technometrics, 37, 41-59.

Nomikos, P., and MacGregor, J.F., (1995b), *Multiway Partial Least Squares in Monitoring of Batch Processes*, Chemometrics and Intelligent Laboratory System, 30, 97-108.

Nomizu, M., Iwaki, T., Yamashita, T., Inagaki, Y., Asano, K., Akamatsu, M., and Fujita, T., (1993), *Quantitative Structure-Activity Relationship (QSAR) Study of Elastase Substrates and Inhibitors*, International Journal of Peptide and Protein Research, 42, 216-226.

Nordahl, Å., and Carlson, R., (1993), *Exploring Organic Synthetic Procedures*, Topics in Current Chemistry, 166, 1-64.

Norinder, U., (1996), *Single and Domain Mode Variable Selection in 3D QSAR Applications*, Journal of Chemometrics, 10, 95-105.


Nouwen, J., Lindgren, F., Hansen, B., Karcher, W., Verhaar, H.J.M., and Hermens, J.L.M., (1997), *Classification of Environmentally Occurring Chemicals Using Structural Fragments and PLS Discriminant Analysis*, Environmental Science and Technology, 31, 2313-2318.

Nyström, Å., Andersson, P.M., and Lundstedt, T., (2000), *Multivariate data analysis of topographically modified α -melanotropin analogues using auto- and cross-auto covariances*, Quantitative Structure-Activity Relationships, 19, 264-269.

Oprea, T.I., (2000), *Property distribution of drug-related chemical data-bases*, Journal of Computer-Aided Molecular Design, 14, 251-264.

Oprea, T.I., and Gottfries, J., (2001), *Chemography: The art of navigating in chemical space*, Journal of Computer-Aided Molecular Design, Submitted for publication.

Phatak, A., and DeJong, S., (1997), *The Geometry of PLS*, Journal of Chemometrics, 11, 311-338.

				7(9)
Uppgjord - Prepared		Datum - Date	Rev A	Dokumentnr - Document no
Godkänd - Approved		File/reference		

Qin, S.J., and McAvoy, T.J., (1992), *Non-Linear PLS Modelling Using Neural Networks*, *Computation and Chemical Engineering*, 16, 379-391.

Ramos, E.U., Vaes, W.H.J., Verhaar, H.J.M., and Hermens, J.L.M., (1997), *Polar Narcosis: Designing a Suitable Training Set for QSAR Studies*, *Environmental Science & Pollution Research*, 4, 83-90.

Rius, A., Ruisanchez, I., Callao, M.P., Rius, F.X., (1998), *Reliability of Analytical Systems: Use of Control Charts, Time Series Models and Recurrent Neural Networks*, *Chemometrics and Intelligent Laboratory Systems*, 40, 1-18.

Rännar, S., MacGregor, J.F., and Wold, S., (1998), *Adaptive Batch Monitoring Using Hierarchical PCA*, *Chemometrics and Intelligent Laboratory Systems*, 41, 73-81.

Sandberg, M., Sjöström, M., and Jonsson, J., (1996), *A Multivariate Characterization of tRNA Nucleosides*, *Journal of Chemometrics*, 10, 493-508.

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S., (1998), *New Chemical Dimensions Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids*, *Journal of Medicinal Chemistry*, 41, 2481-2491.

Savitzky, A., and Golay, M.J.E., (1964), *Smoothing and Differentiation by Simplified Least Squares Procedures*, *Analytical Chemistry*, 36, 1627-1632.

Schaper, K.J., (1991), *QSAR Analysis of Chiral Drugs Including Stereoisomer Combinations*, In: Silipo, C., and Vittoria, A, (Eds.), *QSAR: Rational Approaches to the Design of Bioactive Compounds*, Elsevier Science Publishers, Amsterdam, 1991, pp. 25-32.

Schüürmann, G., Rayasamuda, K., and Kristen, U., (1996), *Structure-Activity Relationships for Chloro- and Nitrophenol Toxicity in the Pollen Tube Growth Test*, *Environmental Toxicology and Chemistry*, 15, 1702-1708.

Sekulic, S., Seasholtz, M.B., Wang, Z., Kowalski, B., Lee, S.E., and Holt, B.R, (1993), *Non-linear Multivariate Calibration Methods in Analytical Chemistry*, *Analytical Chemistry*, 65, 835-845.

Seydel, J.K., Wiese, M., Cordes, H.P., Chi, H.L., Schaper, K.-J., Coats, E.A., Kunz, B., Engel, J., Kutscher, B., and Emig, H., (1991), *QSAR and Modelling of Enzyme Inhibitors, Anticonvulsants and Amphiphilic Drugs Interacting with Membranes*, In: Silipo, C., and Vittoria, A, (Eds.), *QSAR: Rational Approaches to the Design of Bioactive Compounds*, Elsevier Science Publishers, Amsterdam, 1991, pp. 367-376.

Shao, J., (1993), *Linear Model Selection by Cross-Validation*, *Journal of the American Statistical Association*, 88, 486-494.

Shewhart, W., (1931), *Economic Control of Quality of Manufactured Product*, Van Nostrand, Princeton, N.J.

Sjöström, M., Wold, S., Lindberg, W., Persson, J.-Å., and Martens, H., (1983), *A Multivariate Calibration Problem in Analytical Chemistry Solved by Partial Least-Squares Models in Latent Variables*, *Analytica Chimica Acta*, 150, 61-70.

Sjöström, M., Wold, S., and Söderström, B., (1986), *PLS Discriminant Plots*, *Proceedings of PARC in Practice*, Amsterdam, June 19-21, 1985. Elsevier Science Publishers B.V., North-Holland.


Sjöström, M., Eriksson, L., Hellberg, S., Jonsson, J., Skagerberg, B., and Wold, S., (1989), *Peptide QSARs: PLS Modelling and Design in Principal Properties*. In: J.L. Fauchère (ed.): *QSAR - Quantitative Structure-Activity Relationships in Drug Design*. Proc. 7th European Symposium on QSAR, Sept. 1988, Interlaken, Switzerland. Alan R. Liss, Inc., New York, pp. 131-134.

Sjöström, M., Rännar, S., and Rilfors, L., (1995), *Polypeptide sequence property relationships in Escherichia coli based on auto cross covariances*, *Chemometrics and Intelligent Laboratory Systems*, 29, 295-305.


Sjöström, M., Lindgren, Å., and Uppgård, L.-L., (1997), *Joint Multivariate Quantitative Structure-Property and Structure-Activity Relationships for a Series of Technical Nonionic Surfactants*, In: F. Chen & G. Schüürmann (eds.), *Quantitative Structure-Activity Relationships in Environmental Sciences – VII. Proceedings of the 7th International Workshop on QSAR in Environmental Sciences*, June 24-28, 1996, Elsinore, Denmark. SETAC Press, Pensacola, Florida, 1997, pp. 435-449.

Skagerberg, B., Sjöström, M., and Wold, S., (1987), *Multivariate Characterization of Amino Acids by Reversed Phase High Pressure Liquid Chromatography*, *Quantitative Structure-Activity Relationships*, 6, 158-164.

Skagerberg, B., Bonelli, D., Clementi, S., Cruciani, G., and Ebert, C., (1989), *Principal Properties for Aromatic Substituents. A Multivariate Approach for Design in QSAR*, *Quantitative Structure-Activity Relationships*, 8, 32-38.

				8(9)
Uppgjord - Prepared		Datum - Date	Rev	Dokumentnr - Document no
Godkänd - Approved		File/reference		
		A		

- Stork, C.L., and Kowalski, B.R., (1999), *Distinguishing Between Process Upsets and Sensor Malfunctions Using Sensor Redundancy*, Chemometrics and Intelligent Laboratory Systems, 46, 117-131.
- Strouf, O., (1986), *Chemical Pattern Recognition*, John Wiley, New York.
- Stähle, L., and Wold, S., (1987), *Partial Least Squares Analysis with Cross-Validation for the Two-Class Problem: A Monte Carlo Study*, Journal of Chemometrics, 1, 185-196.
- Svensson, O., Josefsson, M., and Langkilde, F.W., (1997), *Classification of Chemically Modified Celluloses Using a Near-Infrared Spectrometer and Soft Independent Modelling of Class Analogy*, Applied Spectroscopy, 51, 1826-1835.
- Taft, R.W., (1956), in: Newman, M.S., (ed.), *Steric Effects in Organic Chemistry*, Wiley, New York.
- Tano, K., (1996), *Multivariate Modelling and Monitoring of Mineral Processes using Partial Least Squares Regression*, Licentiate Thesis, Luleå University of Technology, Sweden.
- Teague, S.J., Davis, A.M., Leeson, P.D., and Oprea, T., (1999), *The Design of Leadlike Combinatorial Libraries*, Angewandte Chemie International Edition, 38, 3743-3748.
- Teppola, P., Mujunen, S.P., Minkinen, P., Puijola, T., and Pursiheimo, P., (1998), *Principal Component Analysis, Contribution Plots and Feature Weights of Sequential Process Data from a Paper Machine's Wet End*, Chemometrics and Intelligent Laboratory Systems, 44, 307-317.
- Teppola, P., and Minkinen, P., (2000), *Wavelet-PLS Regression Models for both Exploratory Data Analysis and Process Monitoring*, Journal of Chemometrics, 14, 383-400.
- Teppola, P., and Minkinen, P., (2001), *Wavelets for Scrutinizing Multivariate Exploratory Models Through Multiresolution Analysis*, Journal of Chemometrics, 15, 1-18.
- Tompson, C.J.S., (1990), *The Lure and Romance of Alchemy - a History of the Secret Link Between Magic and Science*, Bell Publishing Company, New York.
- Topliss, J.G., and Edwards, R.P., (1979), *Chance Factors in Studies of Quantitative Structure-Activity Relationships*, Journal of Medicinal Chemistry, 22, 1238-1244.
- Trygg, J., and Wold, S., (1998), *PLS Regression on Wavelet Compressed NIR Spectra*, Chemometrics and Intelligent Laboratory Systems, 42, 209-220.
- Trygg, J., and Wold, S., (2001), *Orthogonal Projections to Latent Structures, OPLS*, Journal of Chemometrics. In press.
- Trygg, J., Kettaneh-Wold, N., and Wallbäcks, L., (2001), *2-D Wavelet Analysis and Compression of On-line Industrial Process Data*, Journal of Chemometrics, 15, 299-319.
- Tysklind, M., Andersson, P., Haglund, P., van Bavel, B., and Rappe, C., (1995), *Selection of Polychlorinated Biphenyls for use in Quantitative Structure-Activity Modelling, SAR and QSAR in Environmental Research*, 4, 11-19.
- Uppgård, L., Sjöström, M., and Wold, S., (2000), *Multivariate Quantitative Structure-Activity Relationships for the Aquatic Toxicity of Alkyl Polyglucosides*, Tenside Surfactants Detergents, 37, 131-138.
- Wakeling, I.N., Morris, J.J., (1993), *A Test of Significance for Partial Least Squares Regression*, Journal of Chemometrics, 7, 291-304.
- Van der Voet, H., (1994), *Comparing the Predictive Accuracy of Models Using a Simple Randomization Test*, Chemometrics and Intelligent Laboratory Systems, 25, 313-323.
- Van de Waterbeemd, H., (1995), *Chemometric Methods in Molecular Design*, In: Mannhold, R., Krogsgaard-Larsen, P., and Timmerman, H., (Eds.), *Methods and Principles in Medicinal Chemistry*, Vol 2, VCH, Weinheim, Germany.
- Westerhuis, J.A., de Jong, S., and Smilde, A.K., (2001), *Direct Orthogonal Signal Correction*, Chemometrics and Intelligent Laboratory Systems, 56, 13-25.
- Wikström, C., Albano, C., Eriksson, L., Fridén, H., Johansson, E., Nordahl, Å., Rännar, S., Sandberg, M., Kettaneh-Wold, N., and Wold, S., (1998a), *Multivariate Process and Quality Monitoring Applied to an Electrolysis Process – Part I. Process Supervision with Multivariate Control Charts*, Chemometrics and Intelligent Laboratory Systems, 42, 221-231.
- Wikström, C., Albano, C., Eriksson, L., Fridén, H., Johansson, E., Nordahl, Å., Rännar, S., Sandberg, M., Kettaneh-Wold, N., and Wold, S., (1998b), *Multivariate Process and Quality Monitoring Applied to an Electrolysis Process – Part II. Multivariate Time-series Analysis of Lagged Latent Variables*, Chemometrics and Intelligent Laboratory Systems, 42, 233-240.

				9(9)
Uppgjord - Prepared		Datum - Date	Rev	Dokumentnr - Document no
Godkänd - Approved		A		
		File/reference		

- Wikström, P.B., Andersson, A.-C., Forsman, M., (1999), *Biomonitoring Complex Microbial Communities Using Random Amplified Polymorphic DNA and PCA*, FEMS – Microbiology, Ecology, 28, 131-139.
- Wikström, P.B., (2001), *Biomonitoring of Complex Microbial Communities that Biodegrade Aromatics*, Ph.D. Thesis, Umeå University, Umeå, Sweden.
- Winiwarter, S., Bonham, N.M., Ax, F., Hallberg, A., Lennernäs, H., and Karlén, A., (1998), *Correlation of Human Jejunal Permeability (in Vivo) of Drugs with Experimentally and Theoretically Derived Parameters – A Multivariate Data Analysis Approach*, Journal of Medicinal Chemistry, 41, 4939-4949.
- Wise, B.M., Gallagher, N.B, and Martin, E.B., (2001), *Application of PARAFAC2 to Fault Detection and Diagnosis in Semiconductor Etch*, Journal of Chemometrics, 15, 285-298.
- Wold, S., (1978), *Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models*, Technometrics, 20, 397-405.
- Wold, H., (1982), *Soft Modelling. The Basic Design and Some Extensions*. In: Jöreskog, K.G., and Wold, H., (eds.), *Systems Under Indirect Observation*, Vol. I and II, North-Holland, Amsterdam, The Netherlands.
- Wold, S., and Dunn, III, W.J., (1983), *Multivariate Quantitative Structure-Activity Relationships (QSAR): Conditions for their Applicability*, J. Chem. Inf. Comp. Sci., 23, 6-13.
- Wold, S., Albano, C., Dunn, W.J., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W., and Sjöström, M., (1984), *Multivariate Data Analysis in Chemistry*, In: B.R. Kowalski (ed.) *Chemometrics: Mathematics and Statistics in Chemistry*, D. Reidel Publishing Company, Dordrecht, Holland.
- Wold, S., Geladi, P., Esbensen, K., and Öhman, J., (1987), *Multiway Principal Components and PLS-Analysis*, Journal of Chemometrics, 1, 41-56.
- Wold, S., Carlson, R., and Skagerberg, B., (1989a), *Statistical Optimization as a Means to Reduce Risks in Industrial Processes*, The Environmental Professional, 11, 127-131.
- Wold, S., Kettaneh-Wold, N., and Skagerberg, B. (1989b), *Nonlinear PLS Modelling*, Chemometrics and Intelligent Laboratory Systems, 7, 53-65.
- Wold, S., (1992), *Non-linear Partial Least Squares Modelling. II. Spline Inner Relation*, Chemometrics and Intelligent Laboratory Systems, 14, 71-84.
- Wold, S., Johansson, E., and Cocchi, M., (1993a), *PLS*, In: Kubinyi, H., (ed.), *3D-QSAR in Drug Design, Theory, Methods, and Applications*, ESCOM Science, Ledien, pp. 523-550.
- Wold, S., Jonsson, J., Sjöström, M., Sandberg, M., and Rännar, S., (1993b), *DNA and Peptide Sequences and Chemical Processes Multivariately Modelled by Principal Components Analysis and Partial Least Squares Projections to Latent Structures*, Analytica Chimica Acta, 277, 239-253.
- Wold, S., (1994), *Exponentially Weighted Moving Principal Components Analysis and Projections to Latent Structures*, Chemometrics and Intelligent Laboratory Systems, 23, 149-161.
- Wold, S., Kettaneh, N., and Tjessem, K., (1996), *Hierarchical Multiblock PLS and PC Models for Easier Model Interpretation and as an Alternative to Variable Selection*, Journal of Chemometrics, 10, 463-482.
- Wold, S., Antti, H., Lindgren, F., and Öhman, J., (1998a), *Orthogonal Signal Correction of Near-Infrared Spectra*, Chemometrics and Intelligent Laboratory Systems, 44, 175-185.
- Wold, S., Kettaneh, N., Fridén, H., and Holmberg, A., (1998b), *Modelling and Diagnostics of Batch Processes and Analogous Kinetic Experiments*, Chemometrics and Intelligent Laboratory Systems, 44, 331-340.
- Wold, S., Sjöström, M., Eriksson, L., (1999), *PLS in Chemistry*, In: The Encyclopedia of Computational Chemistry, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer III, H. F.; Schreiner, P. R., Eds., John Wiley & Sons, Chichester, 1999, pp 2006-2020.
- Wold, S. and Josefson, M., (2000), *Multivariate Calibration of Analytical Data*, Encyclopedia of Analytical Chemistry, Wiley, pp 1-27.
- Wold, S., Sjöström, M., and Eriksson, L., (2001), *PLS-Regression: A Basic Tool of Chemometrics*, Journal of Chemometrics, Submitted.

