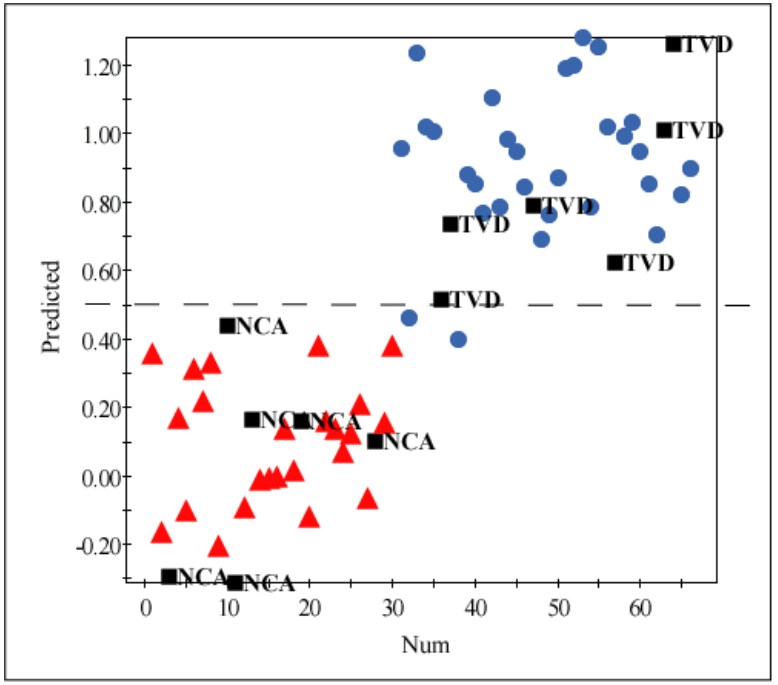
Multivariate data analysis (MVA) - Classification

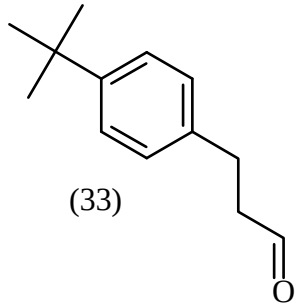
- **Introduction**
- **Methods**
  - -SIMCA
  - -PLS-DA
- **Example**
  - - Archaeologi (detail)
  - - Coronary Heart Disease (diagnosis)
  - - Human exercise study(GC/MS)
- **Conclusions**



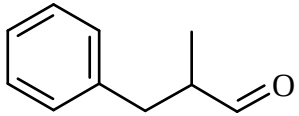
Classification

- Classification exists within all areas of society
- Different classes - different properties – different applications

E.g. Different classes of molecules – suitable for different medical drugs



(33)



(26)

- We need methods for detection of classes and for defining class identity for unknown samples.

Classification - methods

- **SIMCA** (**S**oft **I**ndependent **M**odelling of **C**lass **A**nalogy)

  - PCA of the whole data matrix (overview) gives information about classes, trends, outliers.

  - PCA model for each separate class of samples.

  - Cross validation (CV) decides number of components.

  - Prediction of unknown samples – Do they belong to a class or not ?

  - Class borders in **scores** and **DModX.**

Classification - methods

- **PLS - DA** (PLS - Discriminant Analysis)

  - If PCA doesn't give satisfactory separation between classes.

  - PLS - DA is based on PLS against a "dummy" matrix for separation.

Class A
Class B
Class C

PLS →

A	B	C
1	0	0
1	0	0
1	0	0
0	**1**	0
0	**1**	0
0	**1**	0
0	0	**1**
0	0	**1**
0	0	**1**

Classification - methods

Class A
Class B
Class C

PCA → Projection of ***maximum variation*** in X

Class A
Class B
Class C

PLS -DA →

A	**B**	**C**
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

Projection of ***maximum separation*** between classes in X

PCA, PLS - Classification

Example: Archeology

- Classification of archaeologly interesting samples
  - Previously done by univariate analysis
  - Now multivariate approach.

- 22 samples collected from a archaeology site ouside Mjölby, Sweden

- The samples were characterised by 18 variables, from ICP-AES and test of two concentration pre-treatments for (Fe, Cu, P, Mn, V, Co, Zn, Cr and Ca). This is of interest since untouched soil compared to occupied soiled have got different patterns of trace elements.

- The 22 samples were collected from three different sites in the area. F, S, C

F: Garbage tip, **S:** Occupation site, **C:** Control (untouched soil)

Data table

Two pre-treatments of samples prior to ICP-AES: Total Dissolution(TD) and Nitric Acid(NA)
9 interesting elements (Fe, Cu, P, Mn, V, Co, Zn, Cr and Ca).
Gives a total of 18 variables.

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

PCA – Whole data matrix (overview)

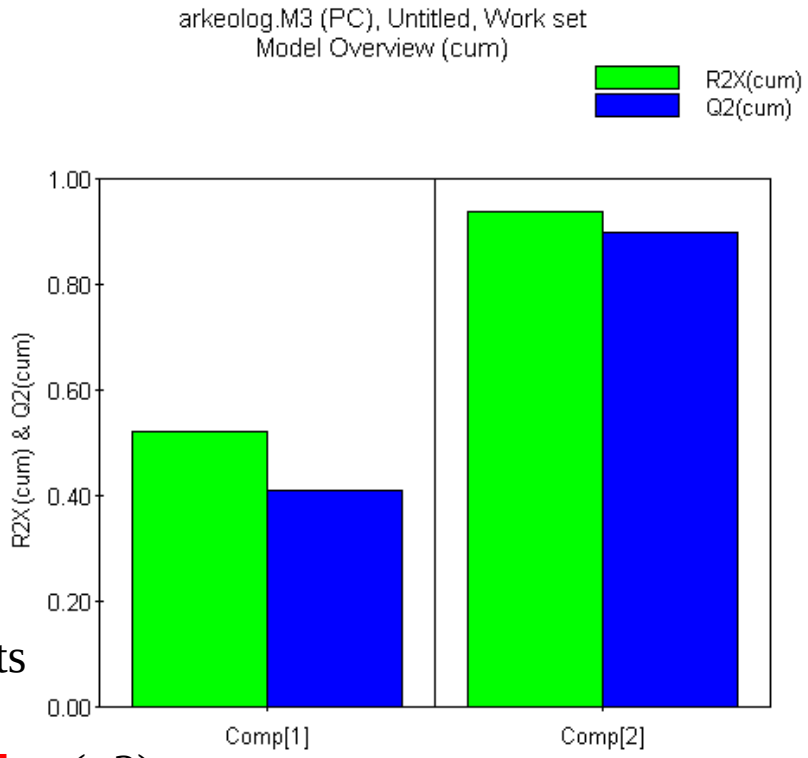
Model M3

Model: M3 Title: Untitled Properties...

Type: PC-X Observations (N) = 19, Variables (K) = 18 Work Set...

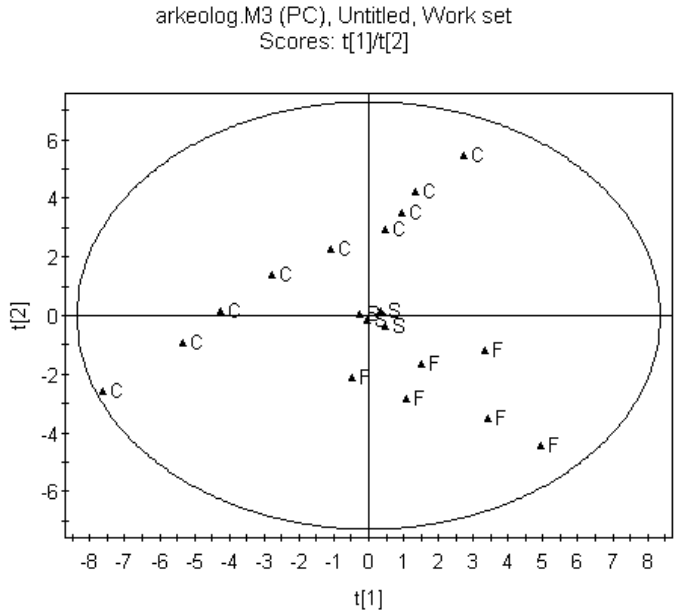
Components:

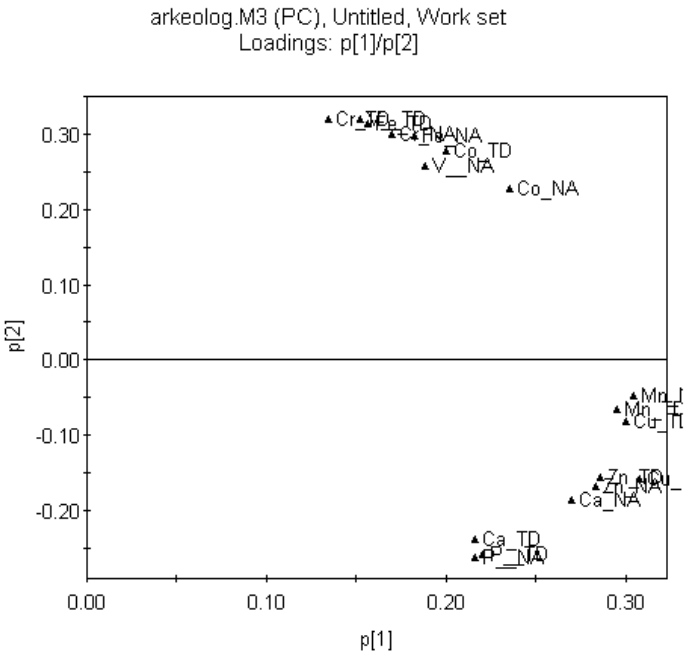
A	R2X	R2X(cum)	Eigen...	Q2	Limit	Q2(cum)	Signifi...	Iterations
00	-	0.000	-	-	-	-	--	--
01	0.523	0.523	9.411	0.409	0.102	0.409	R1	57
02	0.414	0.937	7.457	0.828	0.108	0.898	R1	4



- **Cross validation** gives two significant components
- Two components significant according to **eigenvalue** (>2)
- R2 = 0.93, Q2 = O.90

Interpretation of “scores” and “loadings”

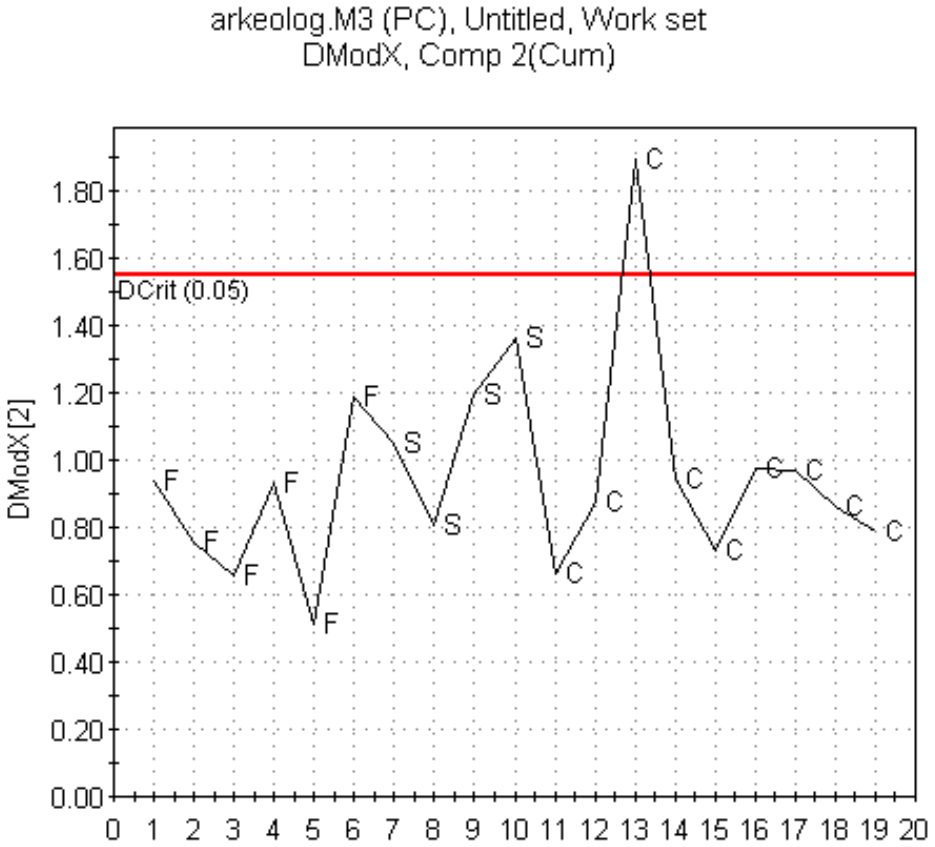




Class separation can be seen in the ‘scores’.

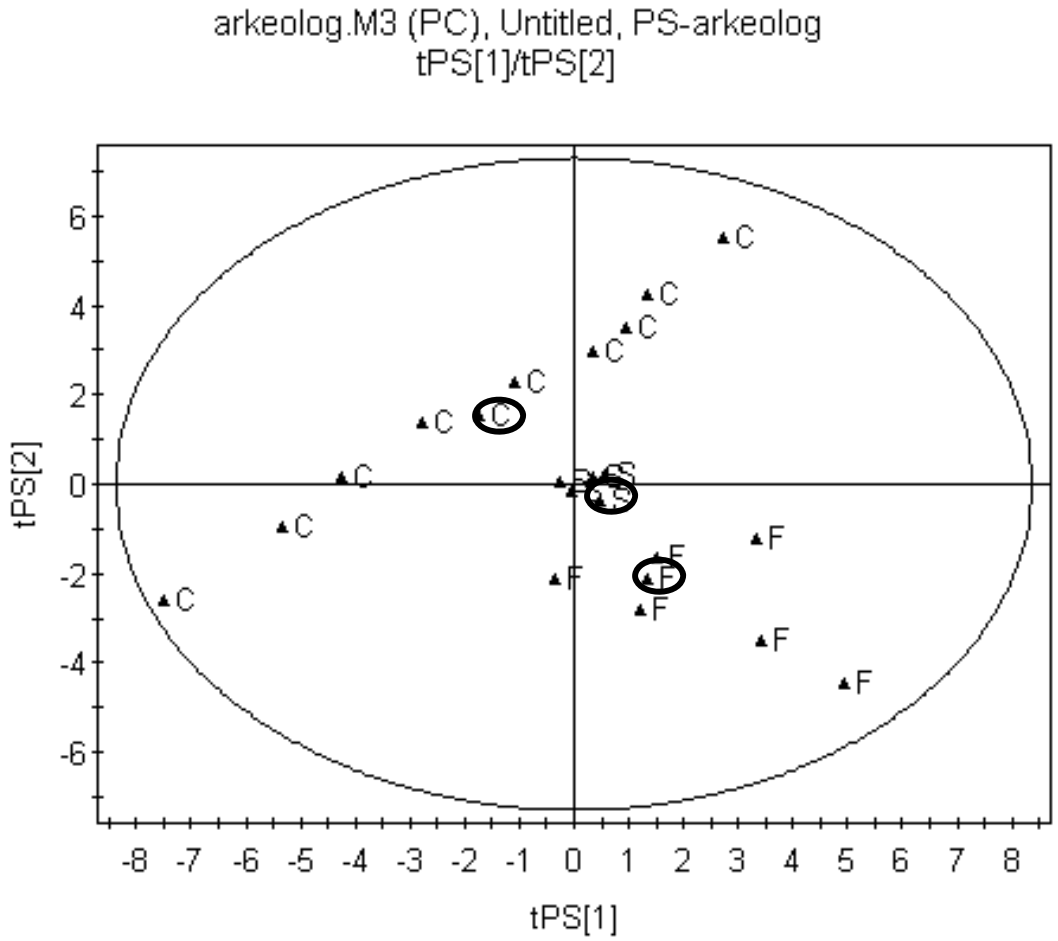
Explanation can be found in “loadings”

DModX



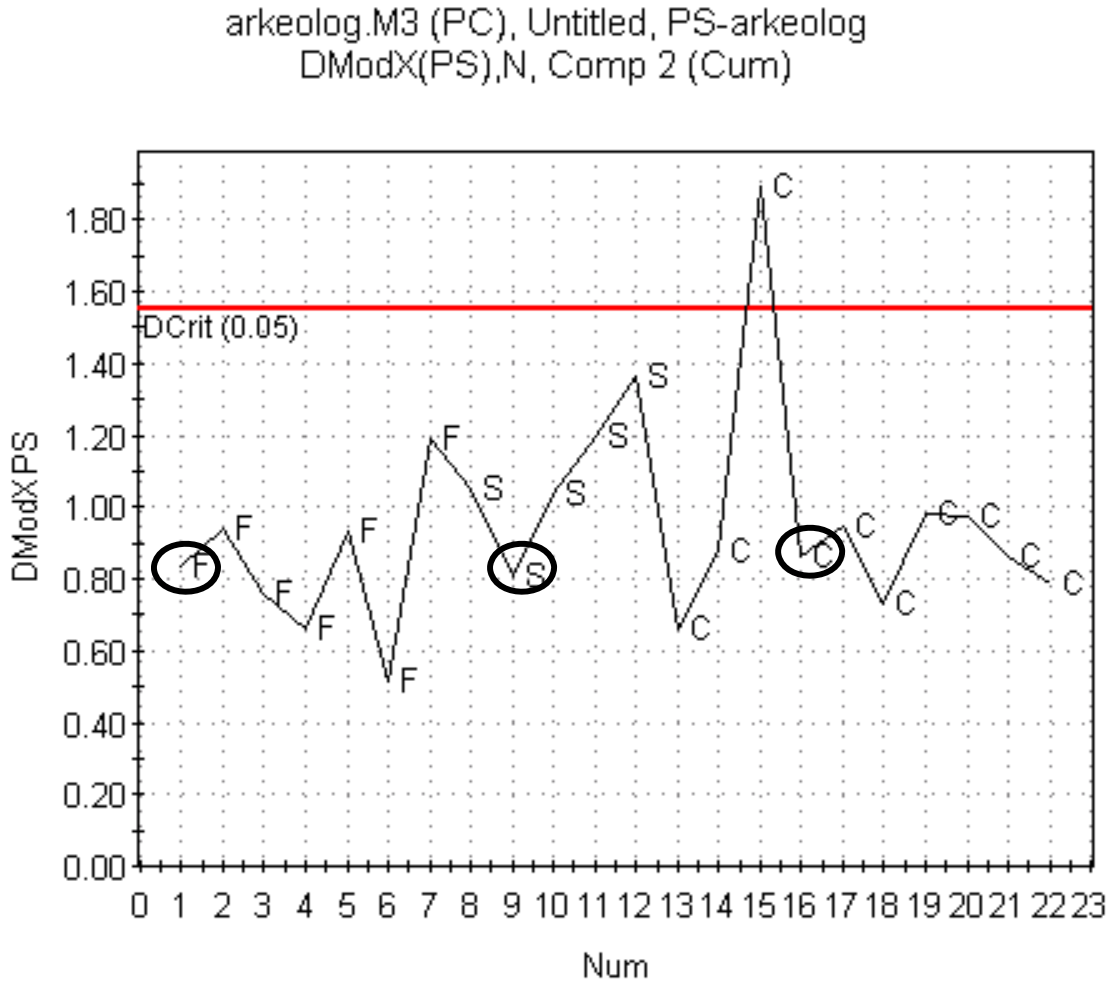
DModX doesn't reveal any extreme outliers.

Prediction of test samples



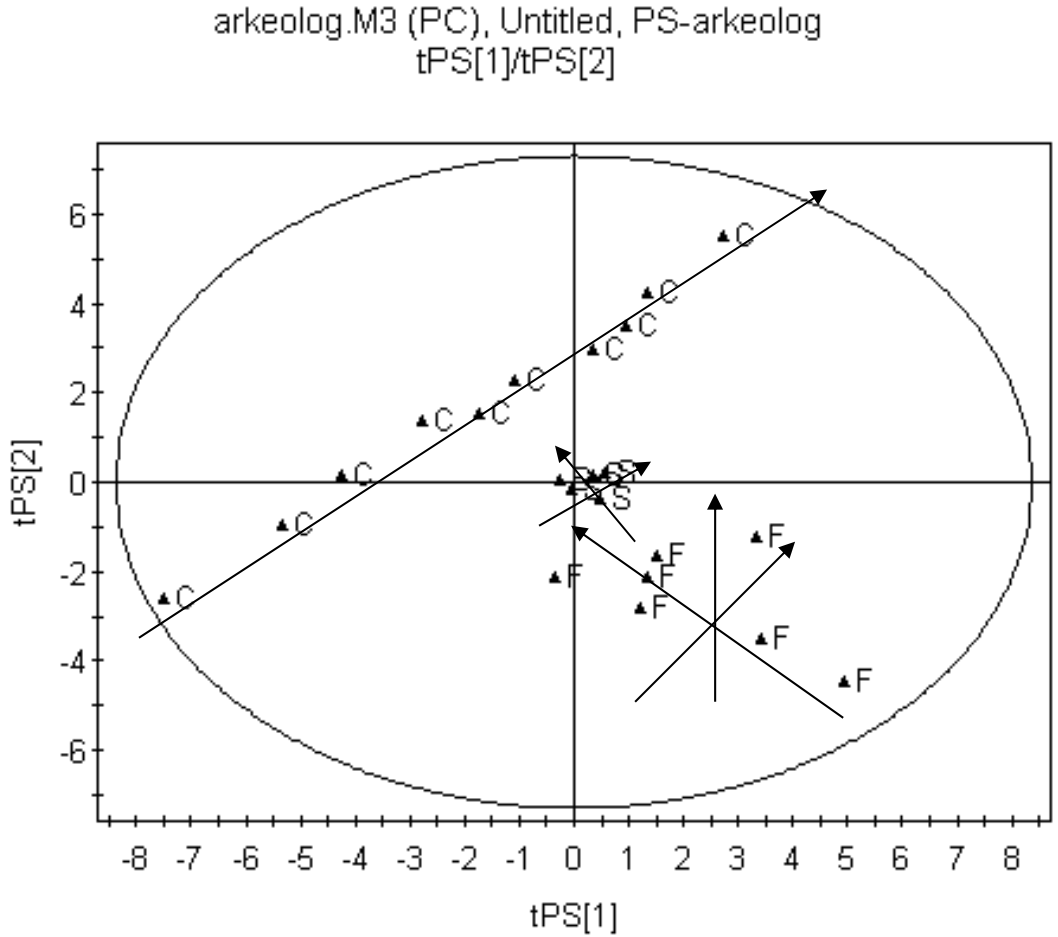
Prediction of three test samples (circled), one from each class, show that our total PCA model can provide information about class discrimination.

DModX test samples



Non of the test samples (circled) show signs of deviating in DModX

SIMCA - One PCA model for each class



By fitting a PCA-modell for each class confidence limits can be calculated for each single class which will lead more reliable predictions of class identity for new samples..

PCA - Class F

Model M4

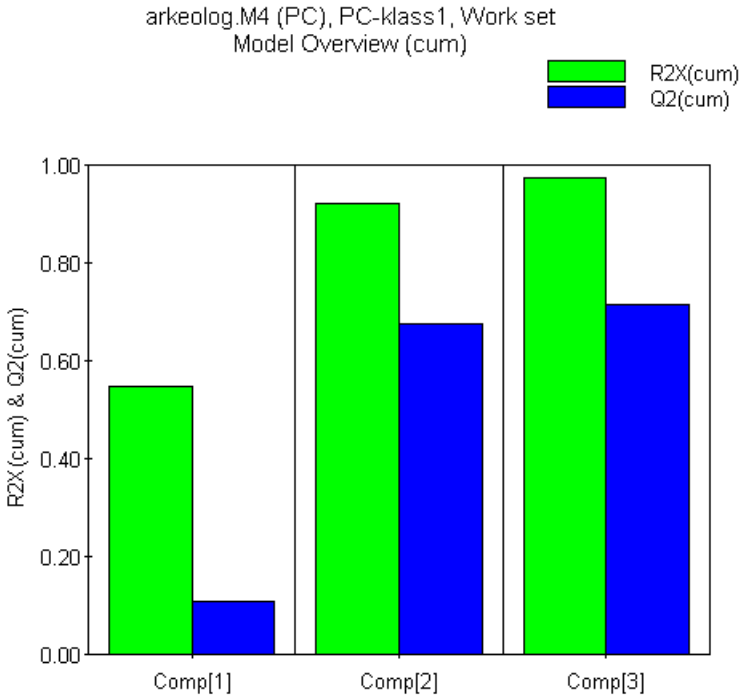
Model: M4 Title: PC-klass1

Type: PC-Cl:1 Observations (N) = 6, Variables (K) = 18

Properties... Work Set...

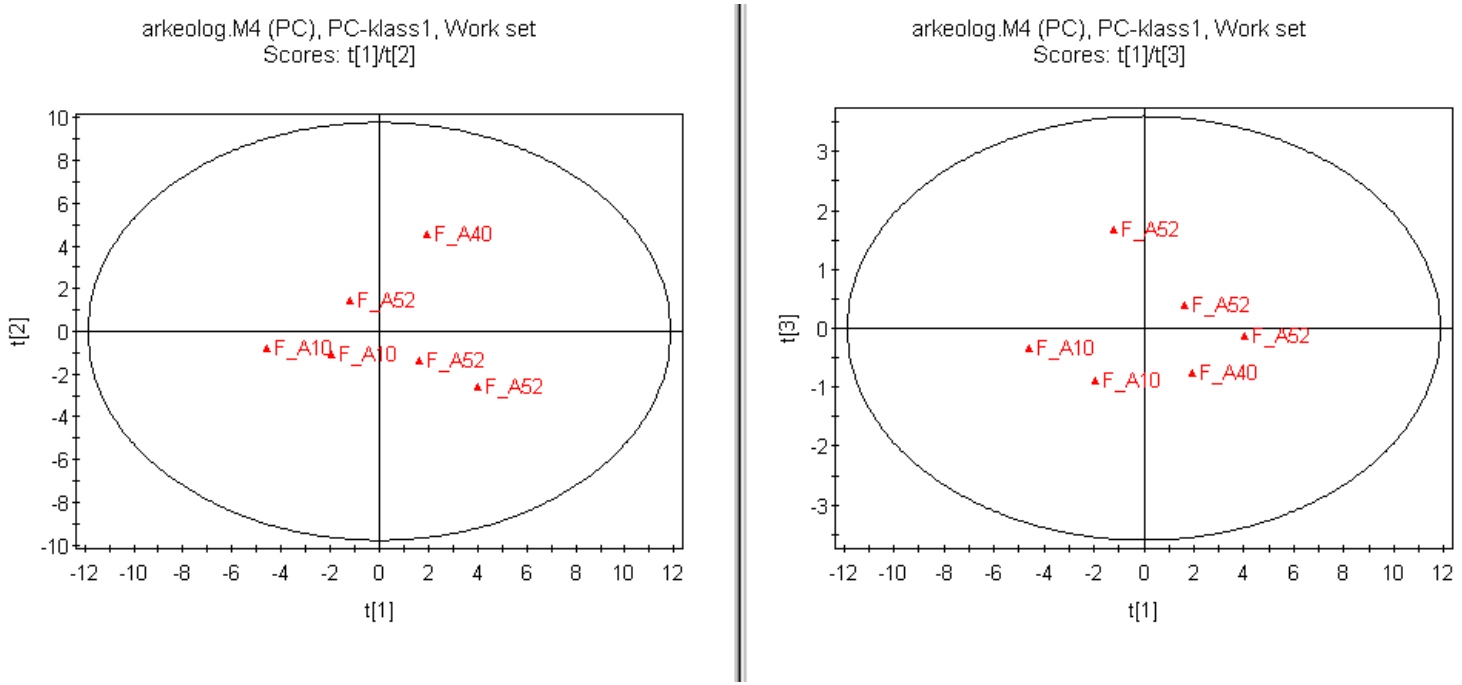
Components:

A	R2X	R2X(cum)	Eigen...	Q2	Limit	Q2(cum)	Signifi...	Iterations
00	-	0.000	-	-	-	-	--	--
01	0.547	0.547	3.281	0.108	0.211	0.108	R2	18
02	0.375	0.921	2.248	0.637	0.244	0.676	R1	9
03	0.051	0.972	0.304	0.118	0.294	0.714	R2	17



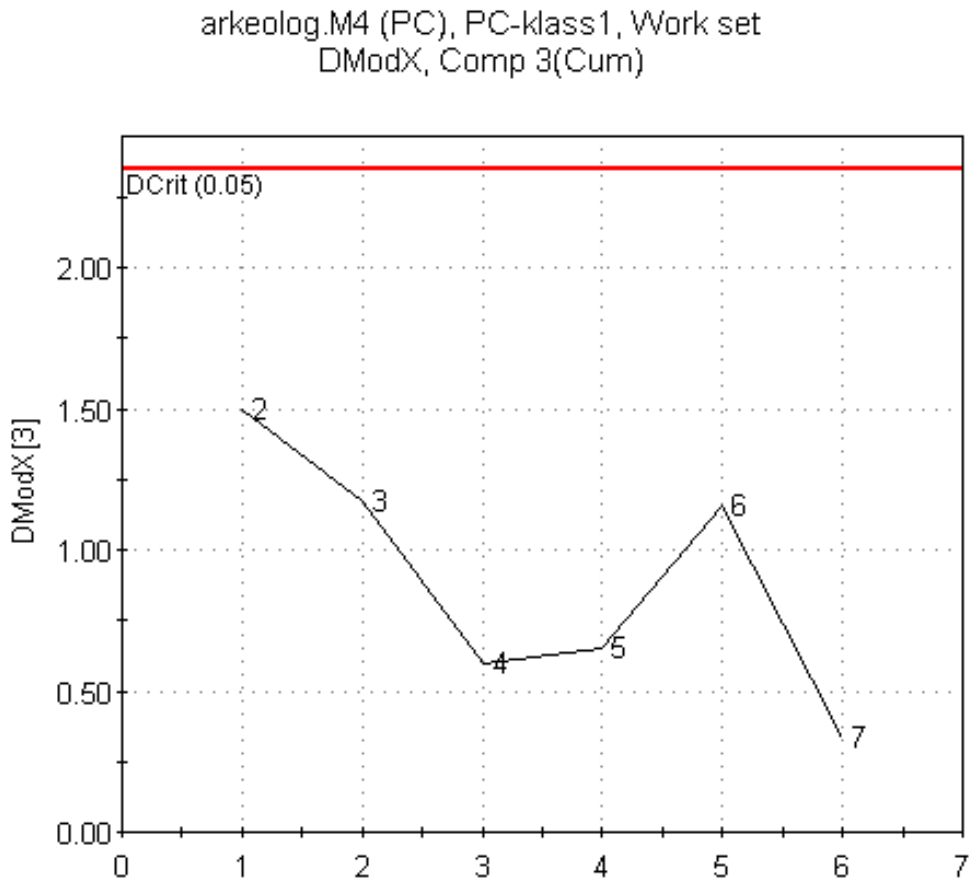
- Cross validation gives three significant components
- R2 = 0.97, Q2 = O.71
- Since the model will be used for prediction of class identity for new samples we leave it to the cross validation to decide on the relevant number of components.

“Scores” - Class F



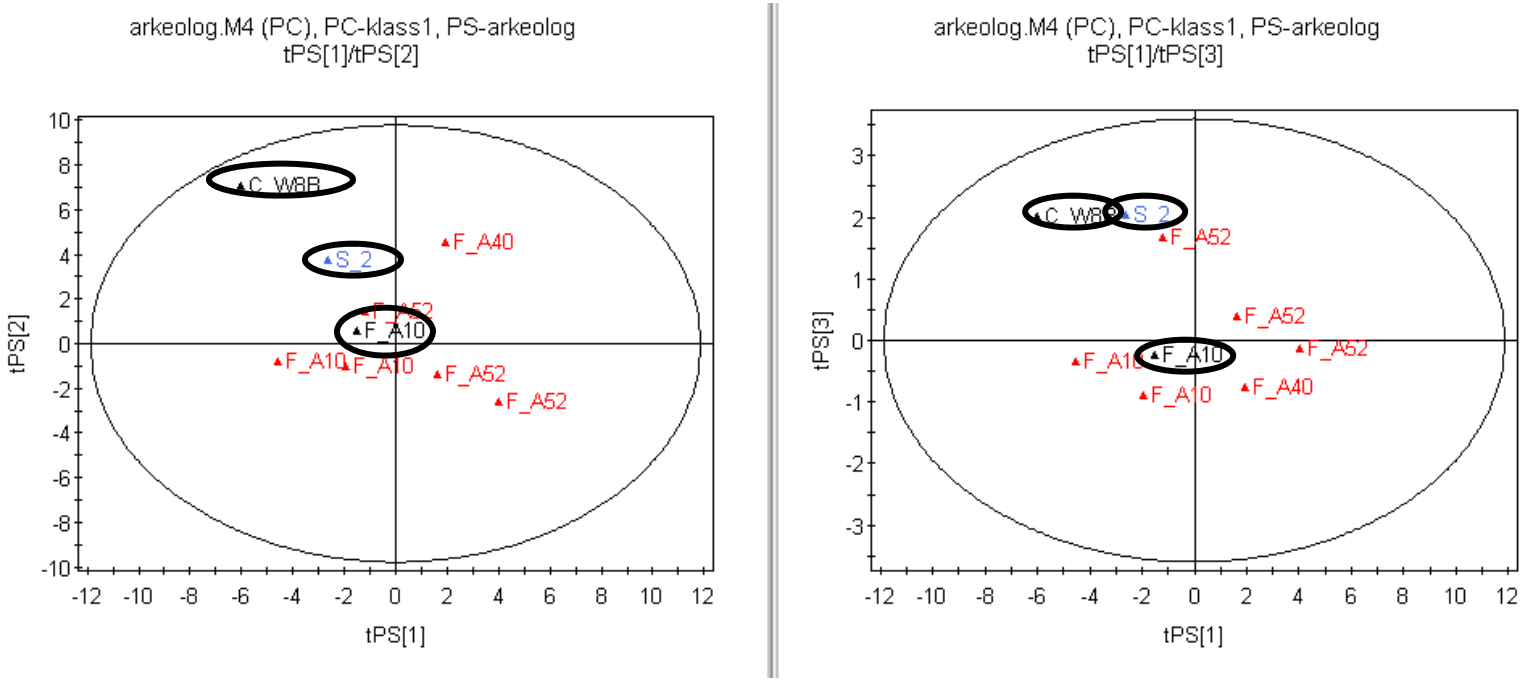
The spread of observations in all three components can be studied in “scores” (t1/t2, t1/t3). The class seem to be homogenous i.e. no outliers or groups within the class.

DModX - Class F



Non of the samples show any signs of deviating in DmodX.

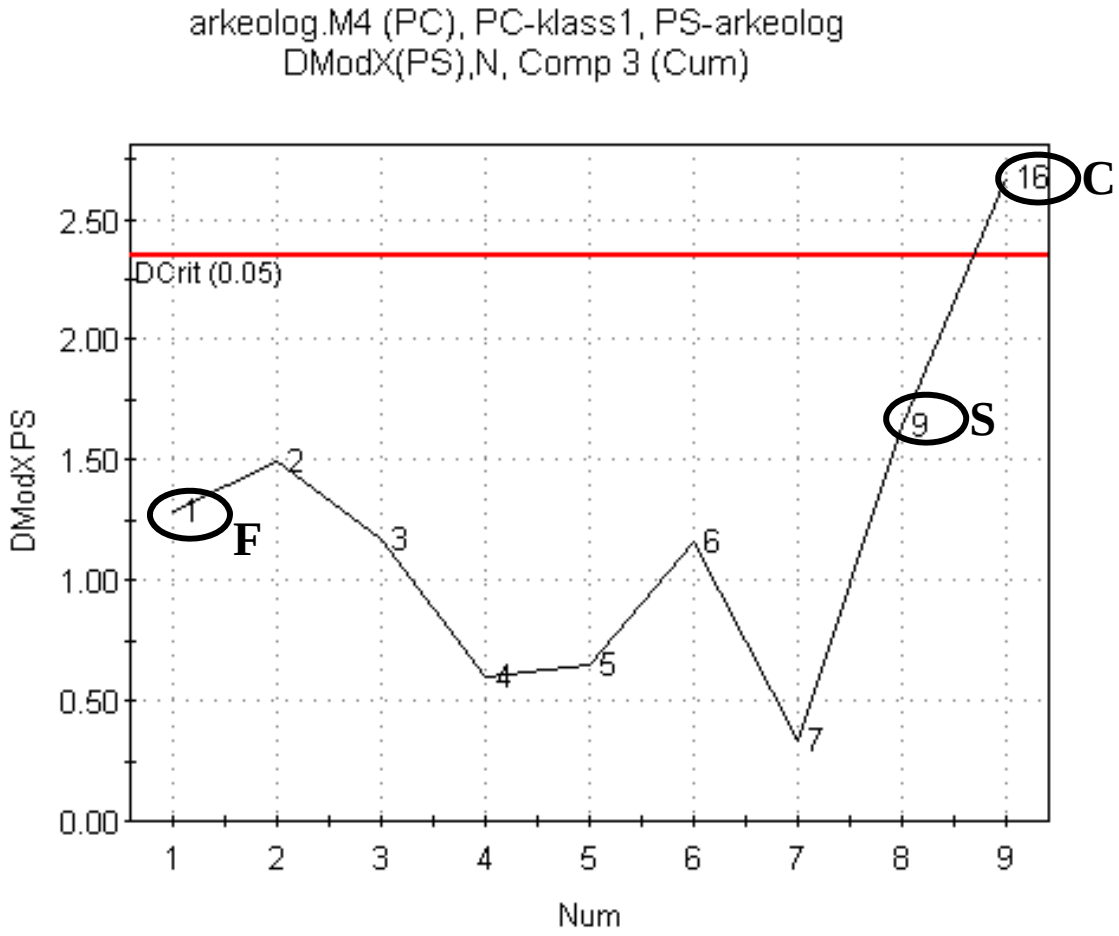
Prediction of test samples - Class F



Non of the test samples (circled) is predicted outside the limits for class F. The sample belonging to class F is predicted in the center of the class, while the samples belonging to class S and C are predicted further away from the observations in class F.

Due to the large spread between the samples in class F the confidence limits become less tight..

DModX predictions - Class F



The sample belonging to class C deviate somewhat in DModX, although not severe!
The sample belonging to class S does not deviate in DModX.

PCA - Class S

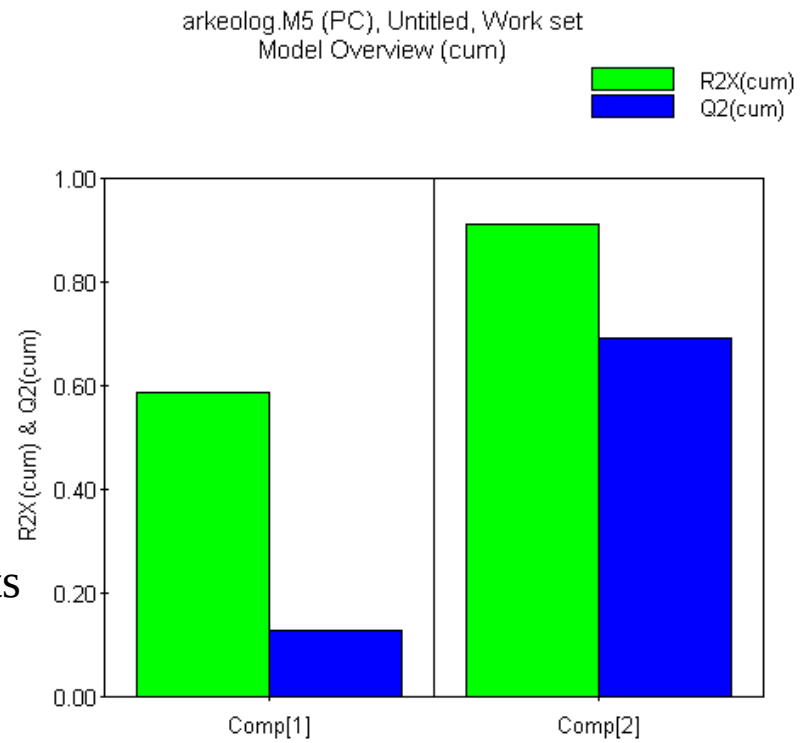
Model M5

Model: M5 Title: Untitled Properties...

Type: PC-CI:2 Observations (N) = 4, Variables (K) = 18 Work Set...

Components:

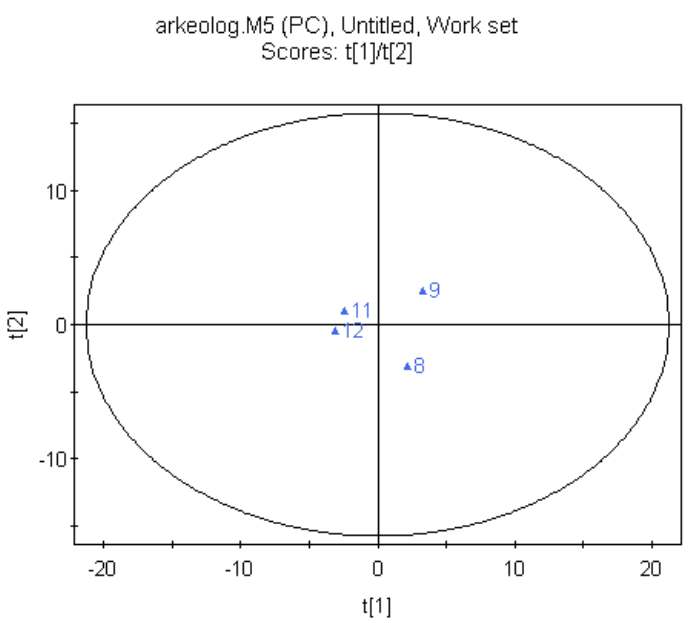
A	R2X	R2X(cum)	Eigen...	Q2	Limit	Q2(cum)	Signifi...	Iterations
00	-	0.000	-	-	-	-	--	--
01	0.588	0.588	2.350	0.126	0.289	0.126	R2	9
02	0.322	0.910	1.289	0.646	0.370	0.690	R1	11

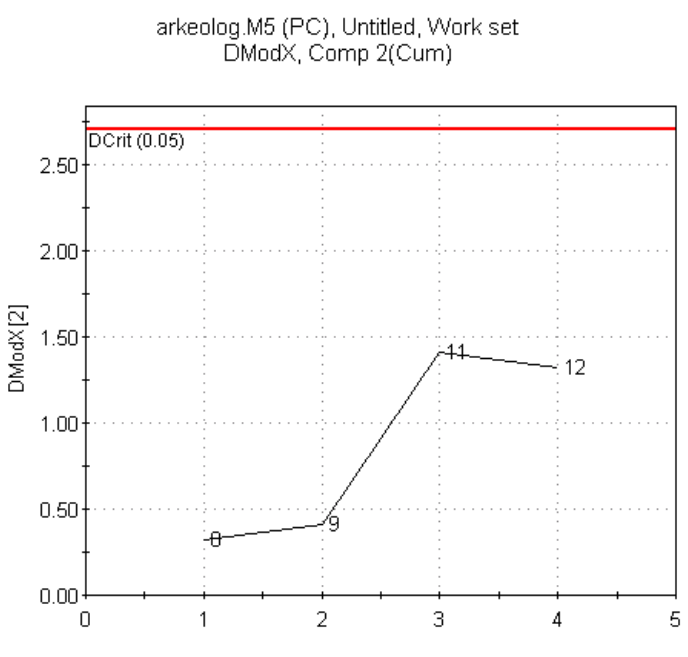


- Cross validation gives two significant components
- R2 = 0.91, Q2 = O.65
- Since the model will be used for prediction of class identity for new samples we leave it to the cross validation to decide on the relevant number of components.

.

“Scores”, DModX - Class S

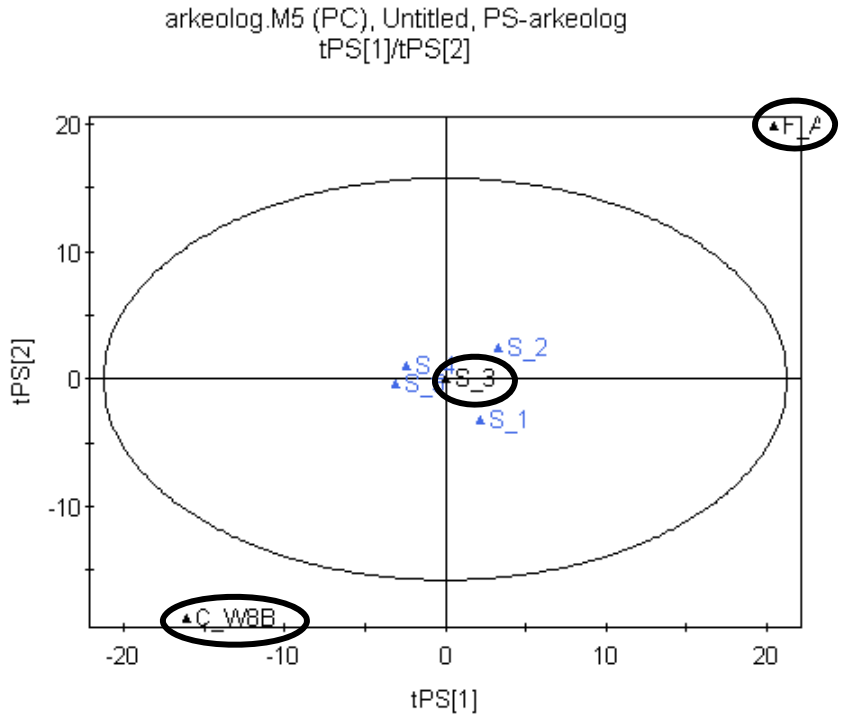




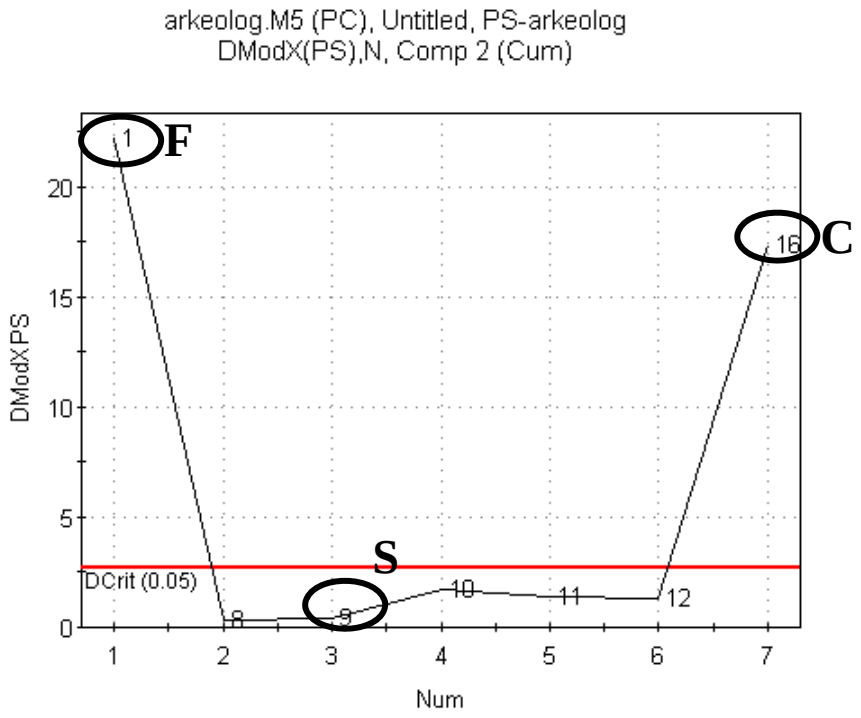
A homogenous group with little spread in scores.
No “outliers”!

No sample shows deviation in DModX!

Prediction of test samples - Class S



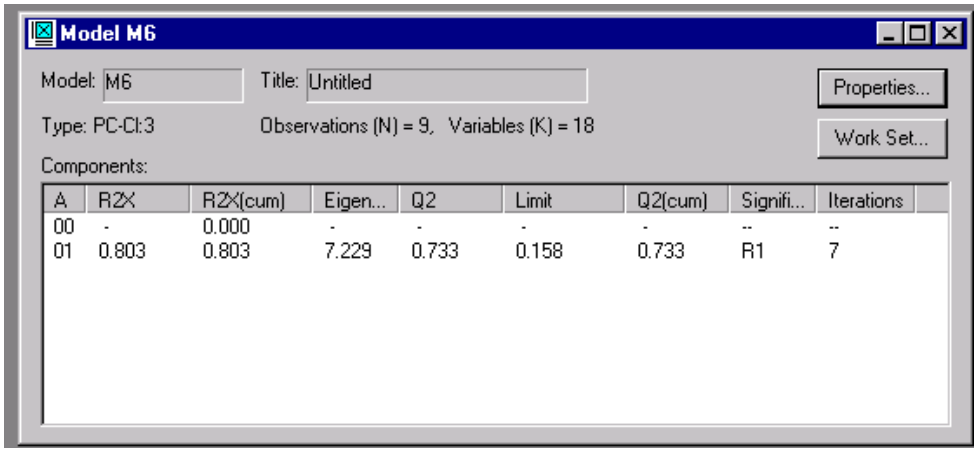
Predictions for the test samples (circled) Show that the samples not belonging to class S are predicted outside the limits for class S. The sample belonging to class S is predicted well within the class borders.

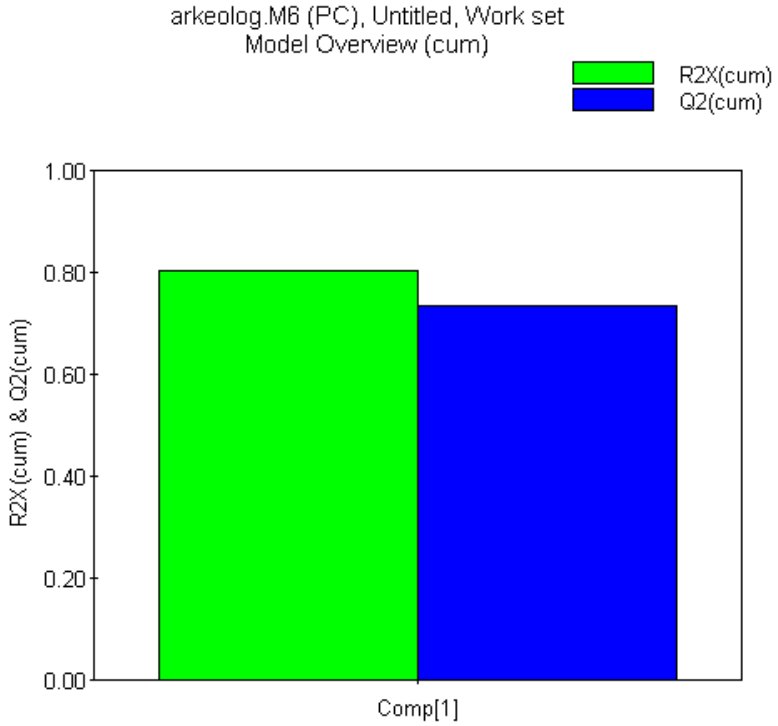


The test samplesbelonging to class F and C show large deviations in DModX.

We can clearly distinguish them from class S. Just as clearly we can also say that the third test sample belongs to class S.

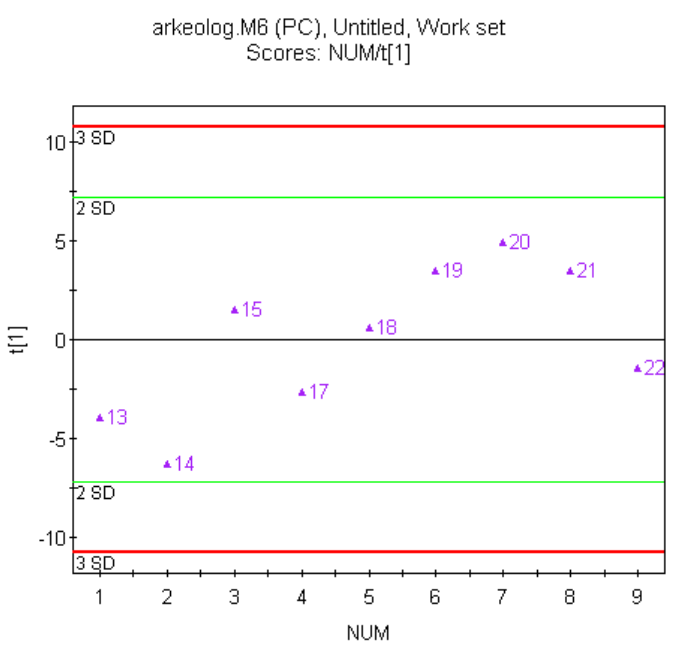
PCA - Class C

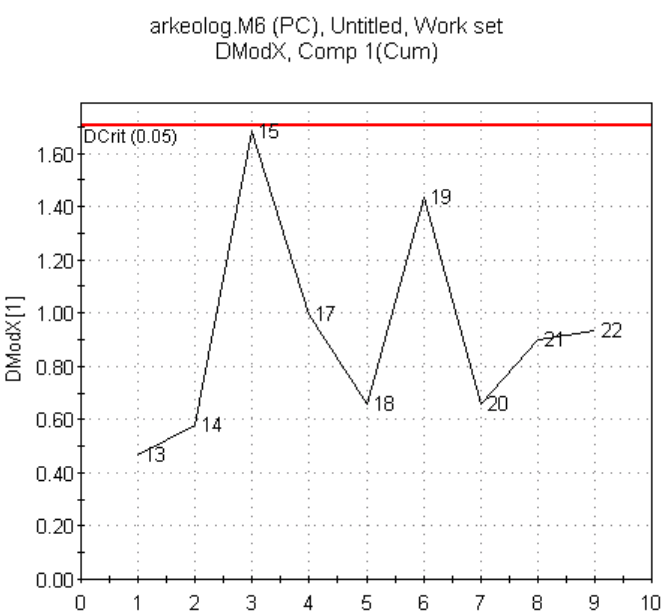




- Cross validation gives one significant component.
- R2 = 0.80, Q2 = O.73
- Since the model will be used for prediction of class identity for new samples we leave it to the cross validation to decide on the relevant number of components.

“Scores”, DModX - Class C

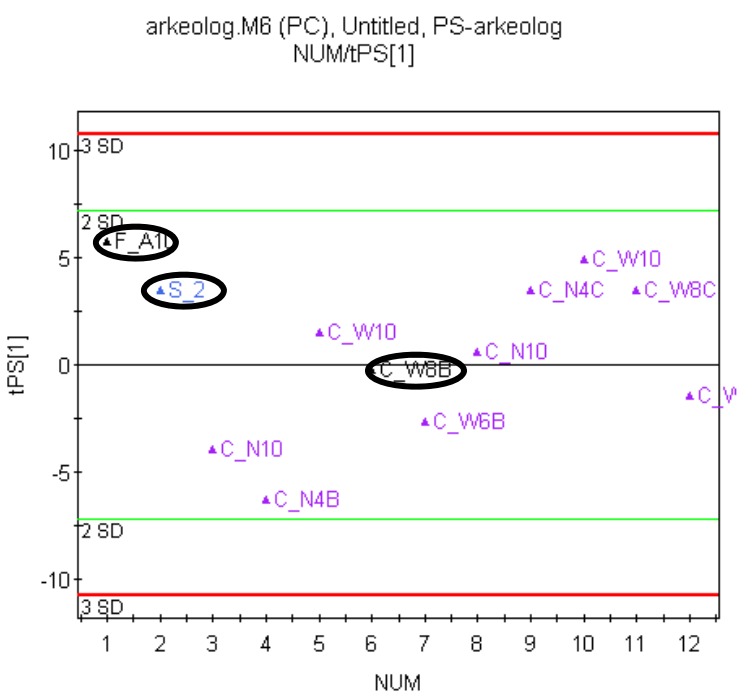


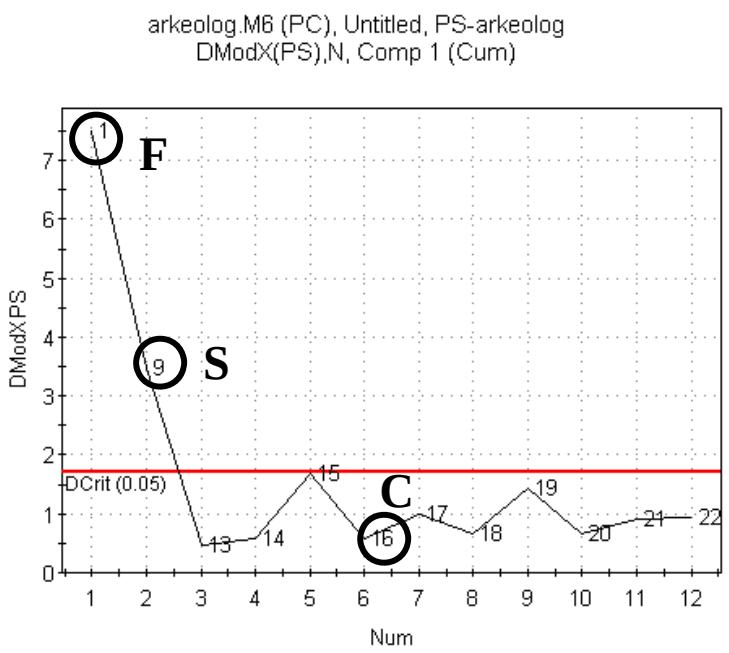


A homogenous group with some spread in scores.
No “outliers”!

Non of the samples show deviation in DModX!

Prediction of test samples - Class C



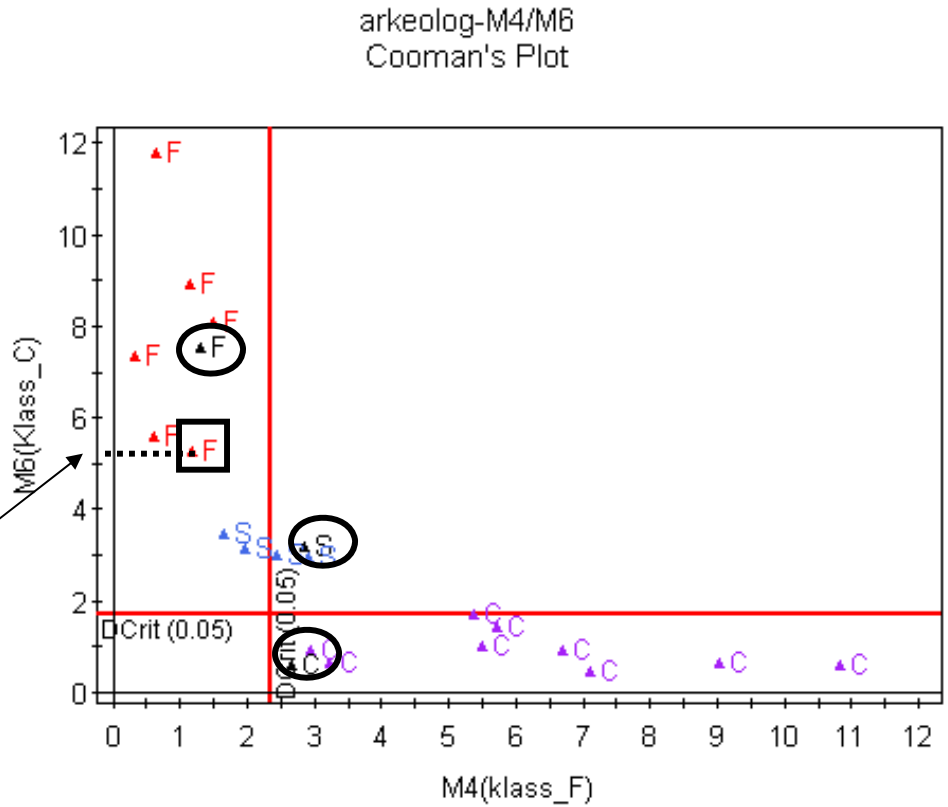


Predictions of the test samples (circled) show that the samples not belonging to class C are predicted within the class limits for class C. The sample belonging to class S is predicted in the center of the class.

The test samples belonging to class F and S do deviate in the predicted DModX.

Hence we can distinguish them from class C. We can also, with high certainty, say that the third sample belongs to klass C.

Prediction of test samples - Cooman's Plot



Distance (DModX) to PCA-model for classC for sample in square.

In the Cooman's plot the distance to a model (DModX) is plotted against the the distance (DModX) to a second model. In this case DModX_classC against DModX_classF. In the plot the samples belonging to klassC have got a small distance to classC (0-2) while the distance to classF varies from 3 to 11. Samples belonging to classF show small distance to classF (0-2) while the distance to classC varies between 5 and 12. Samples belonging to class S are mainly located outside the confidence limits for the two other classes. The test samples (circled) are all predicted into the right class

PLS-DA (three classes)

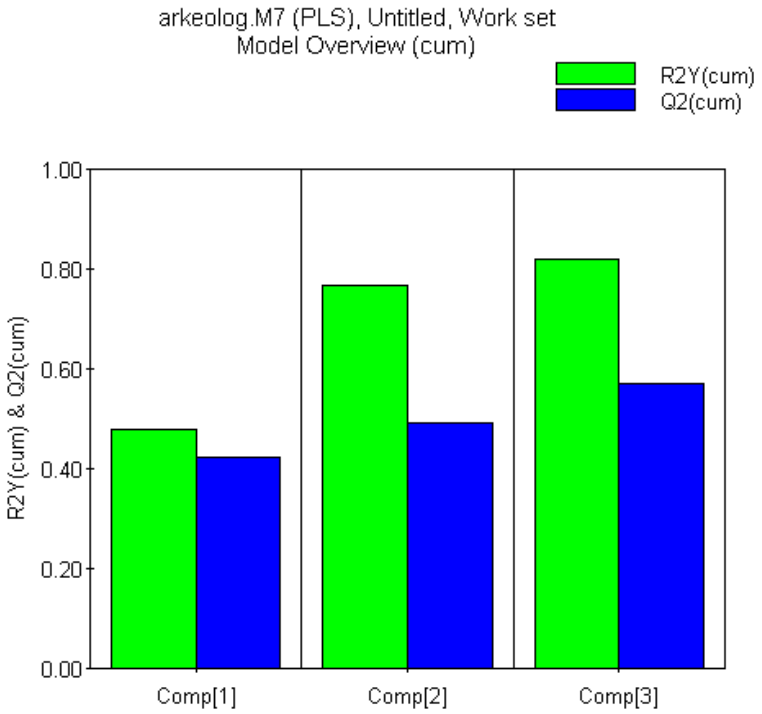
Model M7

Model: M7 Title: Untitled Properties...

Type: PLS-DA Observations (N) = 19, Variables (K) = 21 (X = 18, Y = 3) Work Set...

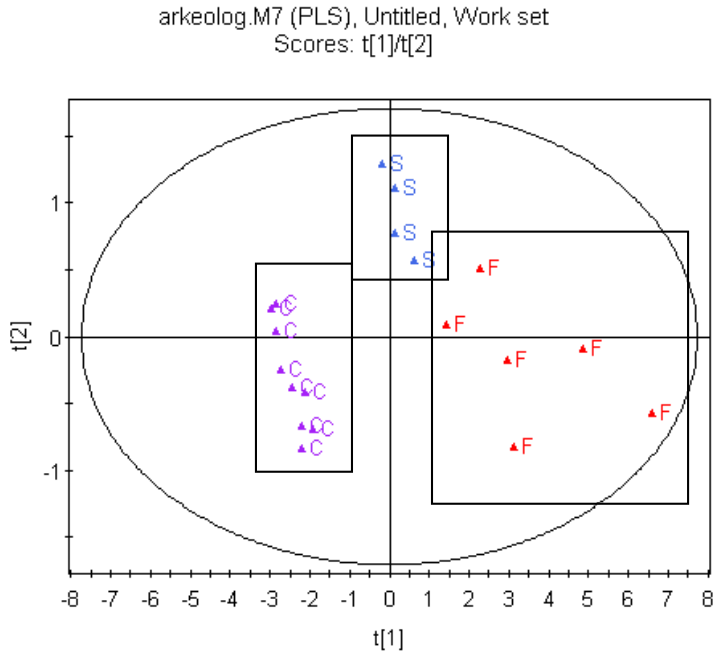
Components:

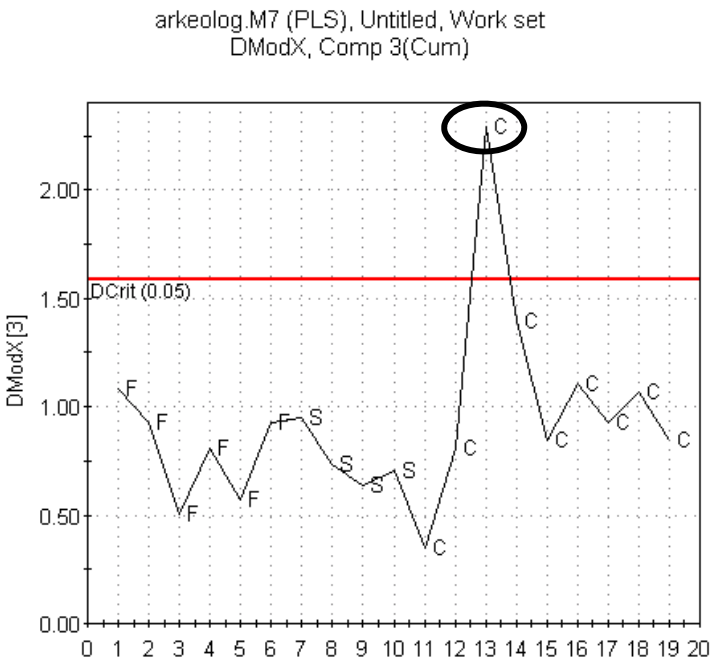
A	R2X	R2X(cum)	Eige...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signi...	It...
00	-	0.000	-	-	0.000	-	-	-	--	--
01	0.466	0.466	8.384	0.478	0.478	0.424	0.097	0.424	R1	4
02	0.085	0.551	1.538	0.289	0.766	0.117	0.097	0.491	R1	14
03	0.407	0.958	7.324	0.052	0.818	0.158	0.097	0.572	R1	3



- Cross validation gives three significan components
- R2X = 0.96, R2Y = 0.8, Q2 = O.57
- Since the model will be used for prediction of class identity for new samples we leave it to the cross validation to decide on the relevant number of components.

“Scores”, DModX - PLS-DA



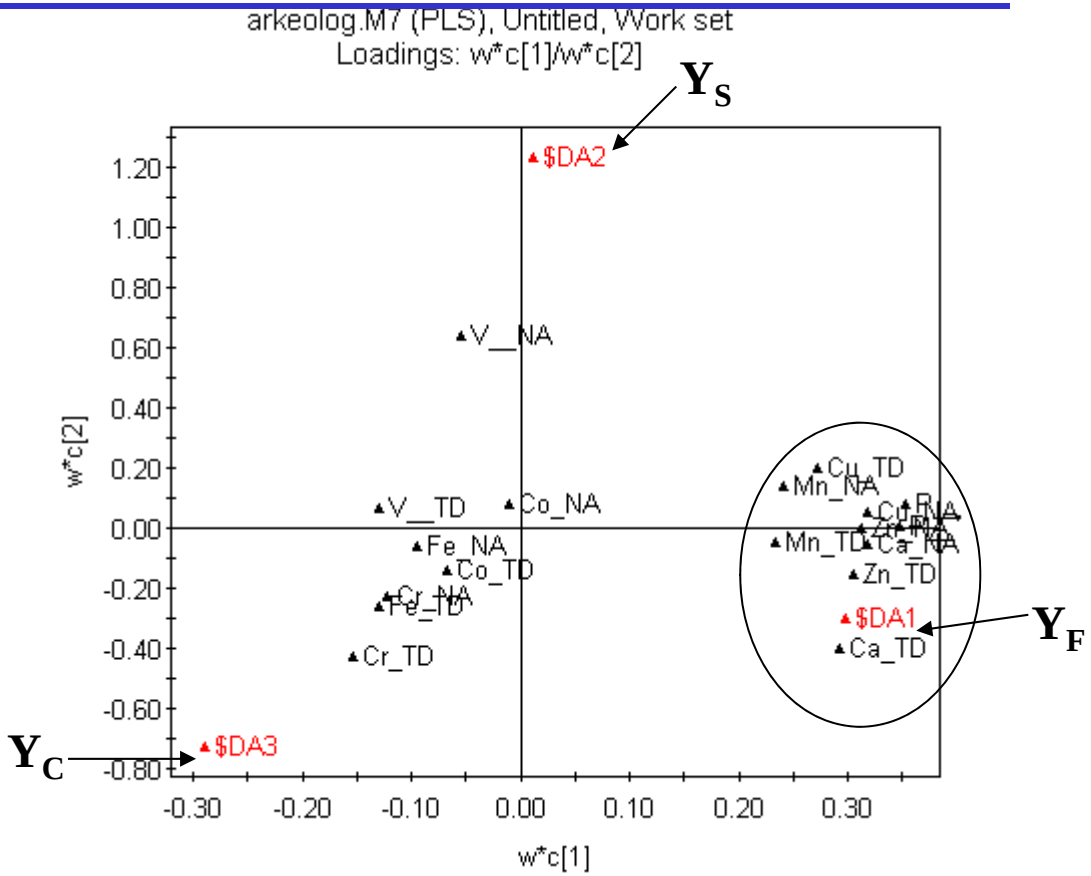


“Scores” (t1/t2) shows that the classes are separated in the first component

DModX show tha one sample from class C is deviating in DModX. (Keep an eye on this sample!)

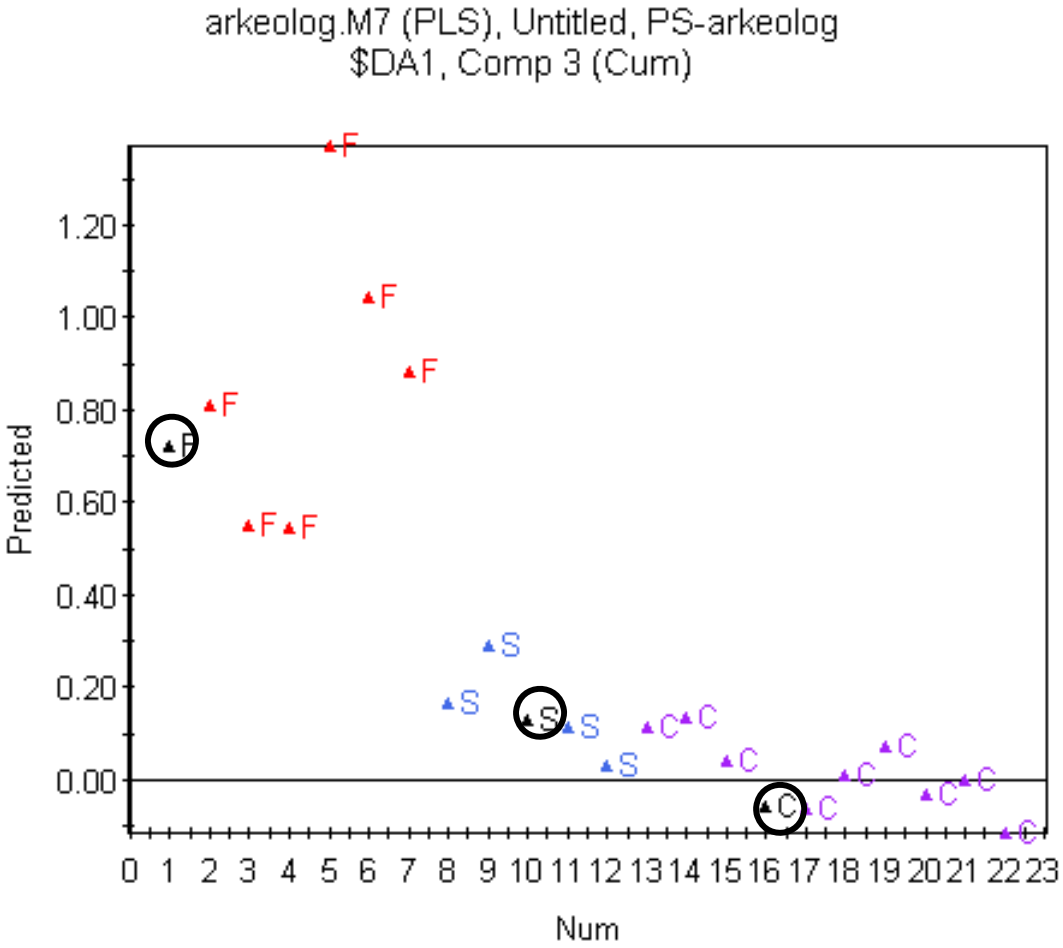
Interpretation of PLS “loadings” (wc1/wc2)

Controll (C) have got high levels Cr, Fe, V and Co and low levels of Mn, Cu, Zn and Ca.
Garbage site (F) has got the opposite pattern. Occupation site (S) is found in between the two extremes in component 1.
The two pre-treatments f _NA and _TA are providing the same information.



The PLS loadings gives us possibilities to interpret which variables that are correlated (x/x, y/y, x/y).
E.g. Y_F positively correlated to circled x-variables. Circled x-variables positively correlated to eachother i.e. are describing the same variation.

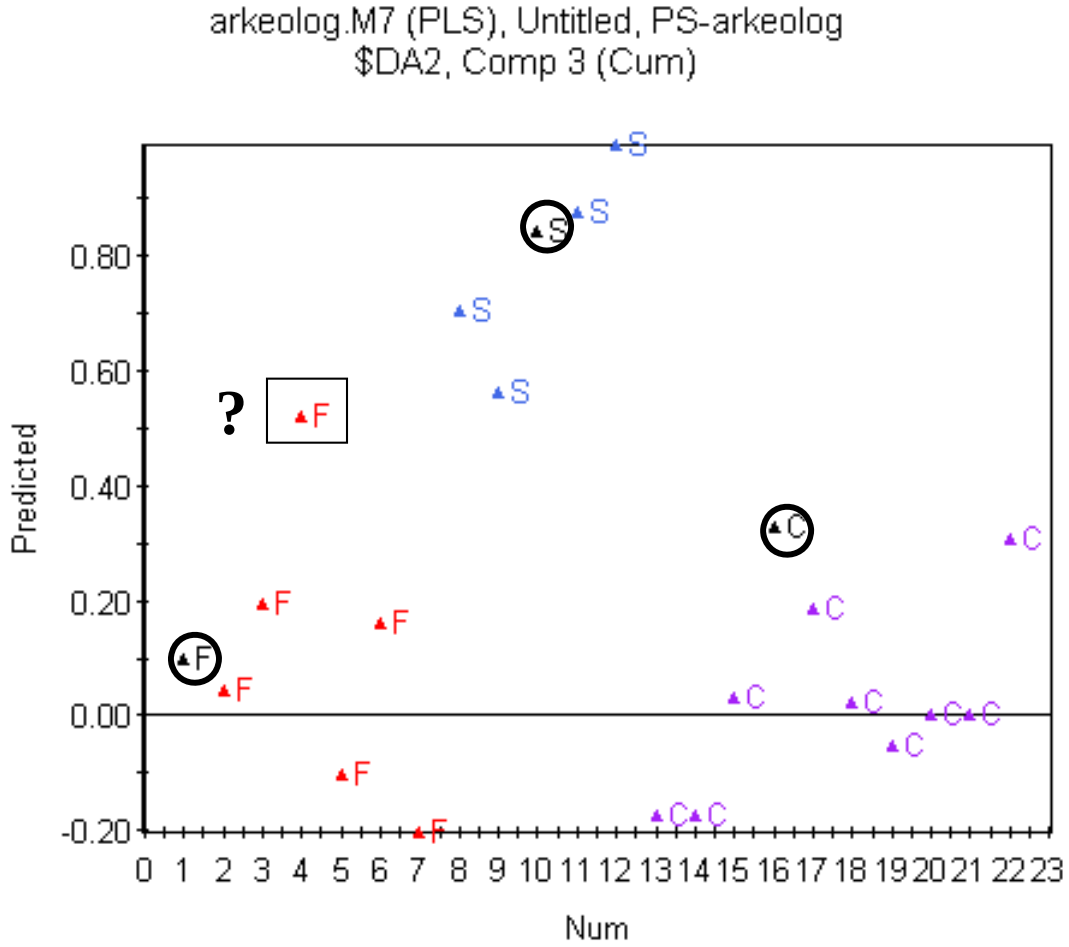
Prediction of test samples - Class F



	Y	
F	S	C
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

Samples belonging to class F are predicted close to 1. Samples not belonging to class F are predicted close to 0. (according to “dummy” Y-matrix).
Test samples (circled). Test sample belonging to F is predicted s F.

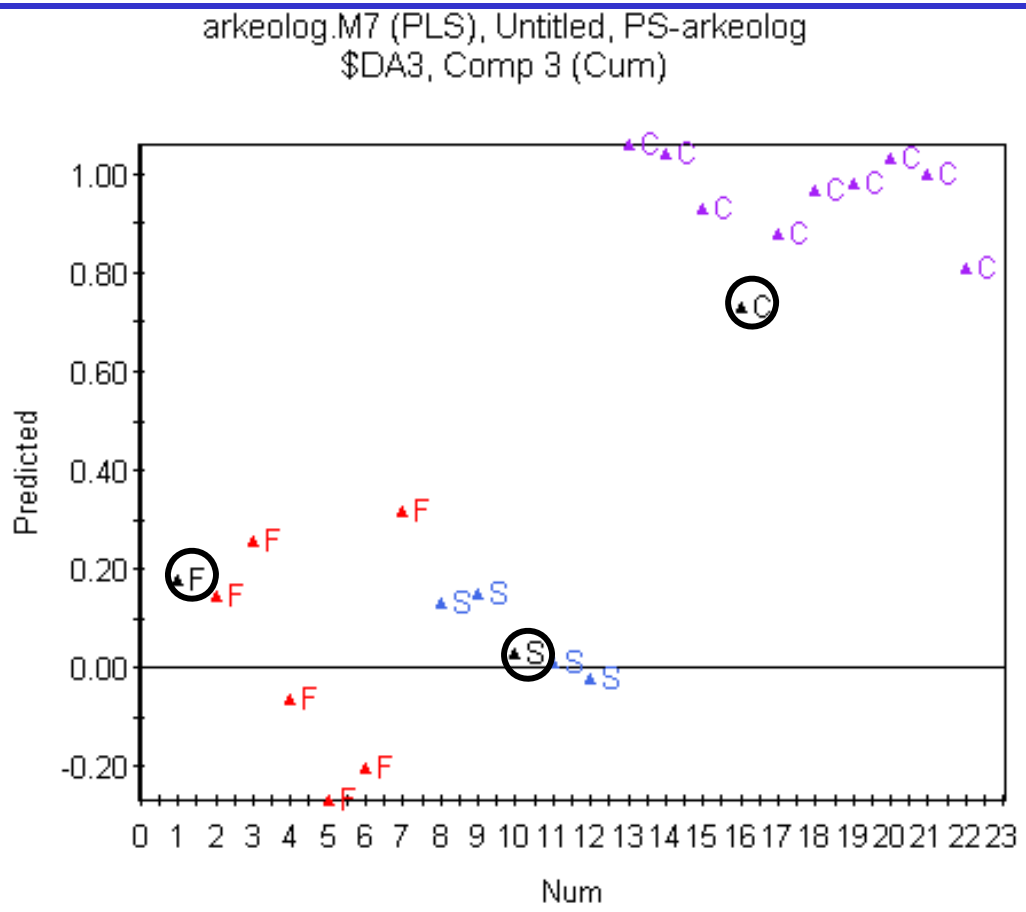
Prediction of test samples - Class S



	Y	
F	S	C
1	***0***	0
1	***0***	0
1	***0***	0
0	***1***	0
0	***1***	0
0	***1***	0
0	***0***	1
0	***0***	1
0	***0***	1

Samples belonging to class S are predicted close to 1. Samples not belonging to class S are predicted close to 0. (acoording to “dummy” Y-matrix).
Test samples (circled). Test sample belonging to S is predicted as S.

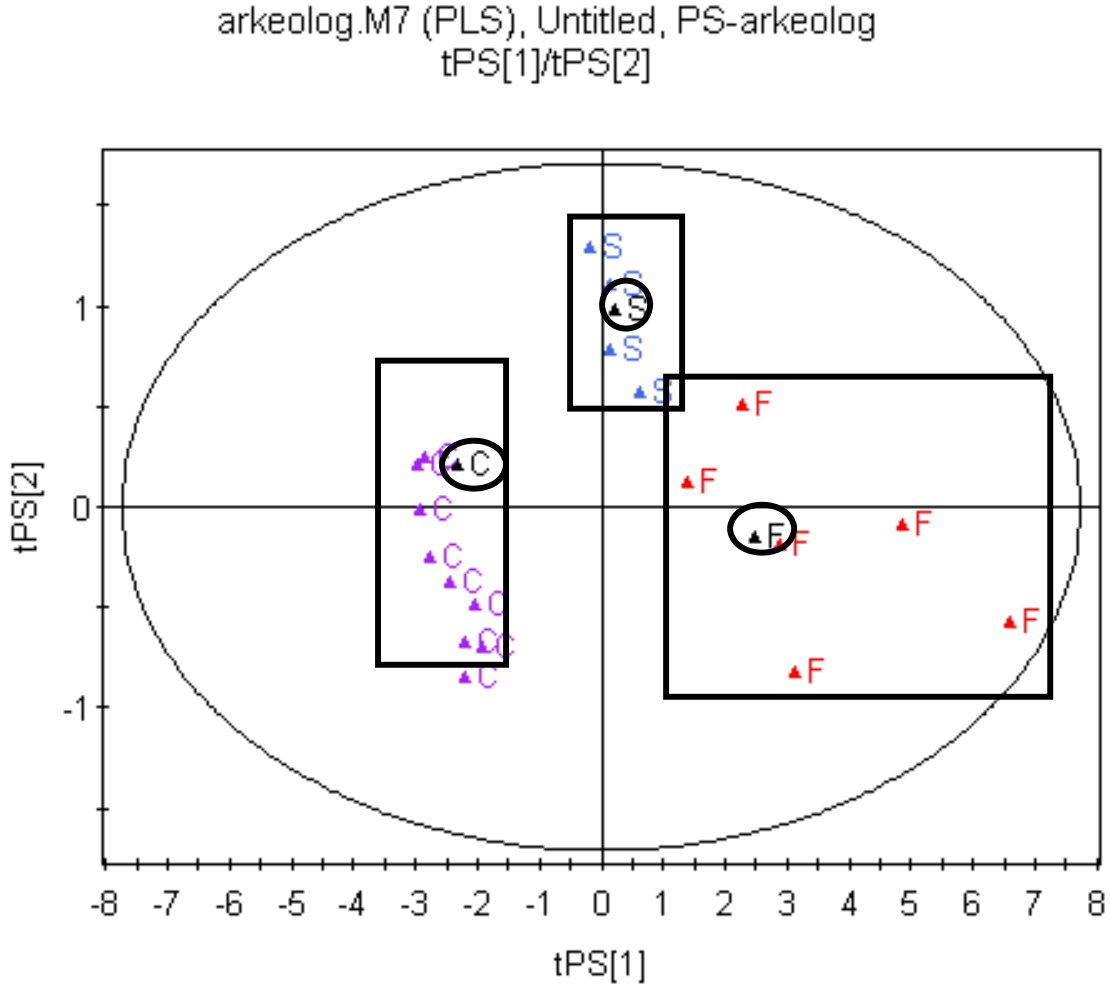
Prediction of test samples - Class C



	Y	
F	S	C
1	0	***0***
1	0	***0***
1	0	***0***
0	1	***0***
0	1	***0***
0	1	***0***
0	0	***1***
0	0	***1***
0	0	***1***

Samples belonging to class C are predicted close to 1. Samples not belonging to class C are predicted close to 0. (acoording to “dummy” Y-matrix).
Test samples (circled). Test sample belonging to C is predicted as C.

Prediction of test samples - in “scores”



Test samples are well predicted into respective class using PLS-DA

Prediction of test samples - table

Obs num	Obs namn	Set	Klass	YF(pred)	YS (pred)	YC (pred)
1	F_A100B	ts	?	**0.722685**	0.097382	0.179933
2	F_A100M	ws	1	**0.811566**	0.043641	0.144793
3	F_A100T	ws	1	**0.549395**	0.195223	0.255383
4	F_A52B	ws	1	**0.543967**	0.519802	-0.06377
5	F_A52M	ws	1	**1.37113**	-0.1037	-0.26744
6	F_A52T	ws	1	**1.04292**	0.159519	-0.20244
7	F_A40	ws	1	**0.883523**	-0.20236	0.318839
8	S_1	ws	2	0.164991	**0.703349**	0.13166
9	S_2	ws	2	0.288444	**0.56056**	0.150996
10	S_3	ts	?	0.131937	**0.839076**	0.028987
11	S_4	ws	2	0.115002	**0.876043**	0.008955
12	S_5	ws	2	0.032711	**0.990776**	-0.02349
13	C_N10B	ws	3	0.113066	-0.1726	**1.05953**
14	C_N4B	ws	3	0.133325	-0.17232	**1.039**
15	C_W10B	ws	3	0.041278	0.030673	**0.928049**
16	C_W8B	ts	?	-0.05623	0.327397	**0.728833**
17	C_W6B	ws	3	-0.0634	0.183637	**0.879764**
18	C_N10C	ws	3	0.010964	0.022503	**0.966534**
19	C_N4C	ws	3	0.072246	-0.0514	**0.979158**
20	C_W10C	ws	3	-0.03228	0.001849	**1.03043**
21	C_W8C	ws	3	0.00087	0.000206	**0.998925**
22	C_W6C	ws	3	-0.11482	0.305011	**0.809807**

	Y	
F	S	C
1	***0***	***0***
1	***0***	***0***
1	***0***	***0***
0	***1***	***0***
0	***1***	***0***
0	***1***	***0***
0	***0***	***1***
0	***0***	***1***
0	***0***	***1***

Example – Coronary Heart Disease (CHD)

- **CHD is a major cause of mortality and morbidity in developed countries**
- **Many risk factors for CHD have been identified from epidemiological studies**
- **The full range of risk factors comprise insufficient density of data to accurately discriminate CHD on an individual basis**
- **Firm diagnosis of CHD needs angiography**

Metabonomic Studies of CHD

- *Angiography* is a x-ray investigation of the hearts artheries.

Angiogram →

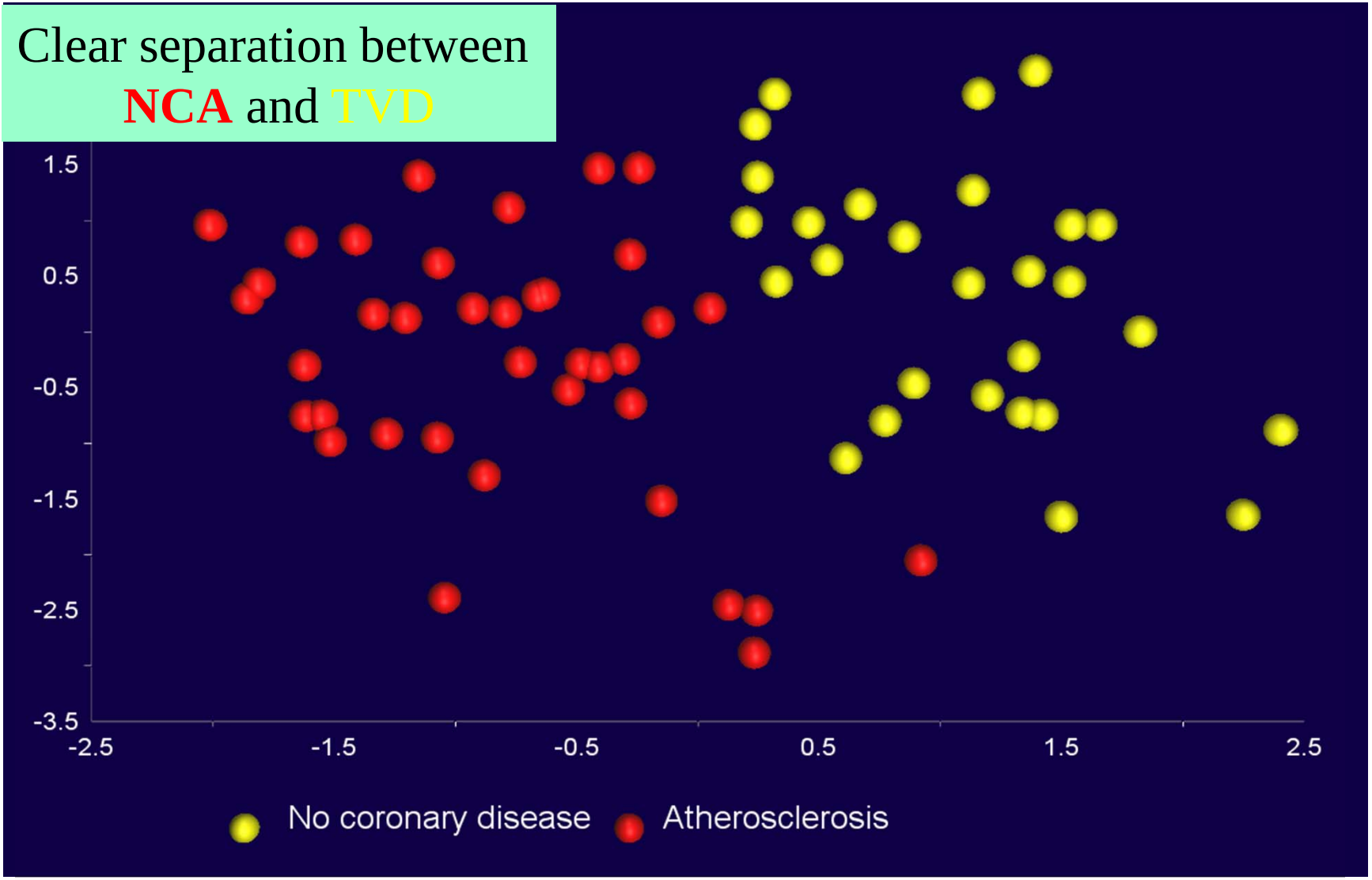
- The pictures reveal severity of clogging of the artheries (number of vessels and grade). These cloggings are the cause of ***CHD.***
- Severe cases of clogging leads to heart attacks if not *"by pass"* surgery is carried out (serious surgical intervention).
- *Could NMR of a simple blood test be indicative of CHD and replace angiography as the diagnostic standard in the clinic?*

CHD: Study details

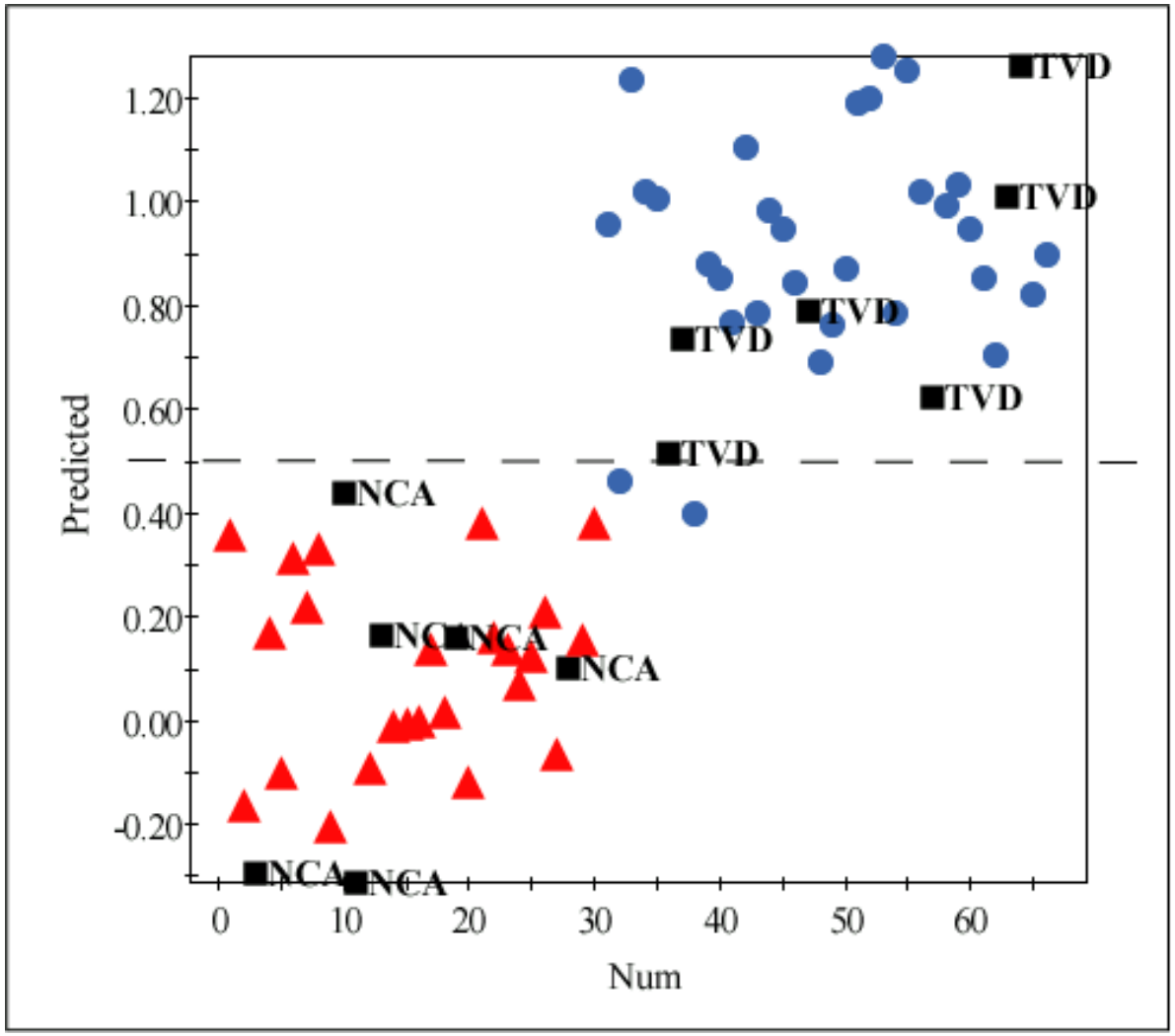
- **Serum samples were collected from patients with normal coronary arteries (NCA; n = 30) and triple vessel disease (TVD; n = 36)**
- **1D ^{1}H NMR spectra were recorded**
- **Chemometrics (PCA, PLS-DA) were applied to NMR spectral data.**

Coronary Heart Disease: PCA

Clear separation between **NCA** and TVD



Coronary Heart Disease: PLS-DA



= NCA

= TVD

= Test samples

Analysis of blood samples Using NMR+ MVA Might be a potential Alternative to ANGIOGRAFI for diagnosis of CHD.

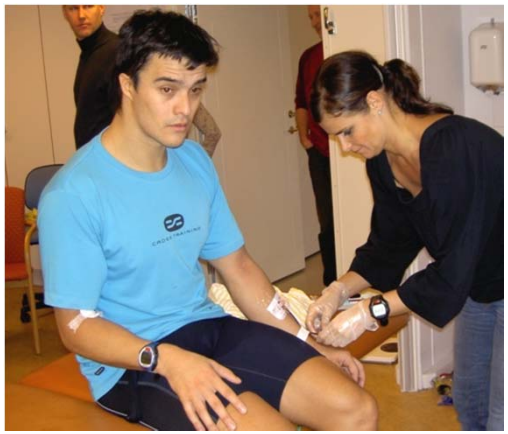
Other diseases? ALS, Alzheimers, Parkinson, Prostate cancer, ...

NUTRITION IMPACT FOLLOWING EXERCISE

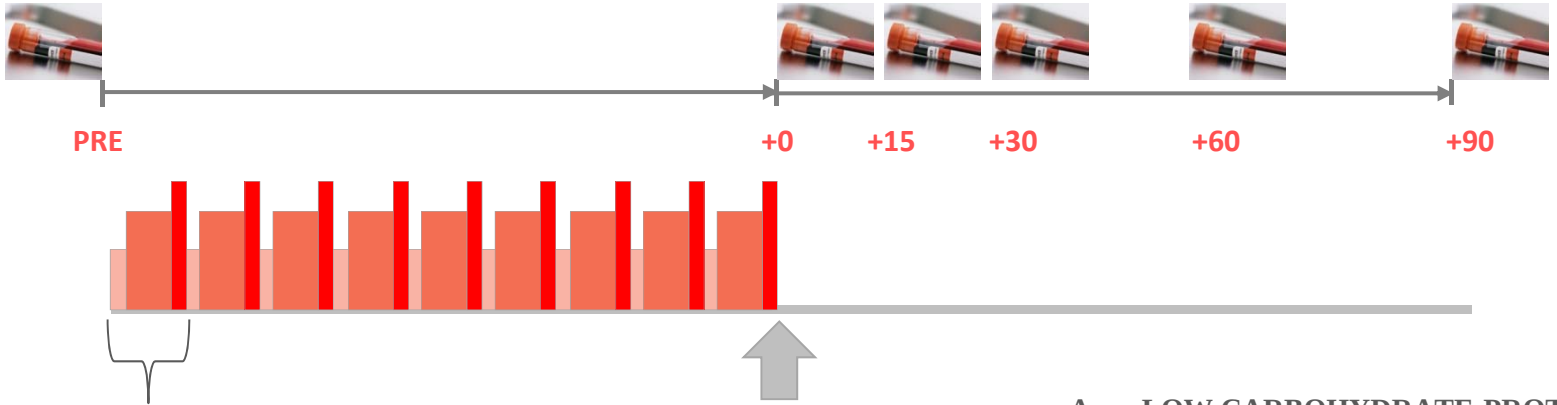
24 healthy male subject

Age: 25.7 y.o
Weight: 77.4 kg
VO_2max: 59.1 ml O_2/kg/min

90 min EXERCISE



90 min RECOVERY



Workload
(percent VO2peak)

2 min, 40%
6 min, 60%
2 min, 85%
x 9 = 90 minutes

A. LOW CARBOHYDRATE-PROTEIN
(1g CHO/kg + 0.5g protein/kg)

B. WATER
(tap water)

C. HIGH CARBOHYDRATE
(1.5g CHO/kg)

D. LOW CARBOHYDRATE
(1g CHO/kg)

SERUM SAMPLES

GC-TOF/MS

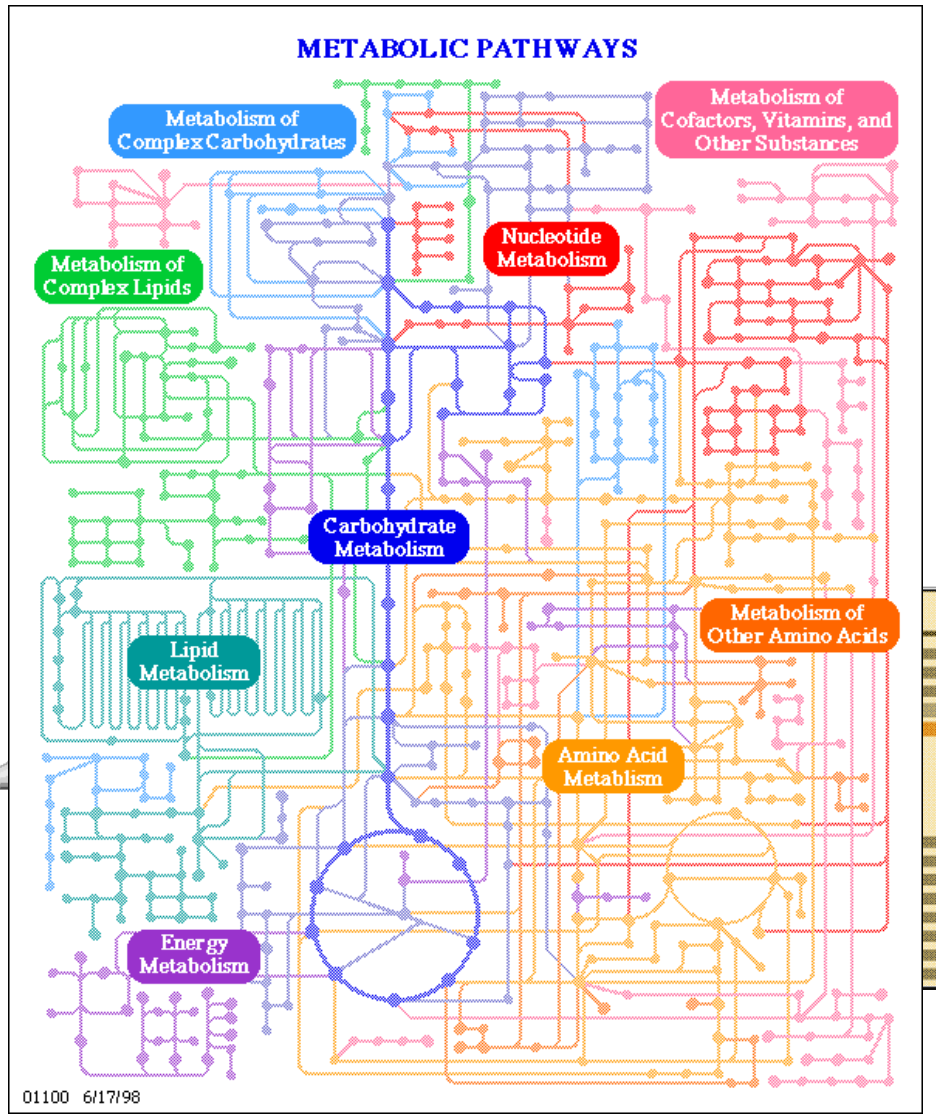
REPRESENTATIVE SAMPLE SELECTION (1)

H-MCR (2,3)

MVA

IDENTIFICATION

BIOLOGICAL EVALUATION

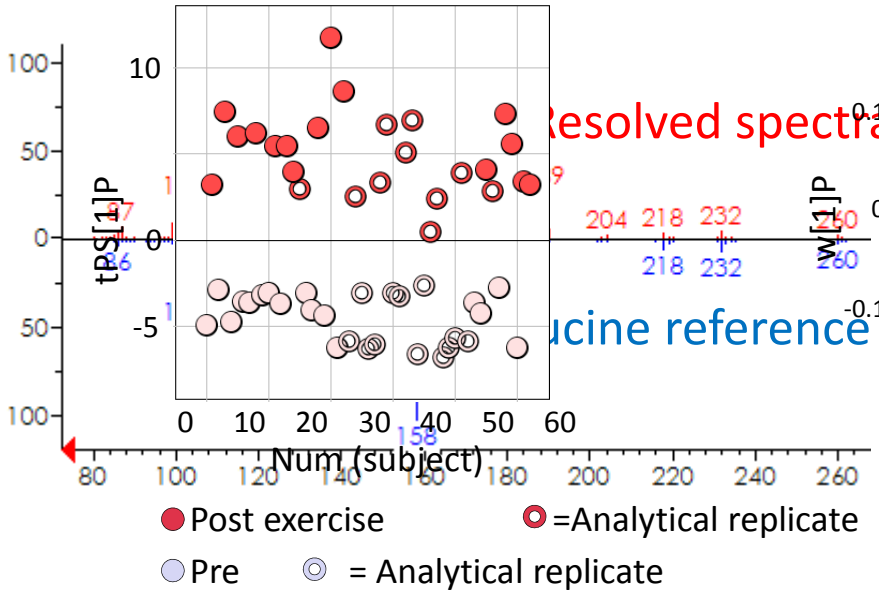


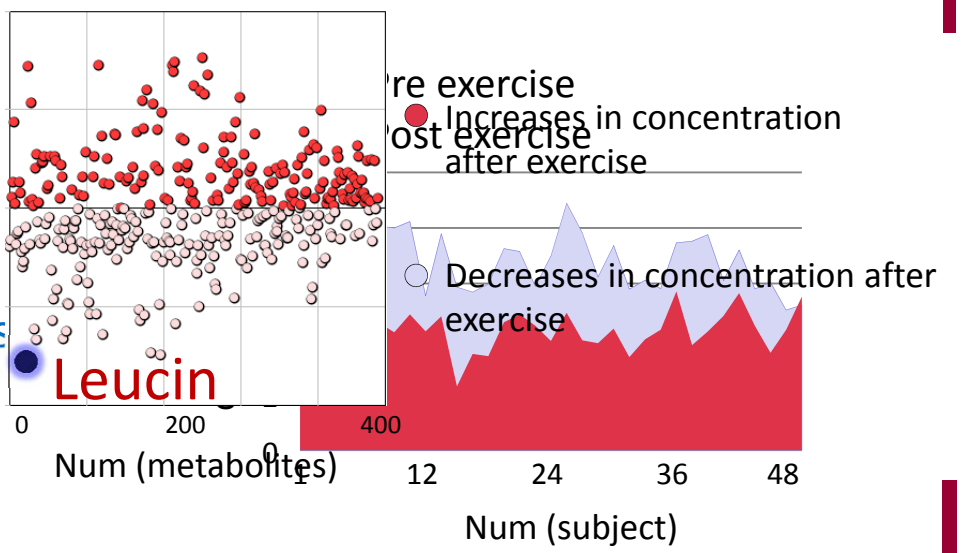
t[2]

1. Thysell E et al submitted to PLoS Computational Biology 2008
2. Jonsson P *et al*, *Analytical Chemistry,* 77, 5635-42, 2005.
3. Jonsson P et al, *Journal of Proteome Research* , 5, 1407-14, 2006.

ACUTE EXERCISE RESPONSE

Predicted sampels in terms of both H-MCR and OPLS





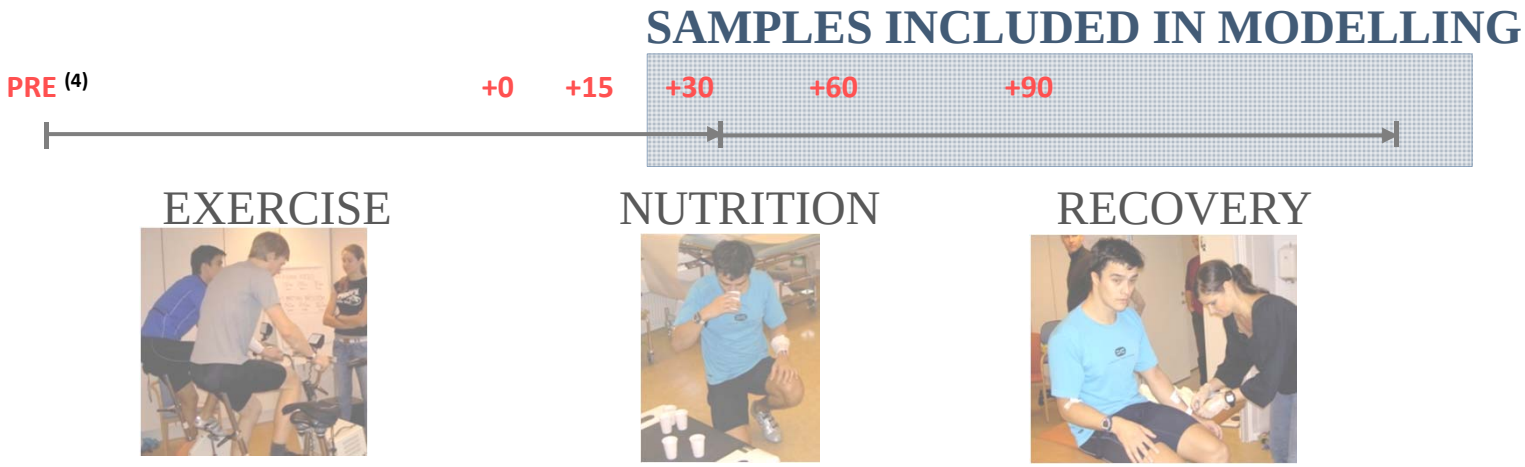
IMPACT OF NUTRITION INTAKE AFTER EXERCISE

DATA

- 218 metabolites from GC/MS analysis was used to describe the samples.

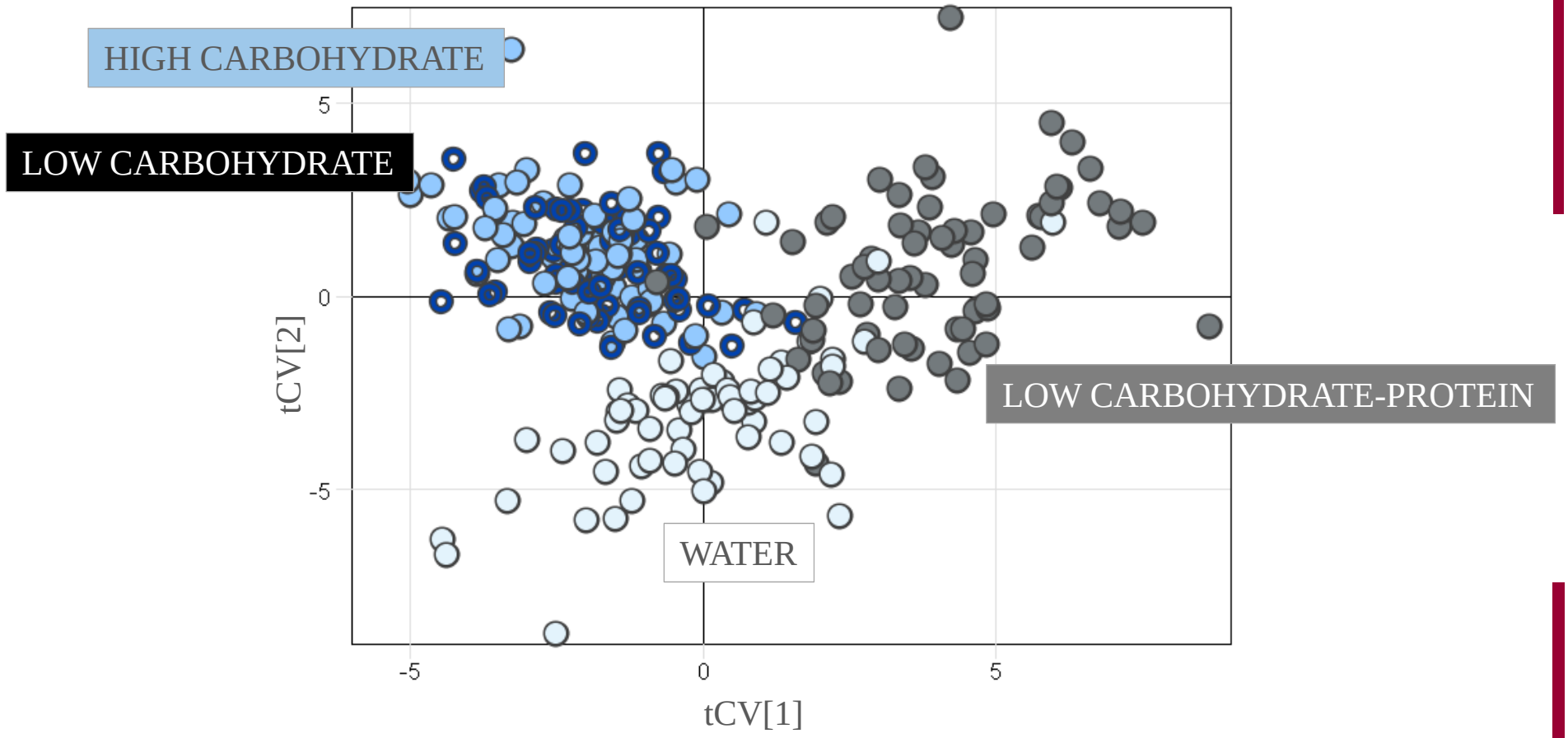
MODELLING

- PLS-DA was used to reveal macronutrient related effects in the recovery phase



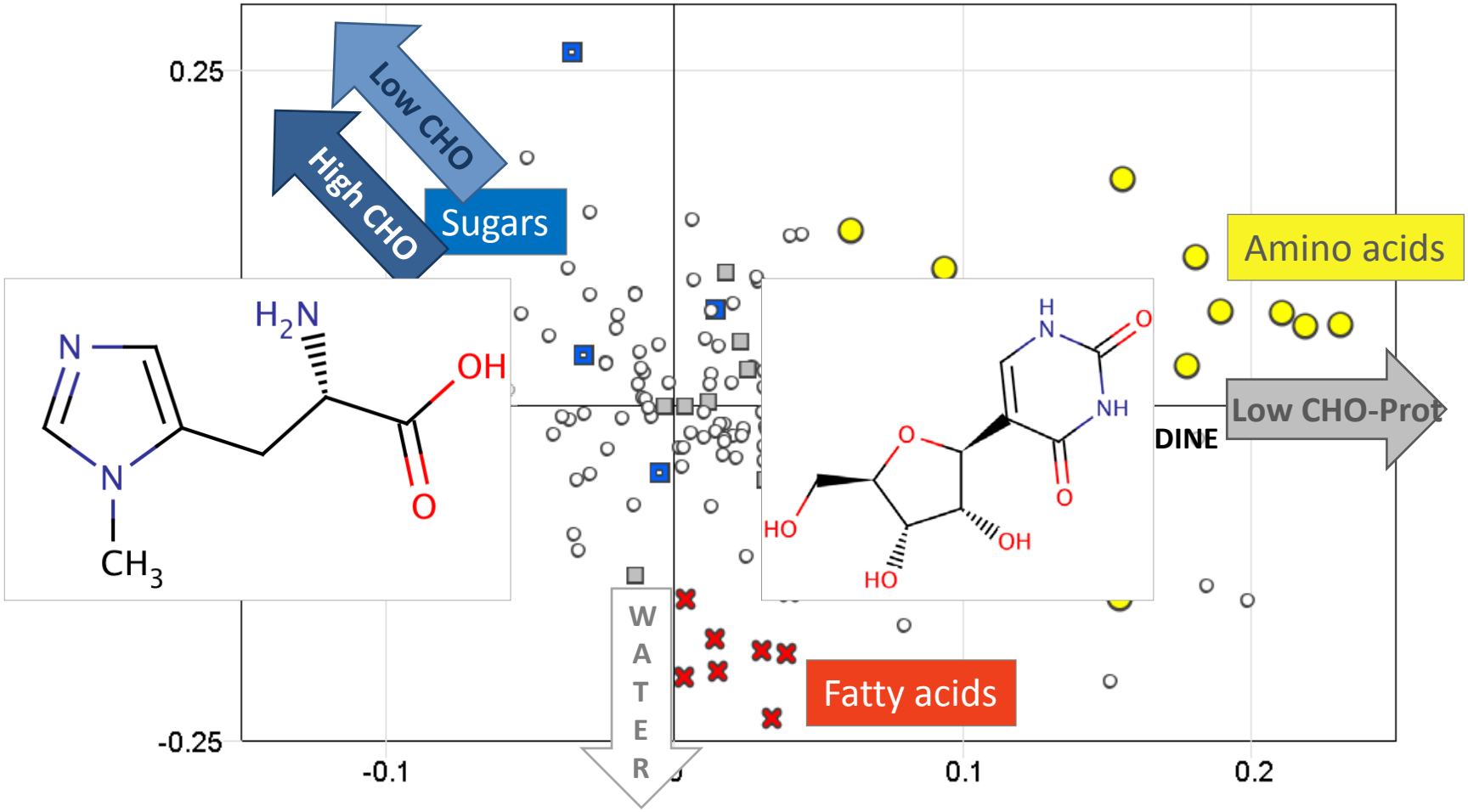
(4) Pohjanen E, et al . *Journal of Proteome Research* **2007**, 6, (6), 2113-20.

PLS-DA SCORES (CROSS-VALIDATED)



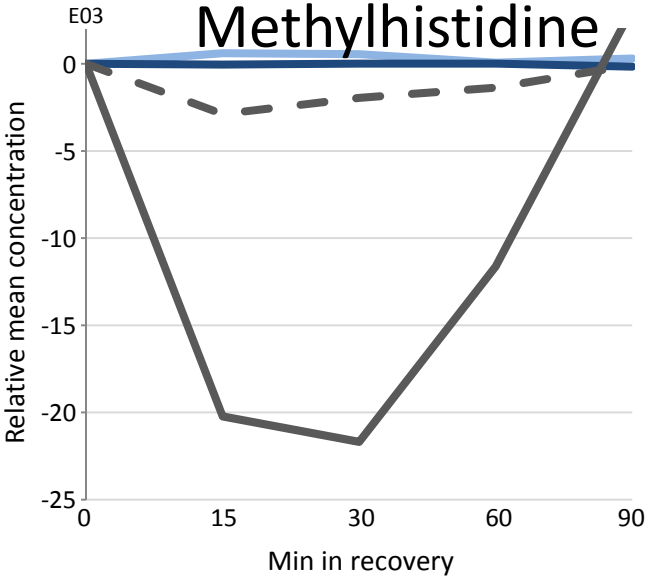
Num of components: 3P+5O
R2X: 0.453
R2Y: 0.662
Q2: 0.521

PLS-DA LOADINGS

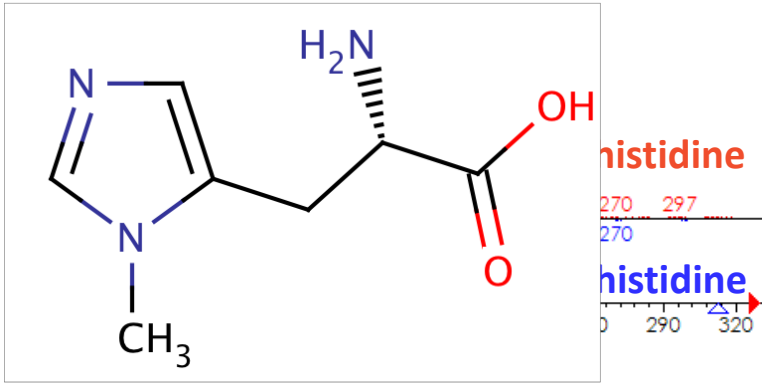


Num of components: 3P+5O
R2X: 0.453
R2Y: 0.662
Q2: 0.521

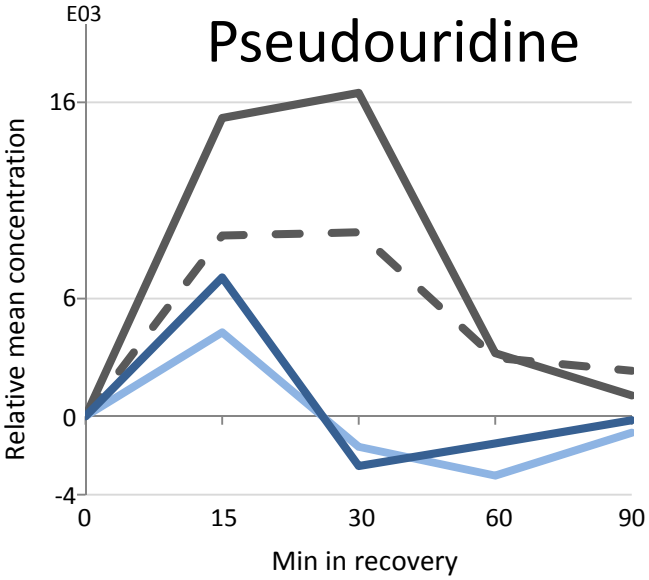
DECREASE IN CONCENTRATION

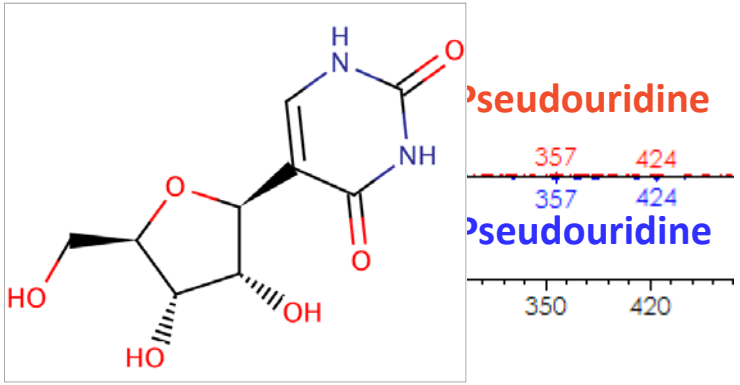


LOW CARBOHYDRATE
HIGH CARBOHYDRATE
LOW CARBOHYDRATE-PROTEIN
WATER



INCREASE IN CONCENTRATION

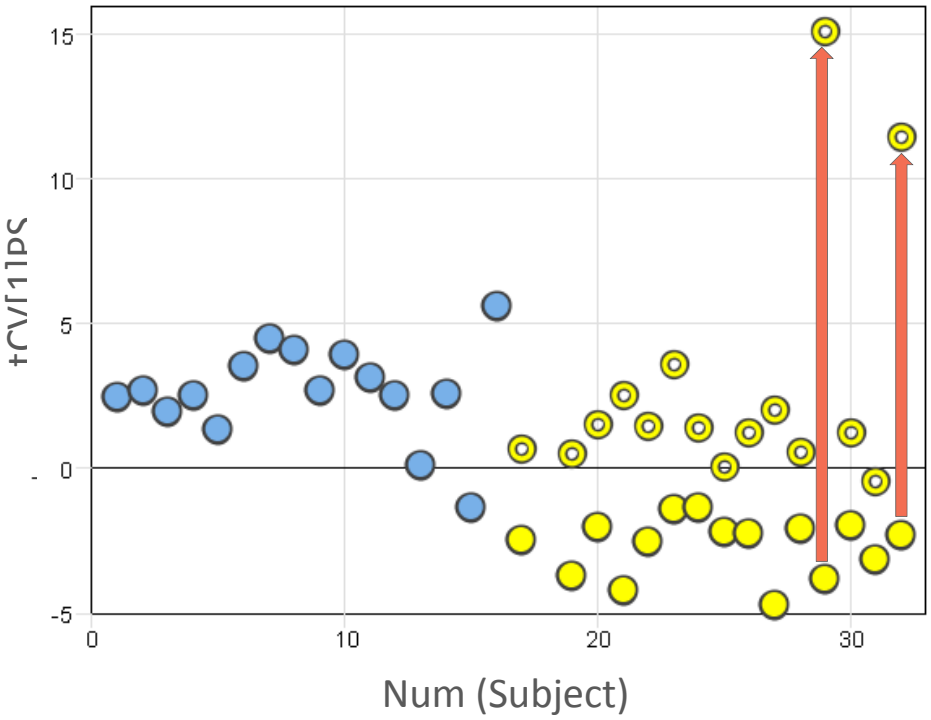


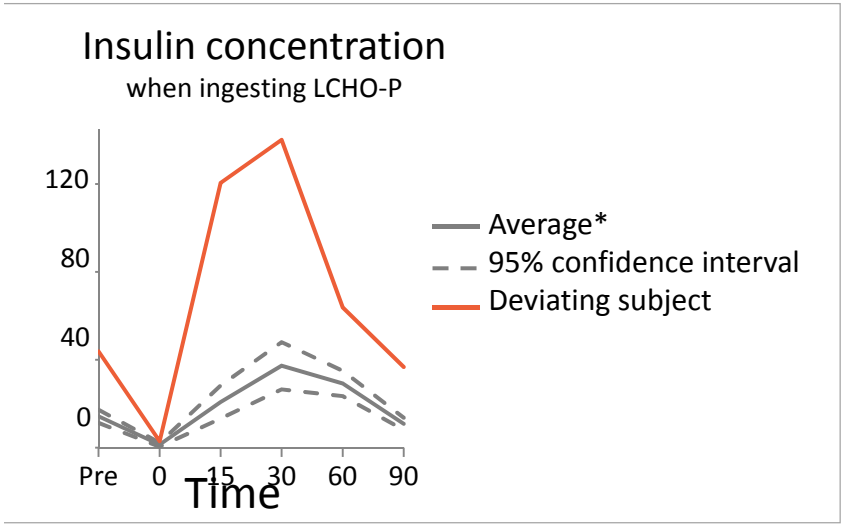


FITNESS STATUS

High fitness level
68.2±2.9mL kg^{-1}min $^{-1}$

Low fitness level
49.4±6.4mL kg^{-1}min $^{-1}$





Num of components: 1P+1O
R2X: 0.228
R2Y: 0.905
Q2: 0.741

Markers for pro-anabolic effect

Individual nutrition modulation

Potential modelsystem for detecting insulin resistance?

Summary - Classification

- **Overview** - detection of classes - PCA of the data table.

- **Methods**
  - SIMCA (PCA of individual classes)
  - PLS-DA (PLS against “dummy” Y, for maximum separation)

- Important to create good models for ***prediction*** of new samples.

- Clasification applications common within many areas.
  - Archaeologi
  - Diagnosis of Coronary Heart Disease (NMR)
  - GC/MS human exercise and nutrition
  - .....