



soluzioni informatiche

Latent variables modeling in metabonomics: a case study

Elena Fioravanzo
S-IN Soluzioni Informatiche

MipTec
Basel, October 13-15, 2009

ABOUT THE COMPANY ...

S-IN, Soluzioni Informatiche:

Italian company that supplies customised computer-assisted solutions in chemistry-related frameworks.

- **Design of experiments and multivariate data analysis**
- **Molecular modelling**
- **ADME profiling**
- **Toxicity predictions (including regulatory framework)**
- **REACH regulation**
- **Physicochemical property predictions**
- **Analysis, storage, and database of analytical and spectral data**
- **Databasing for chemistry-type data**

OUTLINE

INTRODUCTION:

projection approach

APPLICATIONS:

putative markers from LC-MS

CONCLUSIONS

Complicated problems in production and R&D yield complex data sets (the data explosion)



- Data set = table (matrix)
 - N rows (obs, samples, cases,)
 - K variables (properties, tests,)
- Many variables -- large K
- Often few observations ($K \gg N$)
- Missing data
- Noise
- Collinearities, latent variables
- Clusters

"Omics" science

Transcriptomics

Proteomics

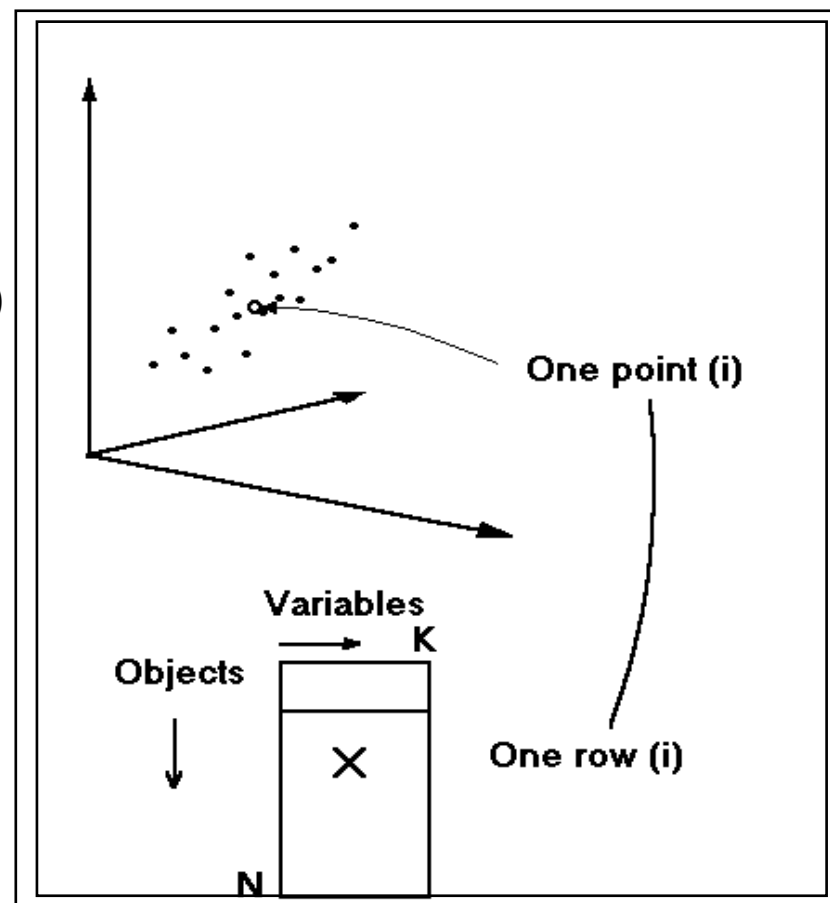
Metabonomics

...

(Systems Biology)

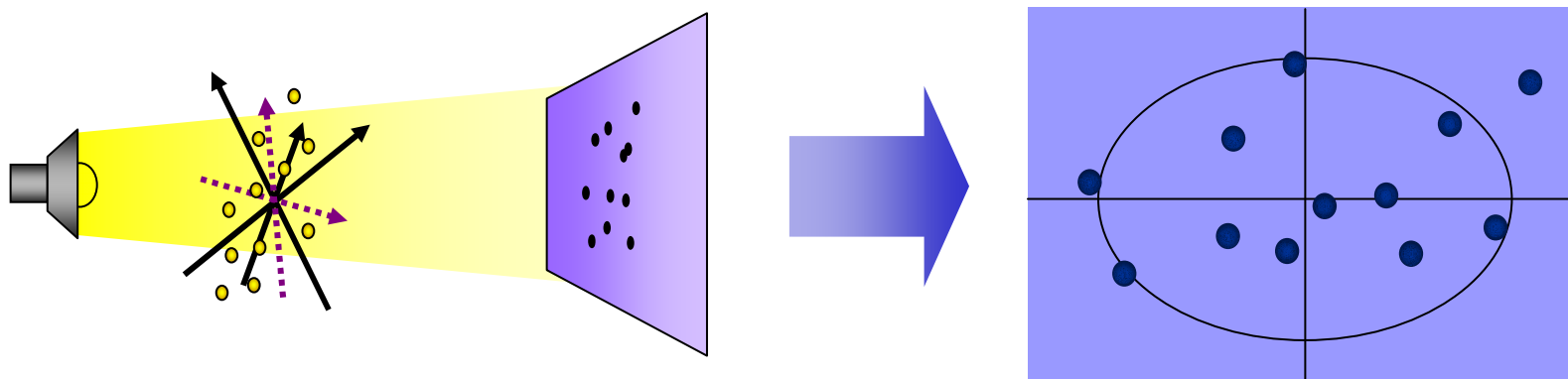
Multivariate analysis by means of projections

- Data shaped as a table, \mathbf{X} ($N \times K$)
- Space with K axes (K -space)
 - K = number of variables (col.s)
 - Each obs. (process time point)
 - is a point in this space
- Multivariate analysis
 - finding structures in K -space
 - describing them (math & stat)
 - using them for problem solving



What is a Projection?

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



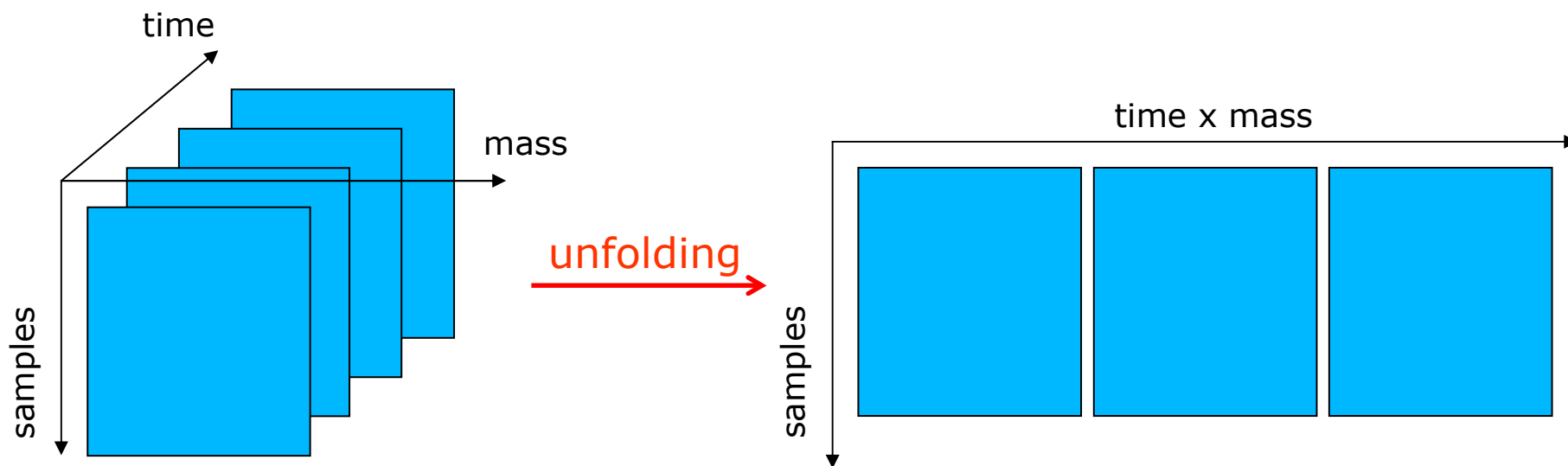
- MVDA methods:
 - PCA
 - Analysis of matrix of data – e.g. pattern recognition
 - PLS
 - Regression models
 - OPLS
 - PLS models where information orthogonal to response has been removed
 - PLS-DA
 - Classification models
 - OPLS-DA
 - Classification methods

CASE STUDY

- OBJECTIVE:
 - Identify putative markers for genetically distinct mice
- DATA:
 - LC-MS experiment on urine samples
- OBSERVATION:
 - 29 mice
 - 3 classes: black, white, nude

3-way dataset reorganization

Every time a chromatographic peak is detected a set of masses is produced so that each variable is a time followed by a mass i.e. 3.2_245.67 (time x mass). The data consist of 29 observations and 4145 variables.



Dataset

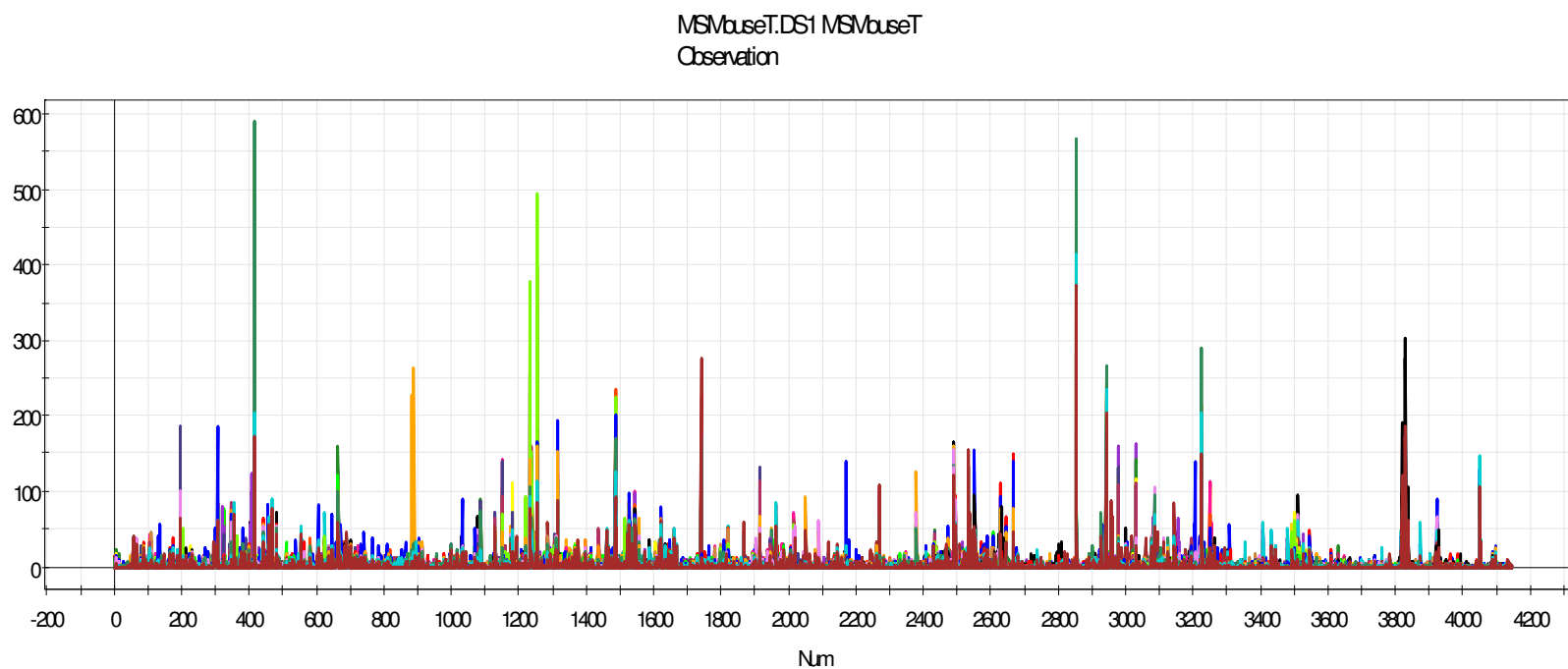
29 observations

Three genetically distinct strains of mice:

9 White

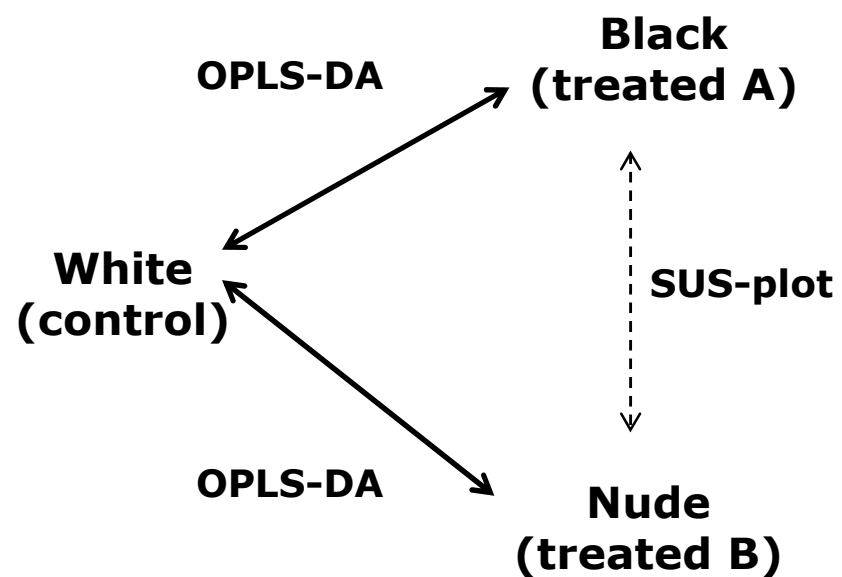
10 Black

10 Nude



SIMCAP+ 12 - 2008-10-24 18:05:05 (UTC+1)

Strategy of analysis



SCALING

Mean centering

$$X_{ki}^{\text{CTR}} = X_{ki} - m_i$$

$$m_i = \sum_k X_{ki} / n$$

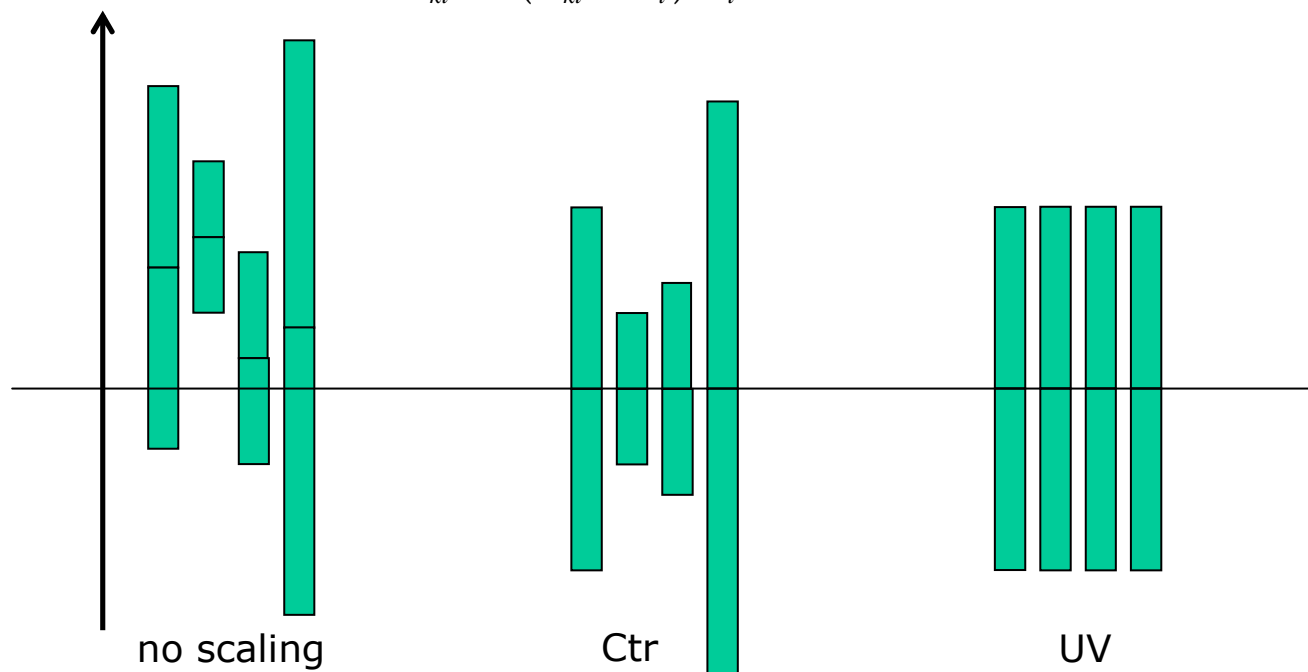
Pareto

$$X_{ki}^{\text{PAR}} = (X_{ki} - m_i) / \sqrt{s_i}$$

$$s_i = \left[\sum_k (X_{ki} - m_i)^2 / (n-1) \right]^{1/2}$$

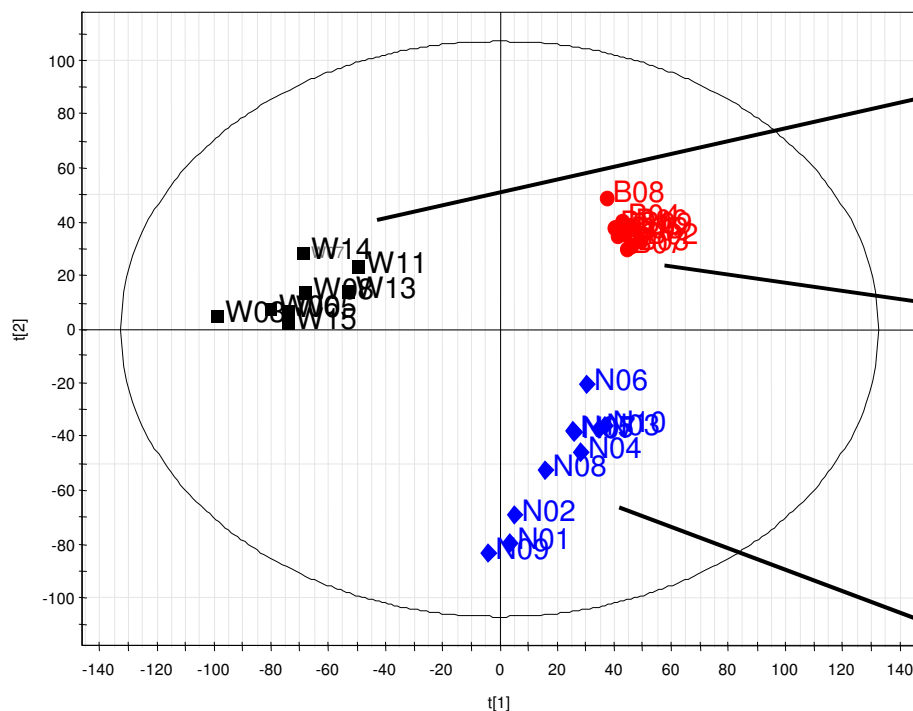
Unit variance

$$X_{ki}^{\text{UV}} = (X_{ki} - m_i) / s_i$$



Is it possible to identify clusters?

PCA Par score plot

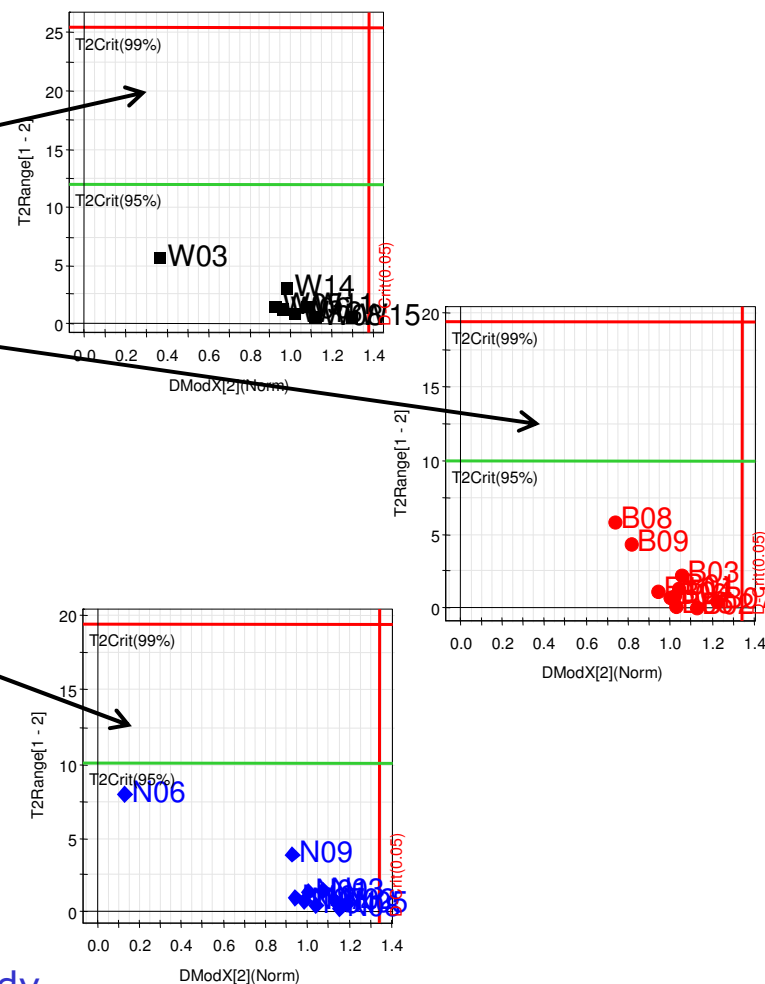


$R^2X[1] = 0.244914$

$R^2X[2] = 0.158829$

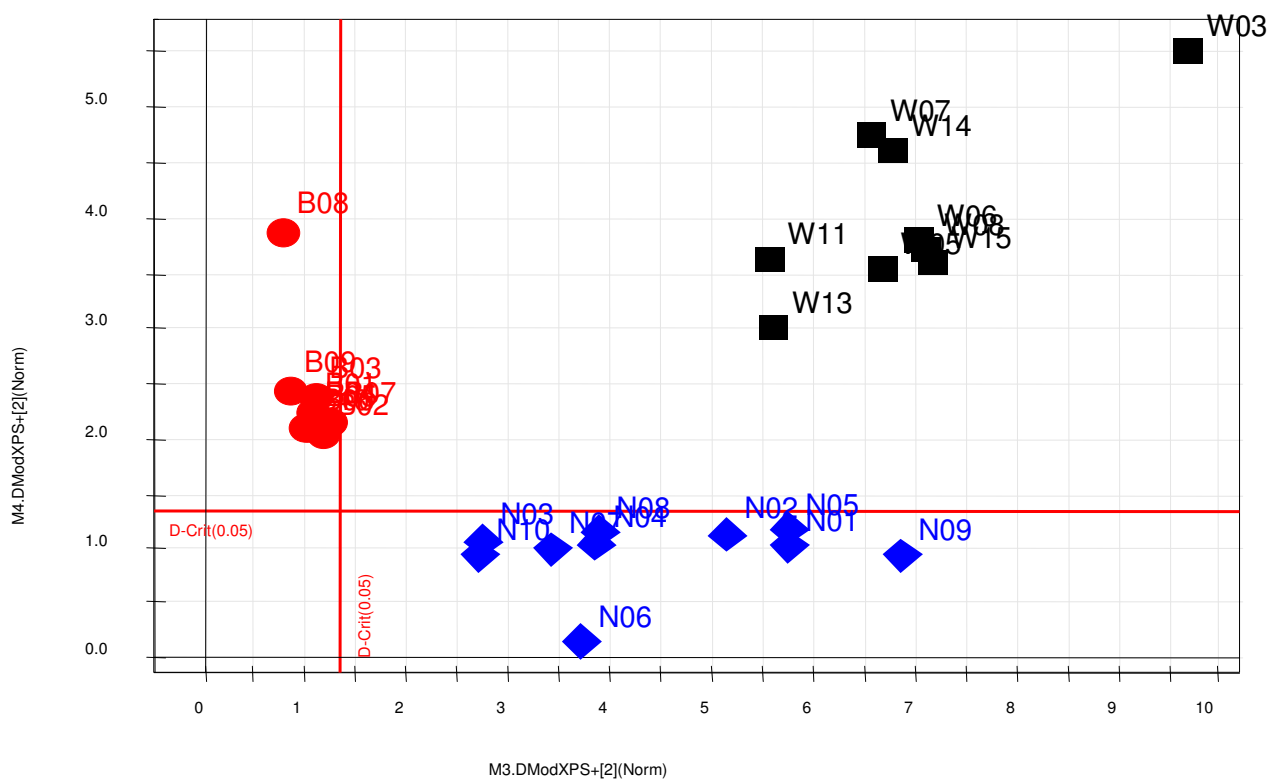
3 components $R^2 = 0.48$ $Q^2 = 0.28$

T2/DModX plot



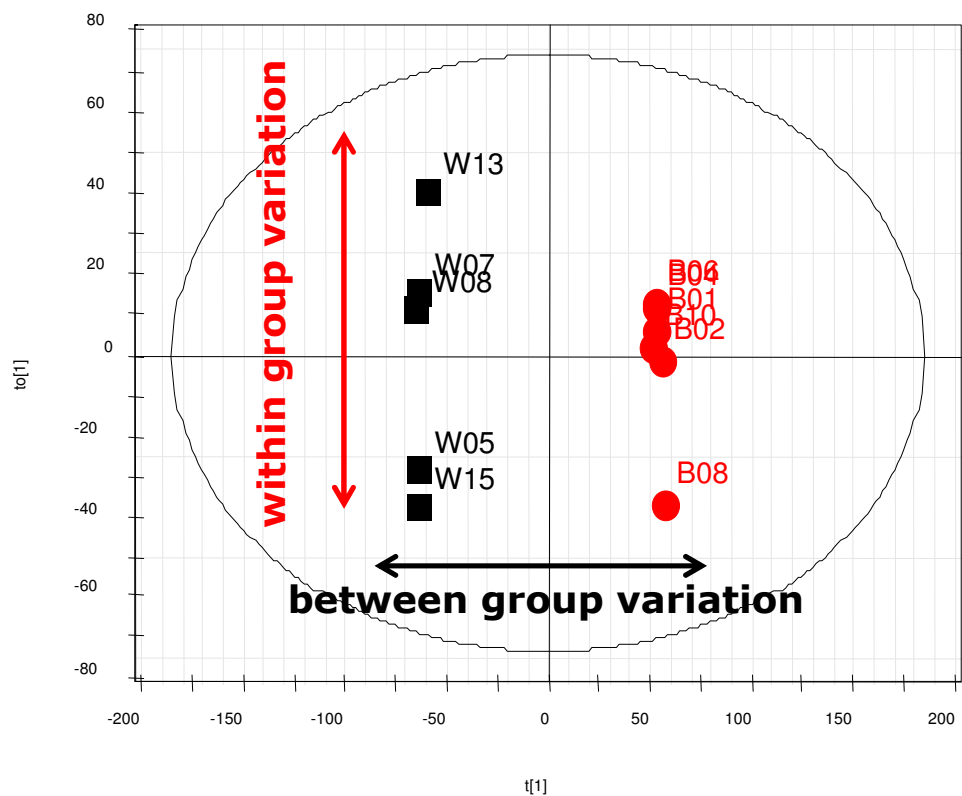
SIMCA (Soft Independent Modeling of Class Analogy)

Cooman's plot

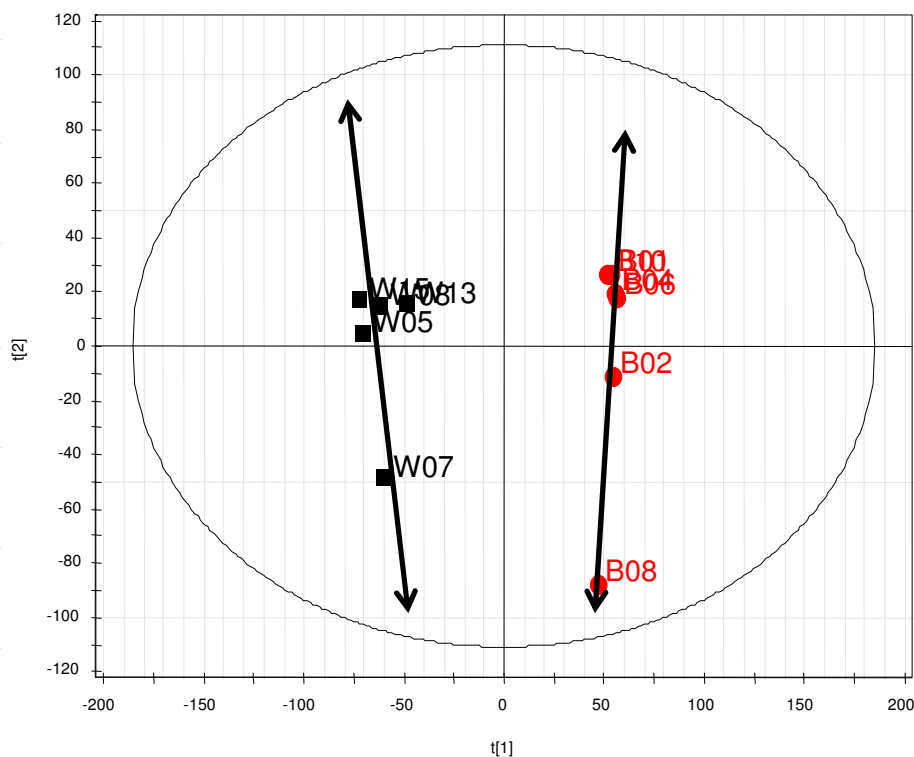


White and Black classes: OPLS-Discriminant Analysis

OPLS-DA Par score plot

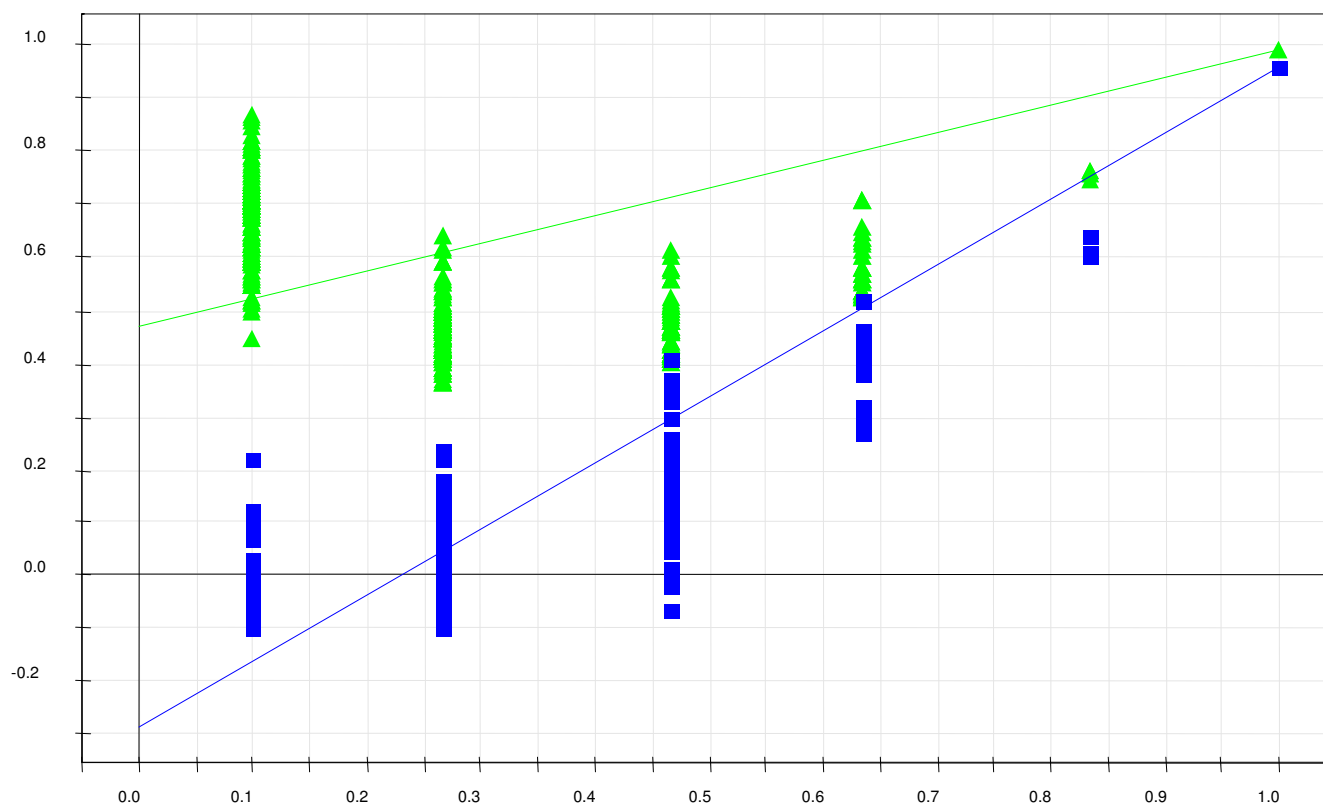


PCA Par score plot

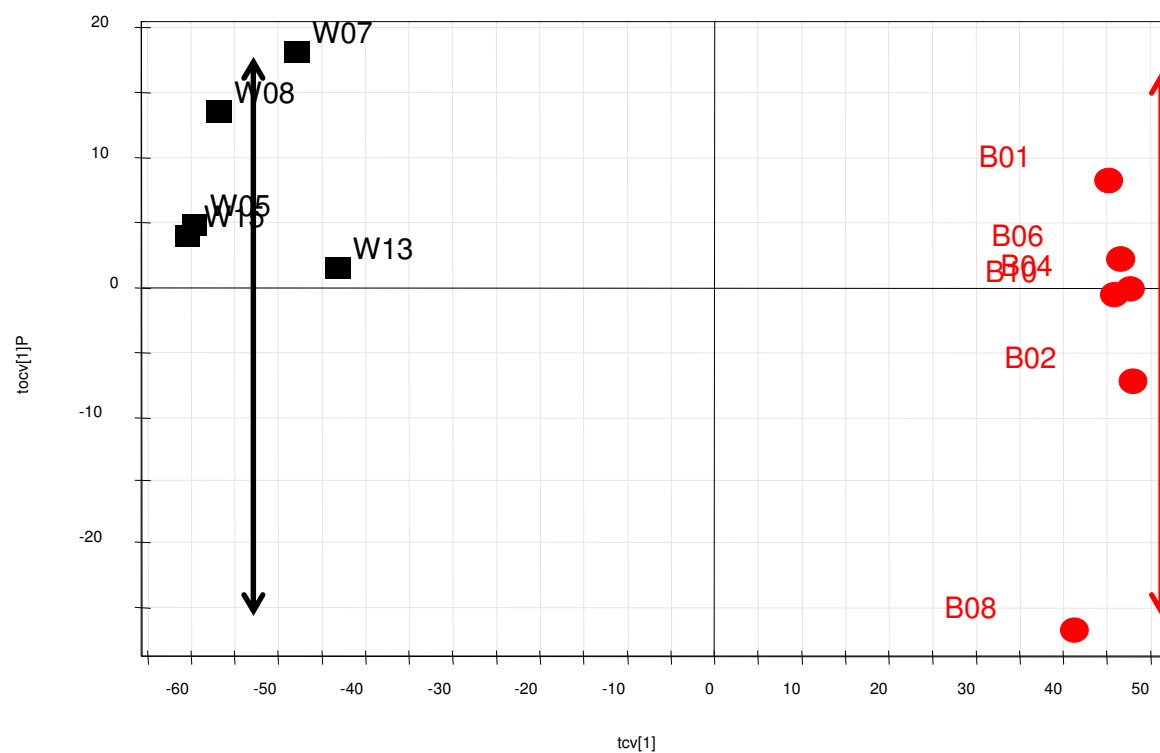


1 + 1 components $R^2 = 0.99$ $Q^2 = 0.97$

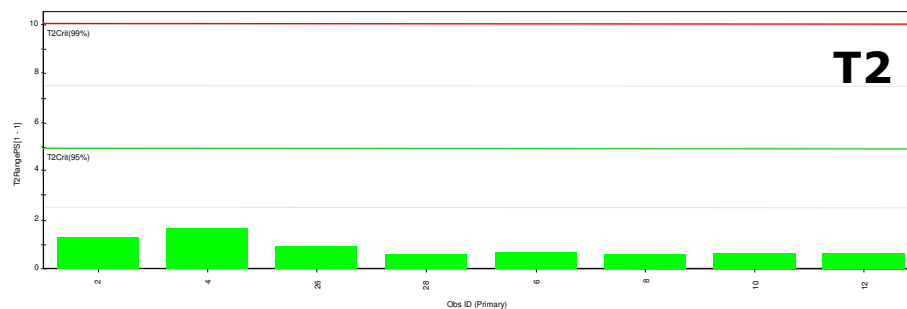
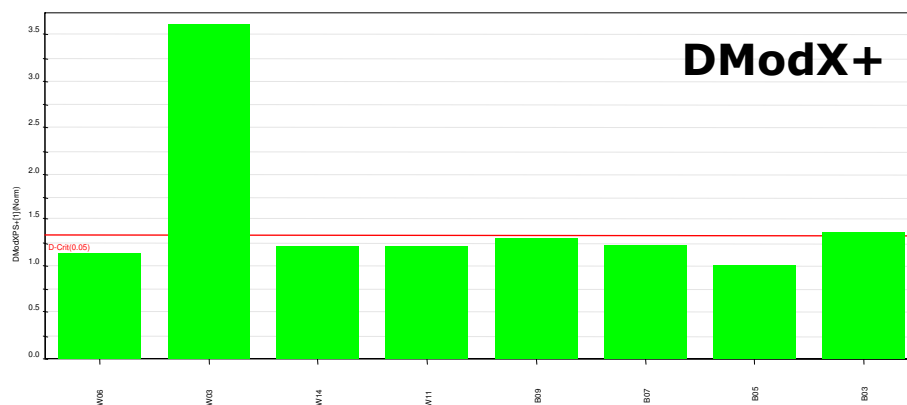
Permutation test



OPLS-DA Par cross-validation score plot



Are the new samples in the Model Space?



Predictions

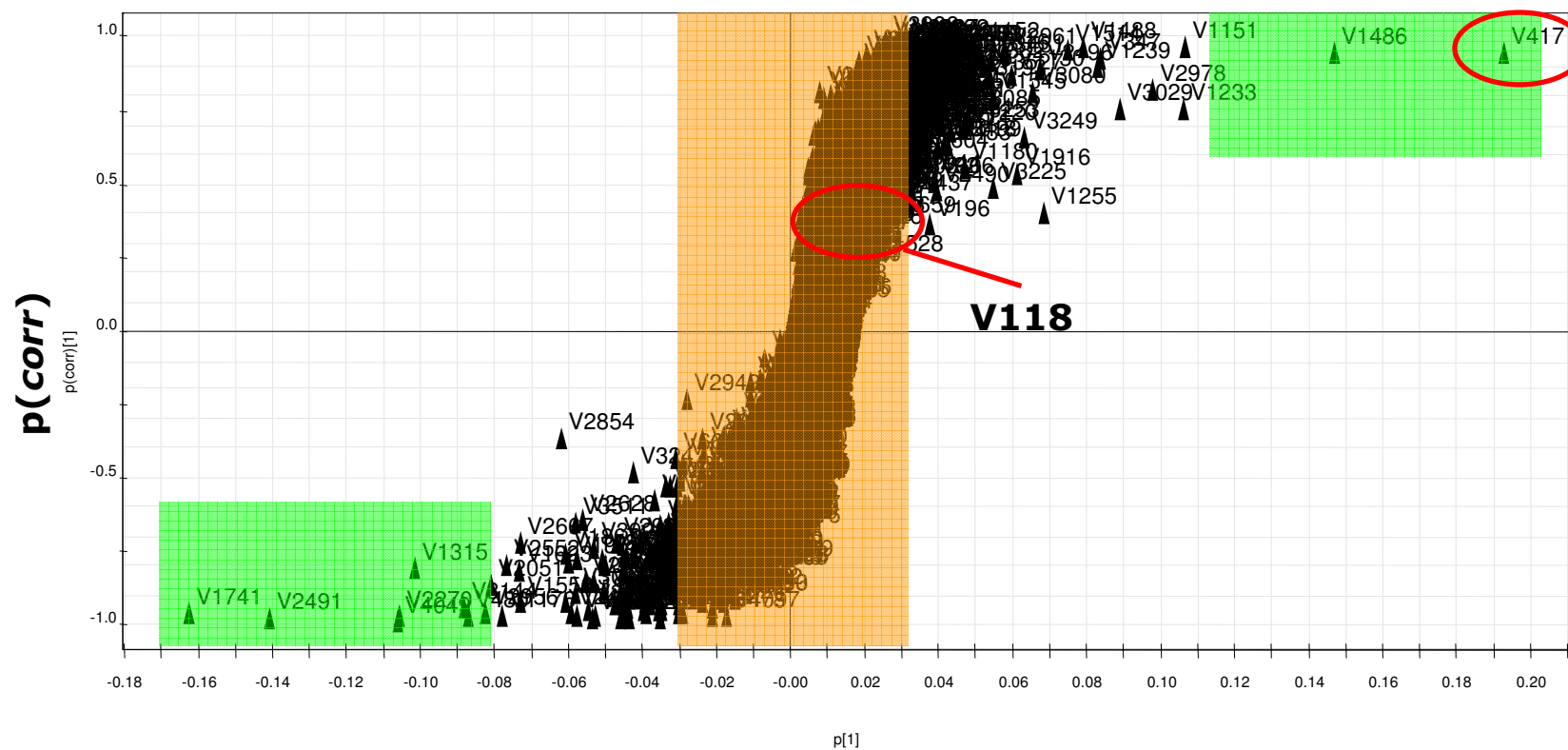
ID	Y_pred[1]	Y_pred[2]
W06	1.039	-0.039
W03	1.130	-0.130
W14	0.952	0.048
W11	0.850	0.150
B09	0.034	0.966
B07	0.067	0.933
B05	0.050	0.950
B03	0.050	0.950

S-plot

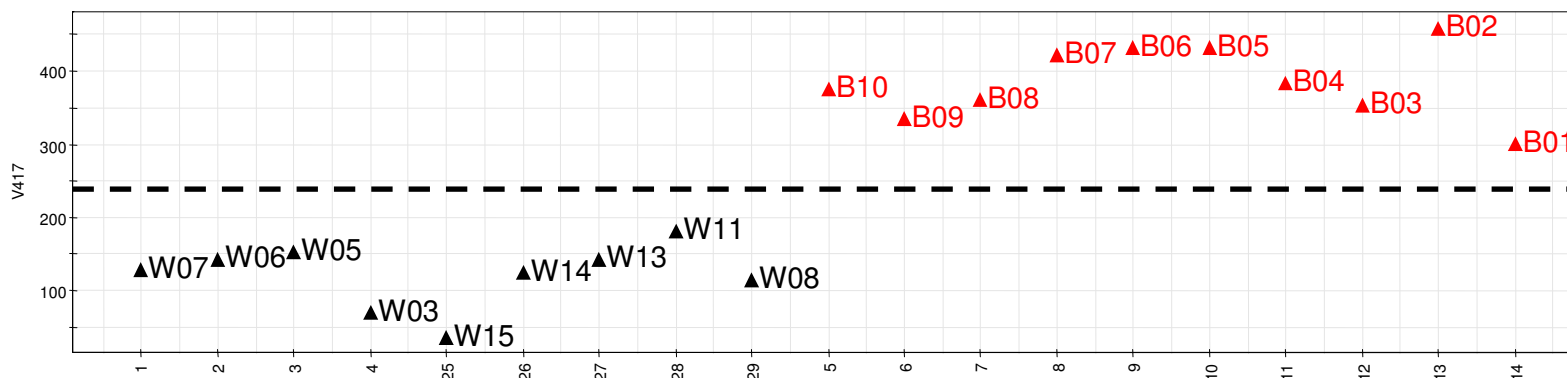
Wiklund S. et al., Anal. Chem. 80 (2008) 115-122

$$\text{Cov}(\mathbf{t}_1, \mathbf{X}) = \frac{\mathbf{t}_1^T \times \mathbf{X}}{N-1} = \mathbf{p}[1]$$

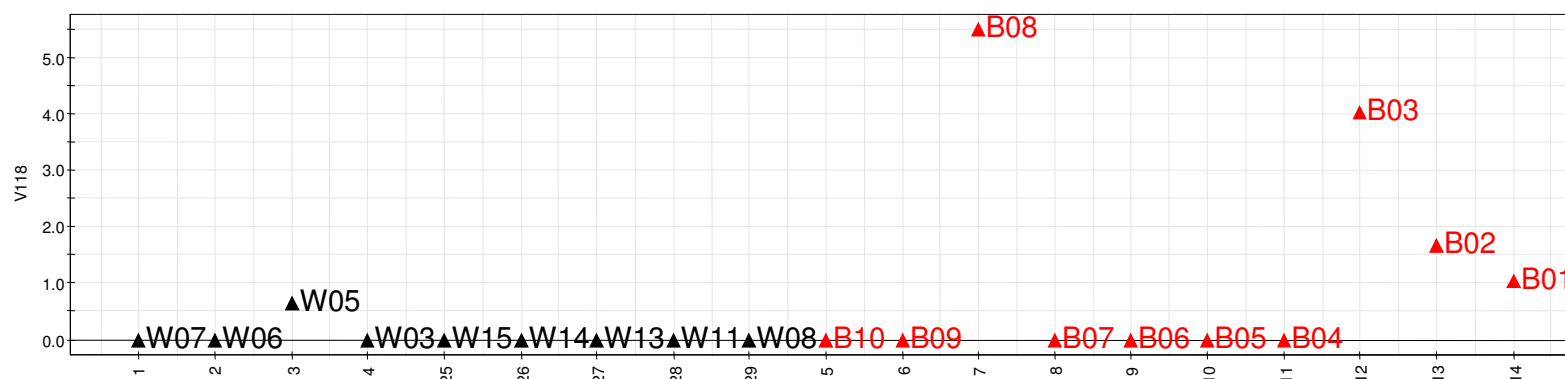
$$\text{Corr}(\mathbf{t}_1, \mathbf{X}) = \frac{\text{Cov}(\mathbf{t}_1, \mathbf{X})}{\sigma_{t_1} \sigma_{\mathbf{X}}} = \frac{\mathbf{p}[1]}{\sigma_{t_1} \sigma_{\mathbf{X}}} = \mathbf{p}(\text{corr})[1]$$



V417 putative marker

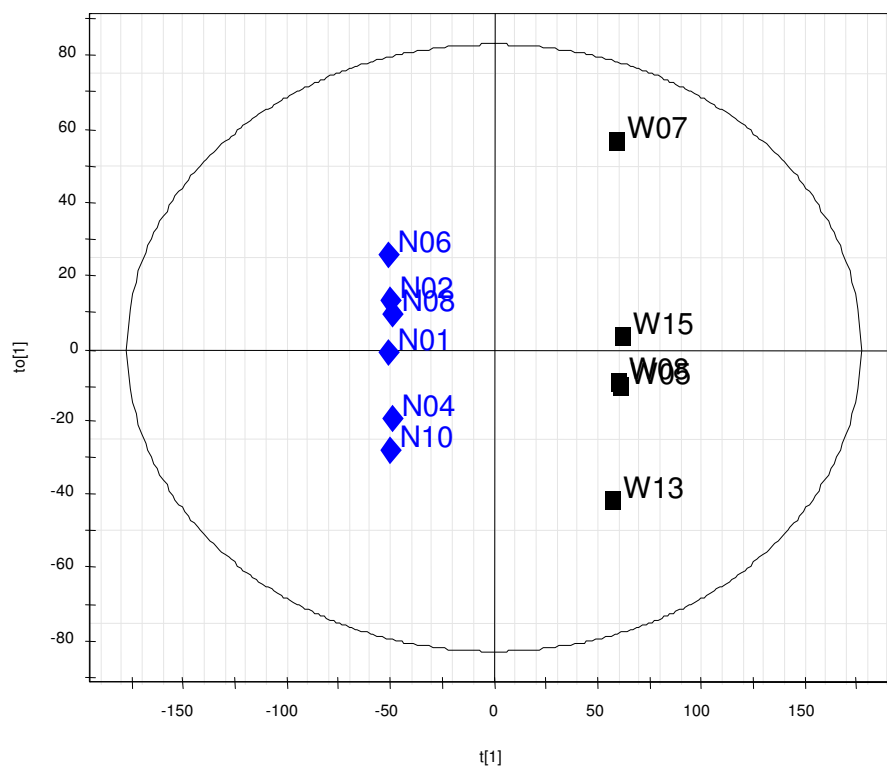


V118



White and Nude classes: OPLS-Discriminant Analysis

OPLS-DA Par score plot

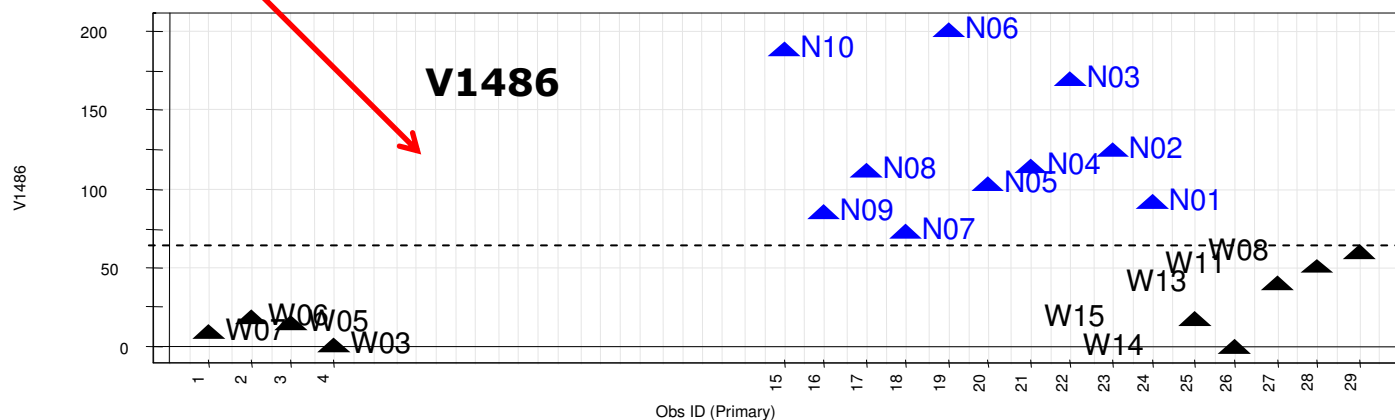
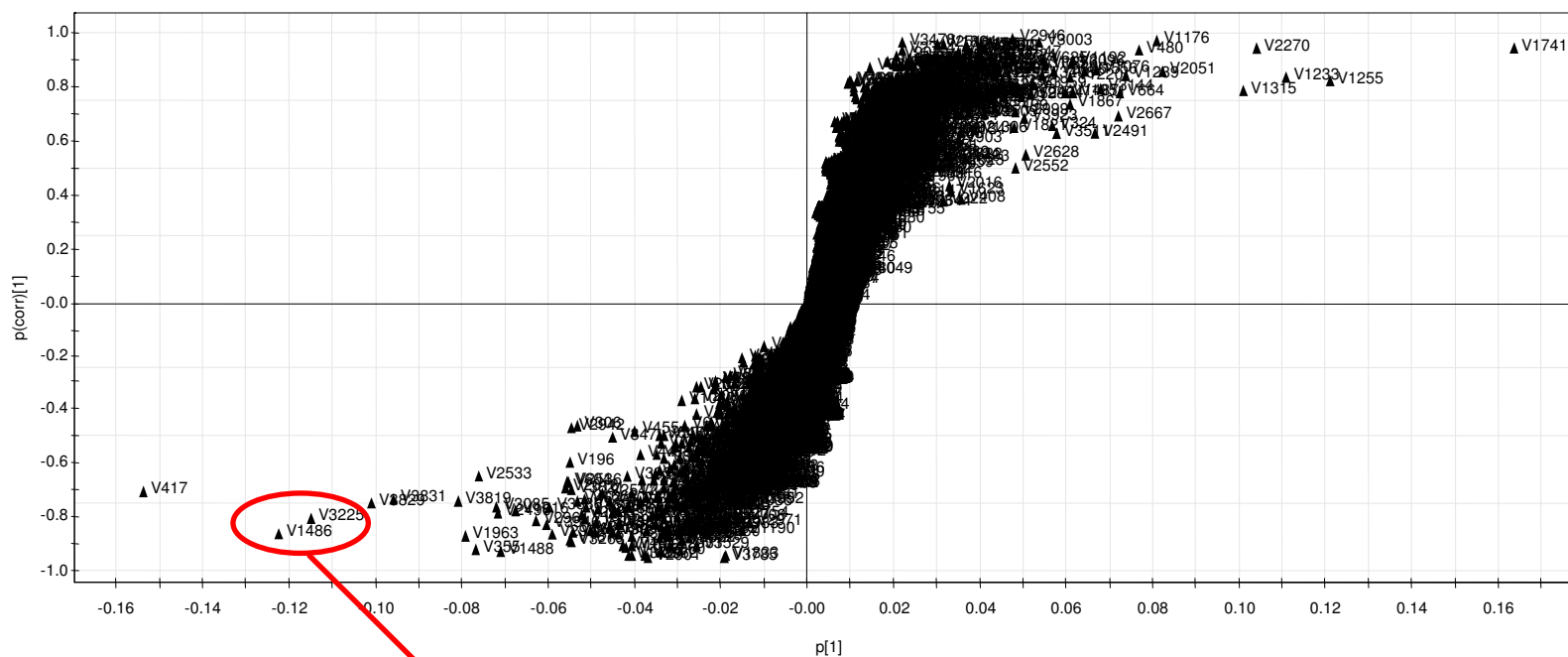


1 + 1 components $R^2 = 0.99$ $Q^2 = 0.95$

Predictions

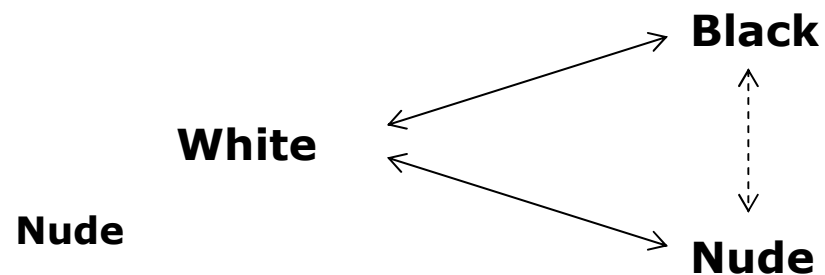
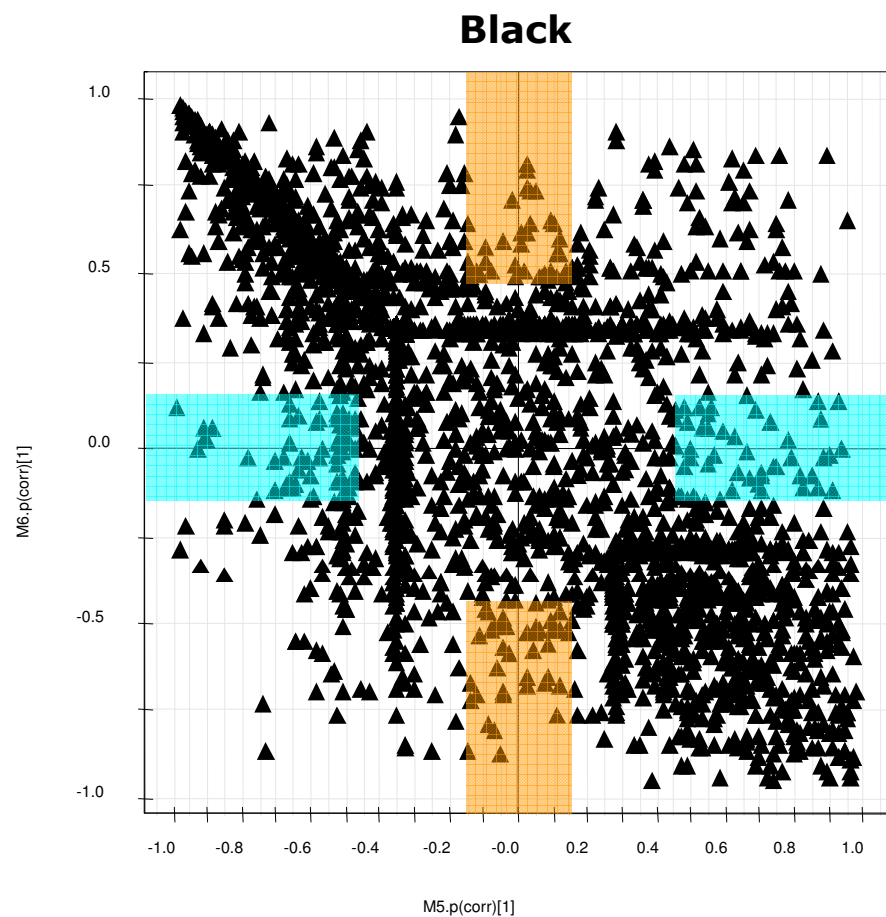
ID	Y_pred[1]	Y_pred[2]
W06	1.010	-0.010
W03	1.075	-0.075
W14	1.002	-0.002
W11	0.912	0.088
N09	0.069	0.931
N07	0.098	0.902
N05	0.064	0.936
N03	0.039	0.961

S-plot



SUS-plot

Wiklund S. et al., Anal. Chem. 80 (2008) 115-122



CONCLUSIONS

- by means of OPLS-DA method good classification models could be generated, able to correctly classify external test sets
- different plots were employed to visually analyze data and results instead of complicated tables of parameters
- S-Plot is a powerful tool to easily identify a putative marker

Statistical analysis was performed by SIMCA P+ 12 (MKS Umetrics)

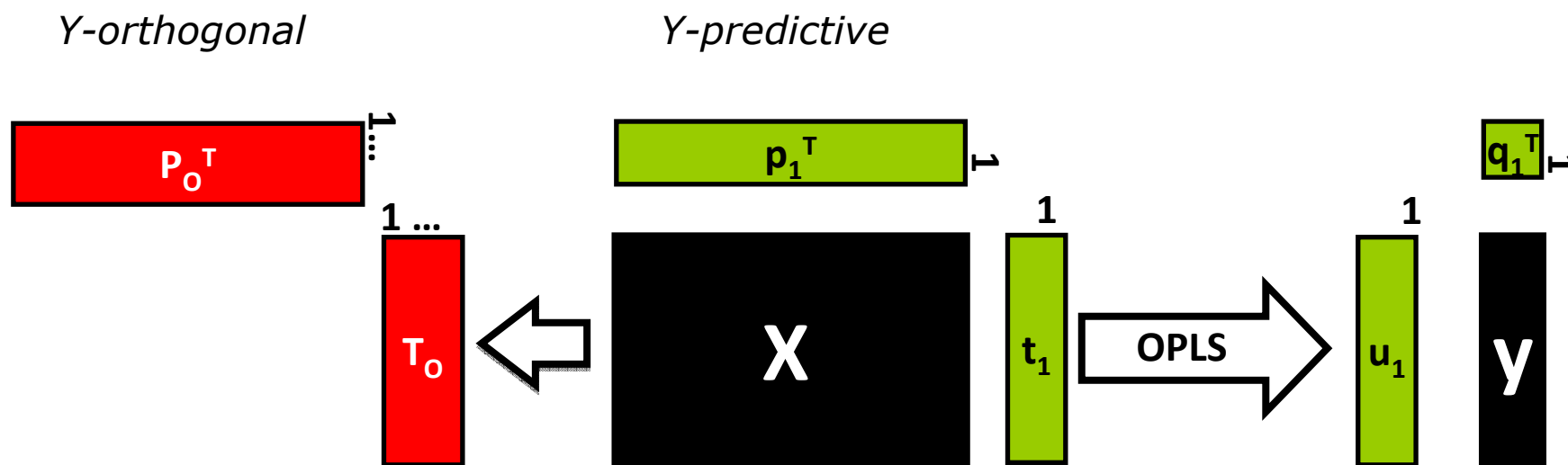


Acknowledgments

Dr. Matteo Stocchero (S-IN) e-mail: matteo.stocchero@s-in.it

Umetrics (Sweden) for data

OPLS decomposition



$$\text{OPLS Model} \begin{cases} X = t_1 p_1^T + T_0 P_0^T + E \\ Y = t_1 q_1^T + F \end{cases}$$