UMETRICS
value from data

An Introduction to MVDA – Some conventional and some more advanced applications

Ing-Marie Olsson, PhD

Application specialist, Application manager

Umetrics AB

Contents

- Introduction to Multivariate Data Analysis (MVA)
  - Why and when MVA?
- Typical problem types and applications
- MVA by Projection methods
  - Principal component analysis (PCA)
    - Apps of PCA
  - Projections to latent structures (PLS)
    - Apps of PLS and PLS-DA
  - Projections to latent structures- Discriminant analysis (PLS-DA)
    - Apps of PLS-DA
  - Summary of PCA & PLS

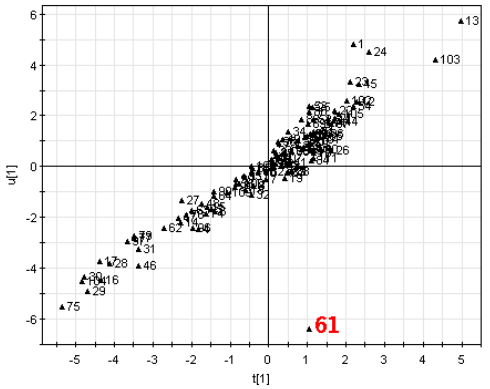
LUNCH

- Method Extensions
  - Multivariate data analysis of ”omics”-data using OPLS
- Manage multiple data sources: Hierarchical modelling
- Multivariate tools in production- in-line monitoring

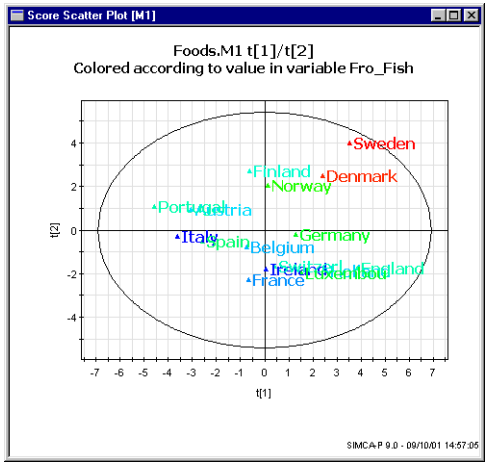
UMETRICS
AN MKS COMPANY

Introduction – Why analyse data at all?

- Data collected from a system
  - may contain noise
  - may be unwieldy or complex
  - may be expensive
  - may contain errors
  - may be wrong

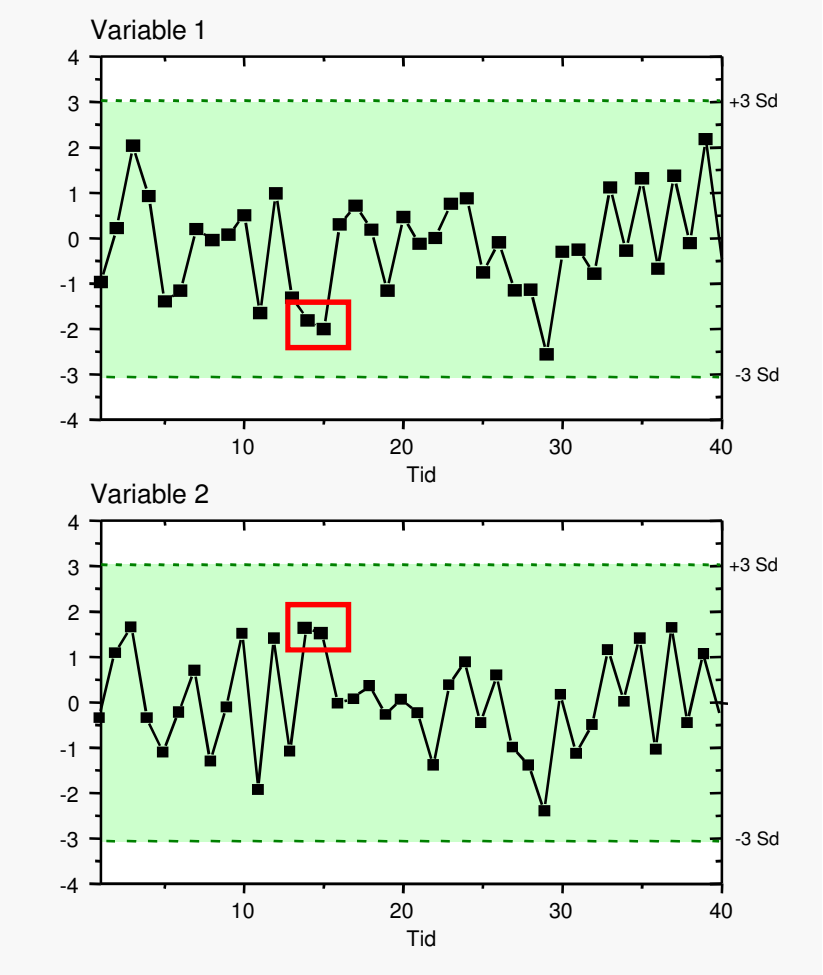


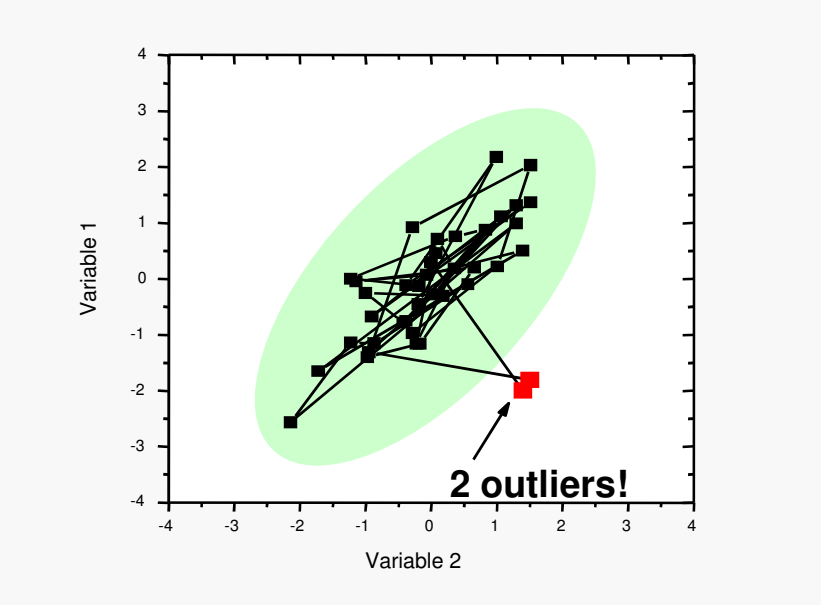
- Information needs to be extracted from that system to
  - Visualise trends
  - Observe patterns
  - Detect errors and outliers
  - Find relationships
  - Make predictions



UMETRICS
AN MKS COMPANY

Data -> information?





- 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!

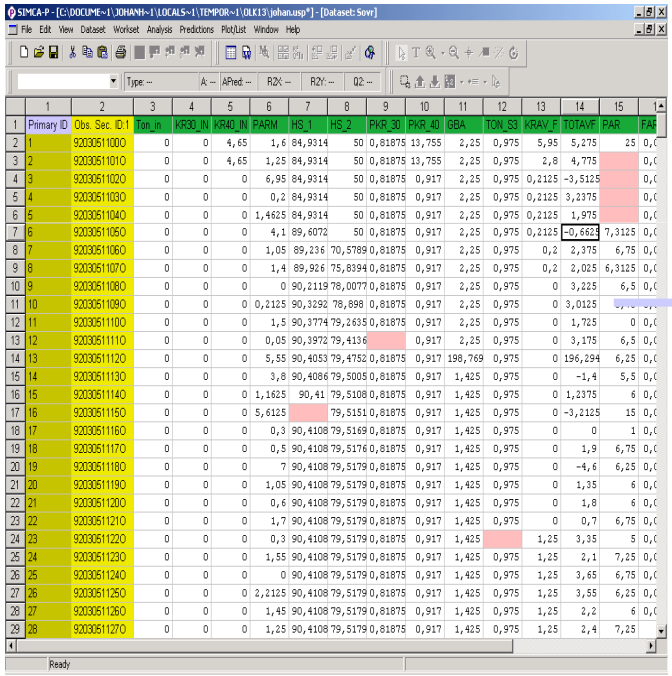
Multivariate Data Analysis (MVA)

Captures the systematic parts in multivariate data sets and visualizes the information in plots and graphs

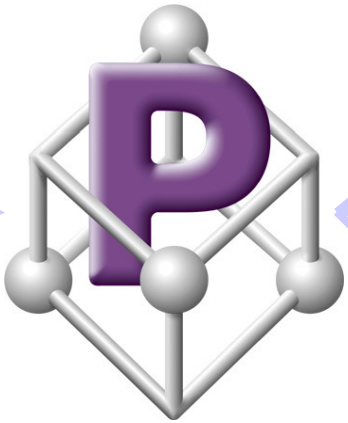
Data

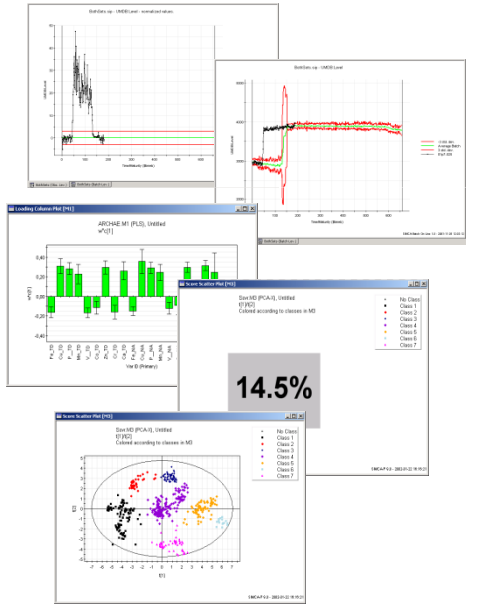
Multivariate Modeling

Information



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	Primary ID	Obs. Sec. ID 1	Ton_in	KR30_IN	KR40_IN	PARM	HS_1	HS_2	PKR_30	PKR_40	GBA	TON_S3	KRAV_F	TOTAVF	PAR	FAR
2	1	92030511000	0	0	4,65	1,6	84,9314	50	0,81875	13,755	2,25	0,975	5,95	5,275	25	0,0
3	2	92030511010	0	0	4,65	1,25	84,9314	50	0,81875	13,755	2,25	0,975	2,8	4,775		0,0
4	3	92030511020	0	0	0	6,95	84,9314	50	0,81875	0,917	2,25	0,975	0,2125	-3,5125		0,0
5	4	92030511030	0	0	0	0,2	84,9314	50	0,81875	0,917	2,25	0,975	0,2125	3,2375		0,0
6	5	92030511040	0	0	0	1,4625	84,9314	50	0,81875	0,917	2,25	0,975	0,2125	1,975		0,0
7	6	92030511050	0	0	0	4,1	89,6072	50	0,81875	0,917	2,25	0,975	0,2125	-0,6625	7,3125	0,0
8	7	92030511060	0	0	0	1,05	89,236	70,5789	0,81875	0,917	2,25	0,975	0,2	2,375	6,75	0,0
9	8	92030511070	0	0	0	1,4	89,926	75,8394	0,81875	0,917	2,25	0,975	0,2	2,025	6,3125	0,0
10	9	92030511080	0	0	0	0	90,2119	78,0077	0,81875	0,917	2,25	0,975	0	3,225	6,5	0,0
11	10	92030511090	0	0	0	0,2125	90,3292	78,898	0,81875	0,917	2,25	0,975	0	3,0125		0,0
12	11	92030511100	0	0	0	1,5	90,3774	79,2635	0,81875	0,917	2,25	0,975	0	1,725	0	0,0
13	12	92030511110	0	0	0	0,05	90,3972	79,4136		0,917	2,25	0,975	0	3,175	6,5	0,0
14	13	92030511120	0	0	0	5,55	90,4053	79,4752	0,81875	0,917	198,769	0,975	0	196,294	6,25	0,0
15	14	92030511130	0	0	0	3,8	90,4086	79,5005	0,81875	0,917	1,425	0,975	0	-1,4	5,5	0,0
16	15	92030511140	0	0	0	1,1625	90,41	79,5108	0,81875	0,917	1,425	0,975	0	1,2375	6	0,0
17	16	92030511150	0	0	0	5,6125		79,5151	0,81875	0,917	1,425	0,975	0	-3,2125	15	0,0
18	17	92030511160	0	0	0	0,3	90,4108	79,5169	0,81875	0,917	1,425	0,975	0	0	1	0,0
19	18	92030511170	0	0	0	0,5	90,4108	79,5176	0,81875	0,917	1,425	0,975	0	1,9	6,75	0,0
20	19	92030511180	0	0	0	7	90,4108	79,5179	0,81875	0,917	1,425	0,975	0	-4,6	6,25	0,0
21	20	92030511190	0	0	0	1,05	90,4108	79,5179	0,81875	0,917	1,425	0,975	0	1,35	6	0,0
22	21	92030511200	0	0	0	0,6	90,4108	79,5179	0,81875	0,917	1,425	0,975	0	1,8	6	0,0
23	22	92030511210	0	0	0	1,7	90,4108	79,5179	0,81875	0,917	1,425	0,975	0	0,7	6,75	0,0
24	23	92030511220	0	0	0	0,3	90,4108	79,5179	0,81875	0,917	1,425		1,25	3,35	5	0,0
25	24	92030511230	0	0	0	1,55	90,4108	79,5179	0,81875	0,917	1,425	0,975	1,25	2,1	7,25	0,0
26	25	92030511240	0	0	0	0	90,4108	79,5179	0,81875	0,917	1,425	0,975	1,25	3,65	6,75	0,0
27	26	92030511250	0	0	0	2,2125	90,4108	79,5179	0,81875	0,917	1,425	0,975	1,25	3,55	6,25	0,0
28	27	92030511260	0	0	0	1,45	90,4108	79,5179	0,81875	0,917	1,425	0,975	1,25	2,2	6	0,0
29	28	92030511270	0	0	0	1,25	90,4108	79,5179	0,81875	0,917	1,425	0,975	1,25	2,4	7,25	

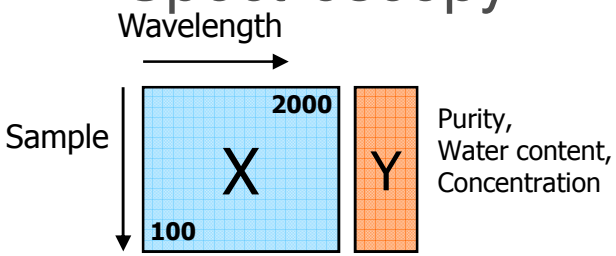




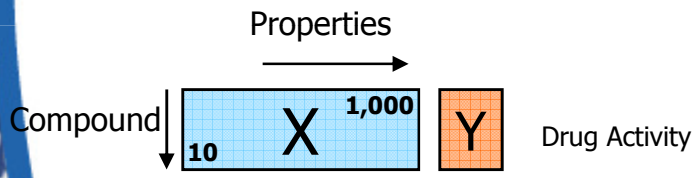
UMETRICS
AN MKS COMPANY

Introduction – When MVA? Typical Application types

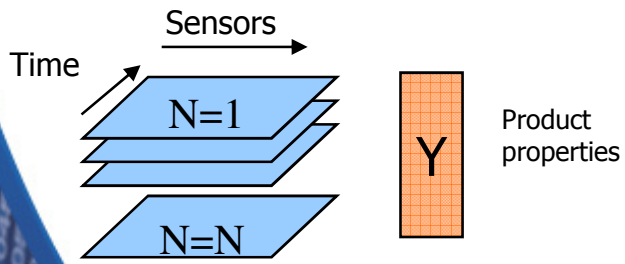
- Spectroscopy



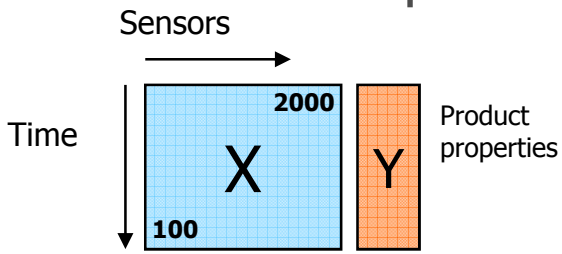
- Drug Design



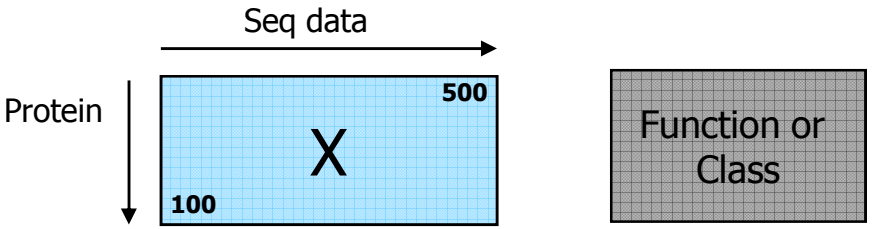
- Batch Processes



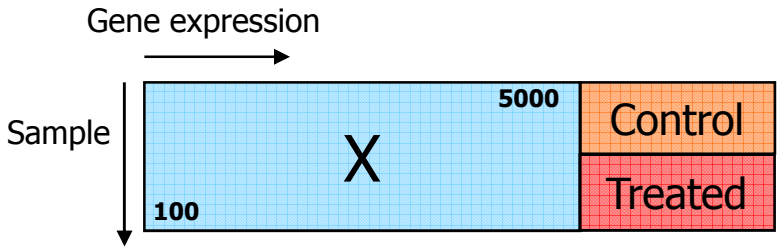
- Continuous processes



- Bio-informatics/Proteomics

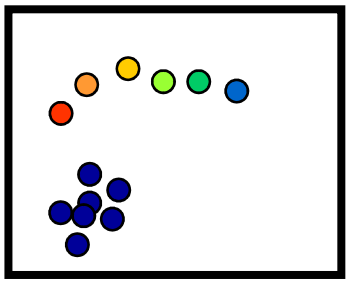


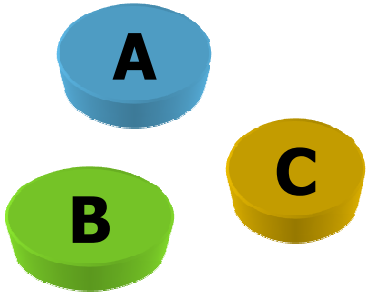
- Genomics

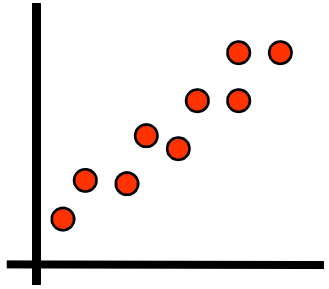


UMETRICS
AN MKS COMPANY

Three Fundamental Data Analysis Questions





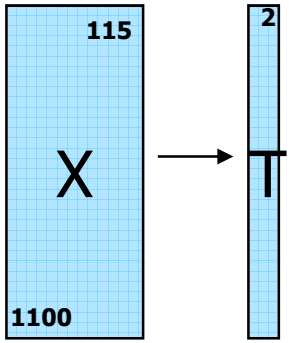


I Overview	II Classification	III Regression
Chemical Property Maps Selection of drug candidates Encoding proteins and DNA sequences Assessing biological variation Trends in quality Process monitoring	Drug Transport Toxicity Mechanisms Control / Treated Classification of raw materials / foodstuffs Genomics and Proteomics Metabonomics	Drug Activity (QSAR) ADMET models Calibration models Online NIR – moisture/ particle size / actives Sensory information Quality prediction Batch Modelling
PCA	**SIMCA / PLS-DA**	**PLS**

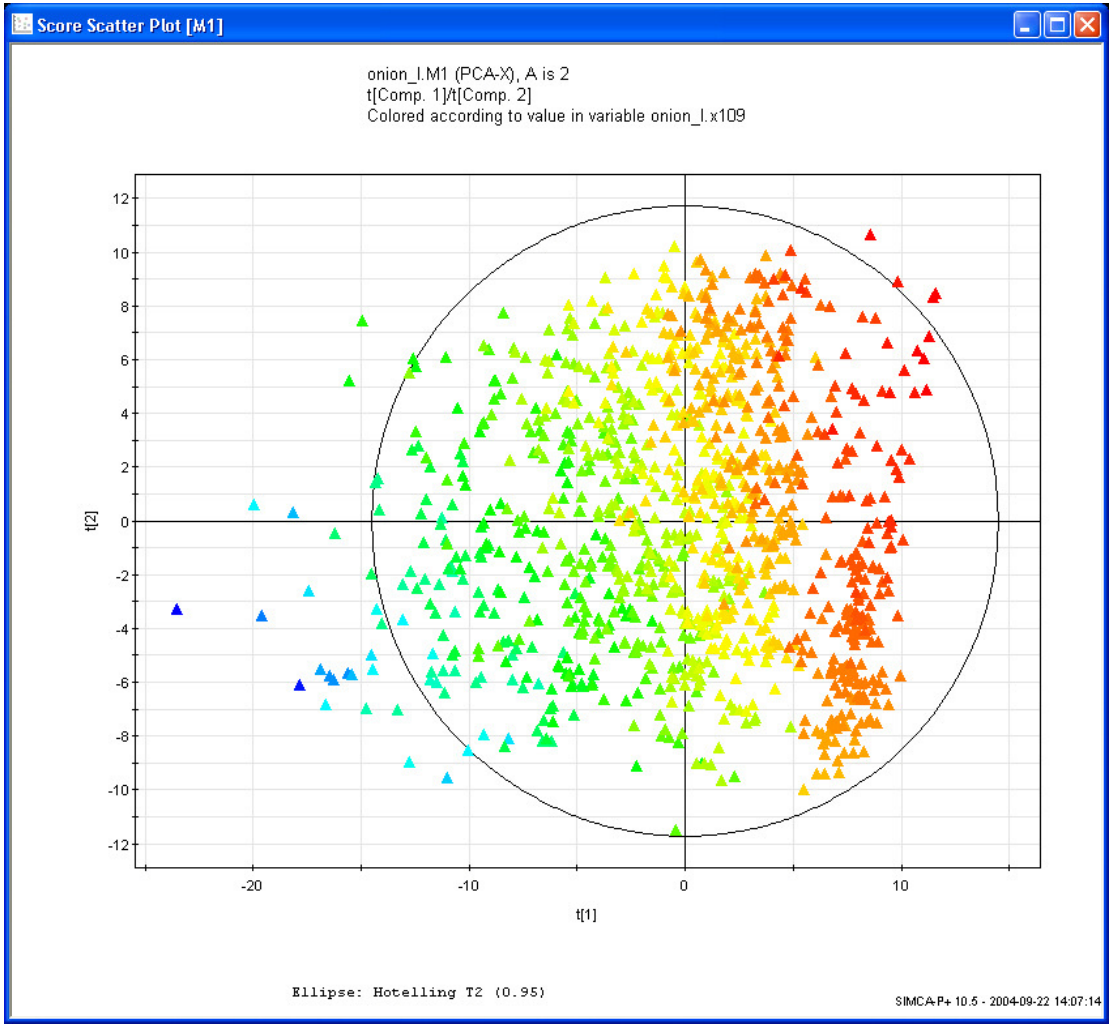
UMETRICS
AN MKS COMPANY

I) Overview: Selection of representative compounds

- 1100+ molecules mapped by 115 variables (B Nordén)

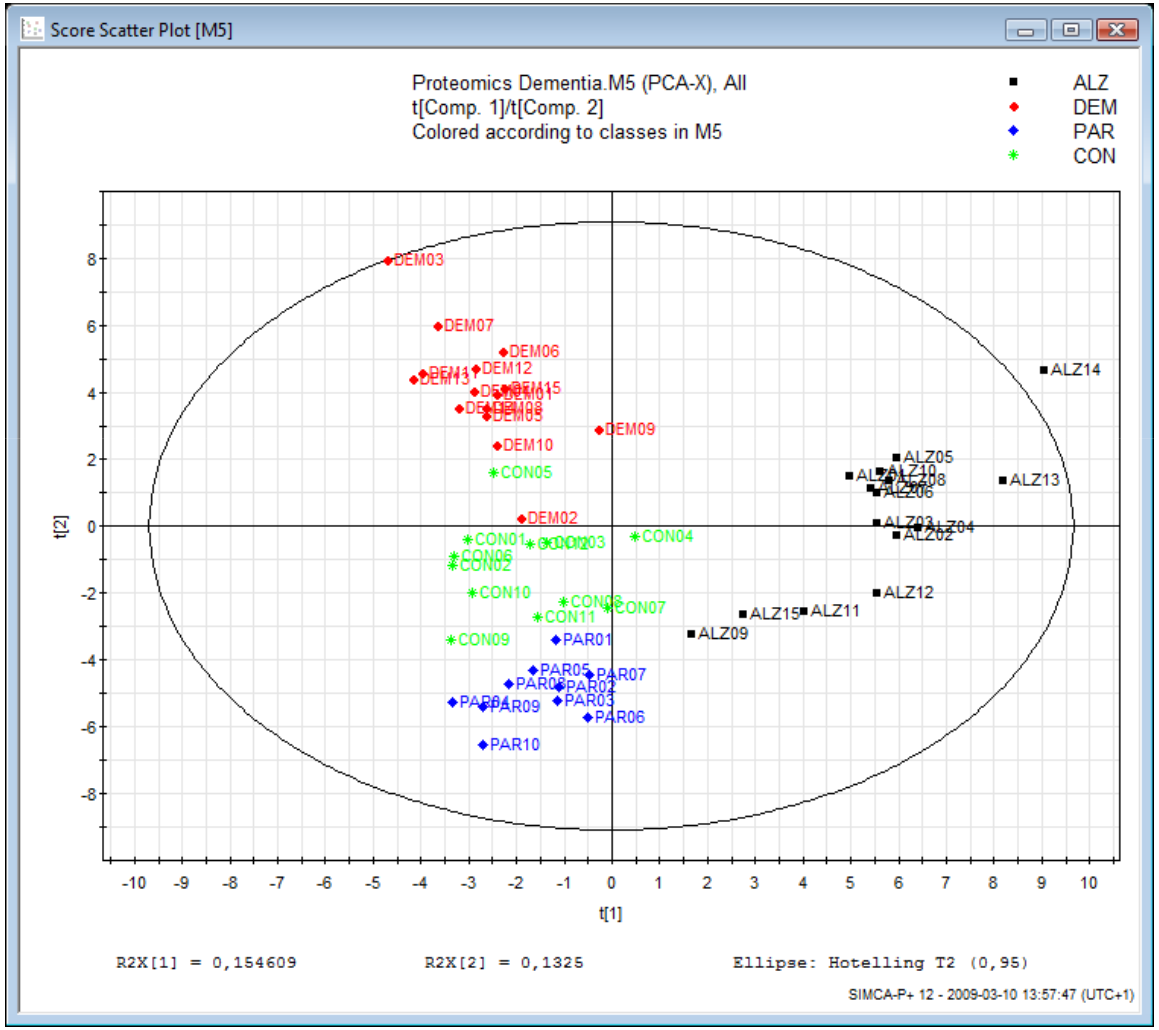


- Two PCs account for 50% of the variance
- Representative and diverse compounds can be seleted (D-optimal design)



II) Classification: Dementia

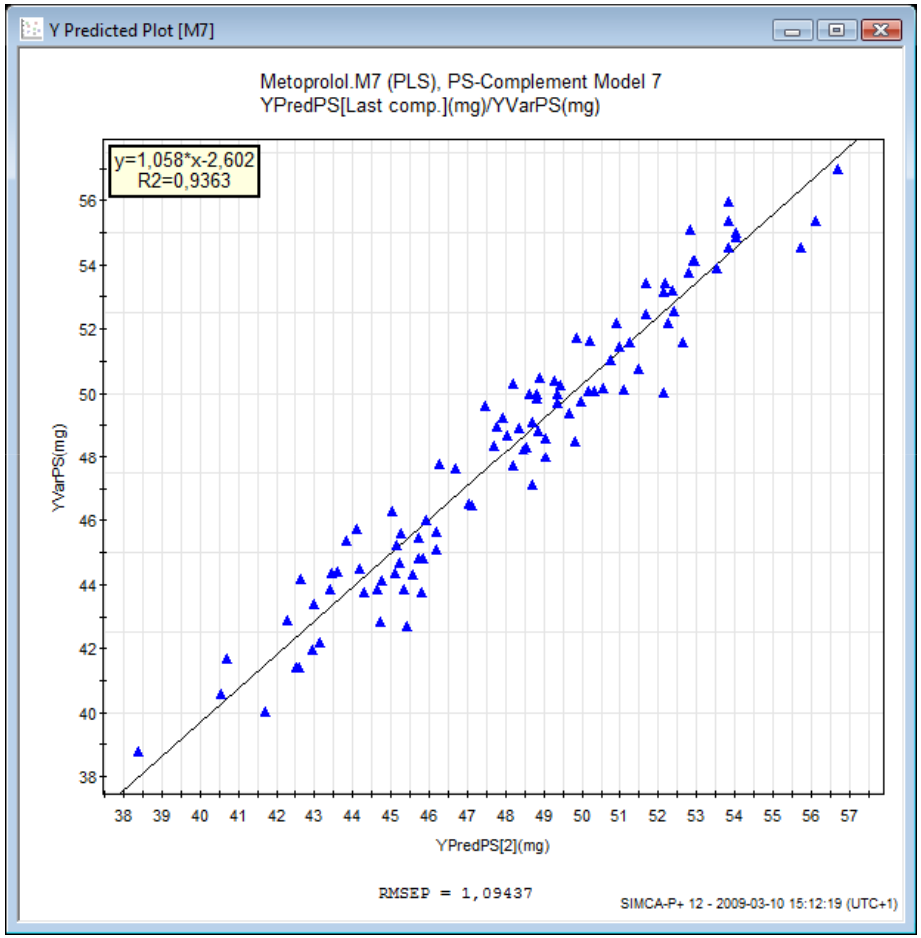
- 52 patients
- Protein level determination
  - 95 proteins
- 4 known classes
  - Frontotemporal demetia
  - Alzheimers
  - Parkinson
  - Control (“healthy”)



III) Regression: Predicting API content in tablets

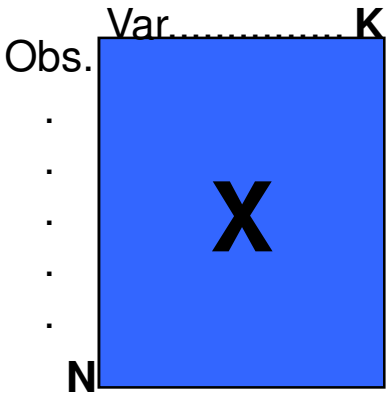
- Metroprolol content in tablets
- NIR measurements
- Non-destructive and fast analysis

- 99 observations for model building
- 99 observations for model validation
- 387 variables



Introduction to Principal Component Analysis (PCA)

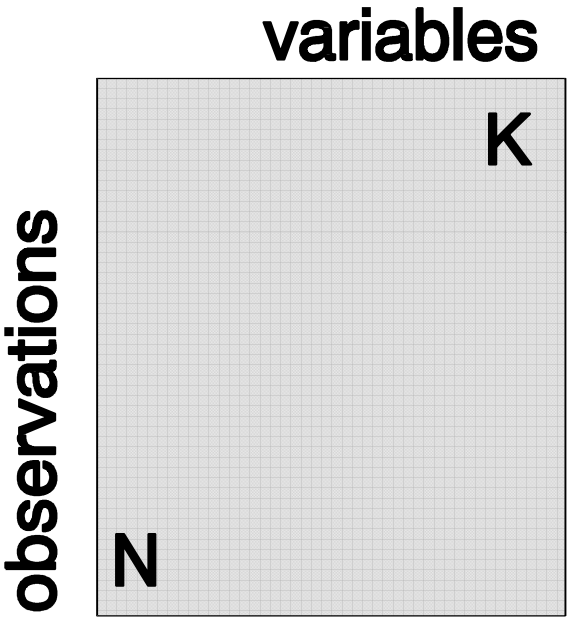
- PCA provides and Summary and overview of a data set **X**
  - Classes, groups among observations?
  - Deviating observations, time-trends?
  - Correlation patterns between variables?
  - Find variables containing unique information
  - ….



UMETRICS
AN MKS COMPANY

Notations

A data table **X** of dimensions N x K



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

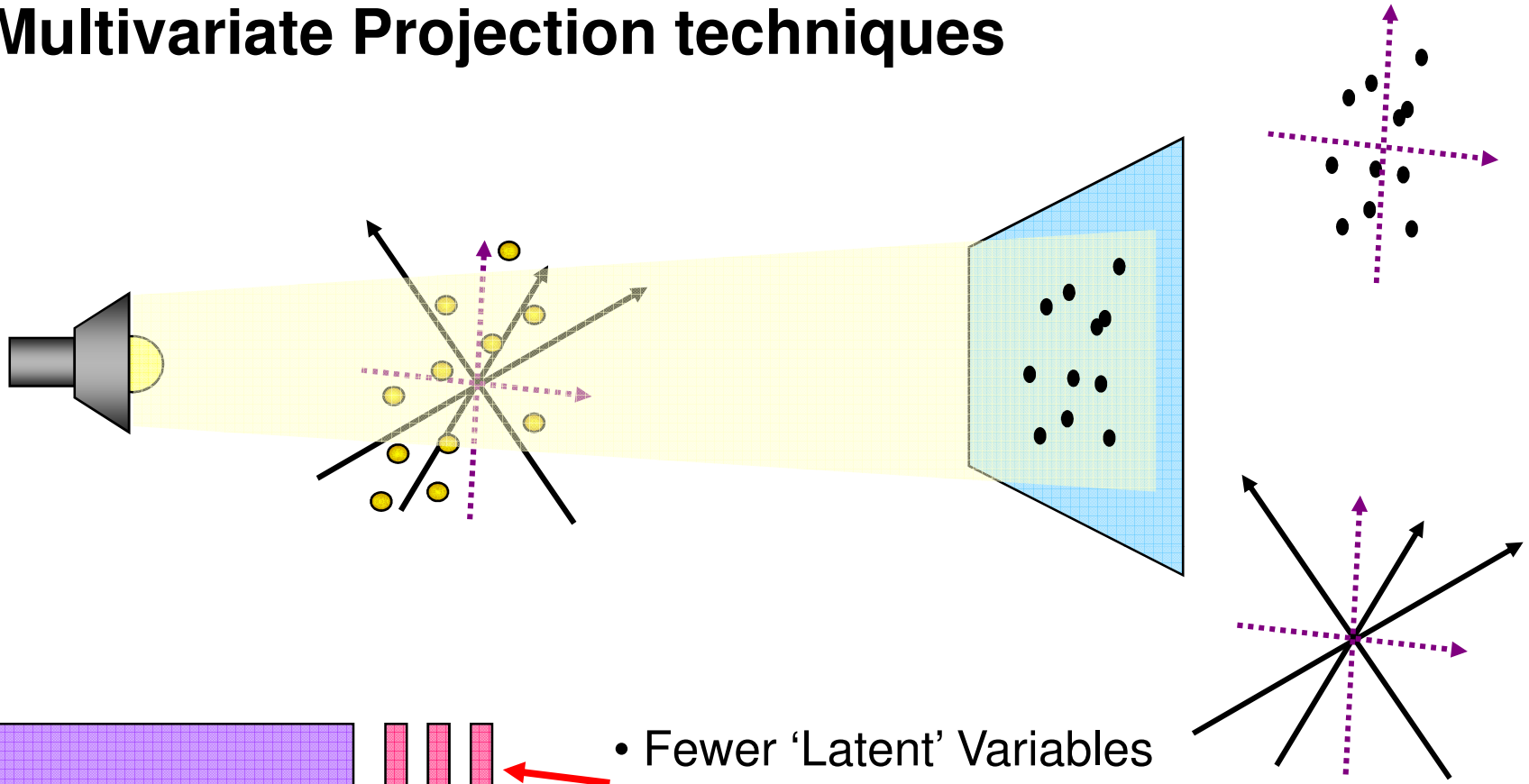
From spectra: NMR, IR, NIR, UV, MS, X-ray, ...

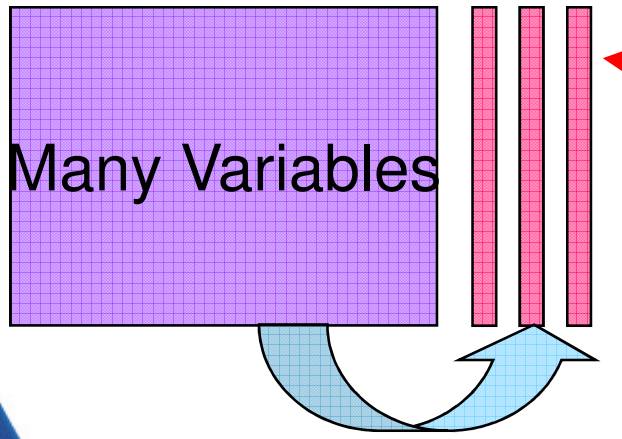
From separation: HPLC, GC, TLC, Electrophoresis

Other: Curve forms, structure descriptors, thermodynamics, quantum mechanics, elemental compositions,..

UMETRICS AN MKS COMPANY

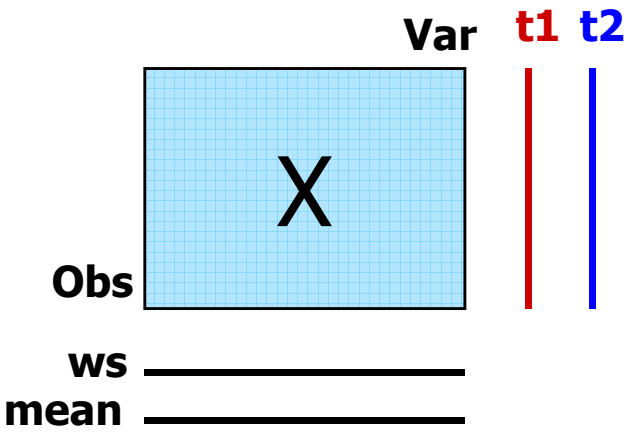
Multivariate Projection techniques

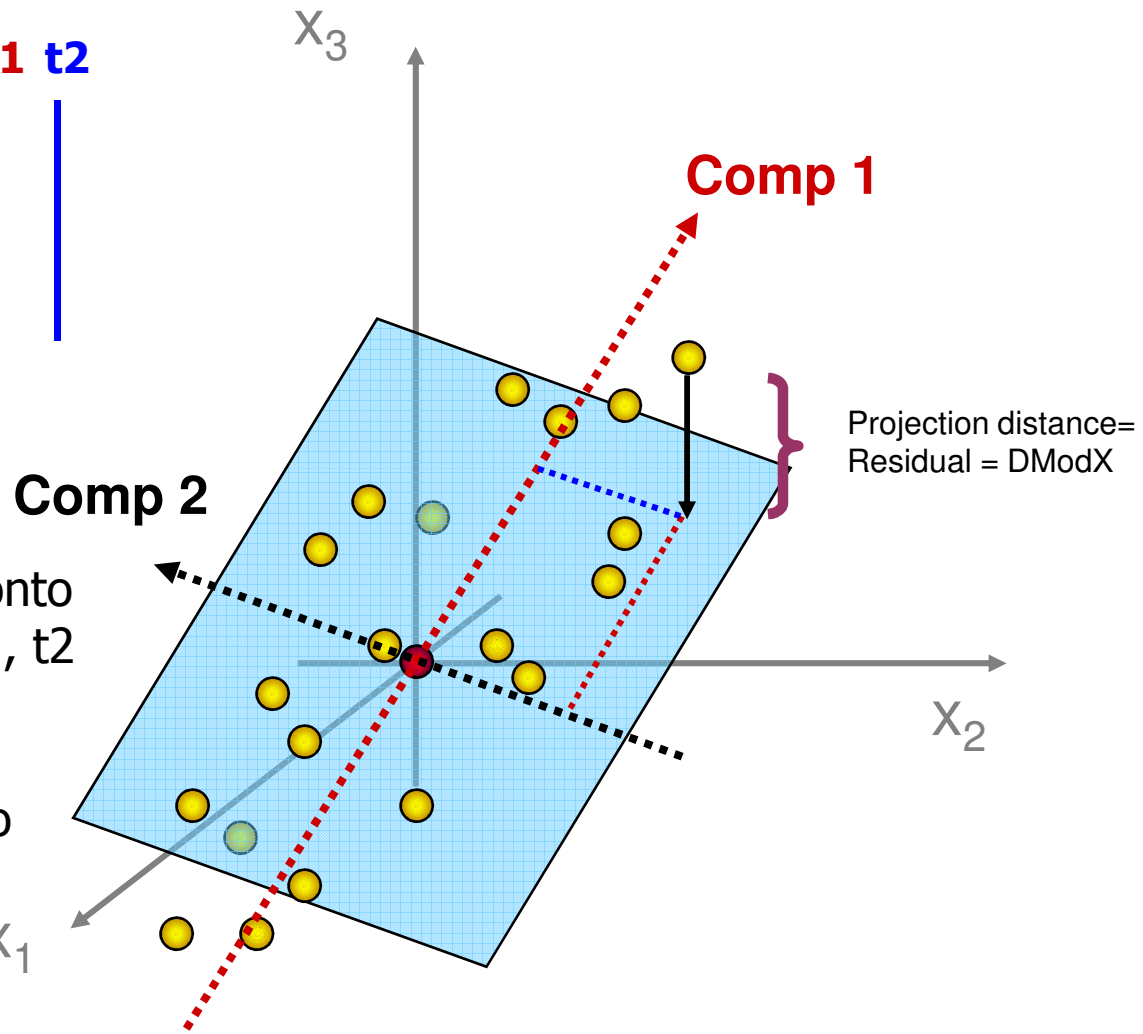




- Fewer ‘Latent’ Variables
- Concise summary of the old
- Finds correlations
  (=systematic variation)
- Leaves noise behind
  (=unsystematic variation)

Projection onto a plane- trends in observations

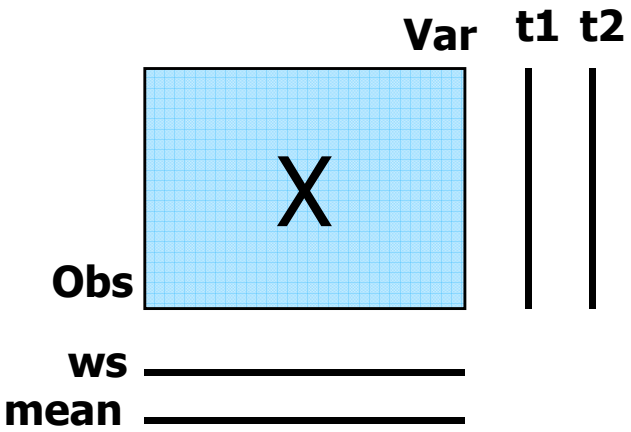




Points are projected down onto a plane with co-ordinates t1, t2

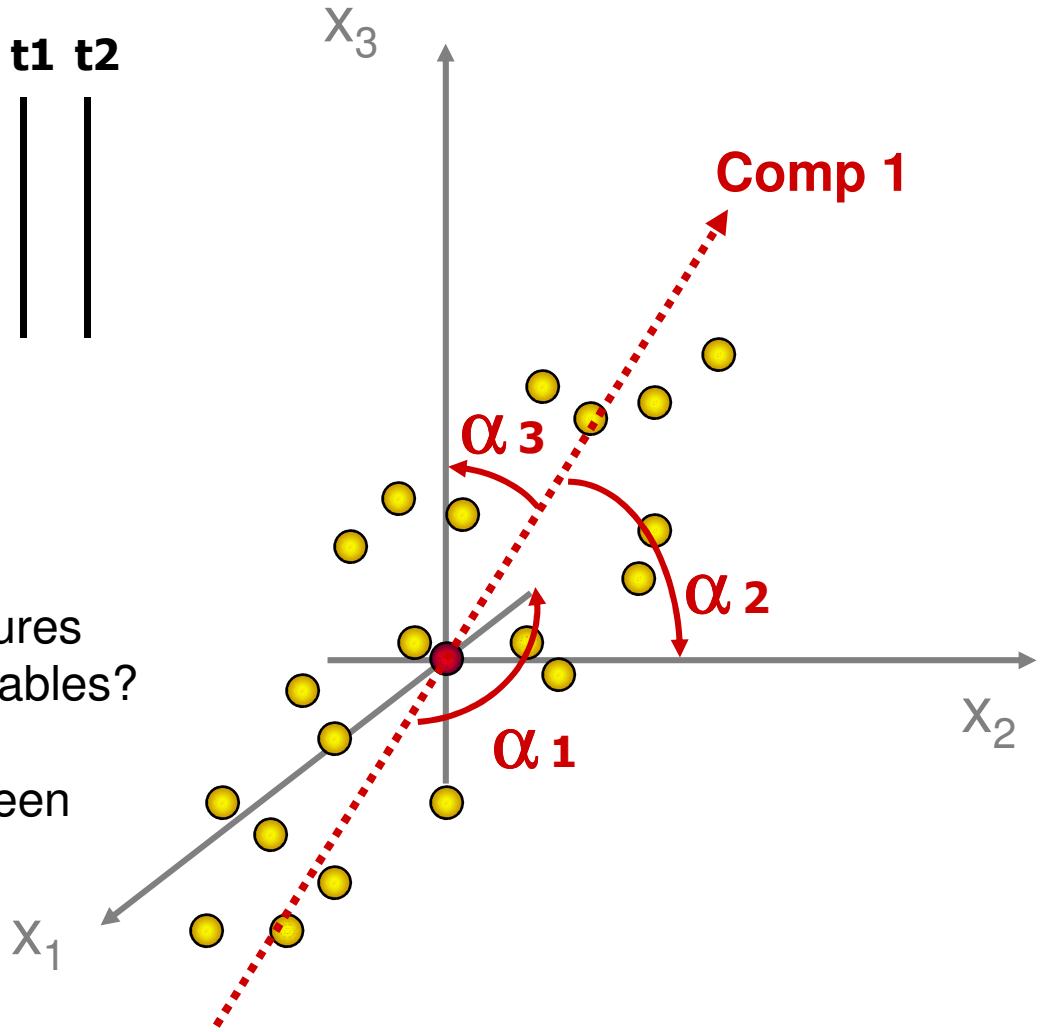
Similar observations close to each other on new plane

What differs between observations?



How do the latent structures relate to the original variables?

Look at the angles between PC and variable axis
Loading p = Cos(α)

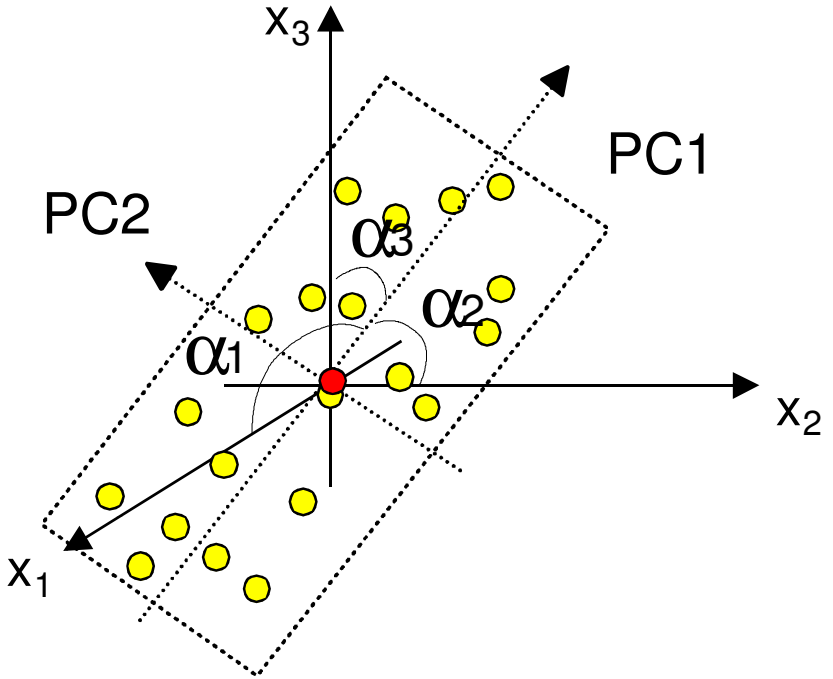


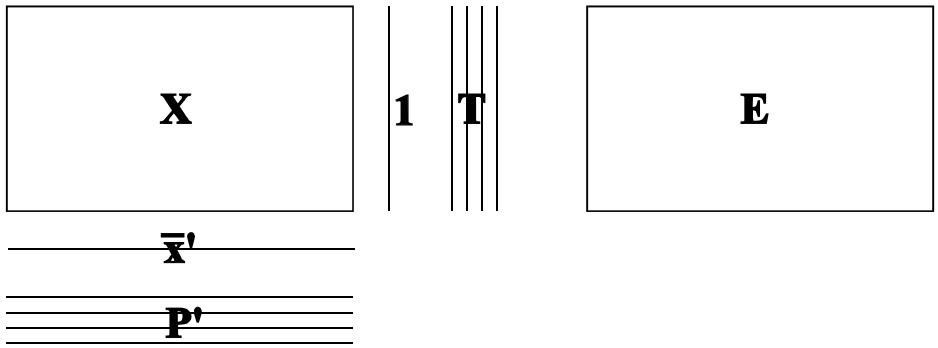
PCA, overview of a data table (data set)

- X is modelled as

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

- 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)





PCA application: The Iris data set

- **Variables:**
  - –Petal width
  - –Petal length
  - –Sepal width
  - –Sepal length
- **Observations:**
  - – 50 speciments of *Iris setosa*
  - – 50 speciments of *Iris versicolor*
  - – 50 speciments of *Iris virginica*.
- Data set known as "The Fisher Iris Data" from 1936
- **Objective:**

Investigate similarities and dissimilarities between the three Iris types

Training data

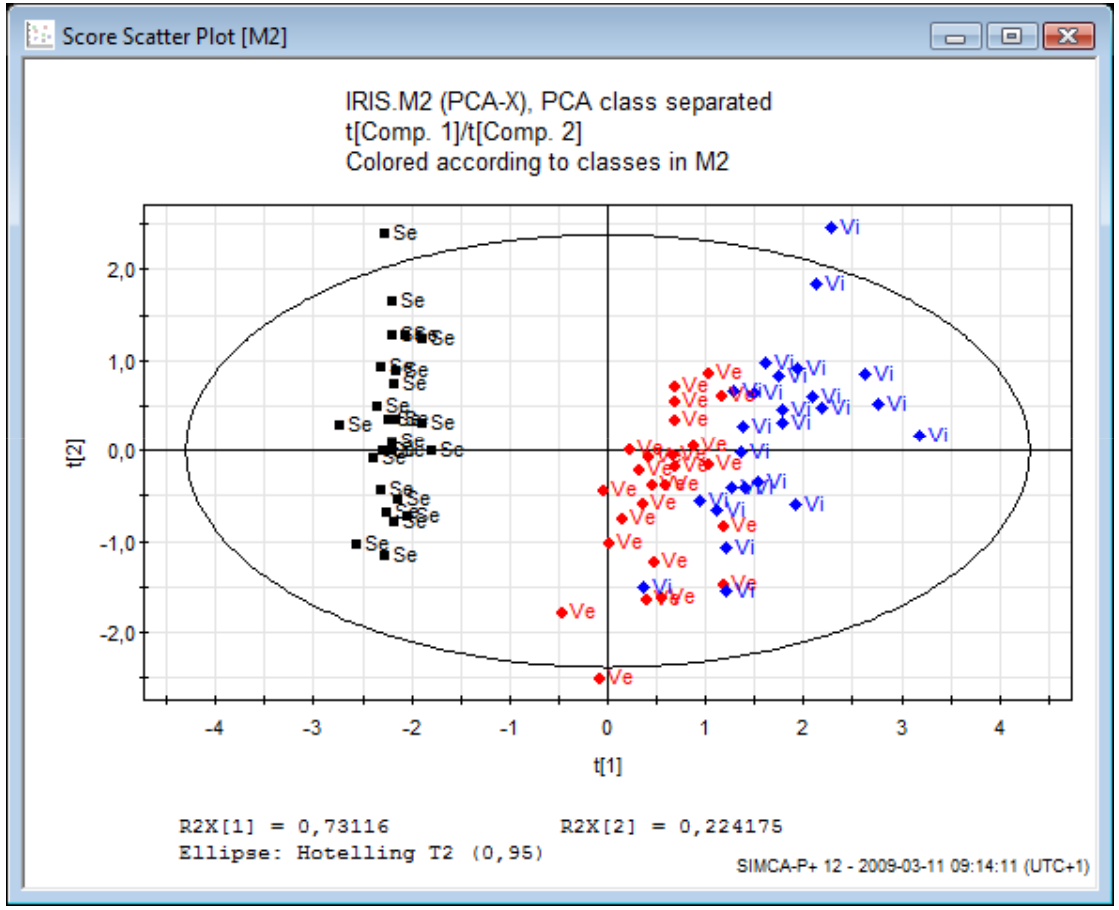
K = 4

N = 75

25 Iris Se. (1 - 25)
25 Iris Ve. (26 - 50)
25 Iris Vi. (51 - 75)

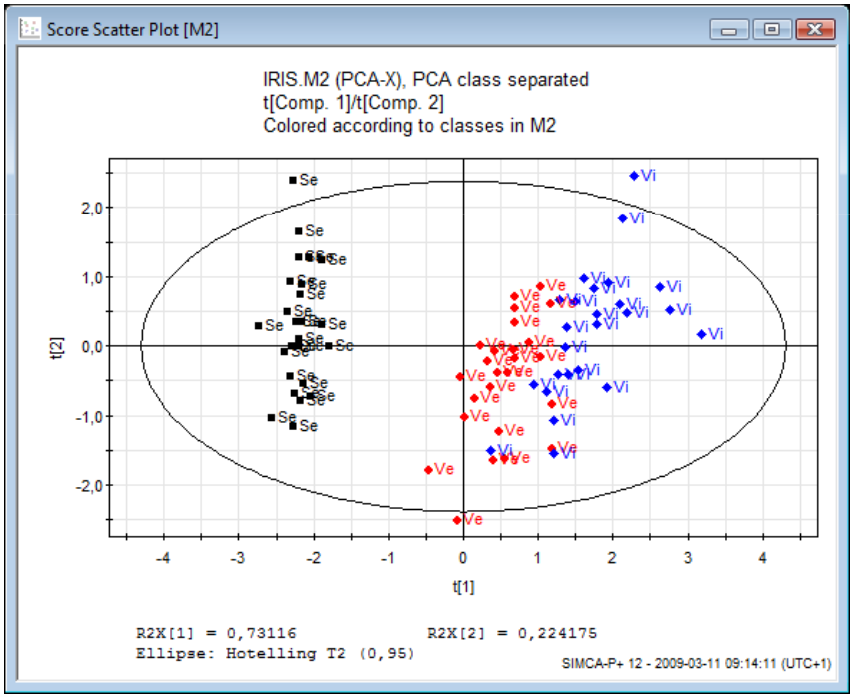
IRIS: Overview of training data (PCA)

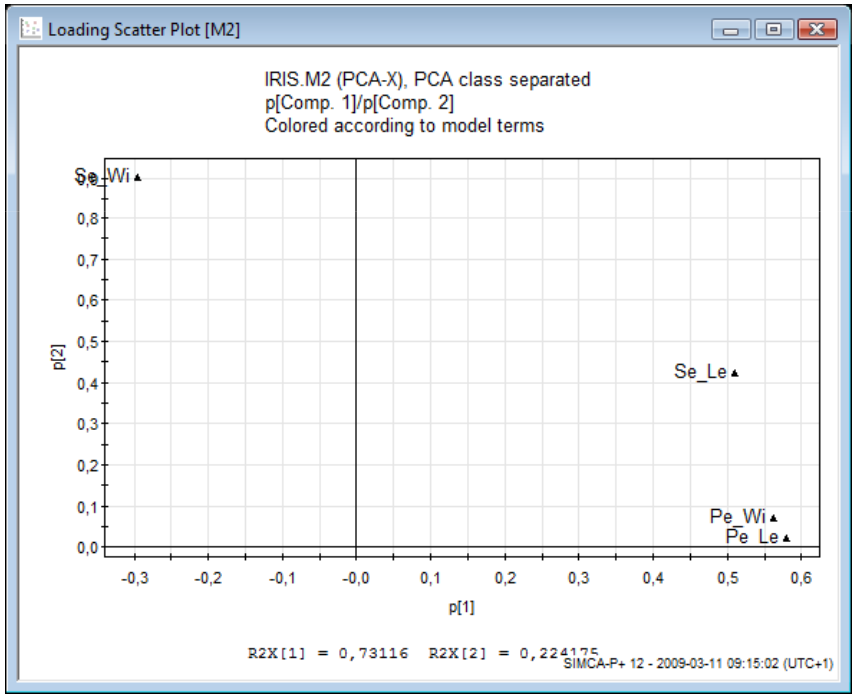
- The PCA score plot shows Setosa well separated from Versicolor and Virginica
- The latter two classes are partly separated
- $R^2 = 0.96$ (A = 2)
- $Q^2 = 0.75$ (A = 2)



IRIS: What is different between the flowers?

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





IRIS: Step 1, A look at the raw data

	Sepal Length	Sepal Width	Petal Length	Petal Width
/ ***Setosa***				
/ Min	4.30	2.30	1.00	0.10
/ Max	5.80	4.40	1.90	0.60
/ ***Versicolor***				
/ Min	4.90	2.00	3.00	1.00
/ Max.	7.00	3.40	5.10	1.80
/ ***Virginica***				
/ 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

PCA application: Clinical Proteomics Data

- Clinically diagnosed dementia patients plus healthy volunteers
- CSF-sampling; quantitative protein arrays (i.e. 95 proteins found in all samples)

Data set (N=52, K=95)

- Healthy volunteers (CON, n=15)
- Alzheimer's disease (ALZ, n=15)
- Frontotemporal dementia (DEM, n=15)
- Parkinson's disease (PAR, n=10)
- Courtesy: J Gottfries, AZ R&D Mölndal

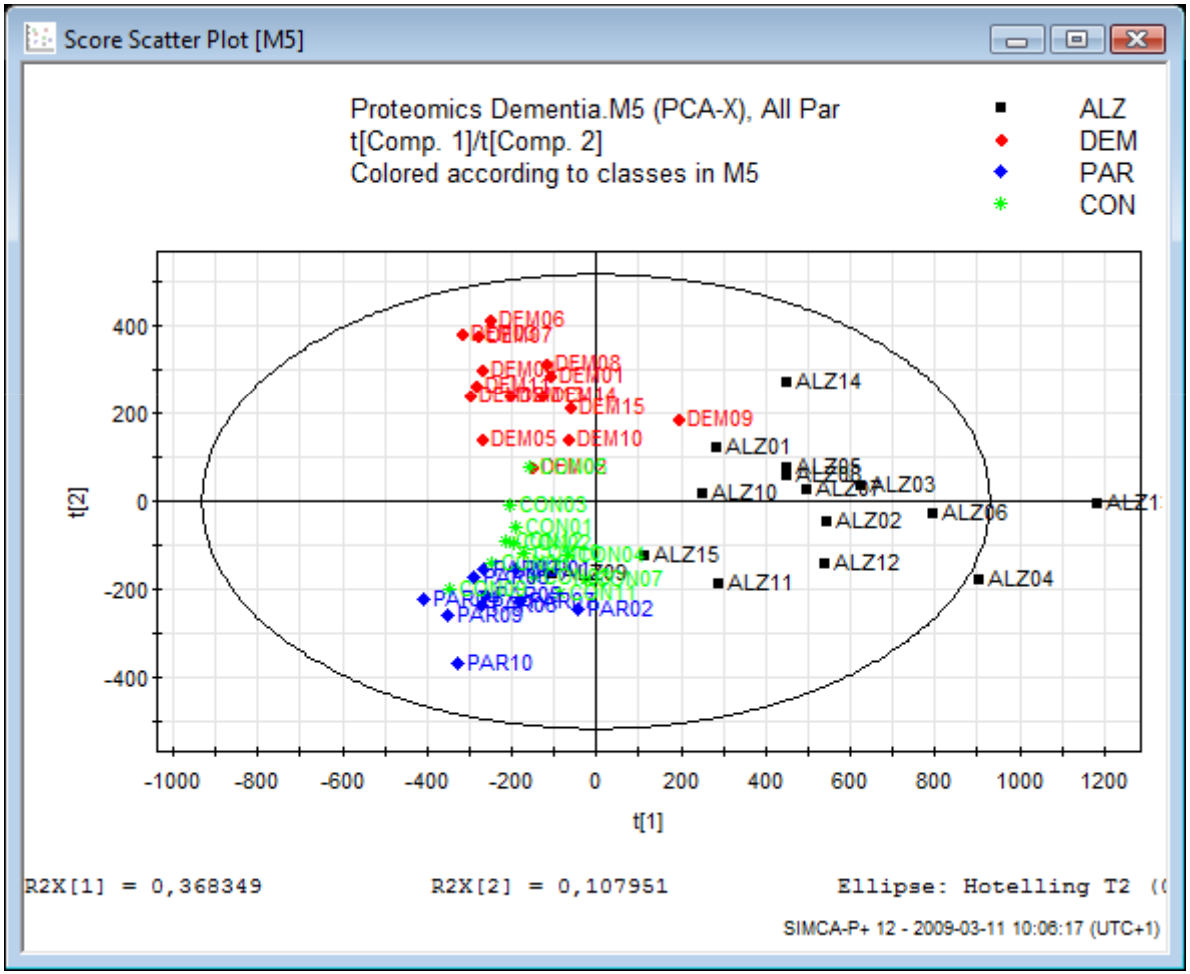
UMETRICS
AN MKS COMPANY

Clinical Proteomics Data- the data

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	Primary ID	Obs. Sec. ID:1	2DGEL_0	2DGEL_0	2DGEL_0	2DGEL_0	2DGEL_0	2DGEL_0	2DGEL_0	2DGEL_0	2DGEL_0	2DGEL_1	2DGEL_1	2DGEL_1	2DGEL_1	2DGEL_1	2DGEL_1	2DGEL_1	2DGEL_1	2DGEL_1	2DGEL_1	2DGEL_1	2DGEL_2	2DGEL_2	2DGEL_2	2DG
2	Var. Sec. ID:1		x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	x21	x22	x23
3	ALZ01	AD 01	1127,89		710,878	1699,08	2908,41	710,878	710,878	7100,98	710,878	1770,19	758,987	1523,31	1920,41	5956,53	1731,63	9095,92	3296,69	20298	15036,3	710,878	710,878	710,878	710,
4	ALZ02	AD 02	2185,78	968,526	771,993	771,993	5305,51	771,993	771,993	7771,99	771,993	7719,93	1053,17	433,21	771,993	3597,81	771,993	4746,13	3917,92	7904,38	14674,6	771,993	771,993	771,993	5528
5	ALZ03	AD 03	1249,39	596,867	1253,42	1253,42	3260,24	1253,42	1253,42	17955,8	1253,42	8526,06	1263,96		2261,21	4625,72	3083,45	1660,9	2338,01	10059,5	10691,9	1253,42	1253,42	1253,42	3092
6	ALZ04	AD 04	1365,65	2358,22	1551,31	1551,31	5787,45	1551,31	1551,31	1551,31	1551,31	2420,48	1158,48	1543,08	1938,25	367,345	690,958	1756,44	3965,28	12099,2	18857,4	1551,31	1551,31	1551,31	6843
7	ALZ05	AD 05	1110,98	1456,79	1092,96	480,422	1751,71	1092,96	1092,96	1092,96	3373,89	6524,25	1641,44	1092,96	4263,75	560,493	4367,56	7276,36	8833,35	9909,14	1092,96	1092,96	1092,96	1092,96	244
8	ALZ06	AD 06	394,797	1448,85	635,119	1357,04	8116,48	1357,04	1357,04	1357,04	1357,04	1357,04	1354,56	2398,44	1342,13	3069,3	1205,43	7891,23	3672,27	8248,74	12546,2	1357,04	1357,04	1357,04	1357
9	ALZ07	AD 07	479,718	5748,57	106,529	1548,55	1216,56	1216,56	4288,82	15897,7	1216,56	6762,37	2123,99	635,968	1496,67	2406,38	4681,2	3343,34	4169,51	12325,8	17927,8	1216,56	1216,56	1216,56	7262
10	ALZ08	AD 08	836,799	1960,16	1282,52	1709,74	5634,62	1282,52	1282,52	14817,3	1092,61	13067,9	3239,24	2761,86	1740,31	2417,32	2502,9	4572,8	3673,29	9109,09	14787,4	1282,52	1282,52	1282,52	2814
11	ALZ09	AD 09	410,78	1184,04	624,228	127,071	5361,96	624,228	2550,5	6626,66	624,228	3086,84	666,357	2069,33	2500,31	2440,1	1740,59	2971,11	3551,54	6447	7763,6	624,228	624,228	624,228	3766
12	ALZ10	AD 10	2359,47	2771,19	938,224	2818,11	1938,28	938,224	3187,55	9105,22	938,224	6560,62	3737,43	1740,66		4880,41	3101,64	3969,41	2878,39	6471,4	12506,5	938,224	938,224	938,224	5120
13	ALZ11	AD 11	1261,73	1077,75	667,421	702,712	1702,46	702,712	702,712	8856,73	702,712	6011,81		2664,3		3679,97	1010,31	5999,22	4188,81	3154,71	11830,3	702,712	702,712	895,864	1072
14	ALZ12	AD 12	1022,14	2582,39	1021,73	336,833	9984,63	1021,73	1021,73	3045,33	1021,73	3424,47	1505,33	804,657		1021,73	4612,75	1279,99	2188,89	8411,48	3633,06	1021,73	1021,73	1021,73	4659
15	ALZ13	AD 13	794,183	2197,6	1879,17	1879,17	12933,5	1879,17	1879,17	29742,5	1879,17	4229,29	1757,39	1124,82	349,658	973,6	1268,18	2056,4	3025,62	3568,68	7437,4	1879,17	1879,17	1879,17	1879
16	ALZ14	AD 14	1695,85	6398,24	117,063	1862,38	11755,9	1535,58	1535,58	25637,3	1880,17	5182,81	2629,61	1535,58	1234,28	4285,82	576,545	8782,35	5327,72	24678	20196,6	1535,58	1535,58	1535,58	6440
17	ALZ15	AD 15	567,759	1380,92	1310,34	2146,02	1310,34	1310,34	1310,34	13509,9	1310,34	4792,16	1139,64	1042,72	2039,82	1425,87	942,421	1965,58	2553,04	6916,53	6322,1	1310,34	1310,34	1310,34	1310
18	DEM01	FTD 01	190,673	130,746	137,85	1832,56	3583,27	3516,77	2170,28	1666,32	1091,61	6230,56	3080,51	1622,66	3769,71	3407,93	4146,2	3909,63	1871,65	6463,84	11236,6	1059,1	1041,4	1013,16	2668
19	DEM02	FTD 02	54,8674	635,192	121,349	586,35	3475,58	2178	1904,27	1816,21	572,375	2829,55	1238,48	121,349	2014,11	1922,81	2537,55	2504,77	1345,09	9317,75	12562,8	778,237	1384,66	1915,48	3958
20	DEM03	FTD 03	101,989	249,086	937,496	1106,79	4867,91	3223,45	2139,75	8349,45	1572,97	4884,42	2155,82	855,093	3738,1	3231,09	1893	6883,66	3688,68	14680,9	12737,4		3081,7	1370,44	2276
21	DEM04	FTD 04	56,4837	304,989	124,272	1136,33	5526,96	3654,99	1570,14	2476,66	2825,54	8115,15	1857,45	124,272	4374,73	2044,59	4114,27	4549,2	2386,08	8145,87	14297,4	1645,26	1831,45	1755,76	2297
22	DEM05	FTD 05		568,967	330,154	1223,37	3145,49	2670,07	2320,01	1620,66	1318,28	3469,48	2479,25	4664,63	887,745	2314,93	3740,67	2357,66	1993,26	7726,36	25743,2	1424,49	1589,42	2435,32	2825
23	DEM06	FTD 06	55,6547	314,935	119,669	276,159	14068,4	2295,96	1119,67	8688,33	2293,45	3873,65	2928,53	1172,36	1945,24	4459,2	1119,67	2542,04	1630,82	16785,8	11716,9	554,058	1029,72	814,033	2053
24	DEM07	FTD 07	136,676	383,77	122,388	858,238	6080,63	2670,46	1978,66	4588,2	1170,81	4606,39	656,769	122,388	3378,01	1954	3771,66	3380,44	2790,49	8509,44	10141,3	812,763	734,256	1218,94	2654
25	DEM08	FTD 08	26,7546	377,086	110,013	110,013	3817,37	1727,67	1003,39	11645,1	3707,23	5869,87	1216,93	704,28	2656,97	1604,35	1458,26	2510,94		9633,91	21970,8		898,751	1027,36	1305
26	DEM09	FTD 09	27,8657	692,734	822,227	1409,3	3199,16	2659,56	1186,48	10905,1	1186,56	8969,38	1186,25	3905,7	992,276	2198,84	3611,35	2237,52	2247,31	8608,59	12570,2	746,004		2497,19	2257
27	DEM10	FTD 10		482,957	144,254	1595,09	1686,88	3544,12	1441,89	10842,7	1096,84	4695,55	3261,65	1680,88	1070,14	4993,2	1894,47	4049,43	2633,43	7581,67	9867,34	700,673	1101,08	918,435	2084
28	DEM11	FTD 11	24,7265	163,415	163,415	163,415	6184,94	2659,58	1524,04	7652,31	1109,52	4069,59	1094,52	1587,69	1512,78	2125,42	2400,3	3095,44	2800,56	14984	11545,9	572,889	764,813	753,153	2598
29	DEM12	FTD 12	602,345	188,766	157,685	1650,97	1157,69	3820,12	2782,3	3414,09	4609,01	3899,51	2451,23	736,443	2667,56	3510,82	2934,49	2835,74	5603,3	8808,11	13420,7	1862,58	2901,93	1252,43	4310
30	DEM13	FTD 13	333,976	108,672	108,672	829,709	1310,71	3149	1008,67	2042,68	5721,64	2033,22	1172,95	1296,04	1945,36	2561,65	1796,46	6657,05	2106,92	7758,78	14552,4	1097,36	684,171	651,534	
31	DEM14	FTD 14	513,274	402,296	826,867	1117,84	1417,39	3222,56	1117,11	8020,05	1034,77	4155,1	790,693	934,982	2409,83	2135,33	2604,17	1103,66	1997,65	15323,9	17849	1697,22	1454,53		1541
32	DEM15	FTD 15	522,585	623,203	437,235	2245,5	2227,95	2131,67	2136,29	5689,56	874,103	5468,65	723,007	991,942	1225,66	1784,92	1414,77	3409,95	2317,57	9699,45	17142,8	1150,14	937,857	3844,81	2423
33	PAR01	PD10	233,342	926,325	414,938	174,527	5092,35	1364,48	1870,24	3775,96	174,527	5513,9	940,887	995,547	1343,55	3131,93	1283,72	3341,87	3082,54	13067,6	14114,1	1200,89	1660,6	2452,55	1809
34	PAR02	PD11	54,9043	394,004	93,4141	249,101	8615,48	2006,49	206,075	3927,94	206,075	4098,68	544,351	1246,8	1415,35	1890,04	1431,17	1945,19	1051,29	11265,1	10236,8	711,931	774,14	621,145	206,
35	PAR03	PD12	864,518	454,859	132,965	975,441	4190,77	1300,02	132,965	3555,91	132,965	4522,84	1188,25	1574,29	1523,26	2126,75	898,867	1275,46	994,489	8133,74	6314,52	677,543	611,343	514,407	2140
36	PAR04	PD13	317,61	280,78	184,533	319,9	2283,77	1154,06	1054,46	1709,98	356,533	2244,01	422,643	2345,75	984,122	1047,67	949,798	865,906	1421,63	3483,46	6196,36	425,026	513,654	391,522	1032
37	PAR05	PD14	396,255	864,846	1366,4	107,37	1878,9	1975,15	1852,31	2249,68	160,614	4958,49	1801,3	2132,65	1423,27	914,591	2986,64	1817,75	1573,04	6916,69	10802,8	2149,17	423,081	1762,96	4230
38	PAR06	PD15	849,23	769,272	625,6	864,279	1202,38	2022,39	1734,44	3673,09	202,909	5319,6	2494,14	2095,5	1427,49	3294,96	1432,06	1962,2	550,2	11730,2	8772,97	202,909		1621,55	2916
39	PAR07	PD16	658,252	781,281	792,426	460,305	2818,41	1475,79	1680,51	3875,25	1044,72	3607,77	1056,31	1771,5	1430,07	2898,22	1423,98	2130,94		6936,92	6714,57	1054,6	425,167	302,458	2596
40	PAR08	PD17	346,893	590,124	175,046	141,574	1141,38	1906,96	141,574	4549,57	141,574	1995,82	1352,21	1563,92	1037,67	2463,83	2596,18	913,851		8943,86	5996,27	141,574	1027,24	1265,1	1689
41	PAR09	PD18	557,792	529,927	305,722		1000,84	1212,47	1525,05	1348,76	100,836	3449,05	693,2	229,679	780,701	889,747	1184,83	1359,39	857,95	4878,9	7099,38	755,008	2000,6	801,977	337
42	PAR10	PD19	497,869	543,41	212,42		1791,53	1501	1206,61	992,453	150,997	3138,69	1512,74	833,604	1050,75	820,687	1241,4	972,92	1084,67	2970,13	5139,88	1524,65	743,512	2455,62	217
43	CON01	Kont 1	954,345	904,192	1235,28	467,678	1146,4	1345,59	1532,48	918,493	1146,79	1504,67	1244,49	146,879	3793,44	1663,96	2836,7	1268,16	1673,61	5574,02	7721,14	288,904	478,875	1340,2	2981
44	CON02	Kont 2	365,966	375,368	129,234	848,54	1629,2	1118,11	1897,08	1631,41	1188,38	1802,97	1210,16	769,393	1201,3	1470,11	1331,49	1694,76	1207,33	5704,57	9899,55	129,234	129,234	710,787	2936
45	CON03	Kont 3	1499,03	584,482	1403,51	163,186	5637,9	834,928	1902,39	1197,47	1580,08	2790,13	1811,54	1303,28	1929,84	2491,24	3168,35	2256,77	2487,16	4094,29	11078	1375,46	730,595	841,46	4795
46	CON04	Kont 4	851,658	1124,41	1271,74	249,857	10280,4	650,011	2271	1271,97	2271	1808,91	1003,98	1351,79	1604,63	673,947	3767,1	4569,3	4672,24	3263,18	10174,2	1454,53	2271	2271	2

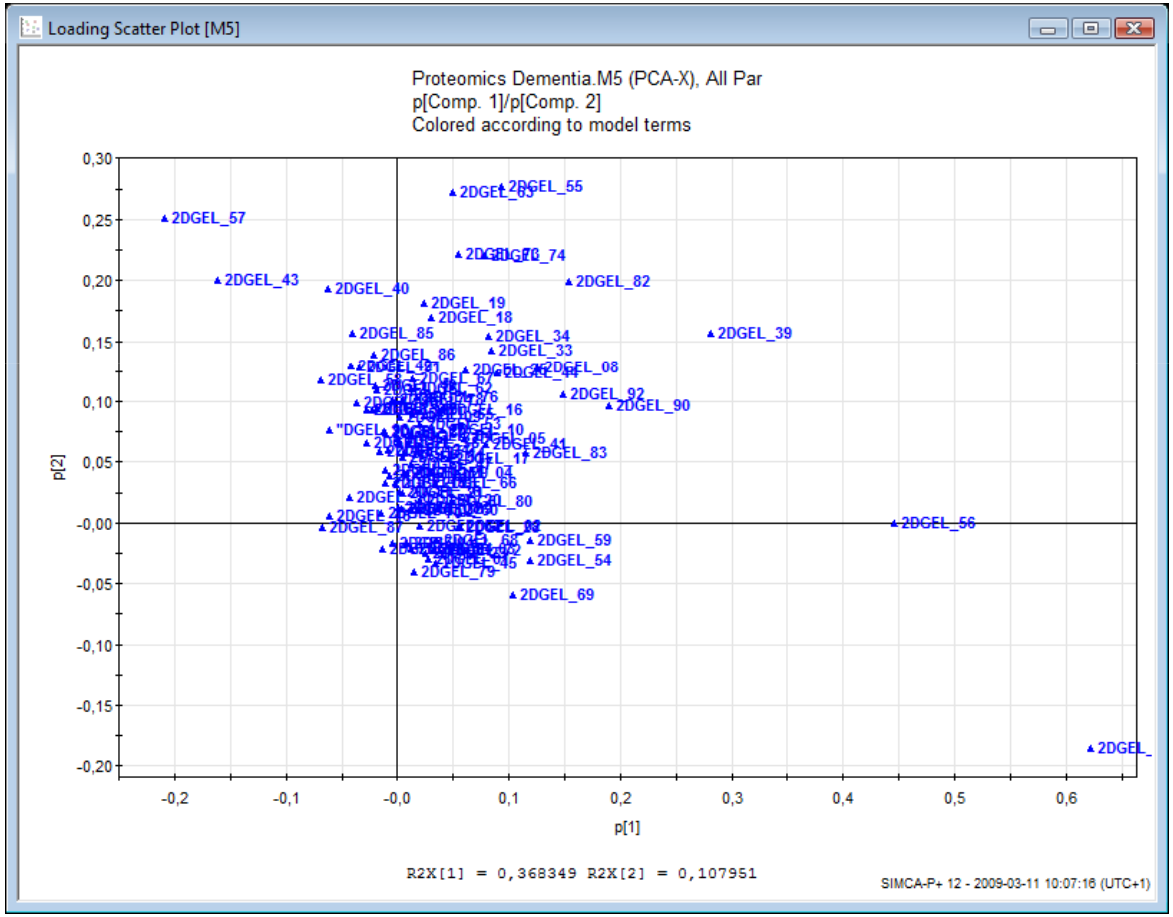
Clinical Proteomics Data

- PCA to overview the 52 by 95 data table
- Score plot: Visualize observations
- Similar observations are close to each other
- No deviating samples
- Clear separation between groups



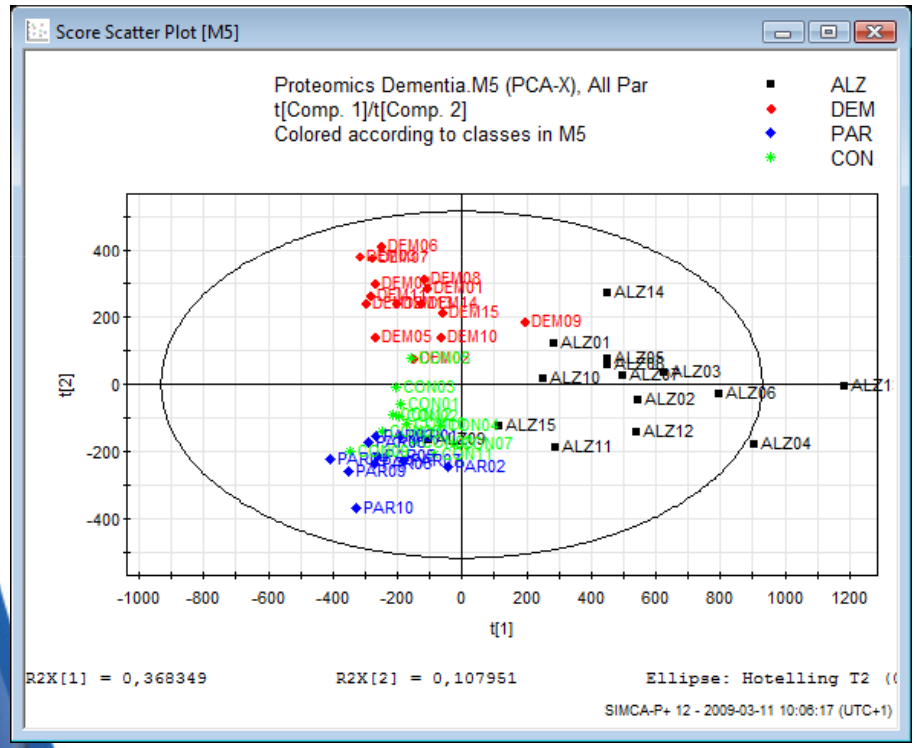
Parameter overview

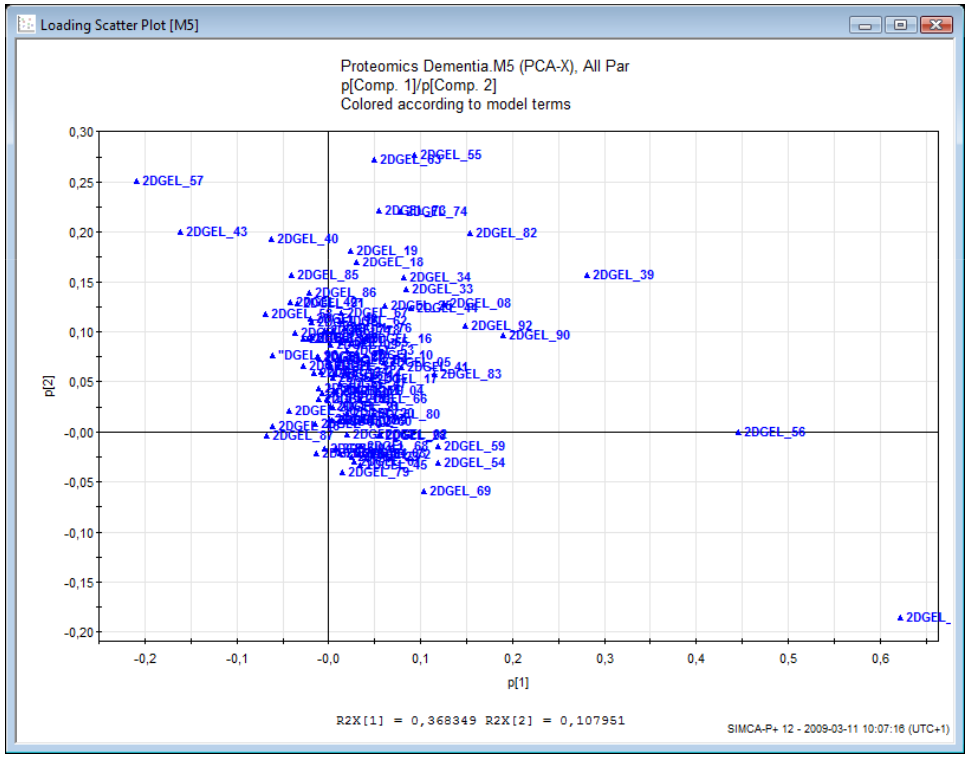
- Loading plot reveals relationship between parameters
- Parameters close to each other *correlates*



What is different between groups?

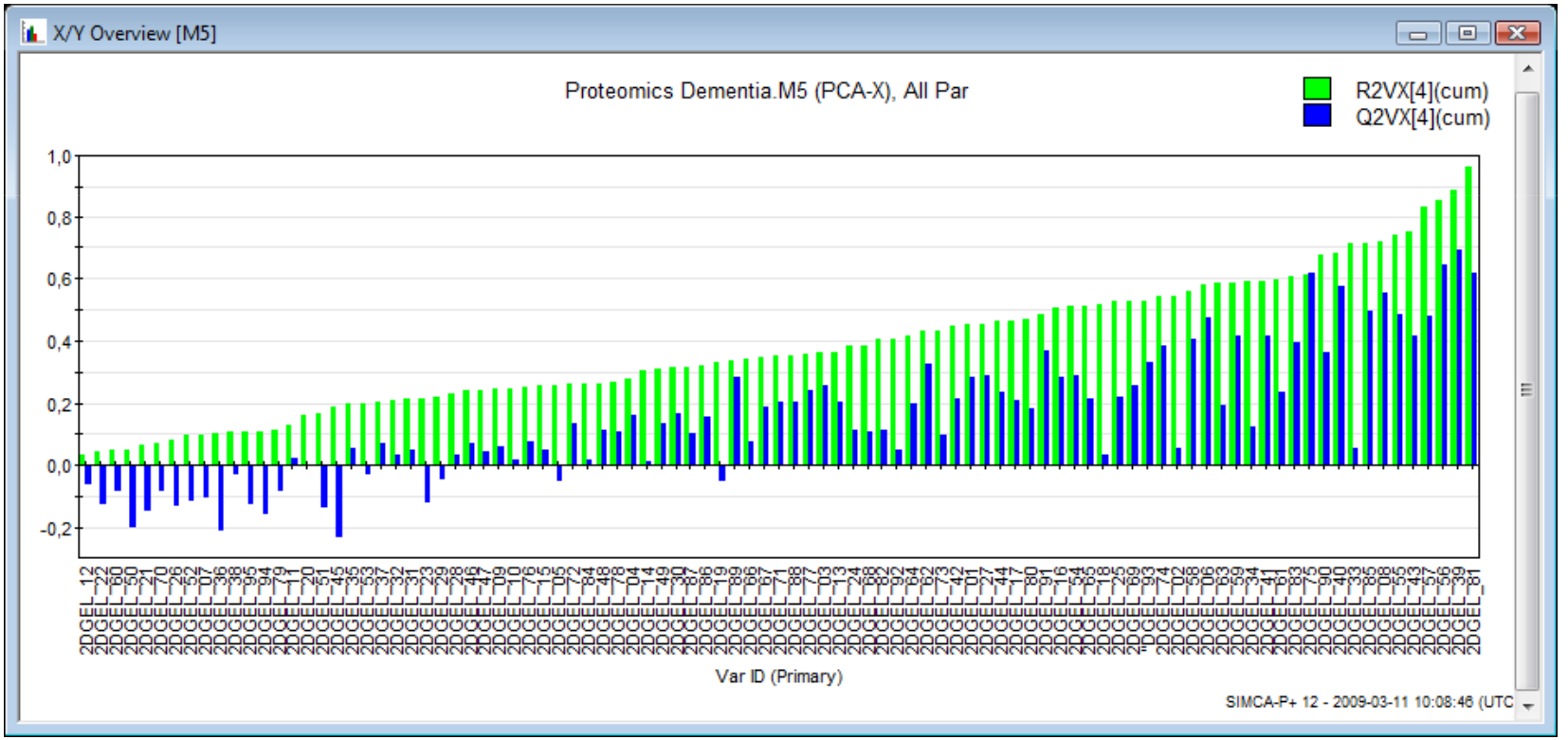
The parameters related to patterns in Score plots are found in Loading plots





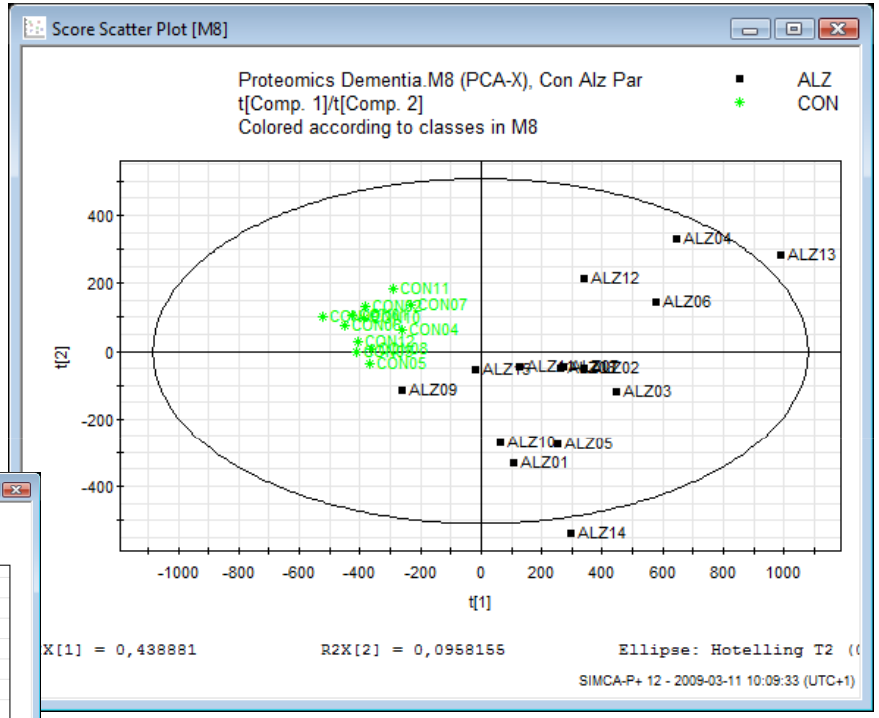
Which parameters are well described by the model?

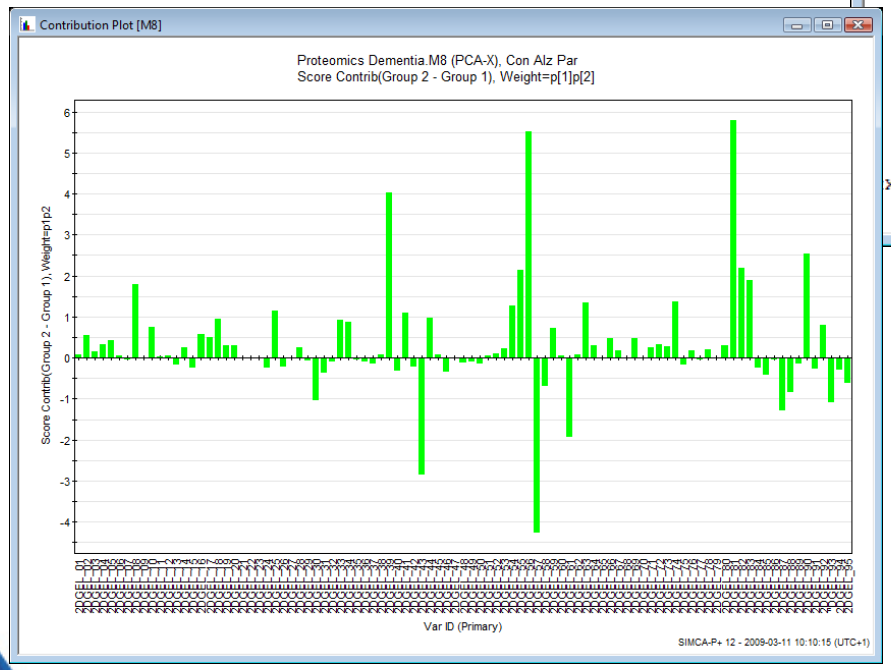
- Right hand: Parameters with *structured variation*
  - Reliable to describe differences in observations
- Left hand: *Noisy* parameters with unstructured/ limited variation



Comparison of groups: Contribution plot

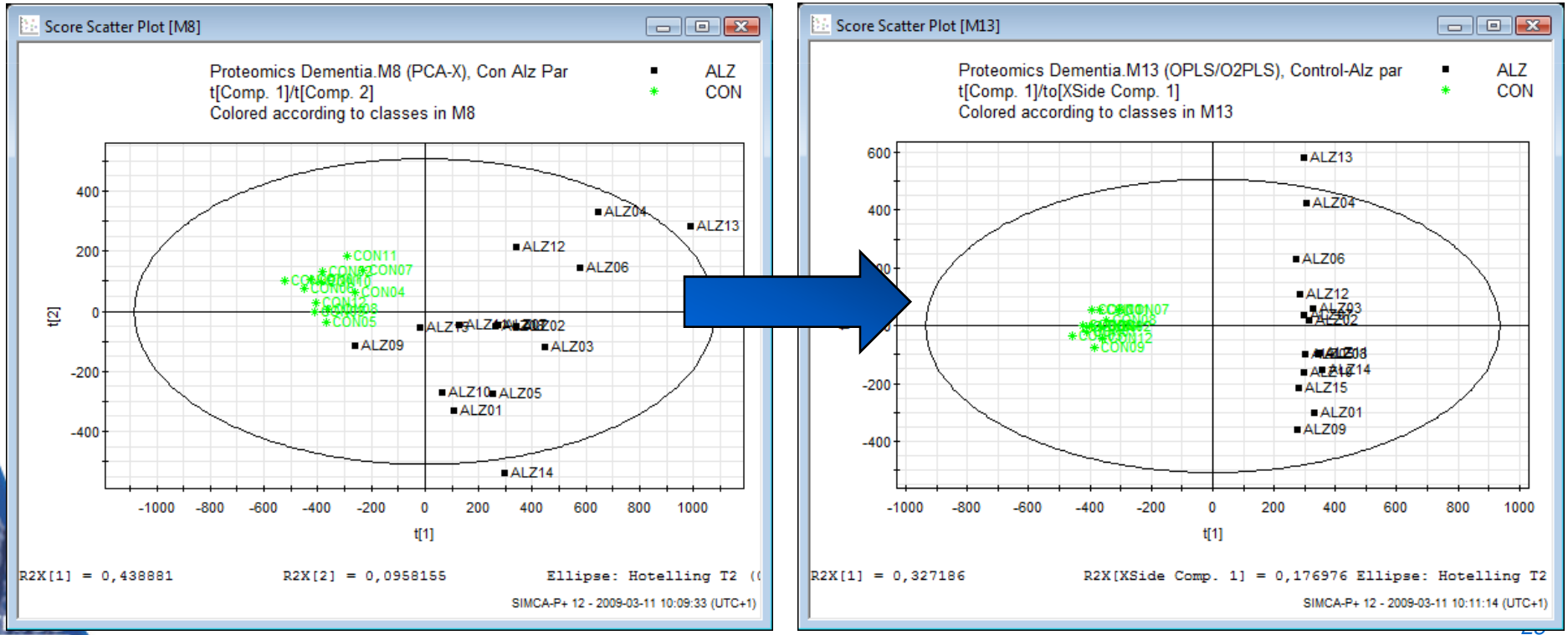
- PCA model for comparison of Control “healthy” patients and Alzheimer patients
- Compare two groups using Contribution plots
  - Systematic differences between two selected sets of observations?





Proteomics extension

- Next step: Compare different dementia groups with control to identify discriminating proteins
  - Focus on changes in protein expression only related to diagnose
- OPLS-DA: A more focused solution



PCA application: Raw material characterization

Particle size distribution of a raw material

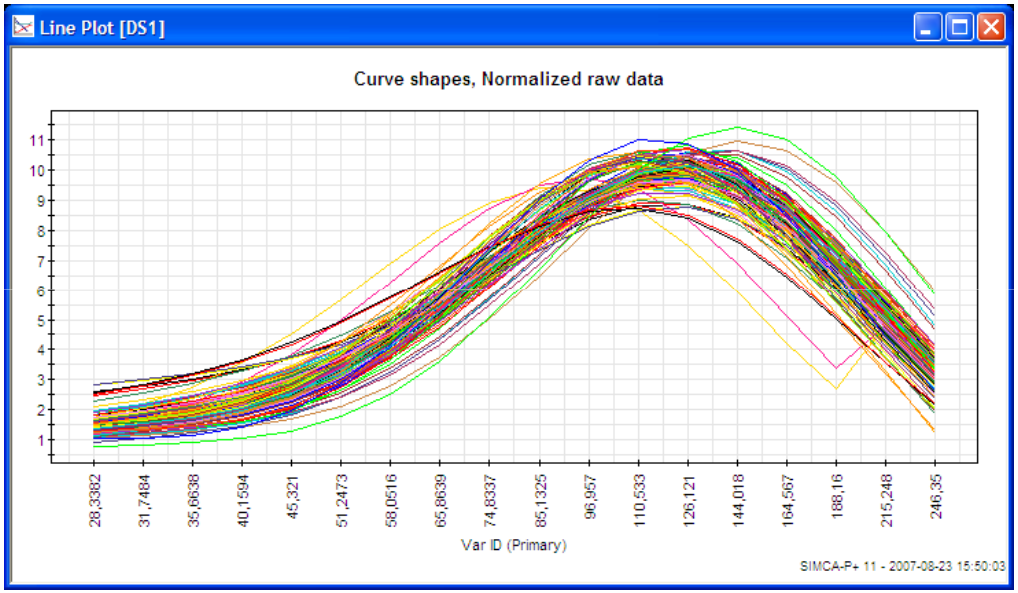
- Raw material from two suppliers have been used for a long time-period
- Customer perform PSD measurements for each delivered batch before approval
- Four new suppliers were evaluated
- 2-4 samples provided by each new supplier

- Data from actual investigation at a mid-sized pharma company
- Historical data from 117 samples

UMETRICS
AN MKS COMPANY

Overview and summary of particle size distribution data

- Historical PSD-data of raw material, 117 samples, 18 variables
- The range of the instrument was 28-246 um
- Similar D50 does not mean similar PSD

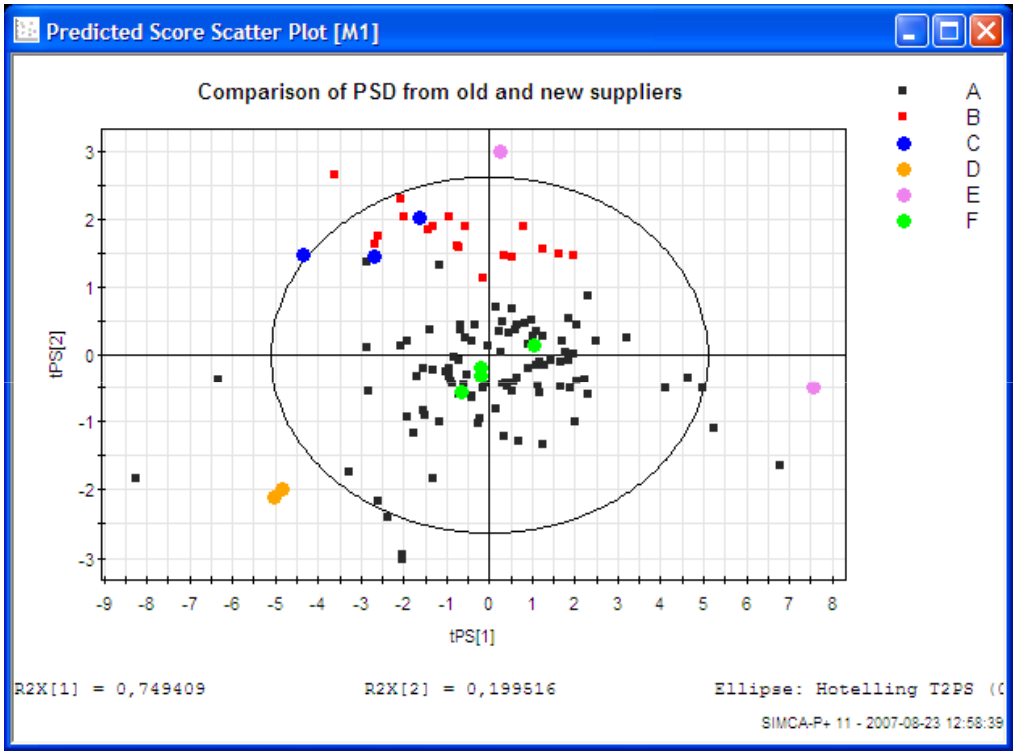


- Model built on known “good” batches

UMETRICS
AN MKS COMPANY

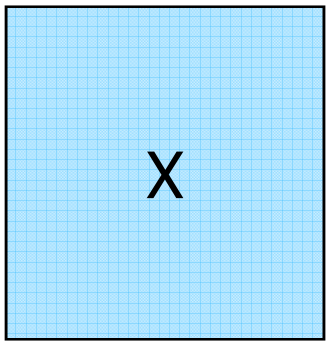
Use models for prediction: How does material from new suppliers fit in?

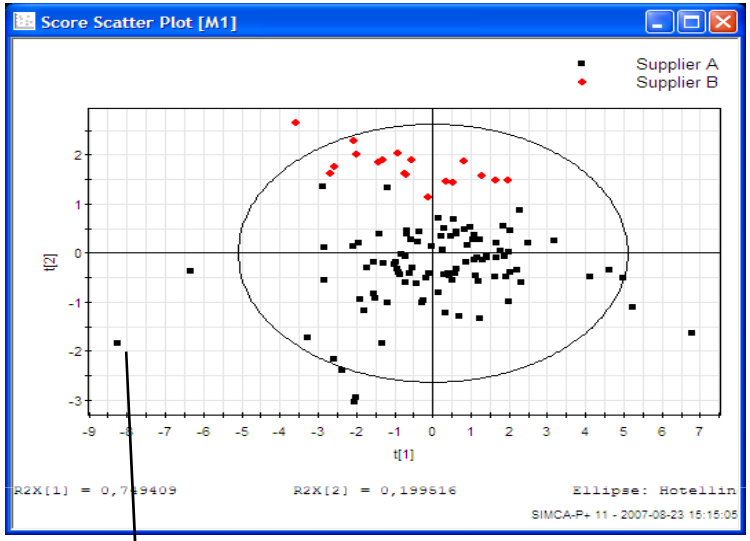
- PSD profiles from material provided by new suppliers were predicted by the model
  - How do they compare to “old” material?
- Supplier C similar to B
- Supplier D slightly different from the rest but low variation
- Supplier E all over the place!
- Supplier F similar to A
- Possibility for at-line automatic reporting

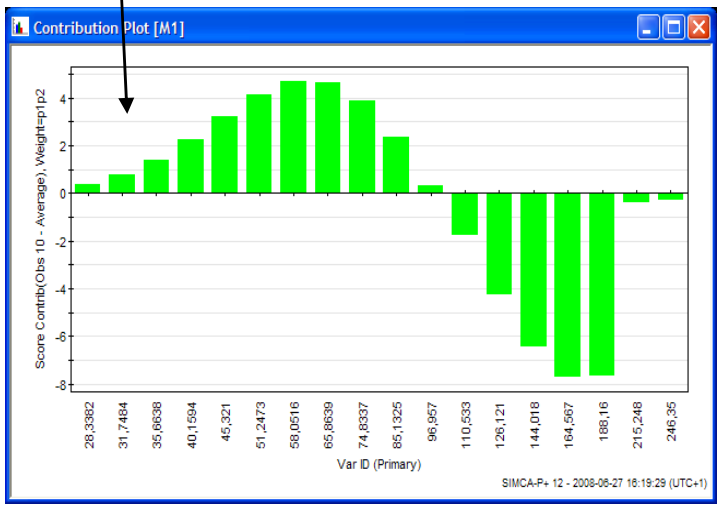


Summary PCA

- PCA provides an unsupervised overview of large data sets
  - Identification of clusters, trends etc
- Why is a sample/ group of sample different?
  - Contribution plots
- Used when no result characteristic is available

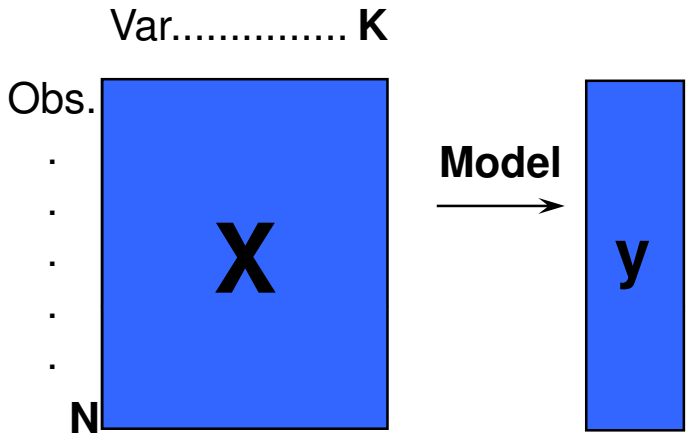




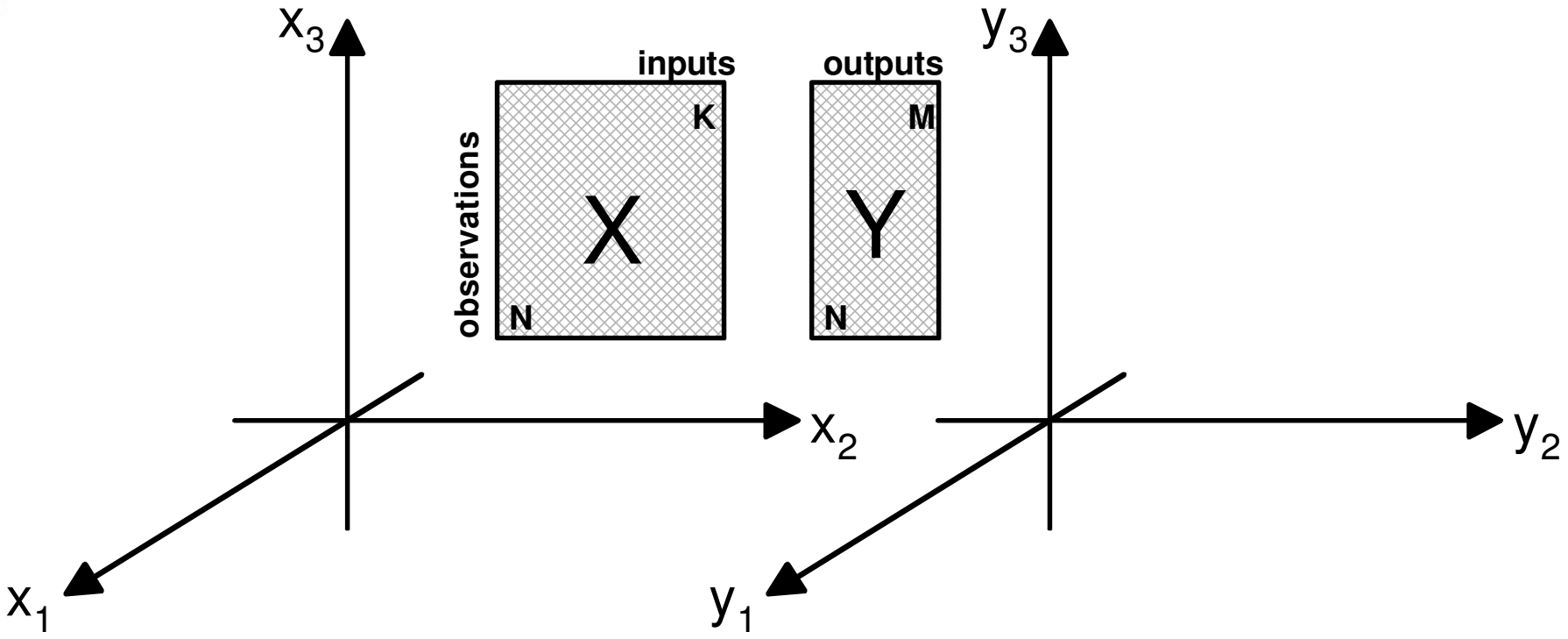


Introduction PLS

- Relationship between **X** and **y**
  - Find a *common/ shared structure* in **X** and **y**.
  - Which variables are related/ not related to response?
  - Classes, groups of observations
  - Deviating observations, time-trends

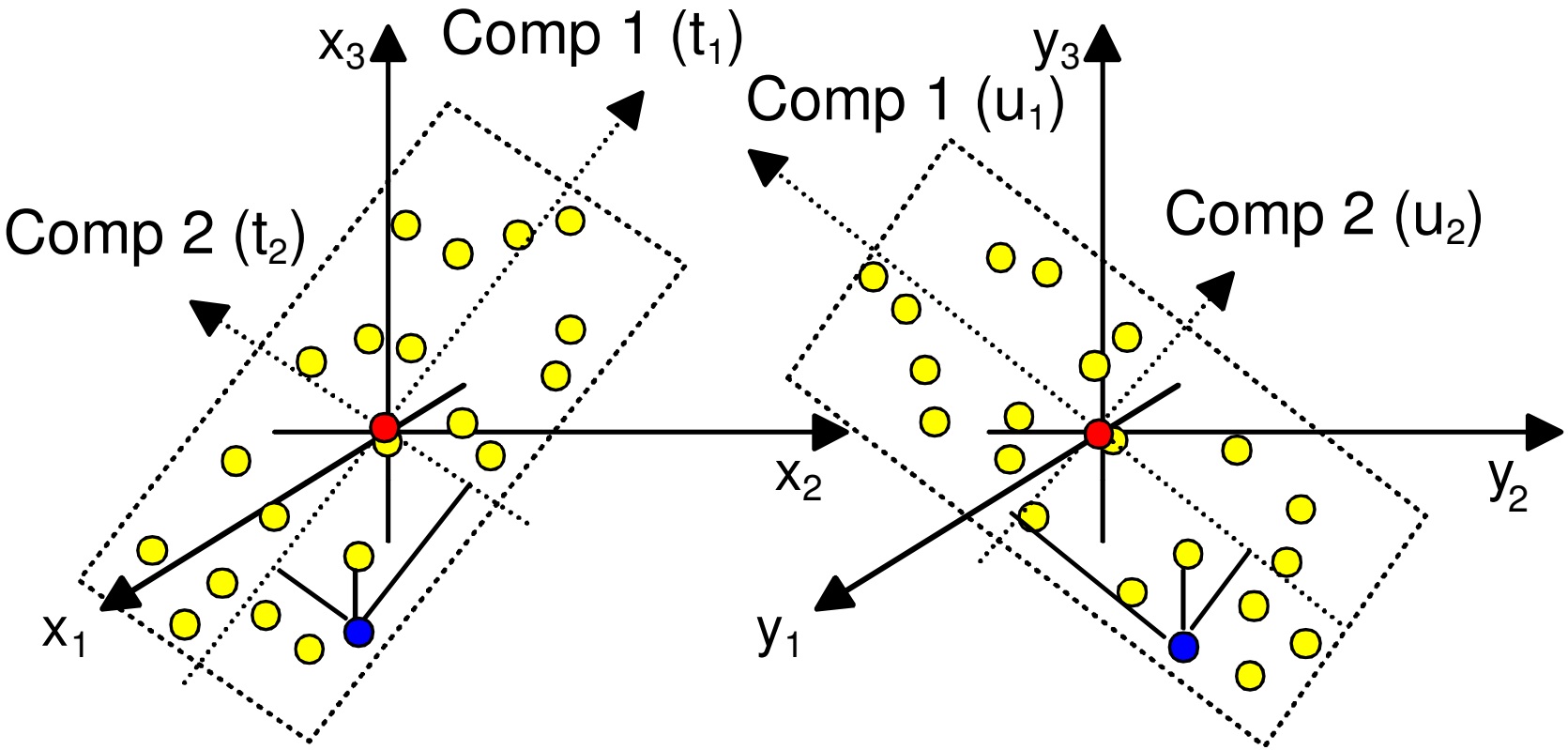


PLS -- Geometric Interpretation, Two spaces



- 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, Two planes



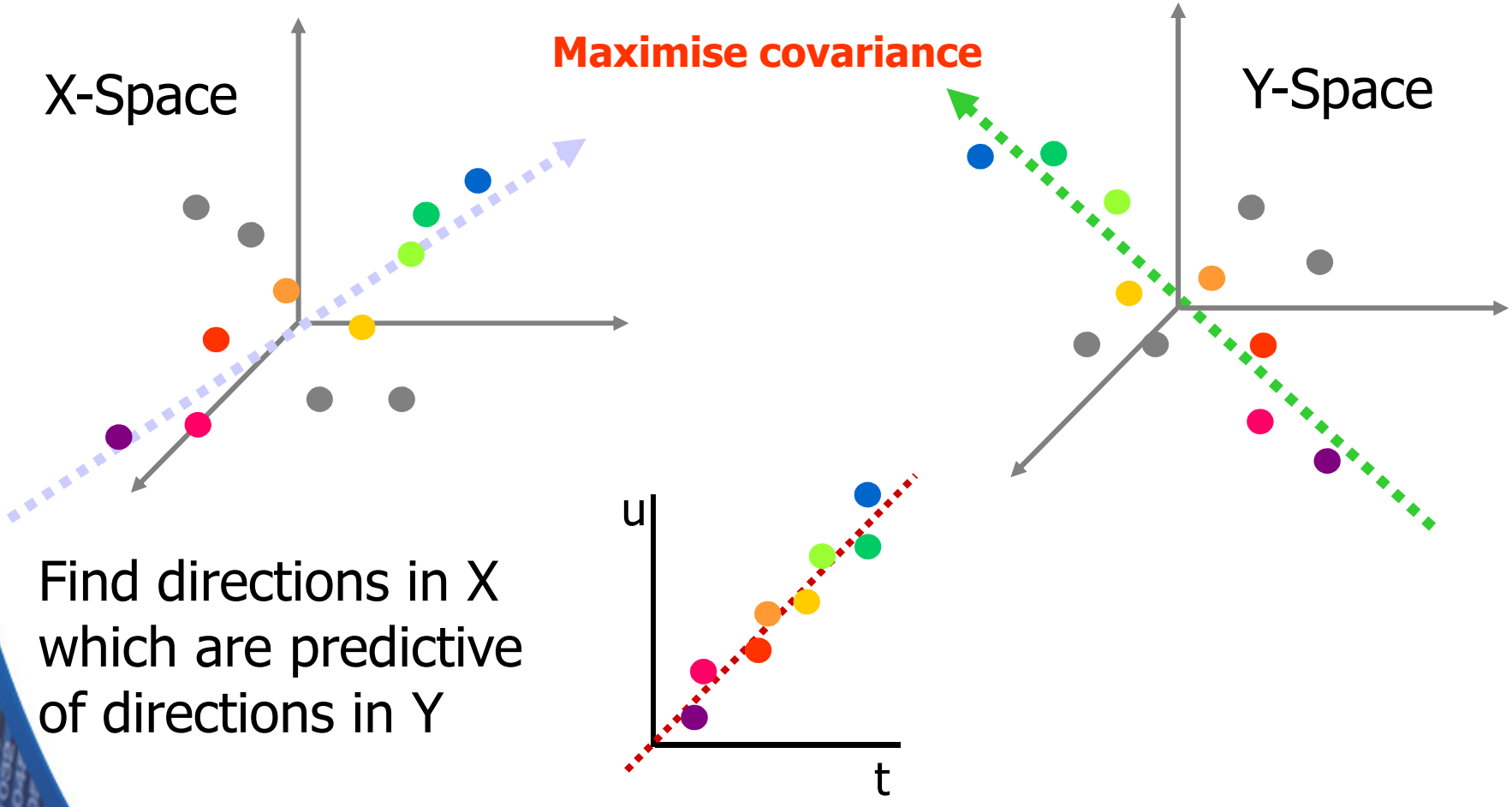
- The PLS components form planes in the X- and Y-spaces
- The variability around the X-plane is used to calculate a **tolerance interval** within which new observations similar to the training set will be located. This is of interest in classification and prediction.

Summary: PLS

Describe structure of X & Y

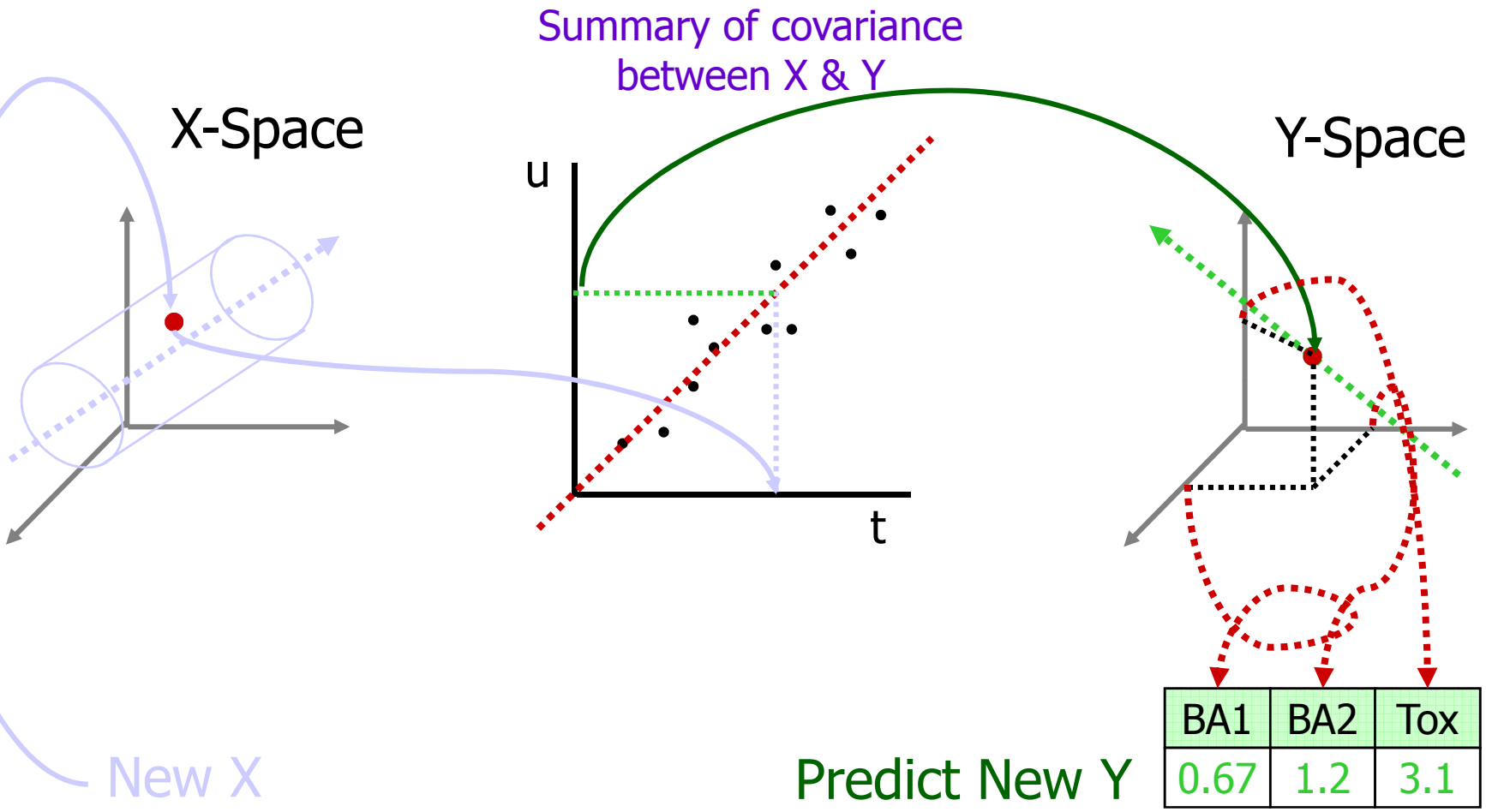
AND

Maximise covariance



Find directions in X which are predictive of directions in Y

Prediction using PLS model



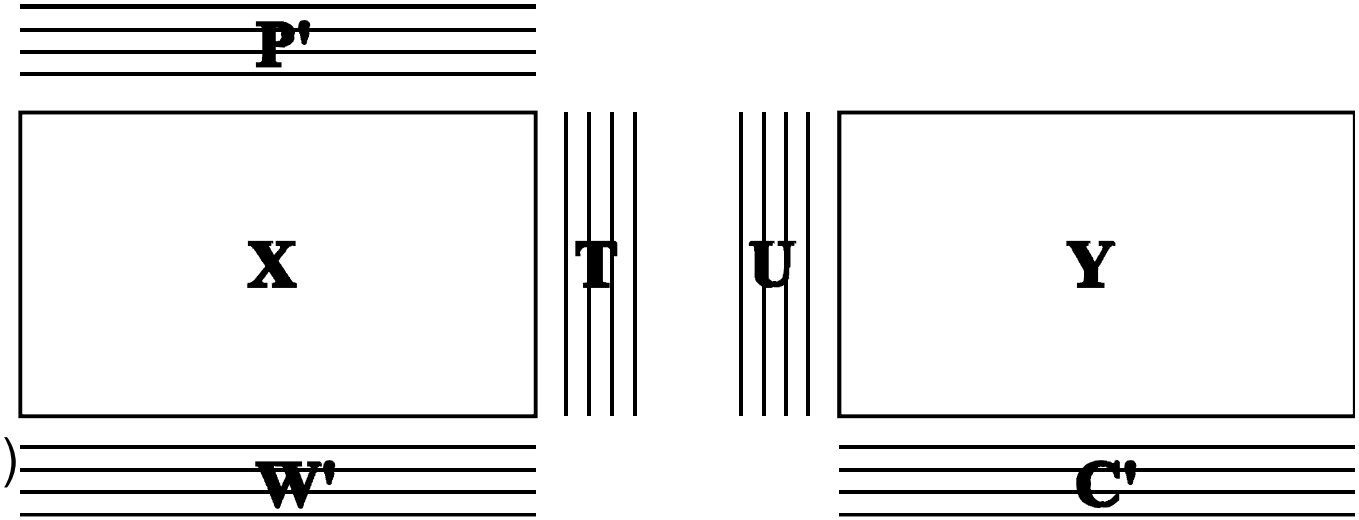
Predict multiple Y's from many X's ***at the same time***

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 with Y

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

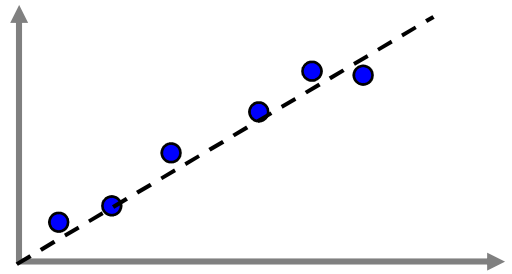
The forms of Y data

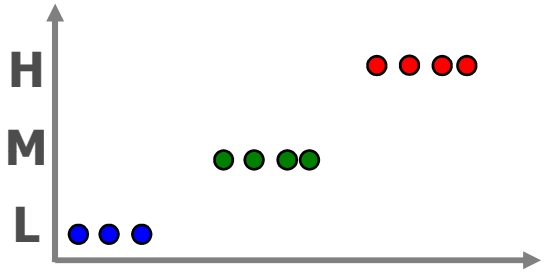
1. Continuous data
   - One–off measurements
   - Averages (May lose information)
   - Curve Fits (IC50)

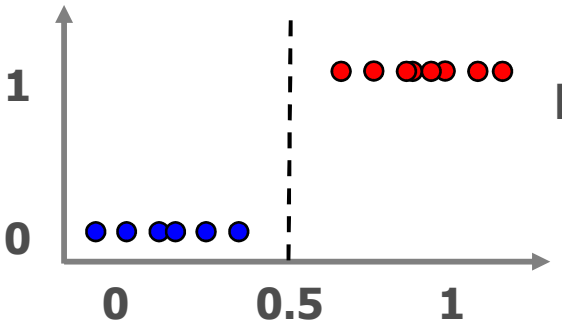
2. Multi-level Qualitative
   - Low / Med / High

3. Qualitative
   - Control / Treated
   - Healthy / Diseased
   - Good / Bad

4. Descriptive
   - Pathology data
   - c.f. Sensory data



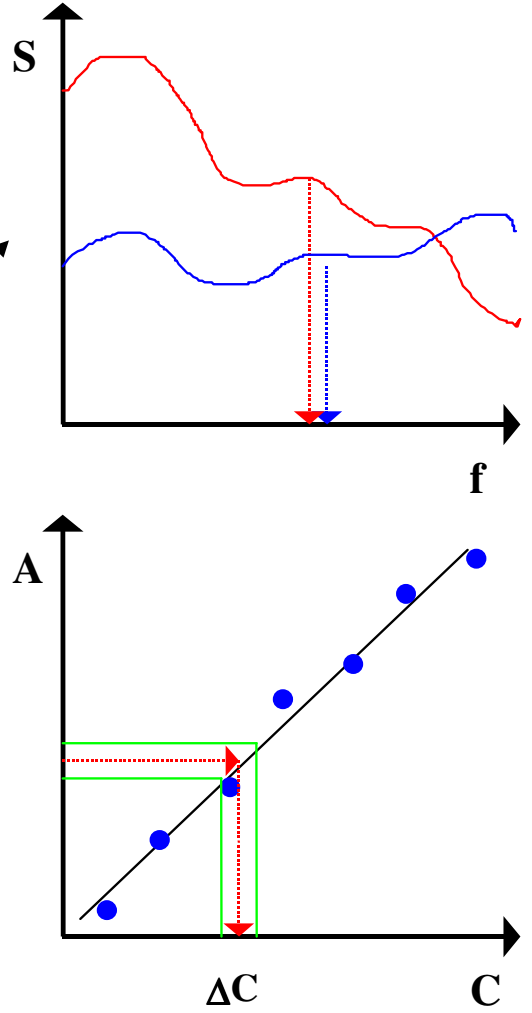




Discriminant Analysis (PLS-DA)

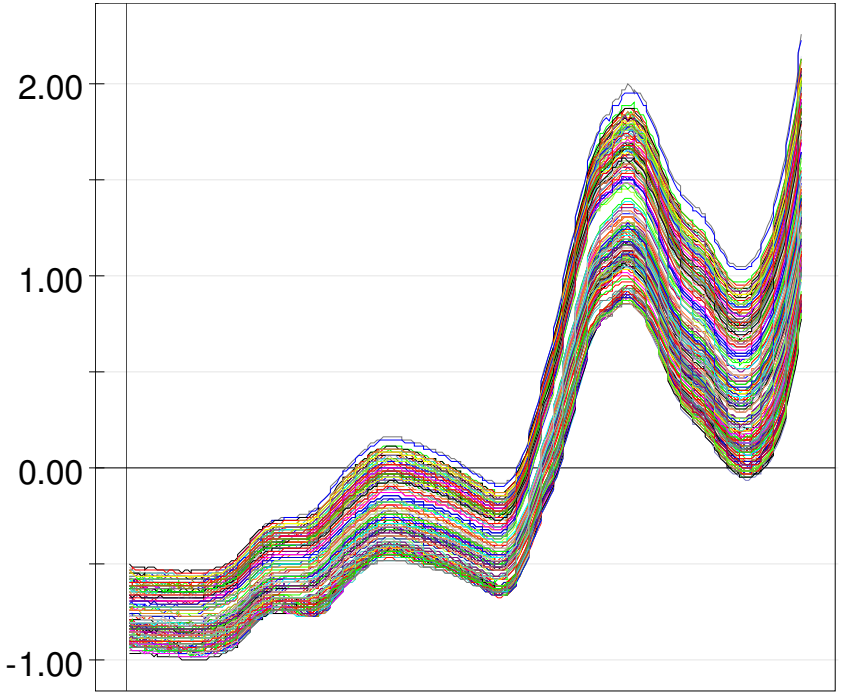
PLS application- Multivariate calibration

- Classical calibration based on single peaks/ frequencies
- Multivariate calibration based on entire spectra
  - Higher selectivity and precision
- **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



PLS Application: Modeling and prediction of Metroprolol content in tablets

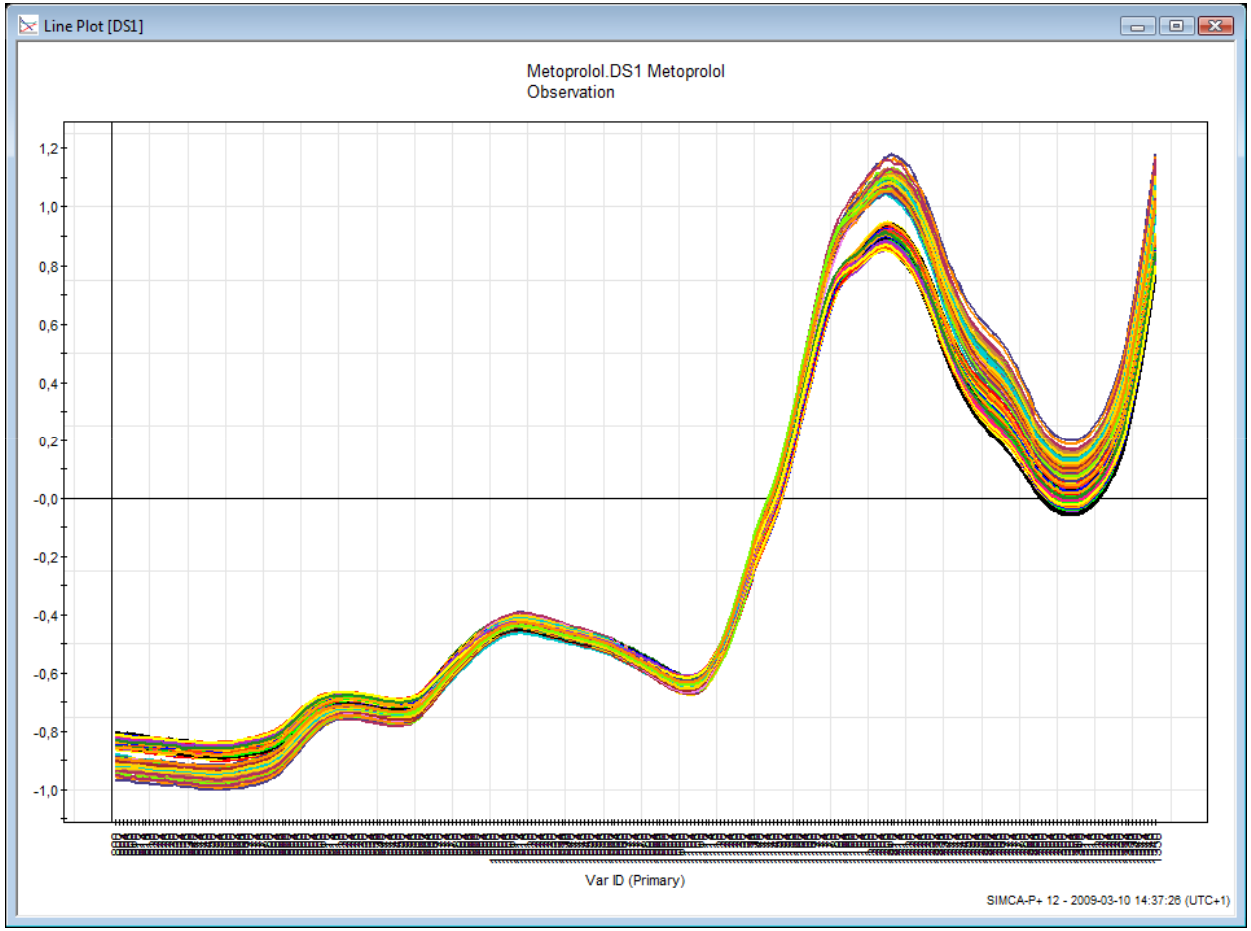
- Very early PAT application
  - In the times of PAC
- Study performed at Astra Hässle (later AstraZeneca), Mölndal, Sweden
- Aim: By NIR predict content of Motroprolol in tablets
- Benefit: Quick and non-destructive analysis technique



UMETRICS
AN MKS COMPANY

Design in constituents and process conditions

- Metoprolol example; NIR 800-1400 nm
- Design consist of 40 samples
  - A and B
- Each sample measured 5 times=> 200 samples
- 387 variables



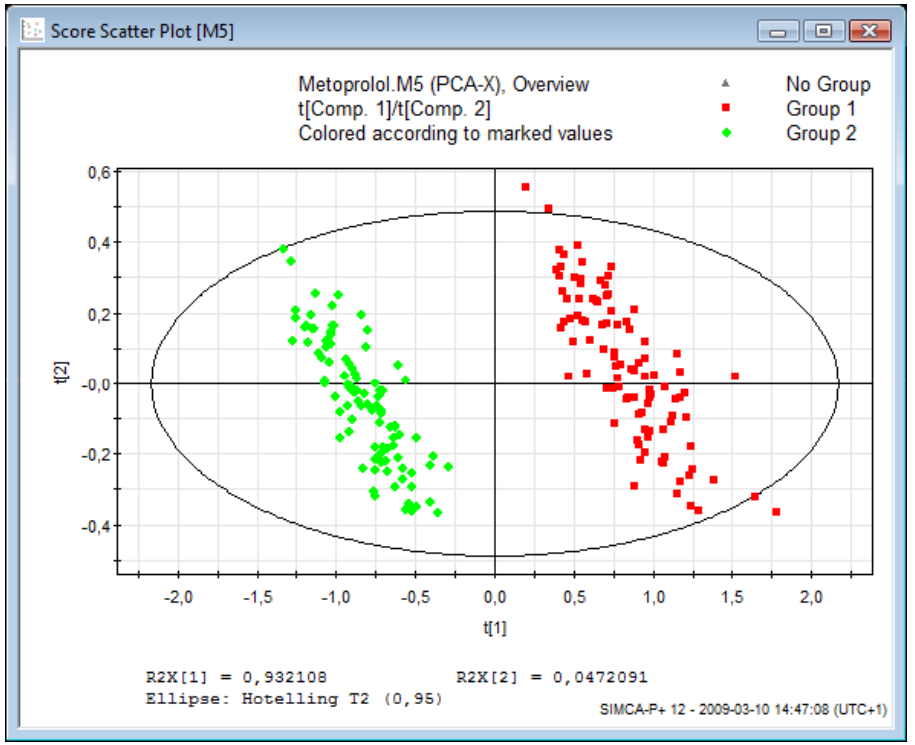
Ref.J. Gottfries, M Josefson, et al., J Pharm Biomed Anal **14** (1996) 1495

Design in constituents and process conditions

Design =

- 5 levels of metoprolol
- 2 tablet heights
- **2 lots of metroprolol-pellets**
- 2 lots of Microcryst. Cell.
- Tablets 3 % water, or “wet”
- 5 tablets per made batch

- PCA for data overview
- Score plot reveals two groups
  - Related to lot of Metroprolol pellets
  - NOT related to amount of Metroprolol in tablet

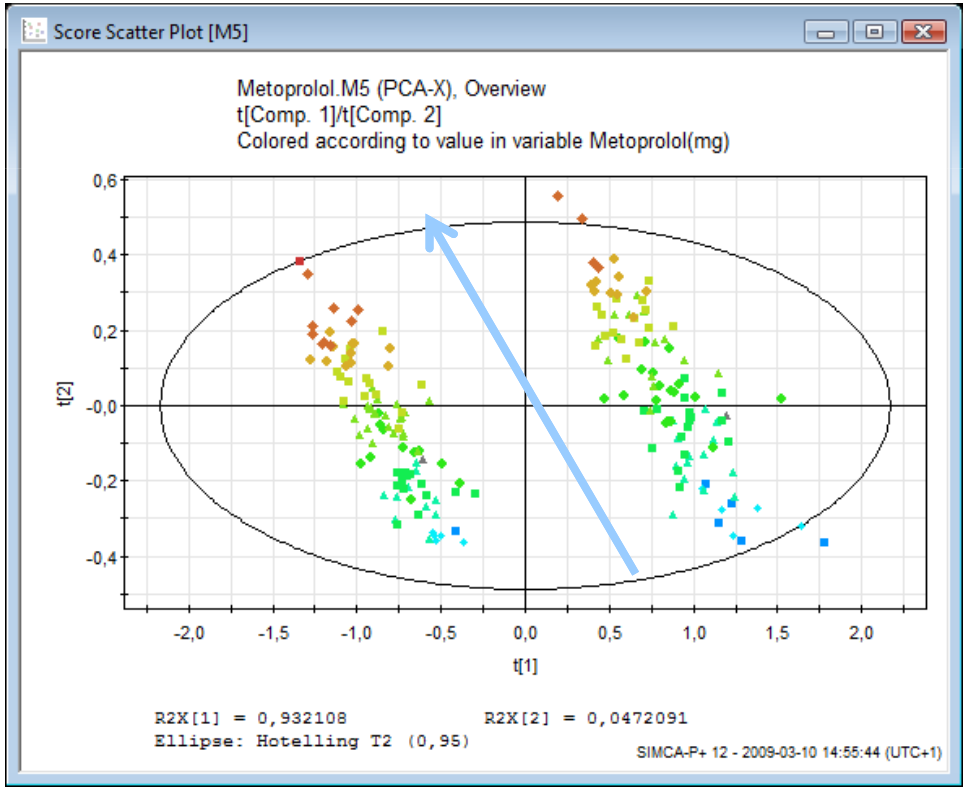


UMETRICS
AN MKS COMPANY

43

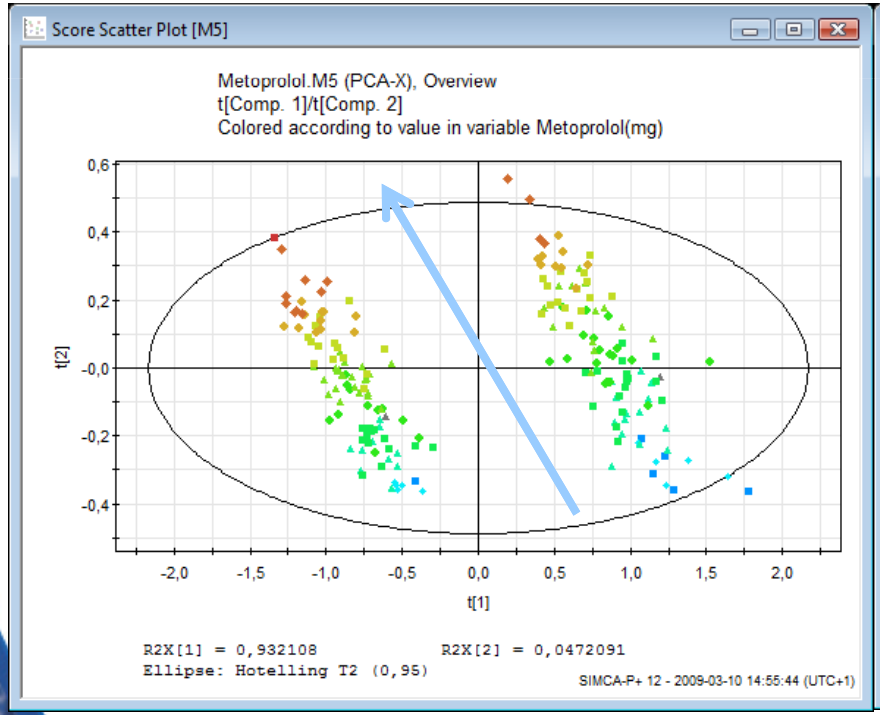
PCA for data overview

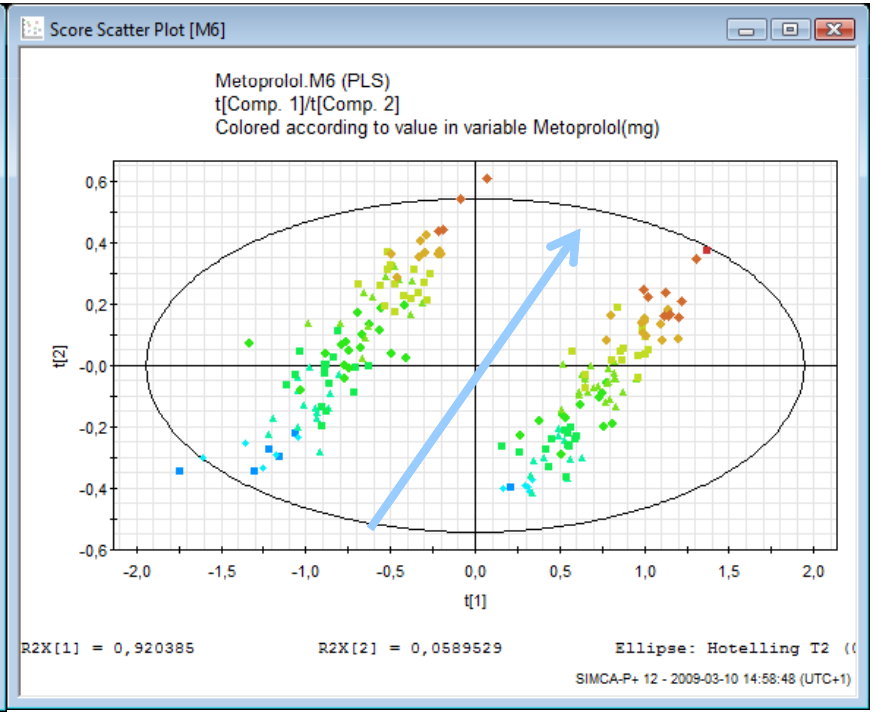
- First component related to lot of Metroprolol
- Second component related to Metroprolol content



PLS modeling of Metroprolol content

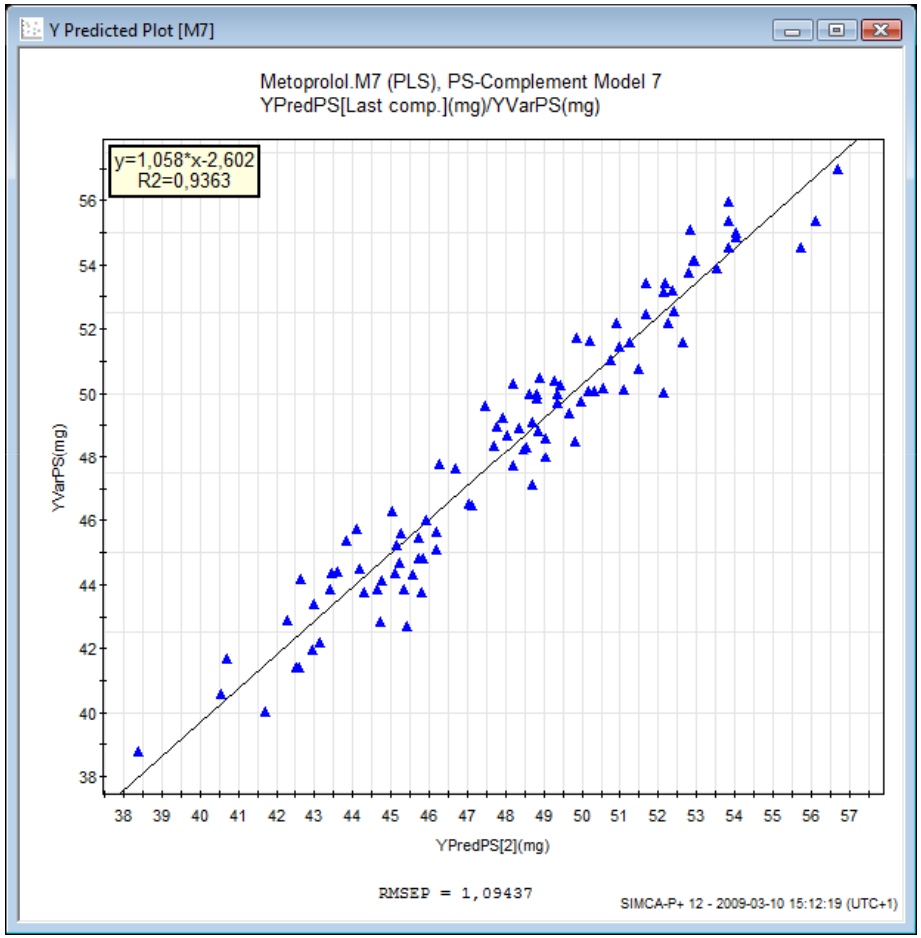
- Rotation of score plot compared to PCA
  - First component still reflects Metroprolol lot
  - Second component describes Metroprolol concentration
- Inner relation!





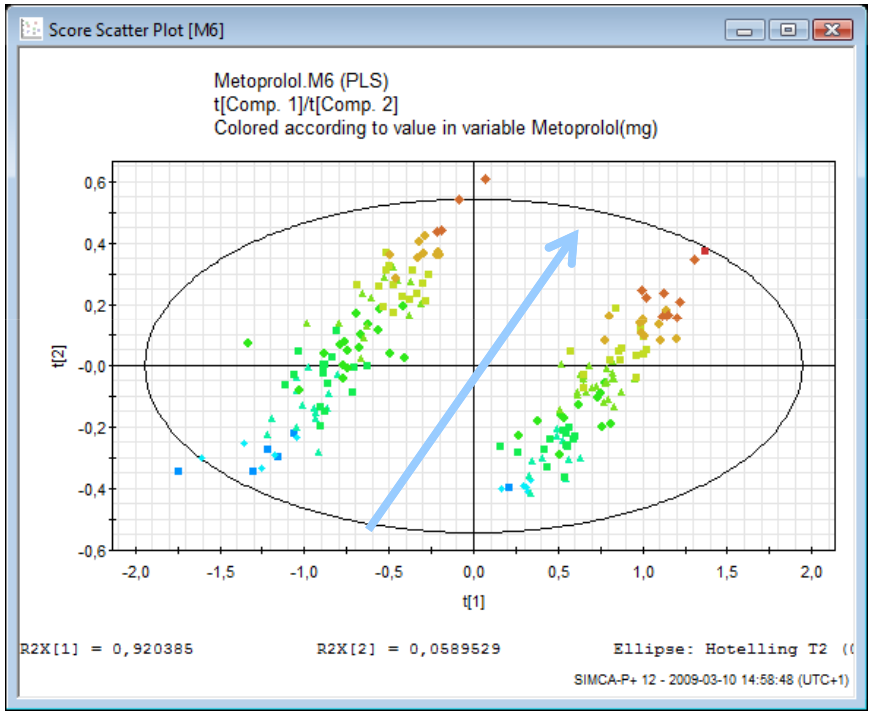
PLS model for prediction

- PLS model was built on half the data set
  - 5 spectra was taken for each tablet
- Remaining spectra were used as prediction set
  - Used to validate the model, how well can it be used for predictions?
- Precision of prediction: 1,09 mg



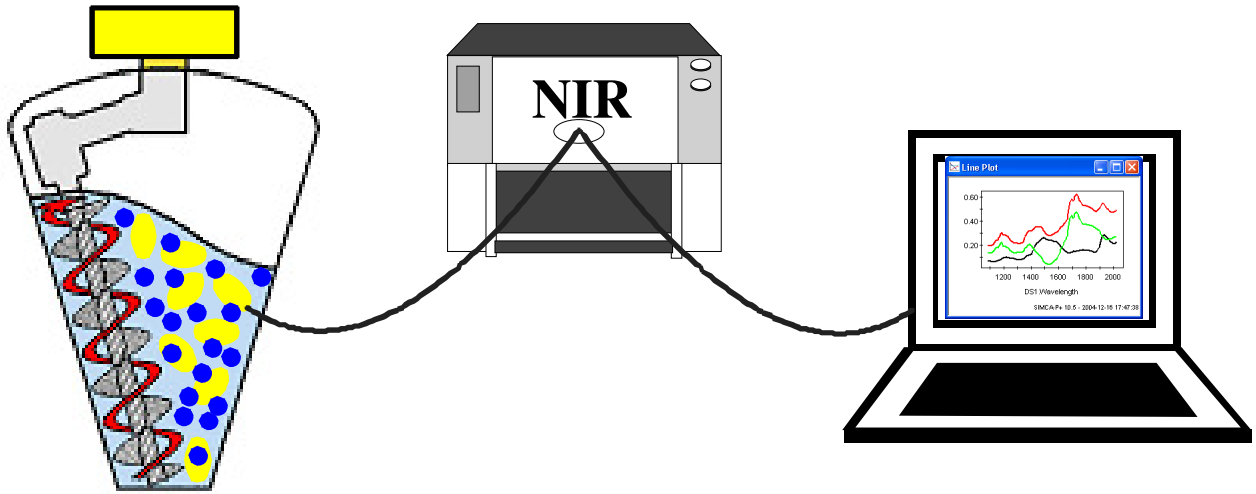
Model for *prediction* or *interpretation*

- The model was found to predict Metroprolol content well
- Precision error on 1.09 mg also includes error in reference method
- How should the model be interpreted?
- Direction of metroprolol content not related to one single component
  - OPLS should be applied for interpretation



In-line example: Monitoring of powder blending

- Objective: To develop a end-point detection model for a blending step
- Vertical cone mixer was used
- NIR spectra were recorded

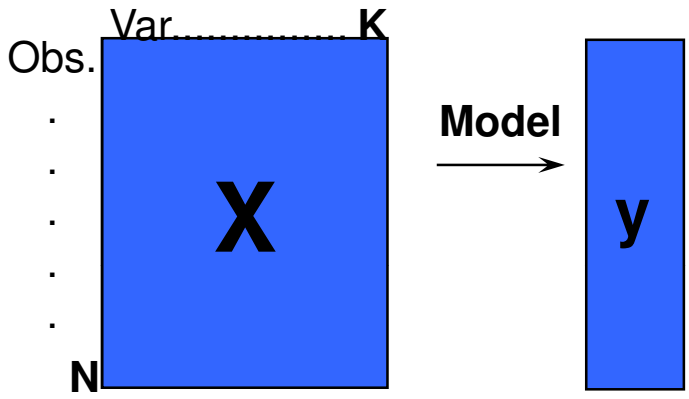


Application of PLS

- Multivariate calibration
- Multivariate Process monitoring
  - Continuous and Batch processes
  - Predict output/ quality
  - Control loop, predict how to run the process
- Sensory data analysis
  - How can a product be improved?
  - Properties of a product appreciated by a specific customer group?
- Stability modeling
  - Modeling of stability test data, improve understanding and predict properties of formulations

Introduction PLS-Discriminant analysis

- Discriminate between classes
  - Identify dissimilarities between known classes in a data set **X**.
  - In PLS-DA **y** represents class belonging.
  - Which variables are related/ not related to class?
  - Miss-classified objects?



UMETRICS
AN MKS COMPANY

Introduction PLS-Discriminant analysis

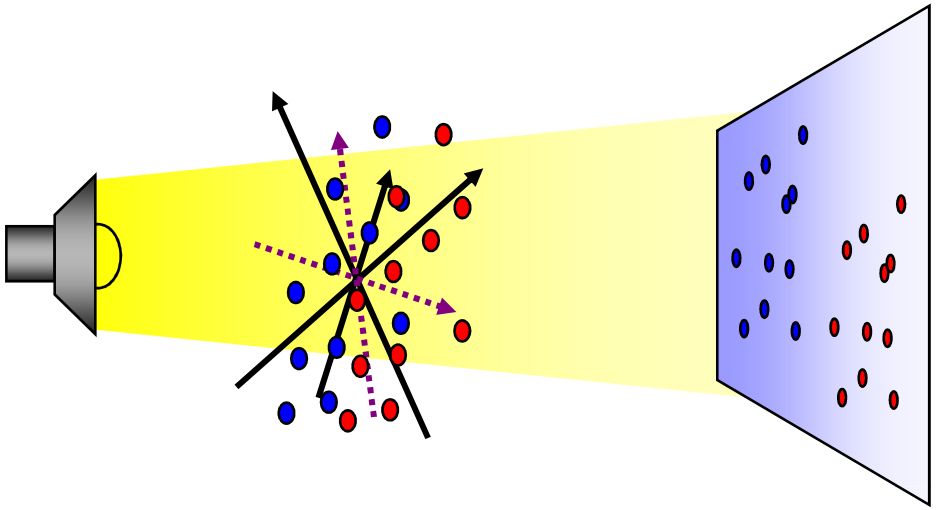
- With PLS the Y-variables are normally continuous
- PLS-DA uses Y=1 or Y=0 to designate class belonging
- Predictions then give value between 0 and 1 depending on membership
- In SIMCA-P, dummy Y-variables are assigned when you define a class ($DA1 or $DA2)

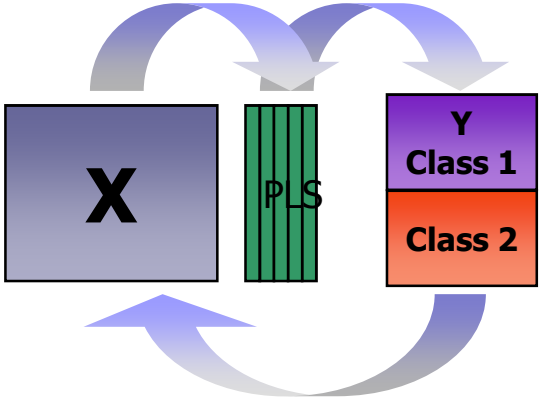
		X data	Y1= Control	Y2= Treated
Obs1	Control	.....	1	0
Obs2	Control	.....	1	0
Obs3	Treated	.....	0	1
Obs4	Treated	.....	0	1

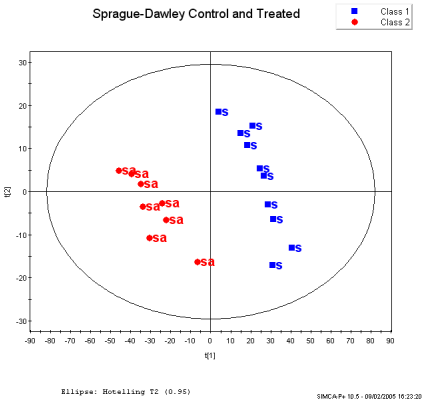
Automatically generated

Separating Groups PLS-DA

- PLS-DA relies on a projection of X as does PCA
- BUT is a Maximum Separation Projection
  - Guided by known class information
  - Easiest to interpret with 2 classes
  - Extendable to more classes
- Advantage:
  - Shows which variables are responsible for class discrimination
- Today: OPLS-DA preferred
- Applications
  - Biomarkers in metabonomics
  - Proteomics and Genomics

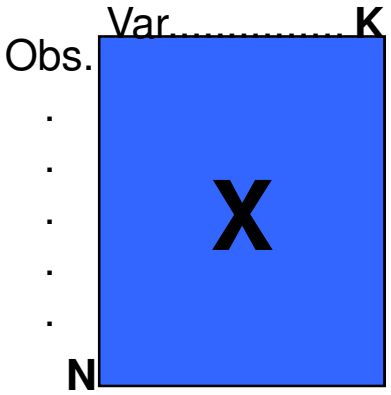




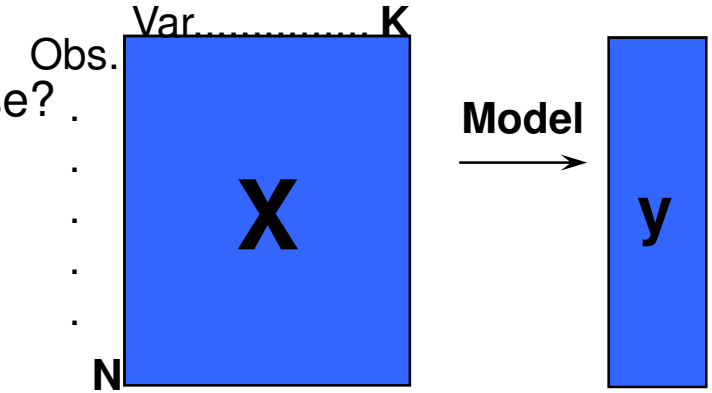


Summary Projection techniques

- Overview of a data set **X** (PCA)
  - Classes, groups among observations?
  - Deviating observations, time-trends?
  - Correlation patterns between variables?
  - Find variables containing unique information
  - ….



- Relationship between **X** and **y** (PLS, PLS-DA)
  - Find a *common/ shared structure* in **X** and **y**.
  - Which variables are related/ not related to response?
  - Classes, groups of observations
  - Deviating observations, time-trends



UMETRICS
AN MKS COMPANY

When DOE when MVA?

DOE

- Designed data
- Uncorrelated variables
- Limited number of variables
- Few experiments/samples
- Cause-and-effect

MVA

- Collected or designed data
- Correlated variables
- Unlimited number of variables
- Many experiments/samples
- Correlation

UMETRICS
AN MKS COMPANY

Application areas MVA

- Research and development
  - Molecular modeling
  - -omics
  - Process development
- Off-line analysis process
  - Fault detection
  - Indications of drift, need for service
  - Long- and short-time trends
  - Historical data to learn about process
- Process supervision- on/ in/at-line
  - Real-time detection of process “mood”
  - Detect deviating behavior at an early stage- Act, not React
  - Example: Injection molding
    - At-line supervision system of machine
    - Cooperation with MKS Instruments
- Or in short: Wherever there is data!

UMETRICS
AN MKS COMPANY

Problem formulation

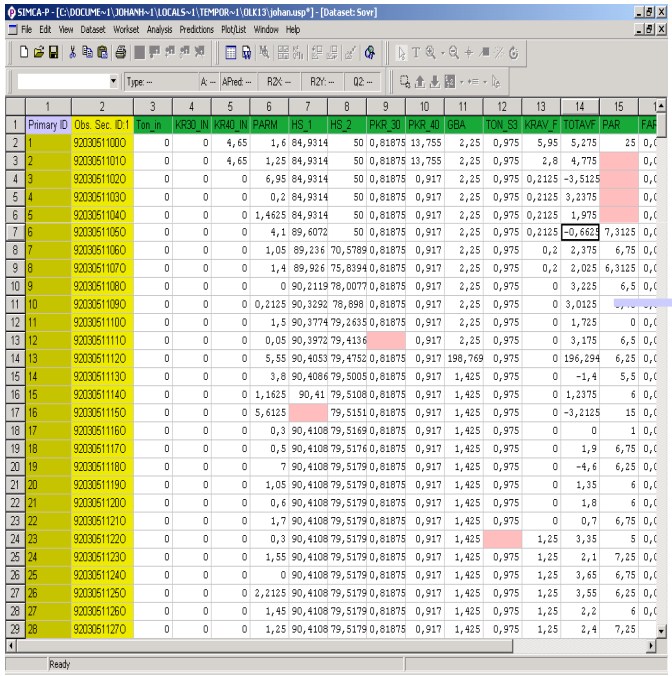
- Identify the question
  - Define goal
    - Preferably quantitative
  - How much do we know?
- Relevant data available?
  - Selection and representation of data
- Expected findings?
- Benefits?
- Reporting and visualization
  - How do I best show my results?

UMETRICS
AN MKS COMPANY

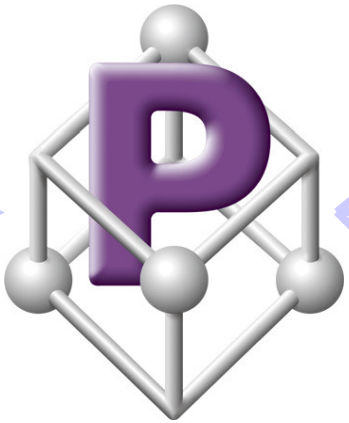
Multivariate Data Analysis in SIMCA-P+

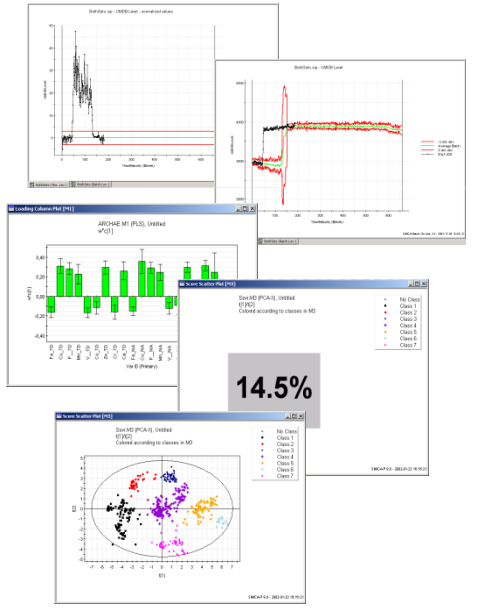
Visual!

Data **Multivariate Modeling** **Information**



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Primary ID	Obs. Sec. ID 1	Ton_in	KR30_IN	KR40_IN	PARM	HS_1	HS_2	PKR_30	PKR_40	GBA	TON_S3	KRAV_F	TOTAVF	PAR
2	1	92030511000	0	0	4,65	1,6	84,9314	50	0,81875	13,755	2,25	0,975	5,95	5,275	25
3	2	92030511010	0	0	4,65	1,25	84,9314	50	0,81875	13,755	2,25	0,975	2,8	4,775	
4	3	92030511020	0	0	0	6,95	84,9314	50	0,81875	0,917	2,25	0,975	0,2125	-3,5125	
5	4	92030511030	0	0	0	0,2	84,9314	50	0,81875	0,917	2,25	0,975	0,2125	3,2375	
6	5	92030511040	0	0	0	1,4625	84,9314	50	0,81875	0,917	2,25	0,975	0,2125	1,975	
7	6	92030511050	0	0	0	4,1	89,6072	50	0,81875	0,917	2,25	0,975	0,2125	-0,6625	7,3125
8	7	92030511060	0	0	0	1,05	89,236	70,5789	0,81875	0,917	2,25	0,975	0,2	2,375	6,75
9	8	92030511070	0	0	0	1,4	89,926	75,8394	0,81875	0,917	2,25	0,975	0,2	2,025	6,3125
10	9	92030511080	0	0	0	0	90,2119	78,0077	0,81875	0,917	2,25	0,975	0	3,225	6,5
11	10	92030511090	0	0	0	0,2125	90,3292	78,898	0,81875	0,917	2,25	0,975	0	3,0125	
12	11	92030511100	0	0	0	1,5	90,3774	79,2635	0,81875	0,917	2,25	0,975	0	1,725	0
13	12	92030511110	0	0	0	0,05	90,3972	79,4136		0,917	2,25	0,975	0	3,175	6,5
14	13	92030511120	0	0	0	5,55	90,4053	79,4752	0,81875	0,917	198,769	0,975	0	196,294	6,25
15	14	92030511130	0	0	0	3,8	90,4086	79,5005	0,81875	0,917	1,425	0,975	0	-1,4	5,5
16	15	92030511140	0	0	0	1,1625	90,41	79,5108	0,81875	0,917	1,425	0,975	0	1,2375	6
17	16	92030511150	0	0	0	5,6125		79,5151	0,81875	0,917	1,425	0,975	0	-3,2125	15
18	17	92030511160	0	0	0	0,3	90,4108	79,5169	0,81875	0,917	1,425	0,975	0	0	1
19	18	92030511170	0	0	0	0,5	90,4108	79,5176	0,81875	0,917	1,425	0,975	0	1,9	6,75
20	19	92030511180	0	0	0	7	90,4108	79,5179	0,81875	0,917	1,425	0,975	0	-4,6	6,25
21	20	92030511190	0	0	0	1,05	90,4108	79,5179	0,81875	0,917	1,425	0,975	0	1,35	6
22	21	92030511200	0	0	0	0,6	90,4108	79,5179	0,81875	0,917	1,425	0,975	0	1,8	6
23	22	92030511210	0	0	0	1,7	90,4108	79,5179	0,81875	0,917	1,425	0,975	0	0,7	6,75
24	23	92030511220	0	0	0	0,3	90,4108	79,5179	0,81875	0,917	1,425		1,25	3,35	5
25	24	92030511230	0	0	0	1,55	90,4108	79,5179	0,81875	0,917	1,425	0,975	1,25	2,1	7,25
26	25	92030511240	0	0	0	0	90,4108	79,5179	0,81875	0,917	1,425	0,975	1,25	3,65	6,75
27	26	92030511250	0	0	0	2,2125	90,4108	79,5179	0,81875	0,917	1,425	0,975	1,25	3,55	6,25
28	27	92030511260	0	0	0	1,45	90,4108	79,5179	0,81875	0,917	1,425	0,975	1,25	2,2	6
29	28	92030511270	0	0	0	1,25	90,4108	79,5179	0,81875	0,917	1,425	0,975	1,25	2,4	7,25

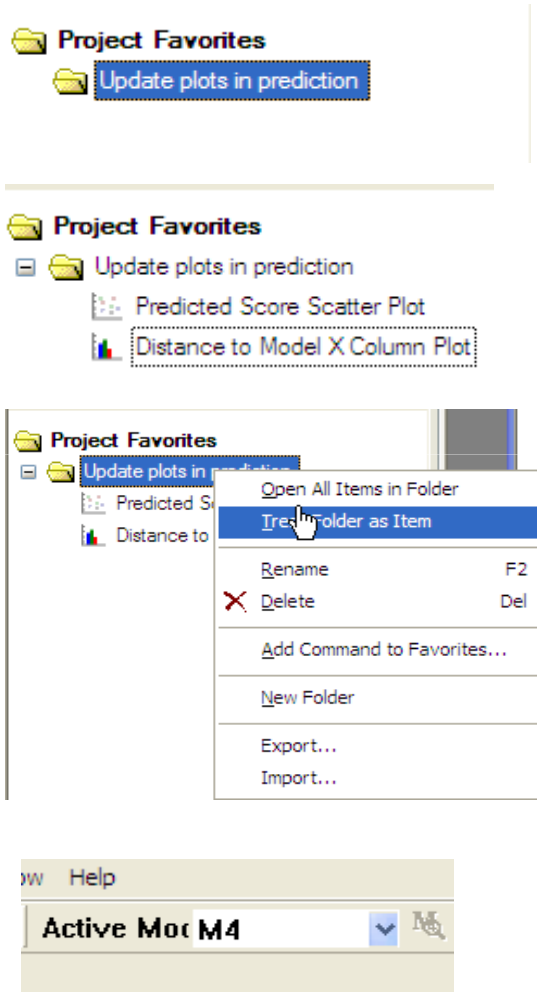




UMETRICS
AN MKS COMPANY

Create Your own Favorites selection

- Create a folder in project favorites
- Set a name on it
- Mark plots and chose ”Add to favorites”
- Set folder to ”Treat Folder as Item”
- Use ”Active Model” to shift model and click on the new favorite



UMETRICS
value from data

Method Extensions

Method Extensions

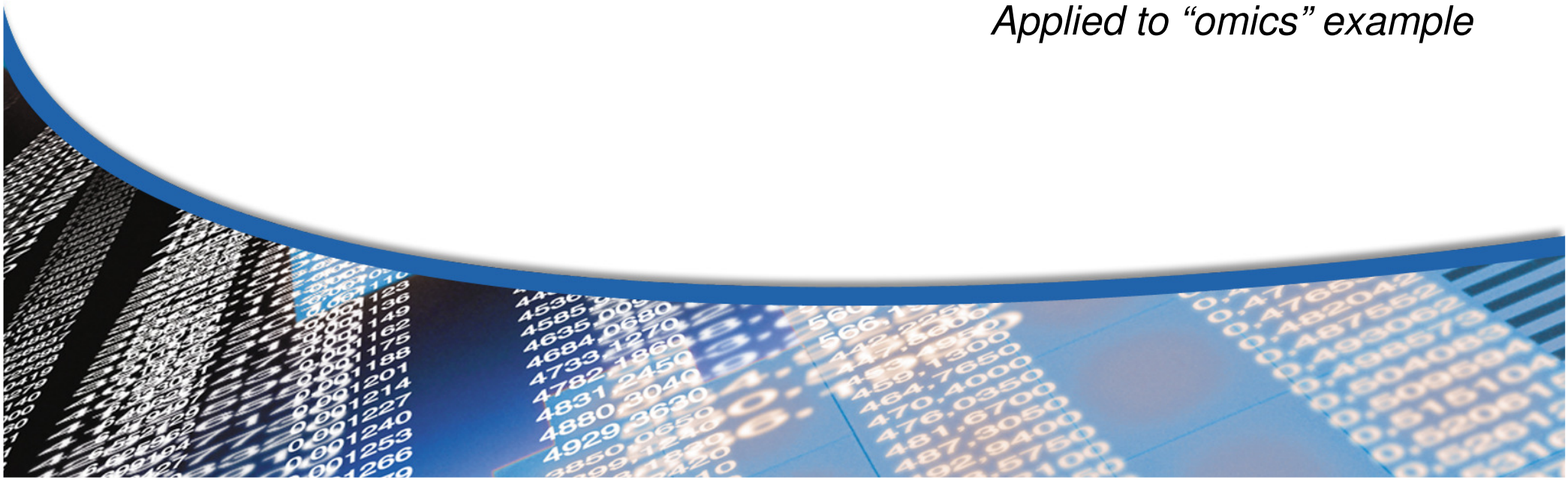
- Orthogonal PLS (OPLS)
  - Concentrates predictive power in the first component (1 Y)
- Non-linear modelling
  - Captures non-linear structure between X´s, between Y´s, between X and Y
- Hierarchical modelling
  - Models are developed in layers, allows zoom-in/zoom-out functionality
- Multivariate batch modelling
  - Unfolding of multi-way data tables to two-way arrays

UMETRICS
AN MKS COMPANY

UMETRICS
value from data

Orthogonal PLS

Applied to “omics” example



Why OPLS?

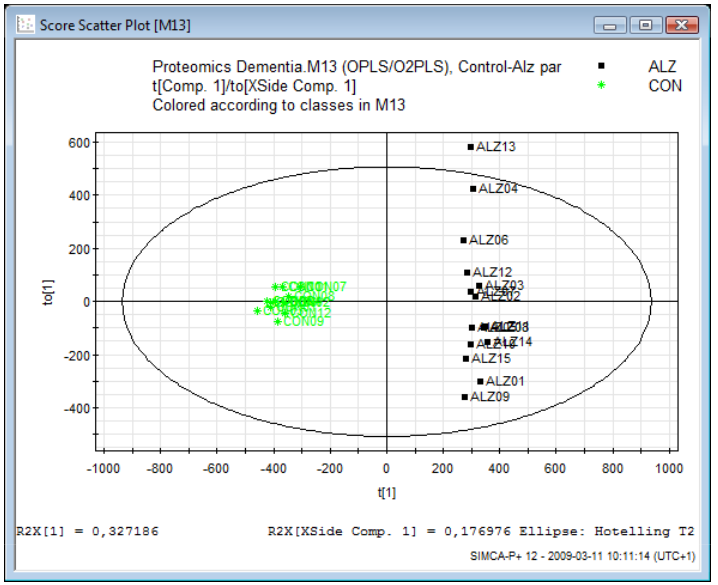
- Improved visualization and interpretation
  - Separates data into predictive and uncorrelated information
  - Improved visualization tools
    - S-plot
    - SUS-plot

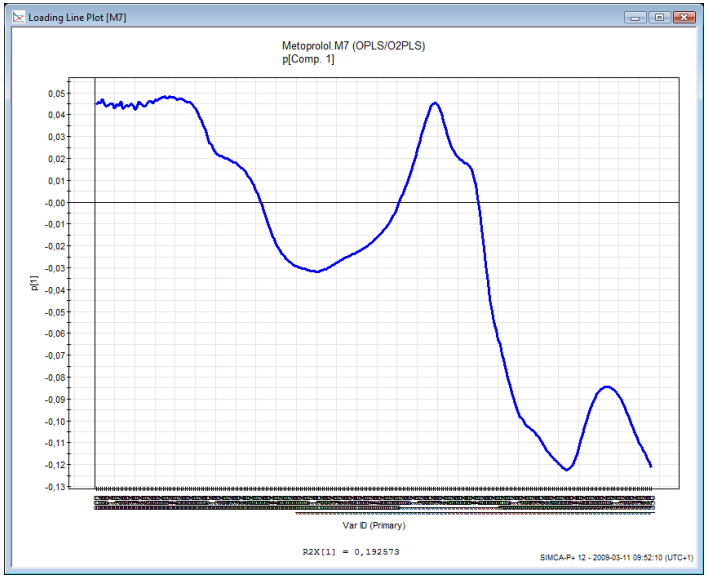
- Concept of uncorrelated information
  - Experimental problem(s)
    - Life style (humans)
    - Growth conditions (plants)
    - Instrument failures

UMETRICS
AN MKS COMPANY

Application areas of OPLS

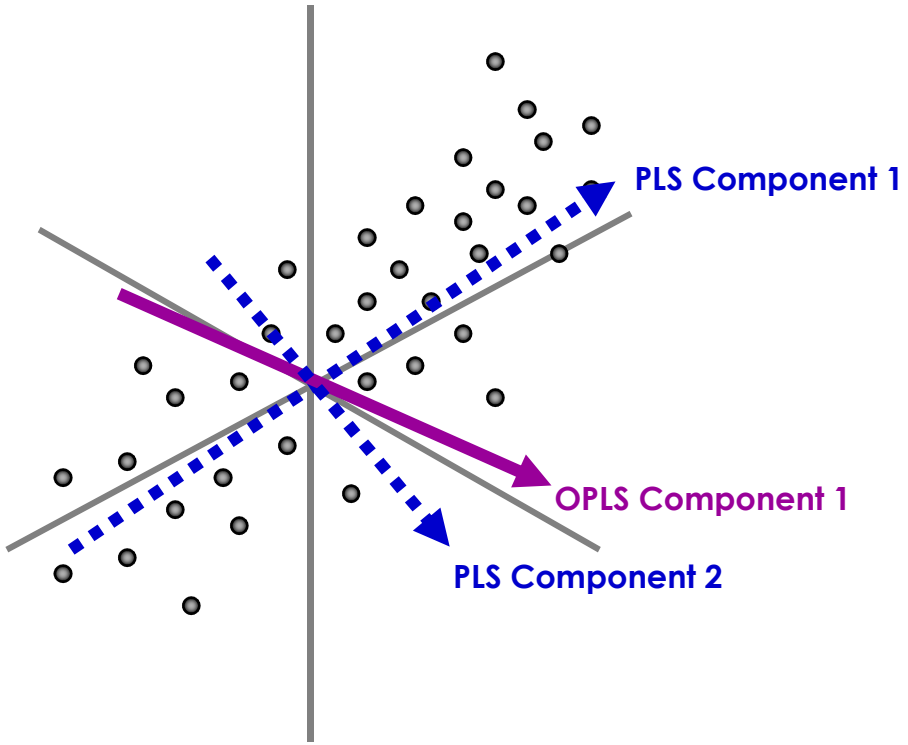
- Anywhere PLS is used
  - For single **y** (one response)
- When interpretation is important
- When there is much noise in data
- OPLS-DA for Omics data analysis
- OPLS for multivariate calibration





O-PLS vs PLS

- PLS compensates for structured noise by extracting successive components
- OPLS 'sees through' structured noise in order to find the Y predictive component

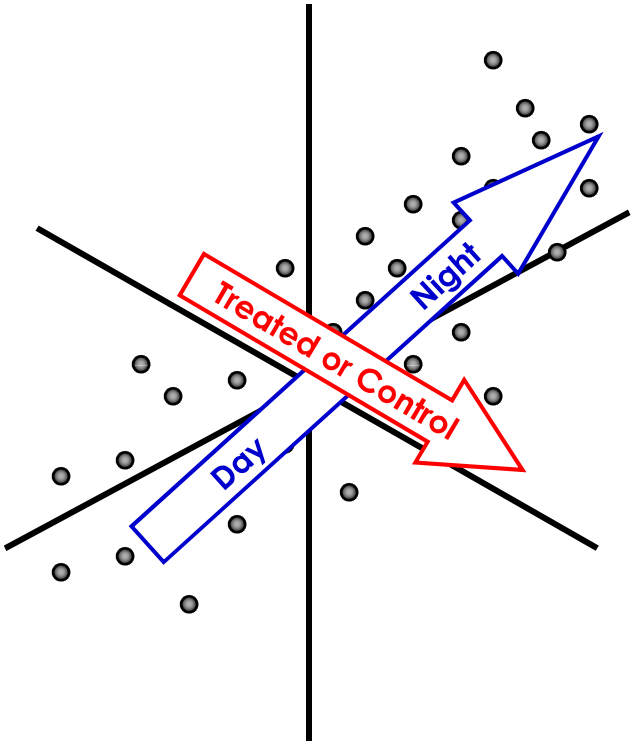


UMETRICS
AN MKS COMPANY

Y-orthogonal variation

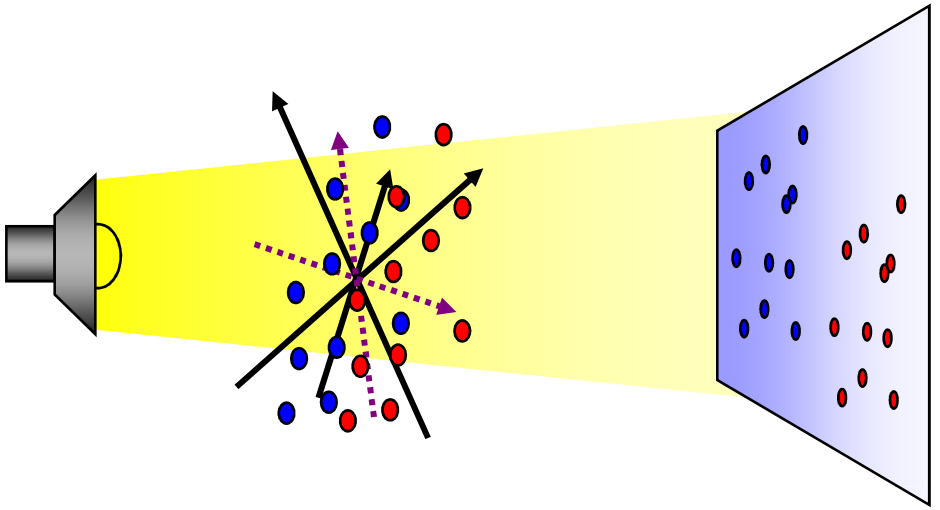
- Also called “Structured Noise” in literature

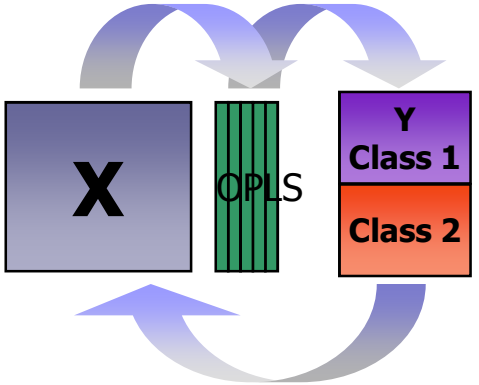
- Examples:
  - Light scattering in spectroscopy
    - but want moisture content
  - Temperature variation in lab
    - but want treated vs. control
  - Diurnal variation in animal metabolism
    - but want healthy vs. diseased
  - Gender variation
    - masks treated vs. control
  - Plate to plate variation
    - masks healthy vs. diseased

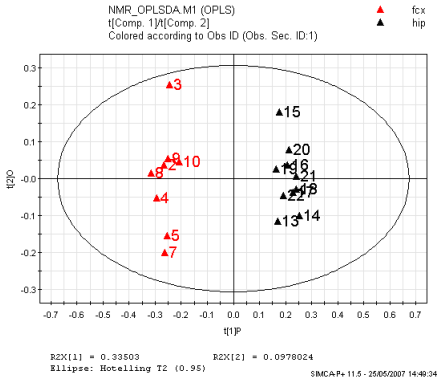


Separating Groups with OPLS-DA

- OPLS-DA relies on a projection of X as does PLS-DA
- BUT is able to differentiate y-related and y-orthogonal variation
- Applications
  - Biomarkers in metabonomics
  - Proteomics and Genomics
  - Transcriptomics

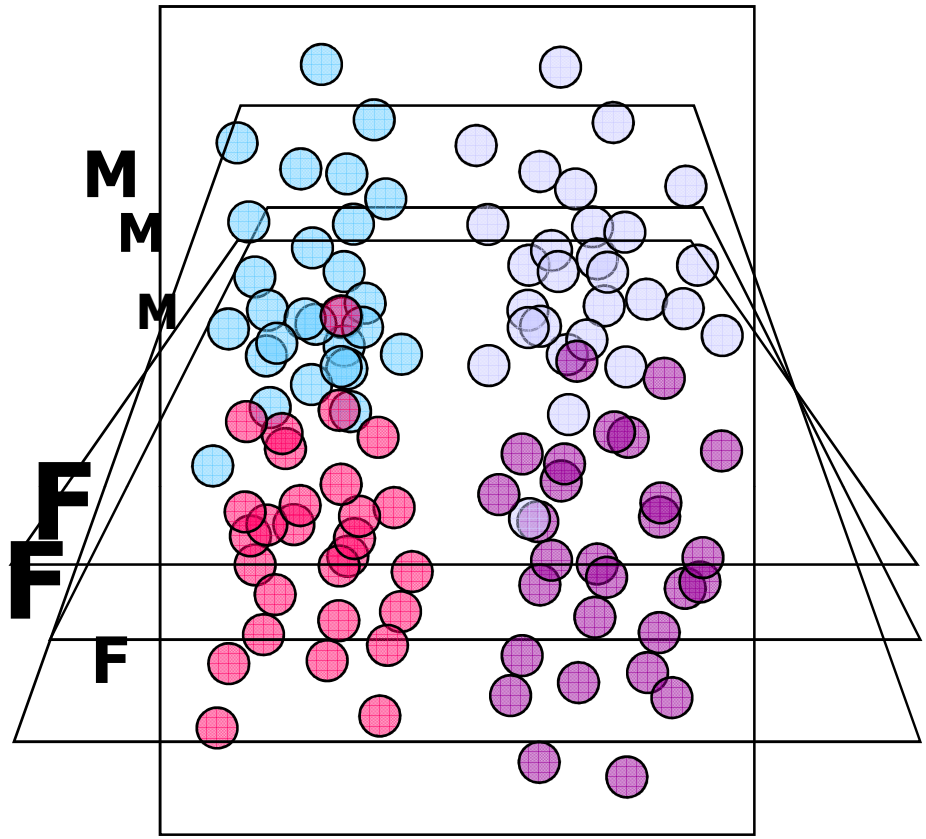






Coping with unwanted variation

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



Control vs Treated

OPLS-DA application: Clinical Proteomics Data

- Clinically diagnosed dementia patients plus healthy volunteers
- CSF-sampling; quantitative protein arrays (i.e. 95 proteins found in all samples)

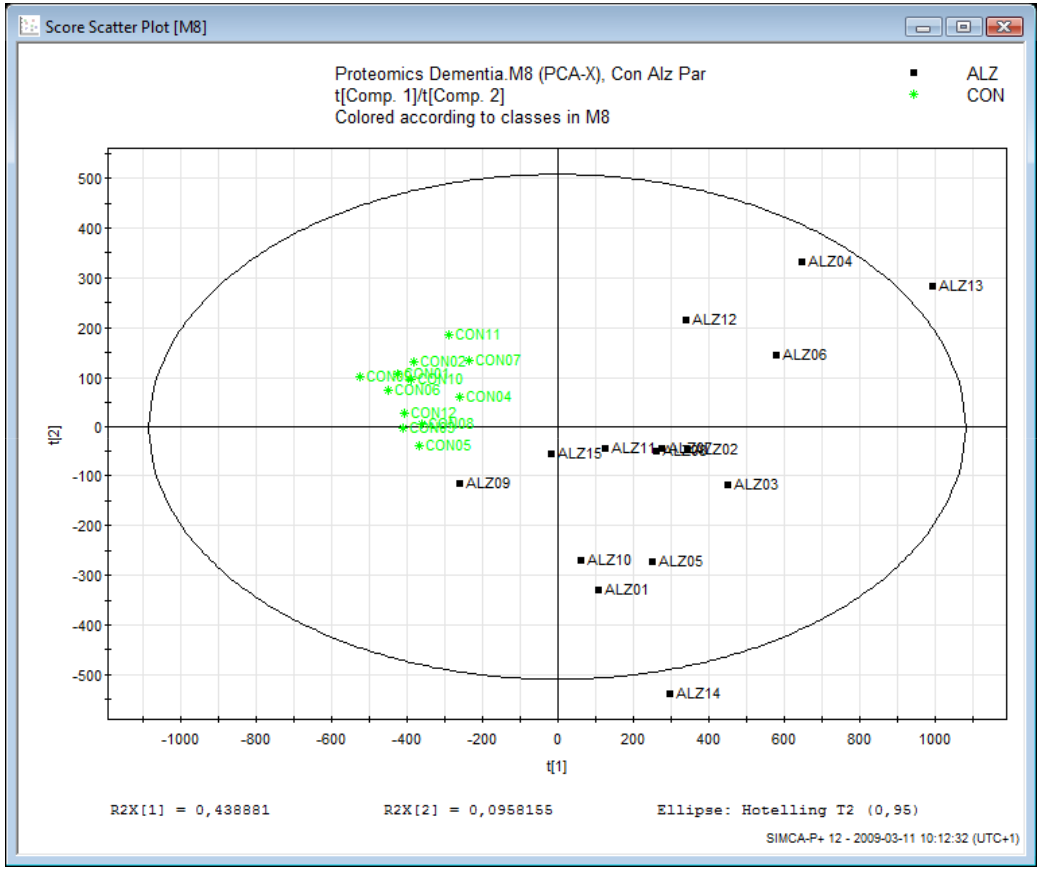
Data set (N=30, K=95)

- Healthy volunteers (CON, n=15)
- Alzheimer's disease (ALZ, n=15)
- Courtesy: J Gottfries, AZ R&D Mölndal

UMETRICS
AN MKS COMPANY

PCA for overview

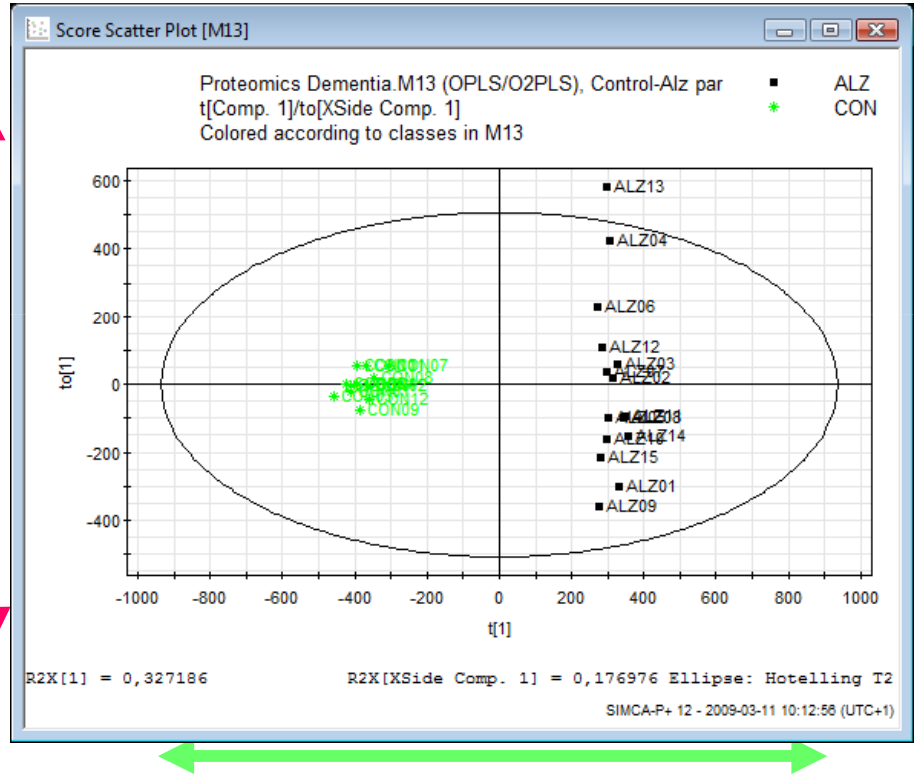
- Always PCA first!
- Deviating samples?
- Groups
  - Expected and unexpected



- OPLS-DA gives *Improved interpretation* when separating classes
  - Class discriminating information (Left to Right)
  - Within class information (Up - Down)
    - Uncorrelated information

- The loadings plot for component 1 is 100% purely related to the separation
  - This is how we find the biomarkers

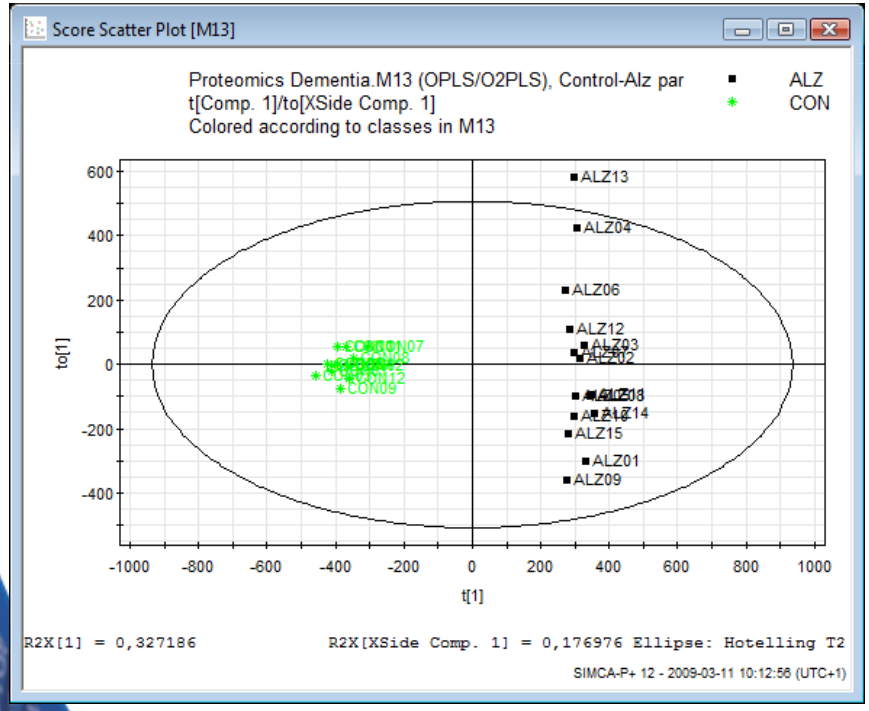
Within group variation

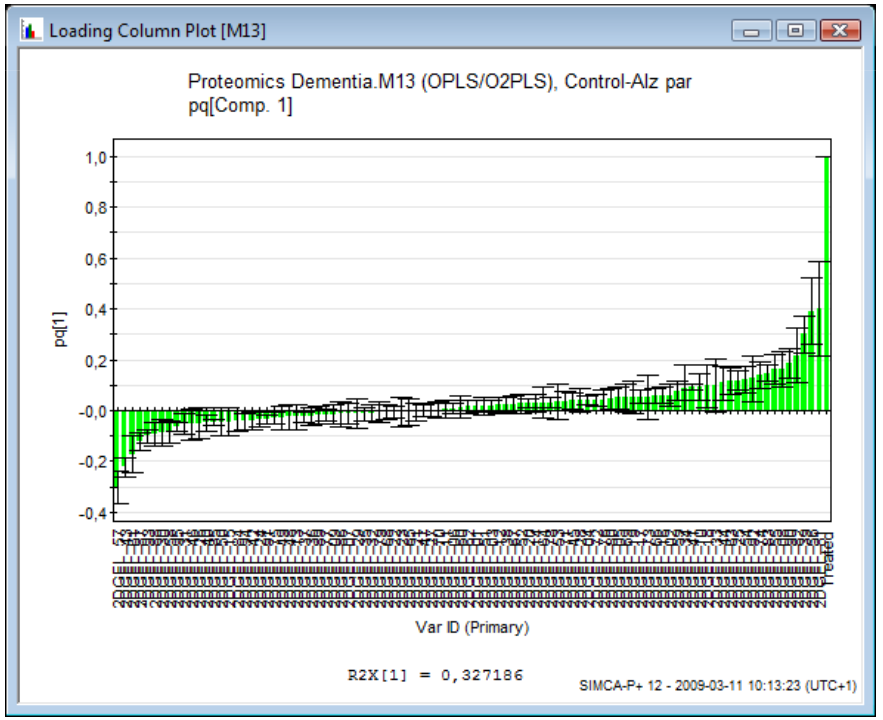


Between group variation

OPLS- group separation

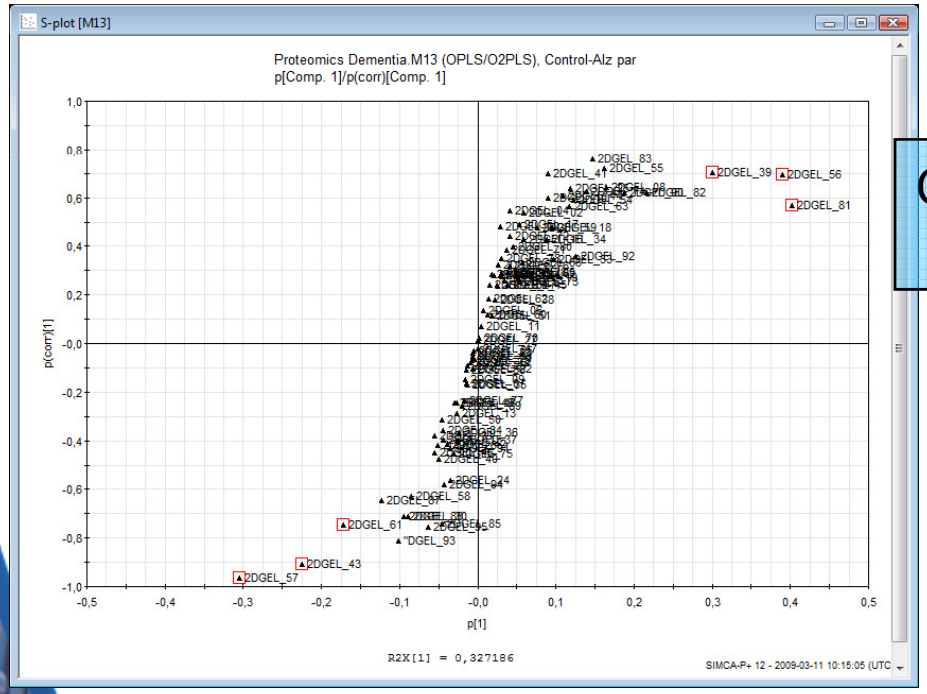
- First component summarize all between group variation
- Interpreted in first loading
  - ONLY related to group!





S-Plot to *identify* treatment related variables

- The ends of the S are the variables which contributes most to the model
- The “top” and “bottom” variables of the S has the highest confidence/reliability
- In OPLS-DA application in ‘Omics’ these are the “potential biomarkers”



Create List

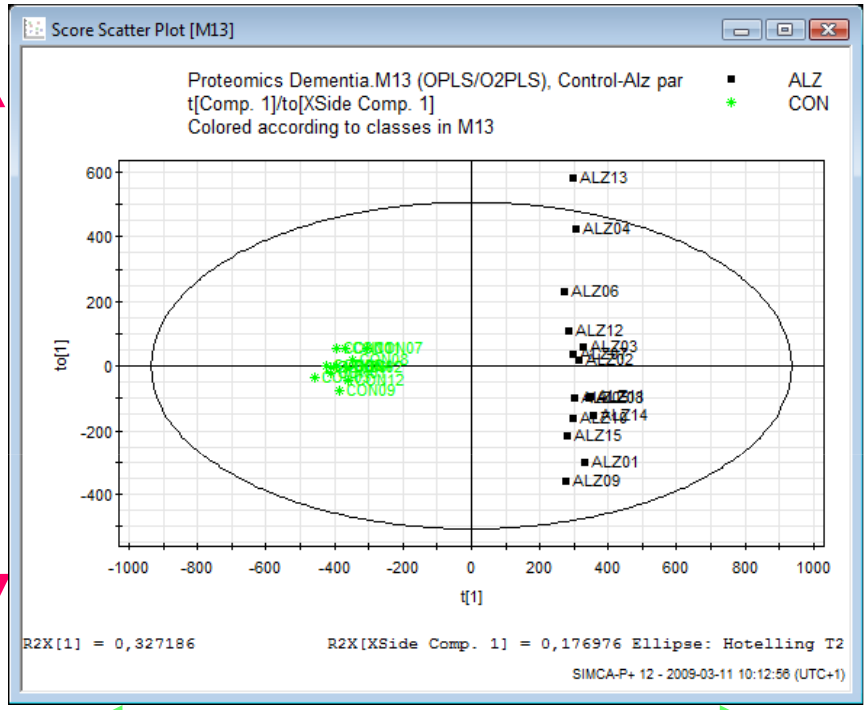
General List [M13]

	1	2	3	4
1	Var ID (Primary)	Var ID (Var. Sec. ID:1)	M13.p[1]	M13.p(corr)[1]
2	2DGEL_39	x39	0,299957	0,701129
3	2DGEL_43	x43	-0,224022	-0,909437
4	2DGEL_56	x56	0,390091	0,69507
5	2DGEL_57	x57	-0,303935	-0,965599
6	2DGEL_61	x61	-0,172102	-0,747149
7	2DGEL_81	x81	0,402474	0,569223

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

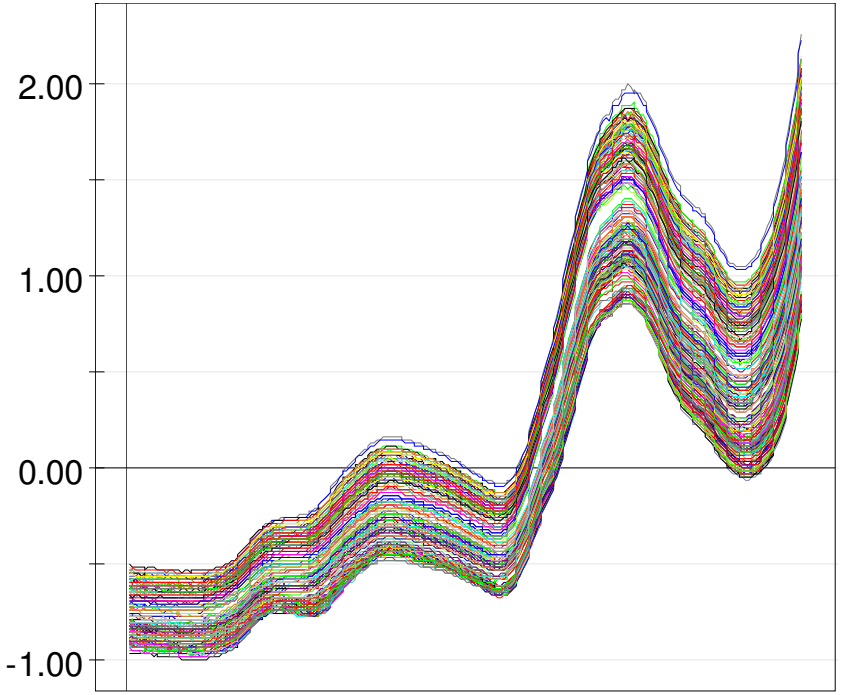
Uncorrelated information / Within group variation

- Causes to uncorrelated information
  - Experimental problem(s)
    - Life style (humans)
    - Growth conditions (plants)
    - Instrument failures
- Interpretation of uncorrelated information
- Main advantage
  - Understanding means that the next series can be performed better.



OPLS Application: Modeling and prediction of Metroprolol content in tablets

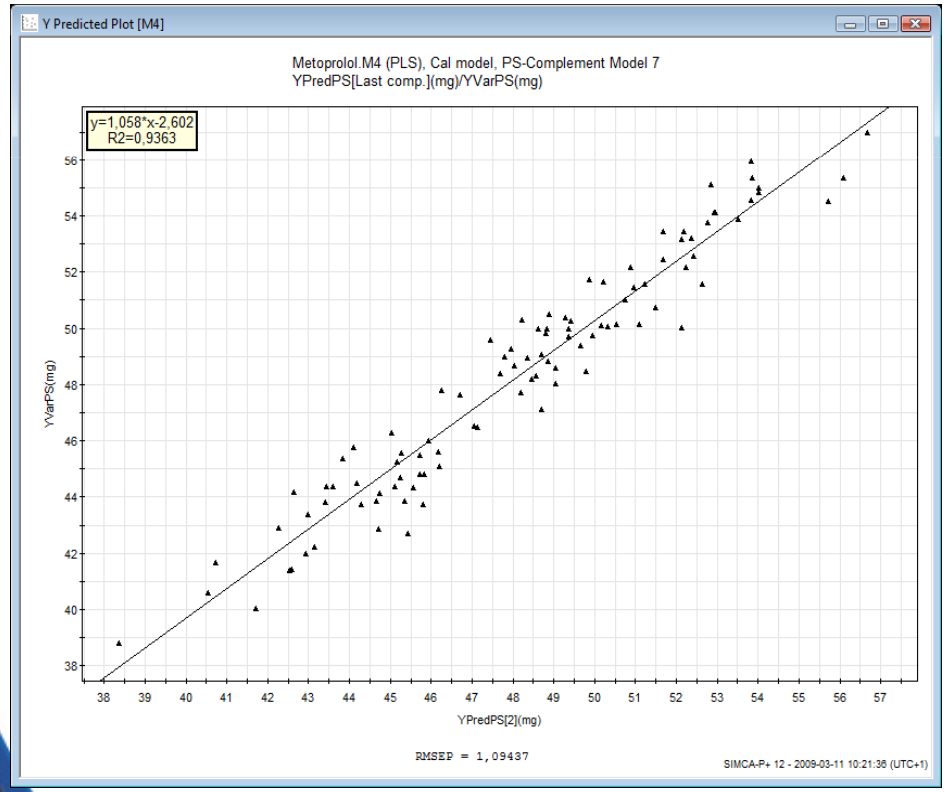
- Very early PAT application
  - In the times of PAC
- Study performed at Astra Hässle (later AstraZeneca), Mölndal, Sweden
- Aim: By NIR predict content of Motroprolol in tablets
- Benefit: Quick and non-destructive analysis technique

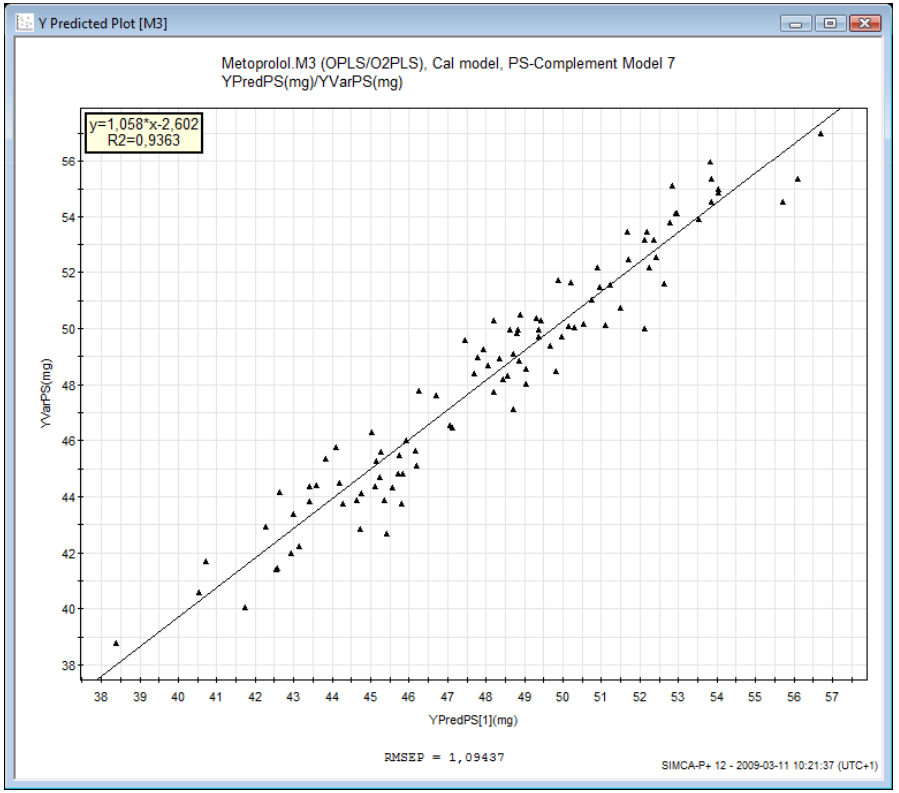


UMETRICS
AN MKS COMPANY

PLS vs OPLS

- Compare predictive ability of models
  - Identical!

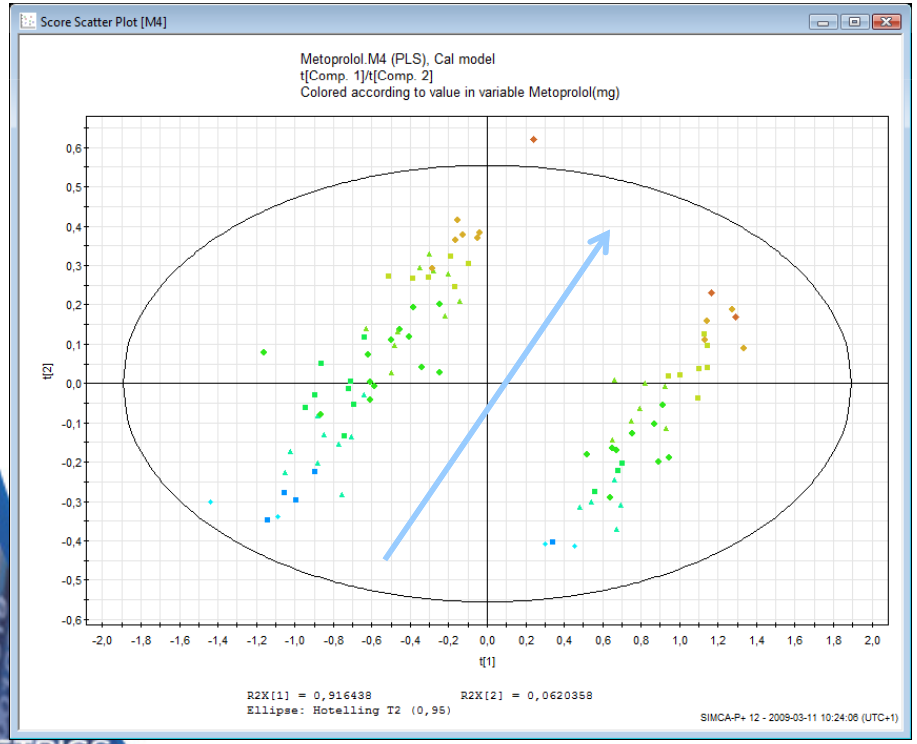


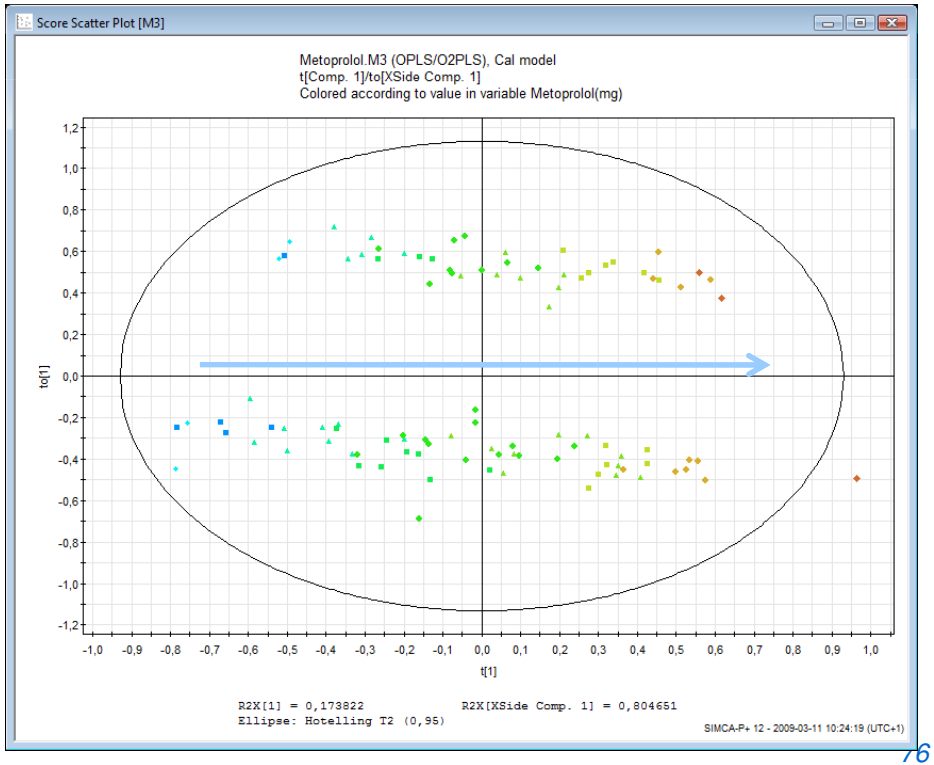


UMETRICS
AN MKS COMPANY

Benefit lies in *interpretation*

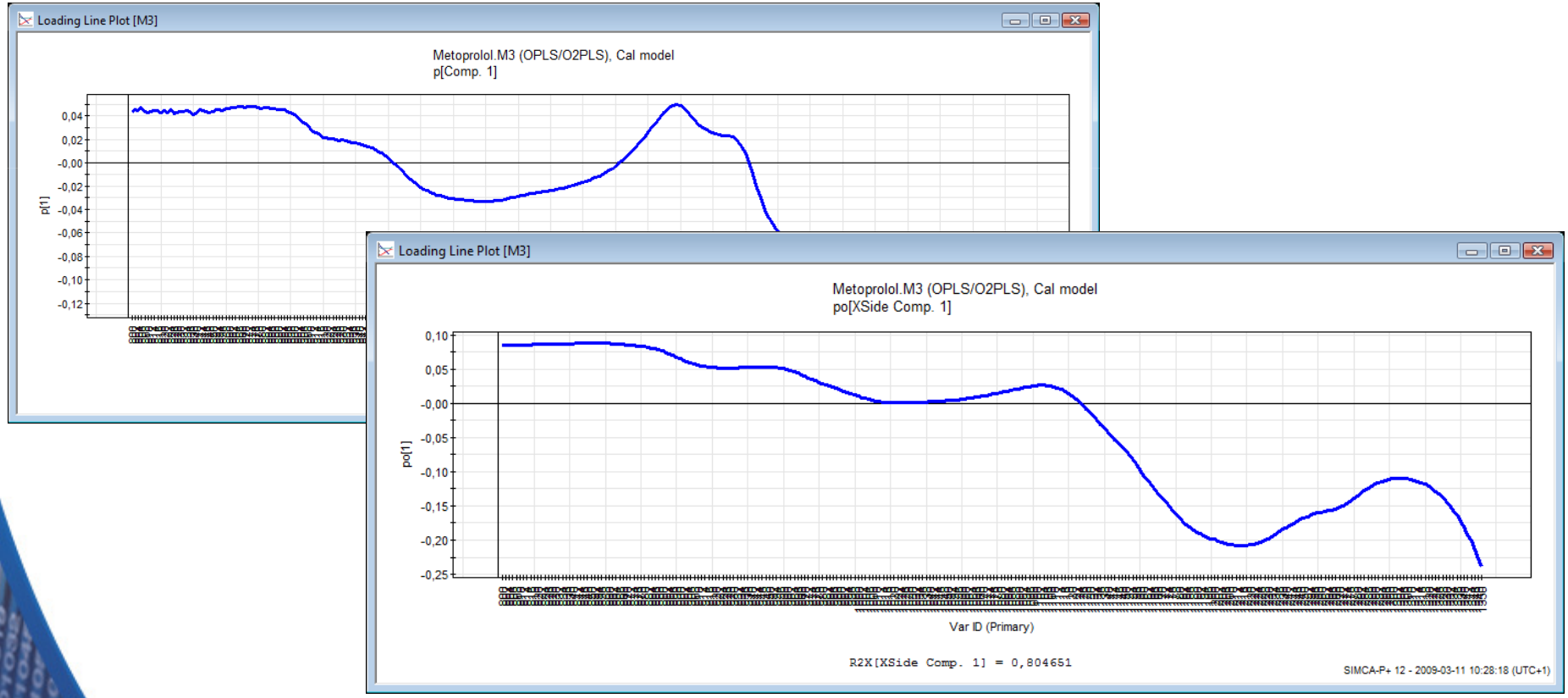
- Interpretation and understanding is today crucial!
- PLS requires 2 components to predict Metroprolol content
  - Which loading should be interpreted?
- OPLS requires 1
  - Interpret first loading





Predictive profile for Metroprolol content

- Represents NIR spectrum
- Spectral shape connected to Metroprolol content
- 2nd component: Scattering etc



Summary, OPLS and OPLS-DA

- OPLS and PLS are identical in predictions

- OPLS focus on interpretation for:
  - Single response cases when PLS gives more than one component

- Easier interpretation, only 1st component used for prediction
  - Straight-forward identification of putative biomarkers
  - More transparent interpretation of model diagnostics

- Identification/interpretation of **y**-uncorrelated information in 2nd – Ath component
  - Used to find experimental or biological effects NOT related to the goal of the experiment
    - Possibility to improve methodology next time
    - Understand system better
- Applications
  - Mainly omics
  - Spectoscopy- PAT
  - …

UMETRICS
AN MKS COMPANY

UMETRICS
value from data

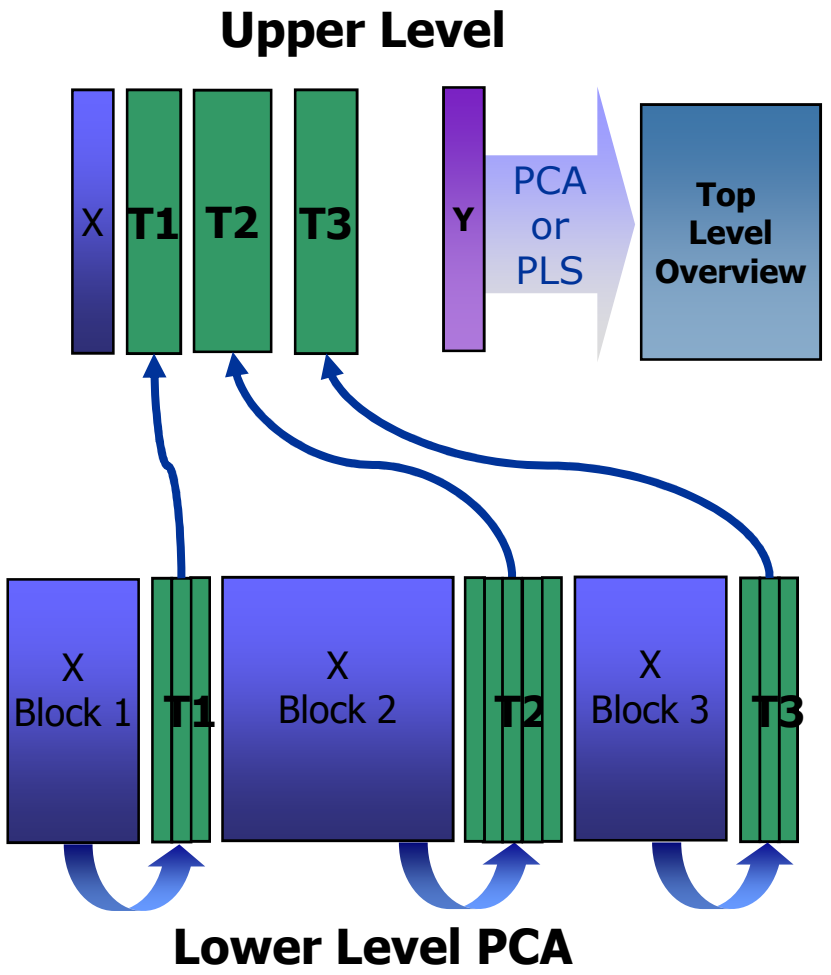
Hierarchical Modelling

To see from two perspectives…

Introduction: Hierarchical Modelling

- Common with multiple data sources
- Data pre-treatment a challenge
  - Various units
  - Various sizes of data tables
  - Various level of information

- Multi-step multivariate modelling
- Aim is simplification and ease of interpretation of large datsets
- Scores (or residuals) from one model used as basis for next
- May be used with PCA, PLS or PLS-DA

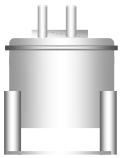


Combining data sources necessary!

Starting material

400 l

2000 l



10000 l

Down Stream

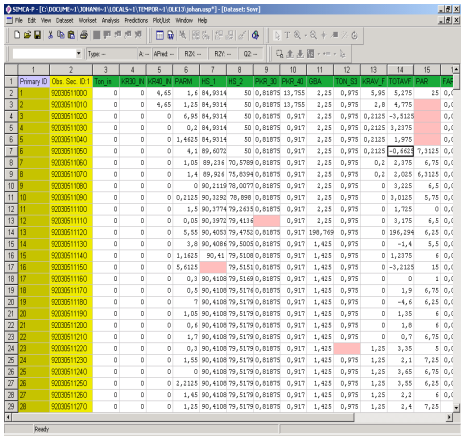
MES

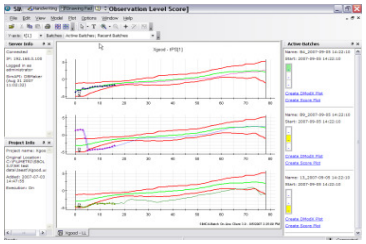
DCS

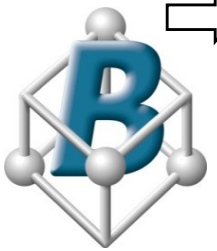
SCADA

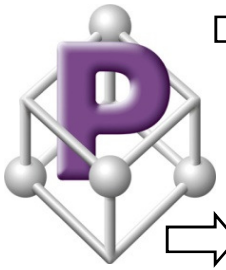
LIMS

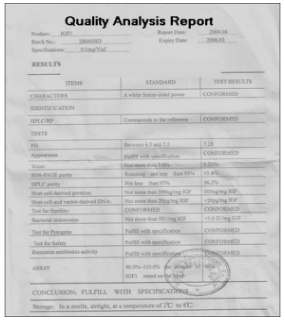
DataBase Historian











Quality Analysis Report

Example: Monitoring (PROC1A from SIMCA-P manual)

- A continuous steady state chemical process
- The process went out of control around time 80 and had to be shut down at time 92
  - Why?
  - When could MSPC detect the issue?
- The data - 33 variables, 92 hourly observations
  - 7 controlled process (feed) variables (x1in-x7in)
  - 18 intermediate (reaction + purification) process variables (x8md-xpen)
  - 8 output variables (y1-y8)
  - y6 = impurity level and y8 = yield are the most important outputs

UMETRICS
AN MKS COMPANY

Hierarchical Models: Arrangement of data

Block X1 (7)	Block X2 (8)	Block X3 (10)
Input and feed (x1 - x7)	Reaction conditions (x8 - x15)	Purification step (x16 - x25)

Block Y1 (6)	Block Y2 (2)
Less important Y´s (y1-y5 & y7)	Important Y´s (y6 & y8)

- Using observations 1 – 79 as a training set, we shall do the following:
  - 1) PLS of block X1 vs block Y2
  - 2) PLS of block X2 vs block Y2
  - 3) PLS of block X3 vs block Y2
  - 4) PCA of block Y1
  - 5) Top level PLS model based on the scores of models 1-4 and block Y2

Summarizing the feed

- PLS of block X1 vs block Y2
- Isolate variability in feed parameters which relates to important responses

HI-PROC - M3

Workset... Options... Title PLS X1 (feed) vs Y2 without obs 1 ...

Type: PLS Observations (N)=78, Variables (K)=9 (X=7, Y=2)

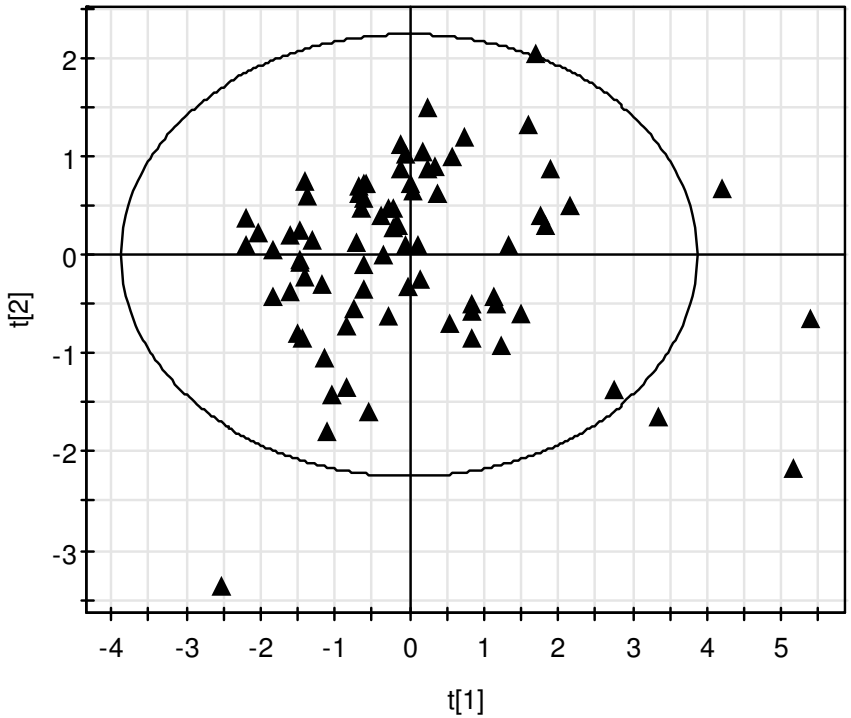
Components:

A	R2X	R2X(cum)	Eigenv...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signifi...	Ite...
0	Cent.			Cent.						
1	0.365	0.365	2.55	0.163	0.163	0.13	0.05	0.13	R1	5
2	0.134	0.499	0.939	0.0359	0.199	-0.0293	0.05	0.104	NS	33
3	0.0842	0.583	0.59	0.0413	0.24	-8.38e-005	0.05	0.104	NS	8
4	0.0544	0.638	0.381	0.0104	0.251	-0.0755	0.05	0.0368	NS	7

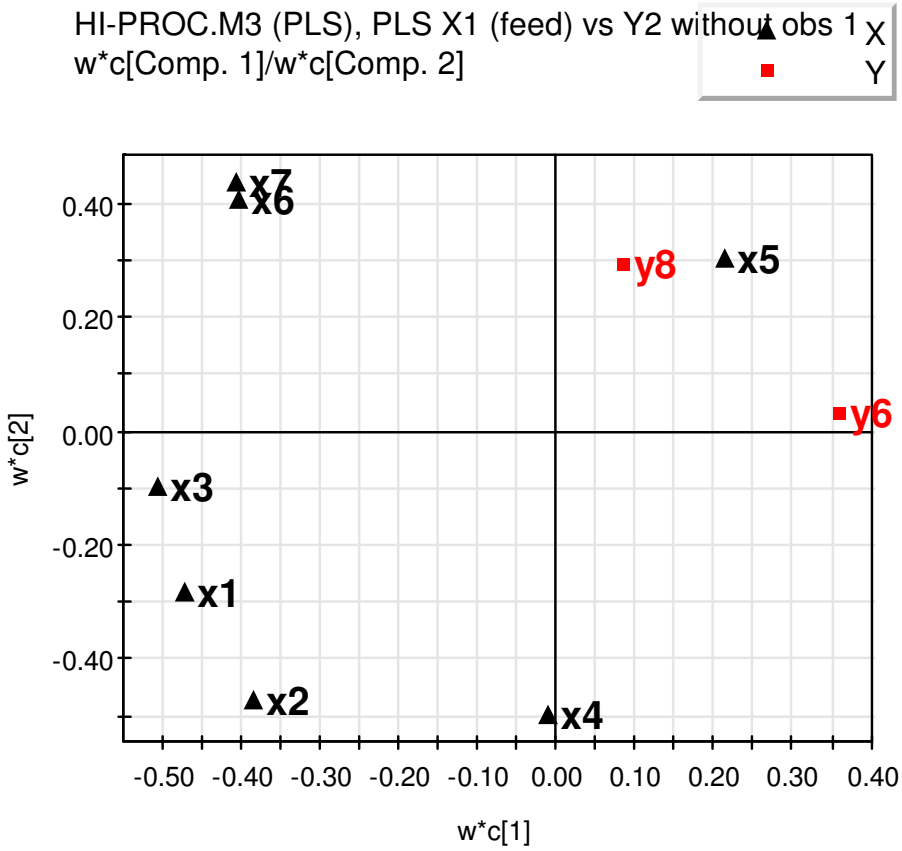
UMETRICS AN MKS COMPANY

Summarizing the feed – scores and loadings

HI-PROC.M3 (PLS), PLS X1 (feed) vs Y2 without obs 1
t[Comp. 1]/t[Comp. 2]



HI-PROC.M3 (PLS), PLS X1 (feed) vs Y2 without obs 1
w*c[Comp. 1]/w*c[Comp. 2]



UMETRICS
AN MKS COMPANY

Summarizing the less important Y´s

- PCA of block Y1 (without observation 1)
- 3 components summarize well (R^2 = 0.75)

HI-PROC - M6

Workset... Options... Title PCA Y1 (less important y´s) ...

Type: PCA-X Observations (N)=78, Variables (K)=6 (X=6, Y=0)

Components:

A	R2X	R2X(cum)	Eigenv...	Q2	Limit	Q2(cum)	Significance	Iterations
0	Cent.							
1	0.364	0.364	2.18	0.116	0.154	0.116	NS	25
2	0.228	0.592	1.37	-0.271	0.177	0.028	NS	27
3	0.157	0.749	0.944	-0.609	0.211	-0.0692	U	125

UMETRICS AN MKS COMPANY

Top level – data arrangement

Block X1 (7)	Block X2 (8)	Block X3 (10)
Input and feed (x1 - x7)	Reaction conditions (x8 - x15)	Purification step (x16 - x25)

Block Y1 (6)	Block Y2 (2)
Less important Y´s (y1-y5 & y7)	Important Y´s (y6 & y8)

X1 (4)	X2 (4)	X3 (4)
scores from model 1	scores from model 2	scores from model 3

Top level PLS model ⇒

Y1 (3)	Y2 (2)
scores from model 4	y6 & y8

HI-PROC

Observations (N) = 92, Variables (K) = 33

Models:

No.	Model	Type	A	R2X	R2Y	Q2(cum)	Date	Title	Hiera...	
1	M1	PCA-X	4	0.554		0.327	2001-11-28	PCA for overview of entire data set		
2	M2	PLS	4	0.654	0.249	0.123	2001-11-28	PLS X1 (feed) vs Y2		
3	M3	PLS	4	0.638	0.251	0.0368	2001-11-28	PLS X1 (feed) vs Y2 without obs 1	B	Model 1
4	M4	PLS	4	0.759	0.46	0.297	2001-11-28	PLS X2 (react cond) vs Y2 obs 1 excl	B	Model 2
5	M5	PLS	4	0.649	0.55	0.431	2001-11-28	PLS X3 (purif. step) vs Y2 obs 1 excl	B	Model 3
6	M6	PCA-X	3	0.749		-0.0692	2001-11-30	PCA Y1 (less important y´s)		
7	M7	PCA-X	3	0.764		-0.0322	2001-11-30	PCA Y1 (less imp Ys) obs 1/65 excl	B	Model 4
8	M8	PLS	4	0.579	0.562	0.454	2001-11-30	Top level PLS model	T	

UMETRICS AN MKS COMPANY

Top level PLS model

There are four significant components explaining 56% of the Y´s.

HI-PROC - M8

Workset... Options... Title Top level PLS model ...

Type: PLS Observations (N)=77, Variables (K)=17 (X=12, Y=5)

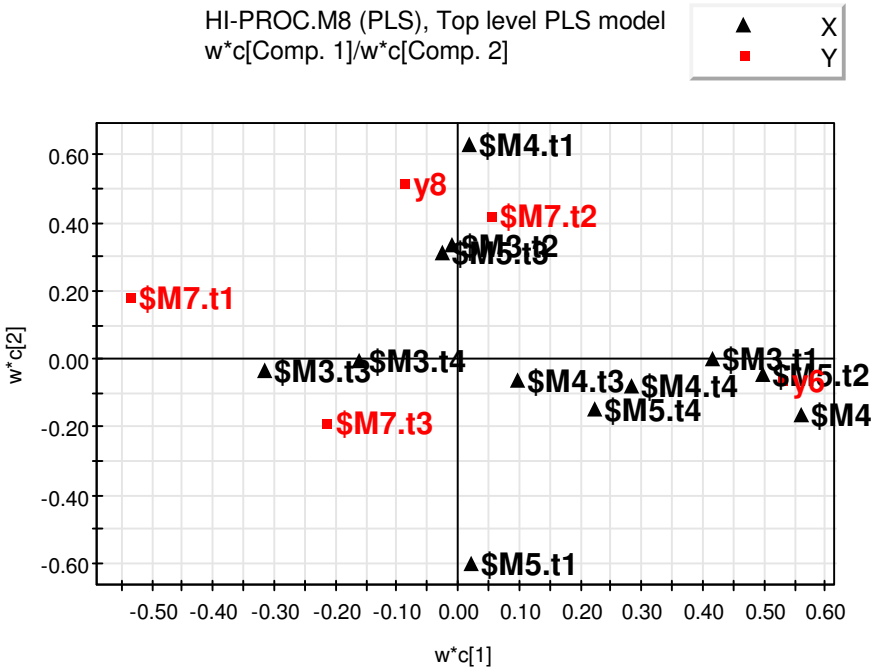
Components:

A	R2X	R2X(cum)	Eigenv...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signifi...	Ite...
0	Cent.			Cent.						
1	0.186	0.186	2.23	0.261	0.261	0.227	0.05	0.227	R1	31
2	0.177	0.362	2.12	0.214	0.475	0.273	0.05	0.437	R1	7
3	0.106	0.468	1.27	0.064	0.539	0.0419	0.05	0.461	R2	11
4	0.111	0.579	1.33	0.0231	0.562	-0.0136	0.05	0.454	R2	20

UMETRICS AN MKS COMPANY

Top level PLS model - Overview Interpretation

- y6 important in the first component.
- Positively correlated with y6 are:
  - Comp.1 of feed (M3)
  - Comp.2 of the reaction conditions (M4)
  - Comp.2 of the purification (M5)
- y6 is negatively correlated with Comp.1 of the the less important Y´s.
- y8 (heavy in comp.2) is negatively correlated to Comp.1 of purification.

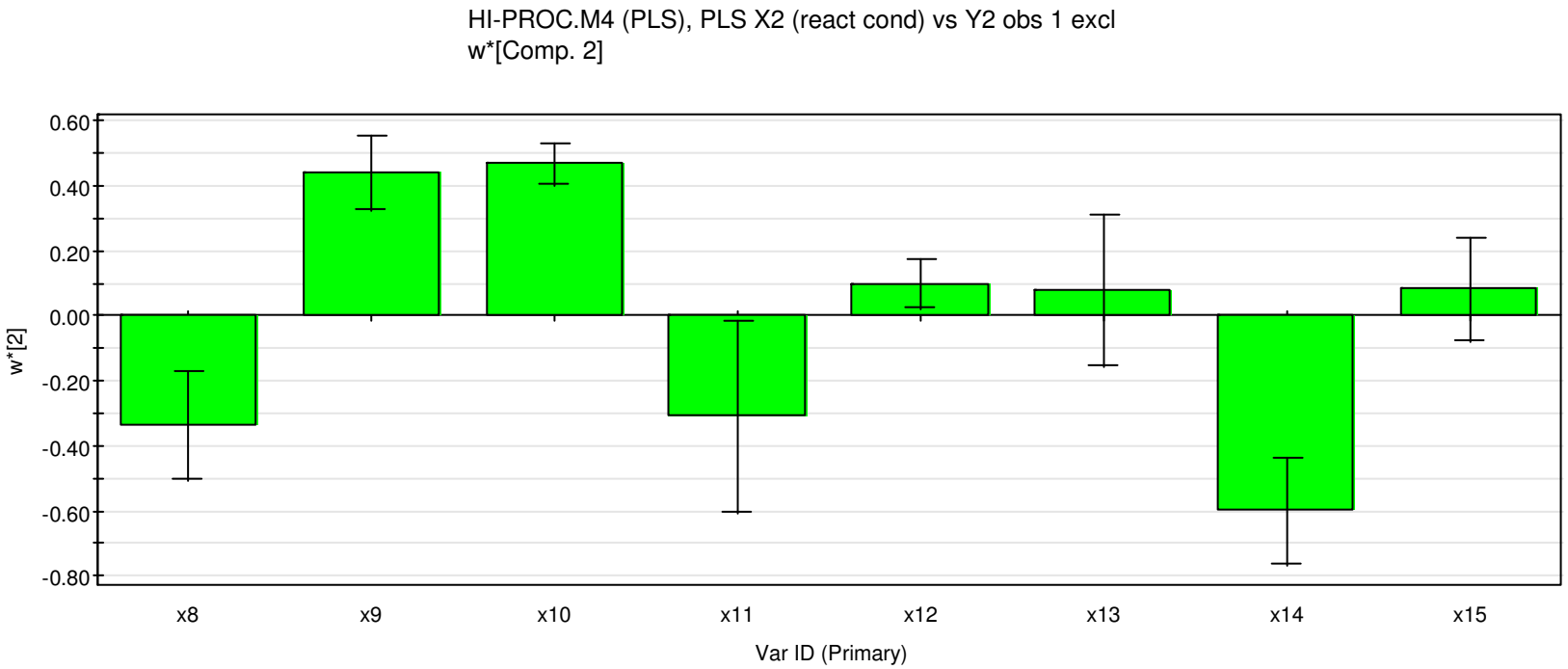


Zoom-in/zoom-out option

- Top level wc-plot: Relationships between y6/y8 and the different sections of the process (feed/reaction/purification)
- With the contribution tool, just double-click on any score variable point…
- A contribution plot uncovers which variables in the lower blocks --feed/reaction/purification -- dominate and the relationship to the Y´s.

Zooming-in via contibution plotting of base-level loadings

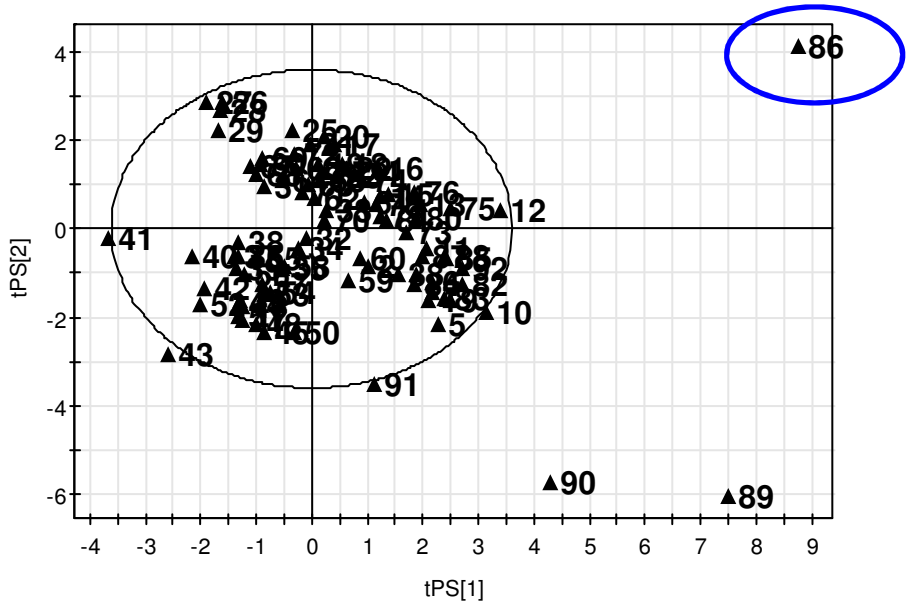
- Example: positive correlation between y6 and second component of reaction conditions (”§M4.t2”)



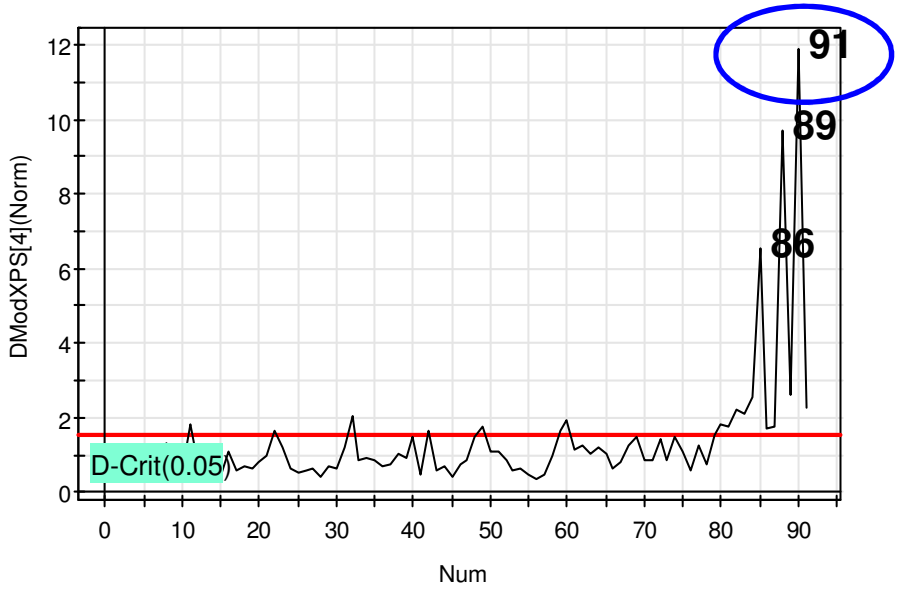
Use hierarchical model for Predictions

- Observations 80-92 predicted
- The scores and DModX for the predicted observations clearly show that the process is going out of control at the end of the sampling period.

HI-PROC.M8 (PLS), Top level PLS model, PS-HI-PROC
tPS[Comp. 1]/tPS[Comp. 2]



HI-PROC.M8 (PLS), Top level PLS model, PS-HI-PROC
DModXPS[Comp. 4]

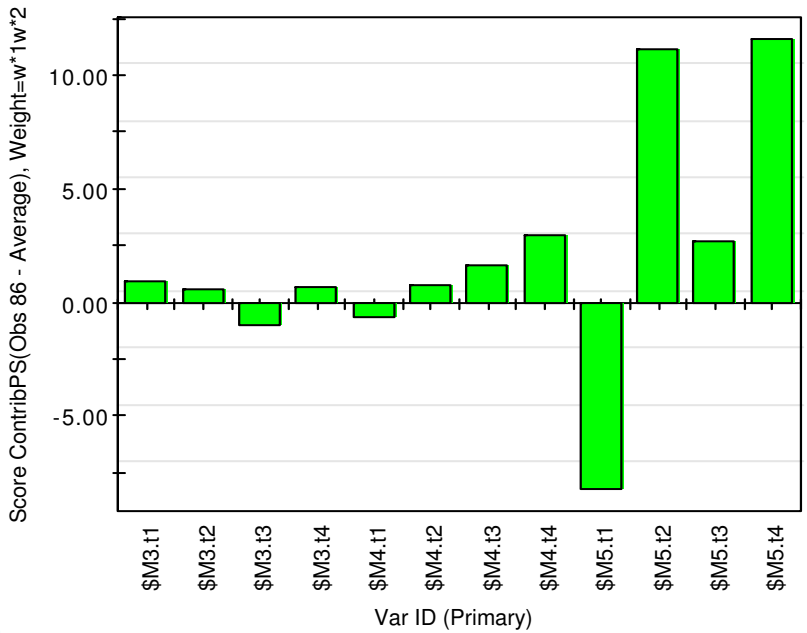


M8-D-Crit[4] = 1.545

Why is observation 86 so different?

- Contribution plot (scores

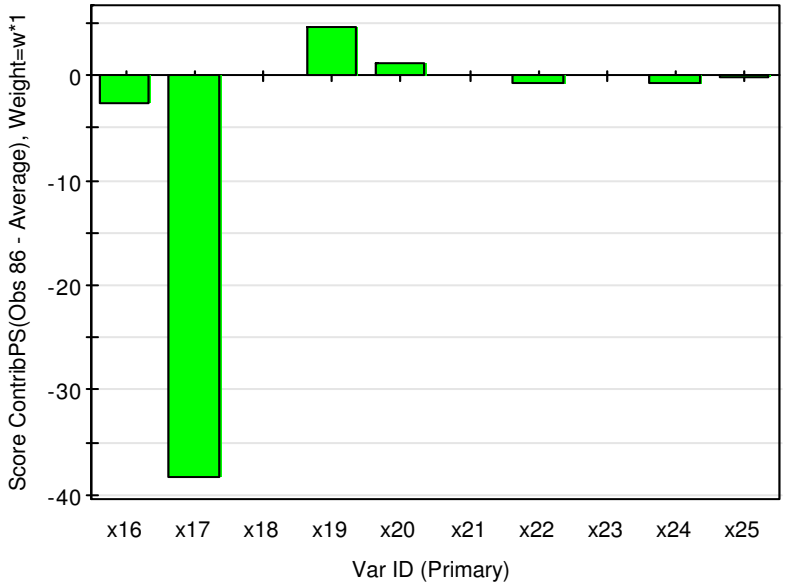
HI-PROC.M8 (PLS), Top level PLS model, PS-HI-PROC
Score ContribPS(Obs 86 - Average), Weight=w*[1]w*[2]



- Problems mainly in the purification step

- Zooming-in on the purif. step

HI-PROC.M5 (PLS), PLS X3 (purif. step) vs Y2 obs 1 excl, PS-HI-PROC
Score ContribPS(Obs 86 - Average), Weight=w*[1]



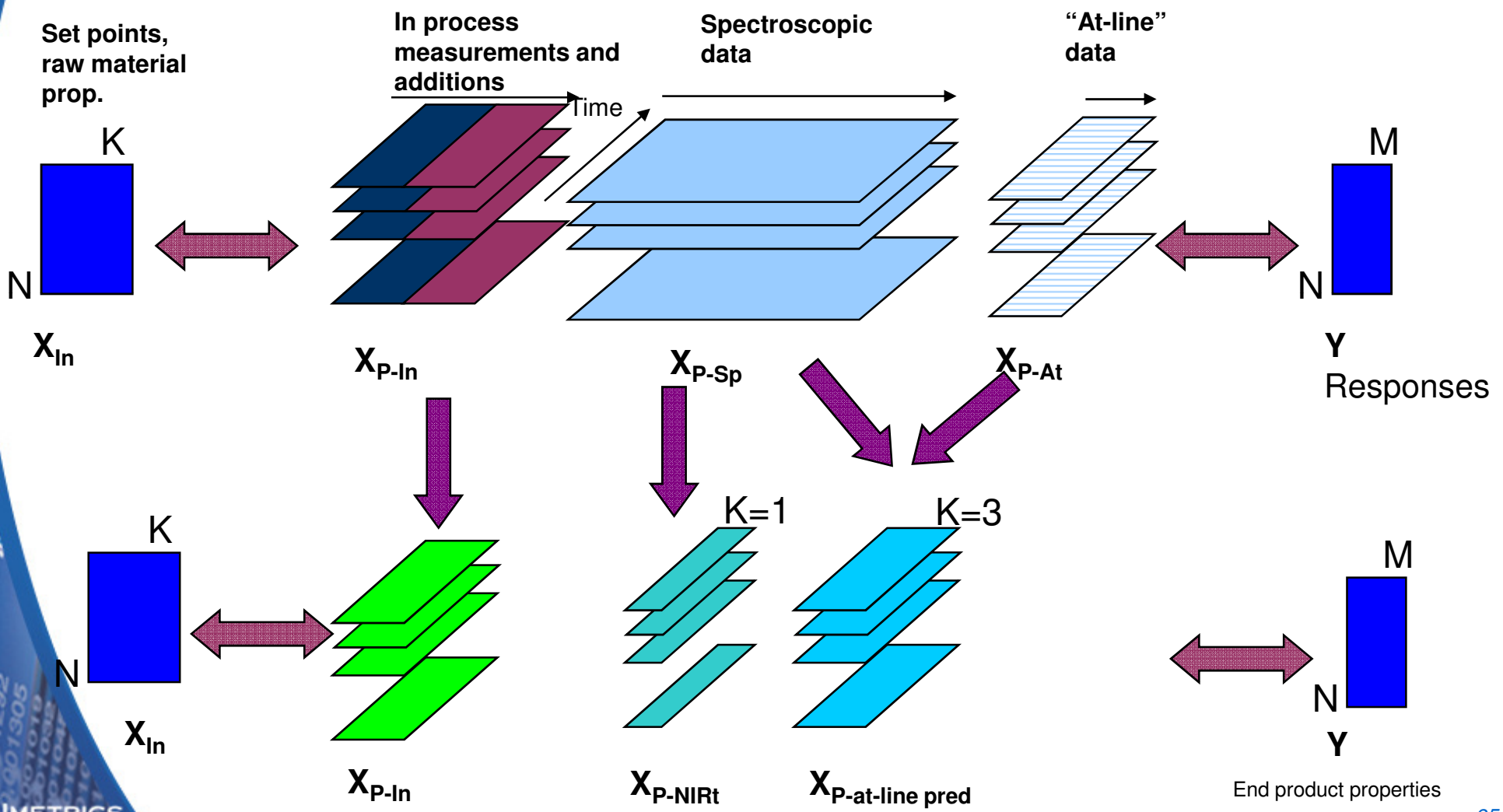
- The automatic contribution plot in the base-level model points to variable x17 (xhnx)

Conclusions – PROC1A

- The hierarchical approach to multivariate analysis:
  - Enable combinations of different data sources
    - Reaction conditions, spectral, environmental….
  - Simplifies the interpretation of complex problems
- The zoom-in/zoom-out capabilities allow us
  - to understand complex relationships in terms of segments of a process
  - to zoom-in on a single segment to look at the individual process variables
- The model identified the process upset after observation 80
- The deviation of observation 86 was mainly due to upsets in purification step

UMETRICS
AN MKS COMPANY

Hierarchical modeling of up-stream fermentation process



UMETRICS
value from data

Multivariate batch modeling

Contents

- Introduction to batch modeling
- Organization of batch data
- Two levels of batch modeling
  - observation level
  - batch level
- Tracing batch evolution
- Diagnosing upsets
- Batch modeling in practice

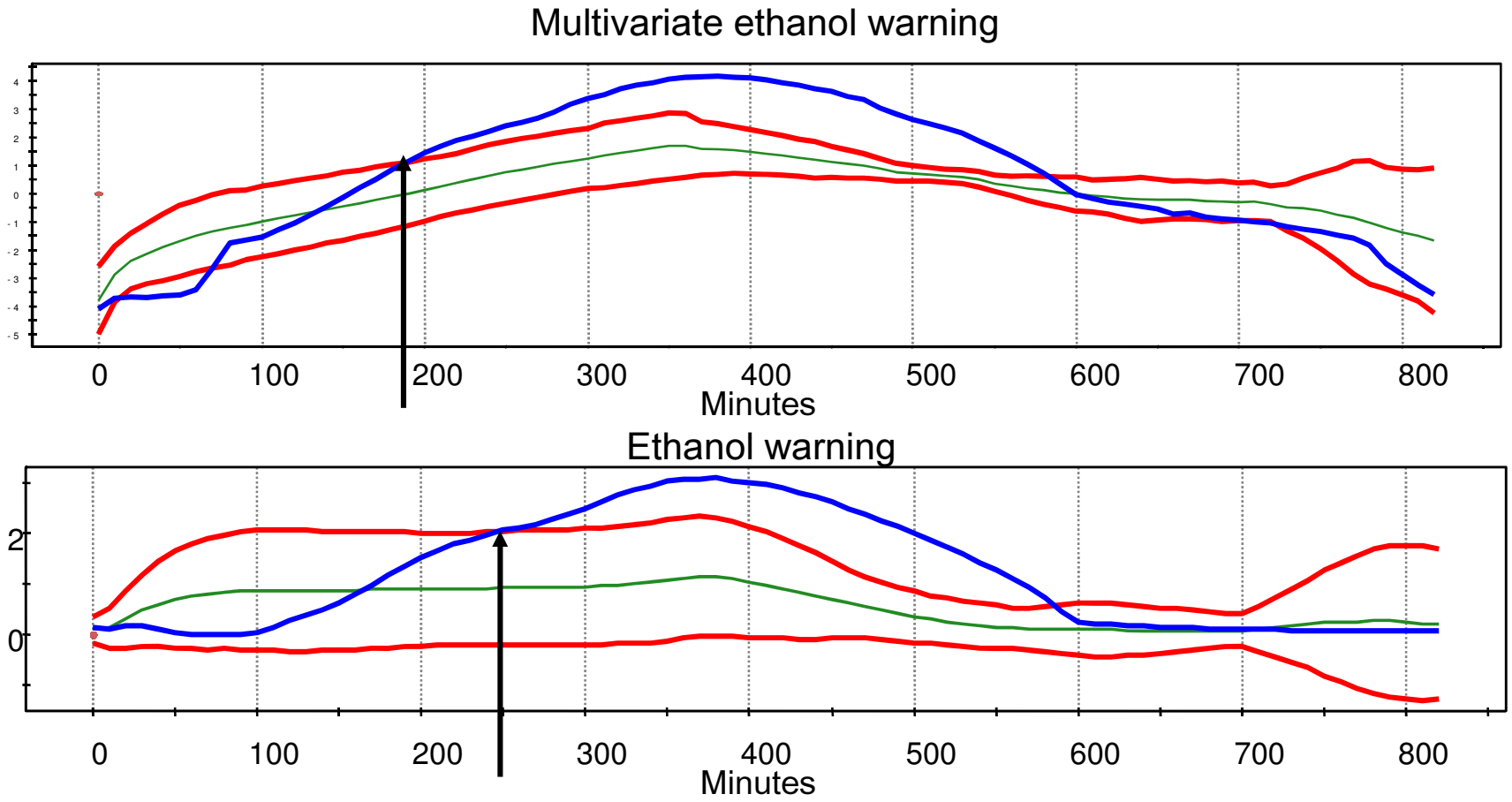
UMETRICS
AN MKS COMPANY

Multivariate Batch modelling

- A batch process is **a finite duration process**
  - Defined START and END
- 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)

UMETRICS
AN MKS COMPANY

Early fault detection



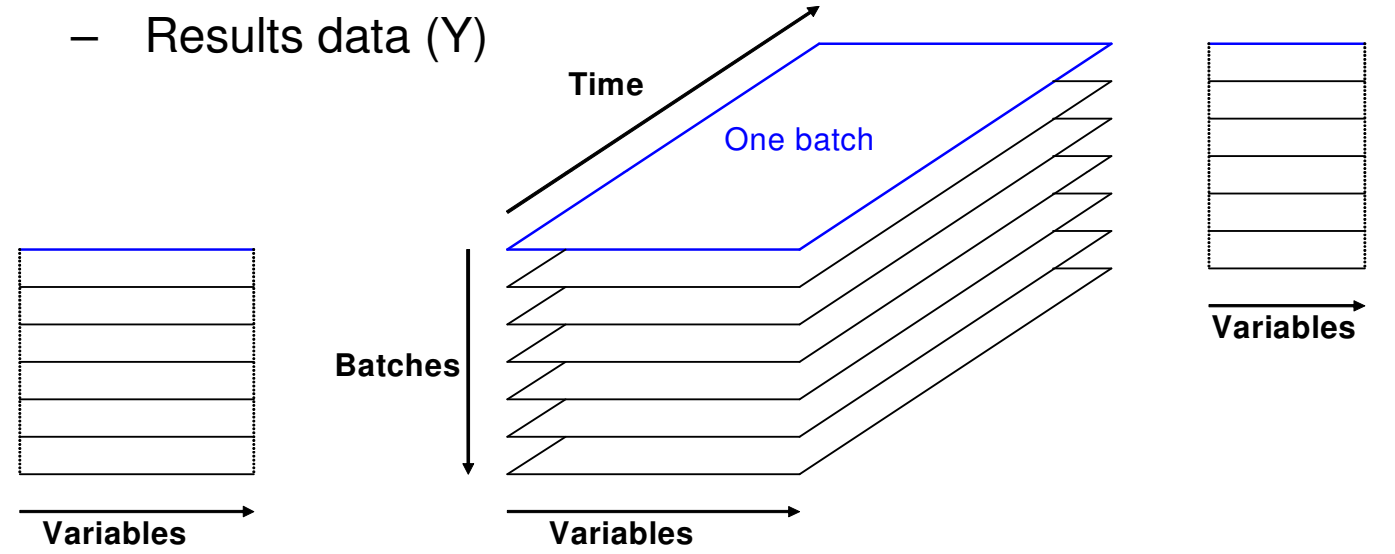
It is important to get an early warning

Multivariate warning occurs 1h before univariate warning

UMETRICS
AN MKS COMPANY

Three blocks of data

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



Initial conditions/
Non dynamic parameters

Z

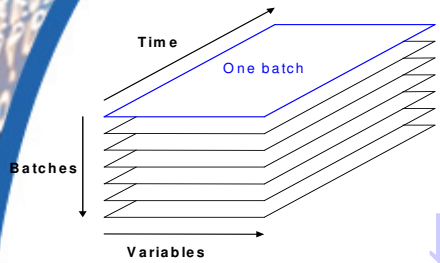
Evolution measurements

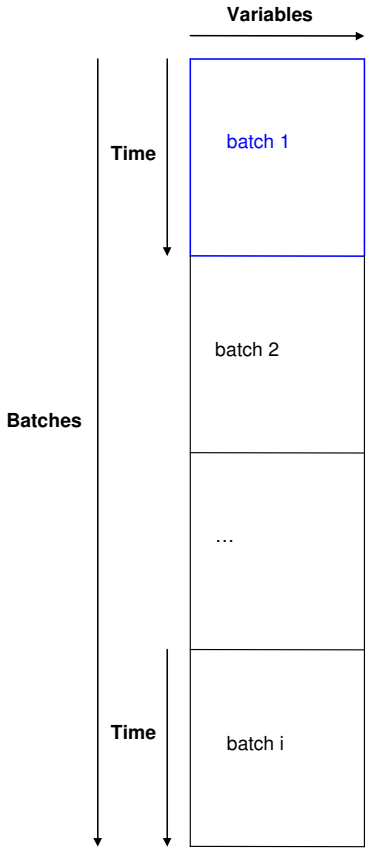
X

Results characteristics

Y

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

Baker's yeast production

- Data come from Jästbolaget AB in Sweden
- The production of the final product took 14 hours
- There were 34 batches, of which 23 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?**

7 variables were monitored

- Ethanol
- Temperature controlled
- Feed of molasses controlled, f(quality of molasses)
- NH3 feed controlled, f(feed of molasses)
- Air flow controlled
- Level in tank
- pH controlled

- Data were sampled every 10 minutes. A batch took 14 hours, resulting in 84 data points per batch

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**

UMETRICS
AN MKS COMPANY

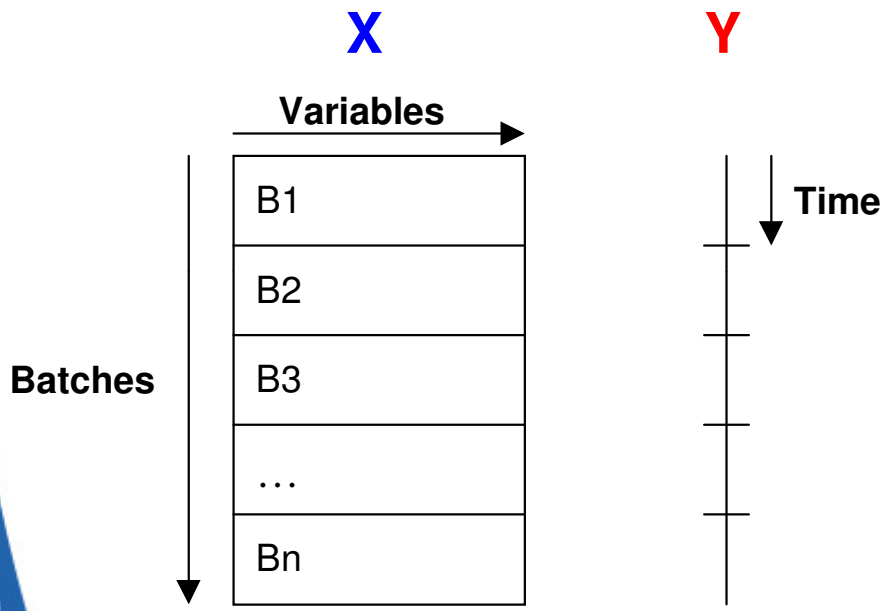
Two levels of batch modelling

- Observation level
  - looks at each individual observation
  - maturity prediction
  - progress monitoring
  - PLS vs Time

- Batch level
  - looks at all available data for the whole batch
  - results prediction
  - PCA: Batch-to-batch variation
  - PLS vs result parameter

UMETRICS
AN MKS COMPANY

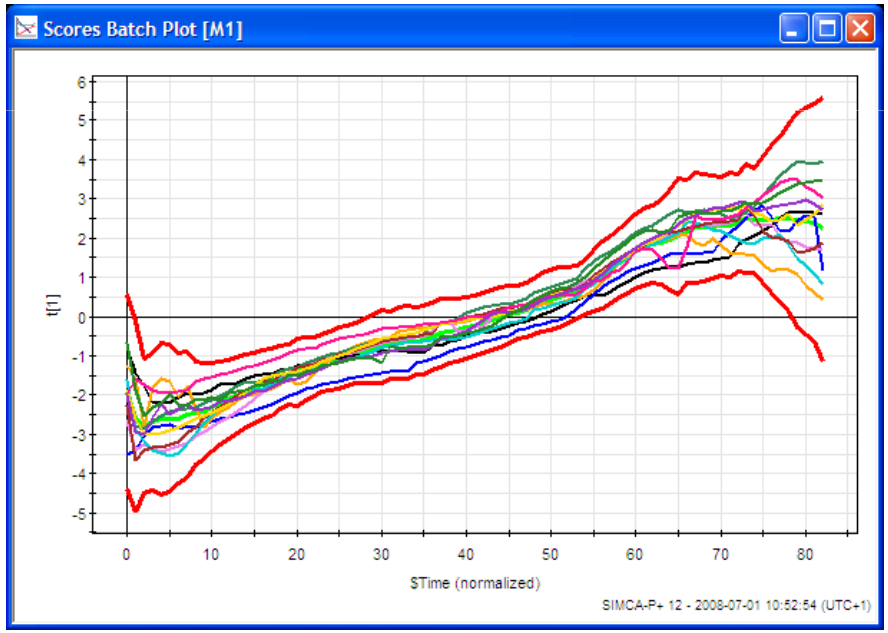
PLS - observation level

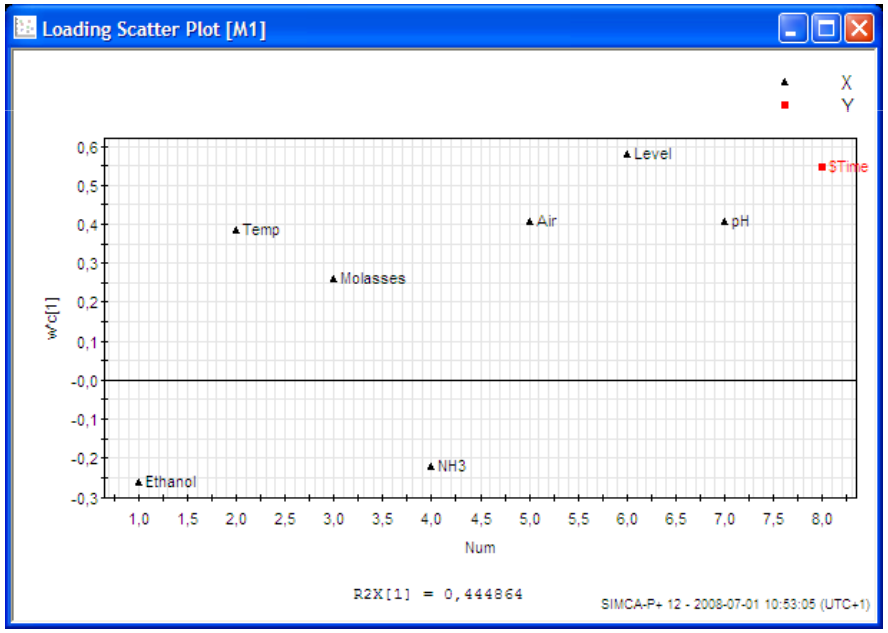


- Each row has the data from a single observation
- The batches follow each other
- Maturity (or time) is used as Y variable
- The resulting scores are new variables that capture
  - t_1: linear relation to Y
  - t_2: quadratic relation to Y
  - t_3: cubic relation to Y

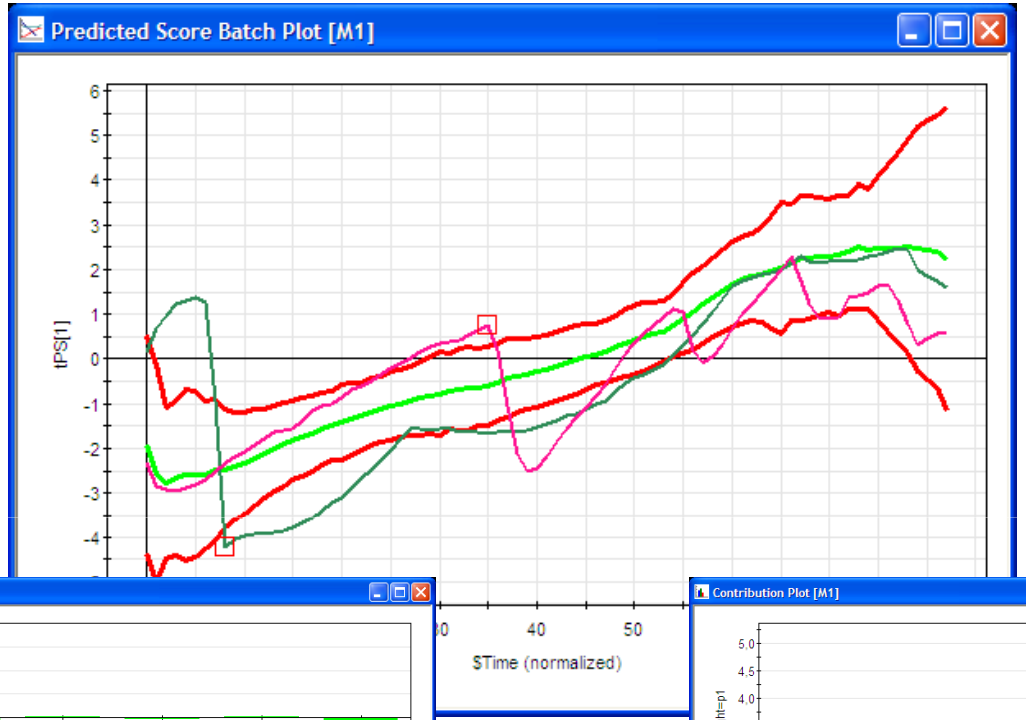
Scores and loadings of observation level PLS model

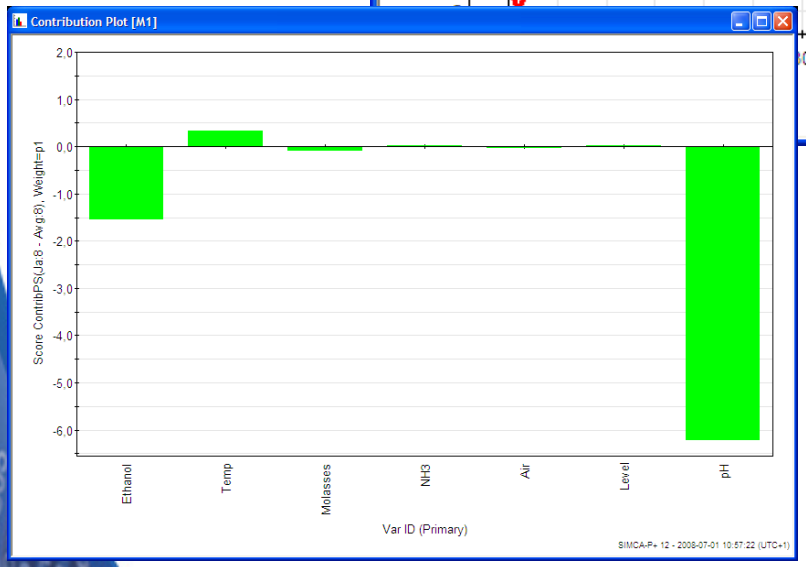
- Local batch time is positively correlated with level in tank, air flow, pH, and temperature. The response variable is little correlated with feed of molasses, ethanol content, and feed of NH_3.

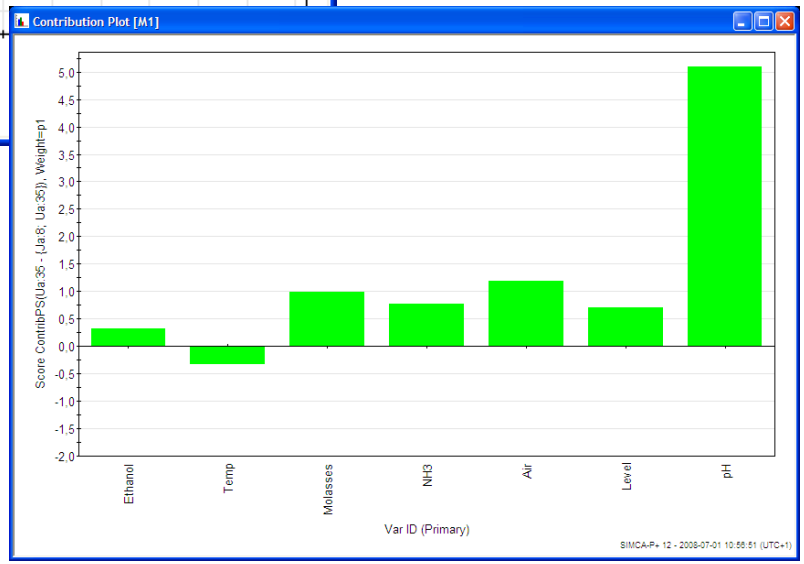




Prediction of new batches, what went wrong?

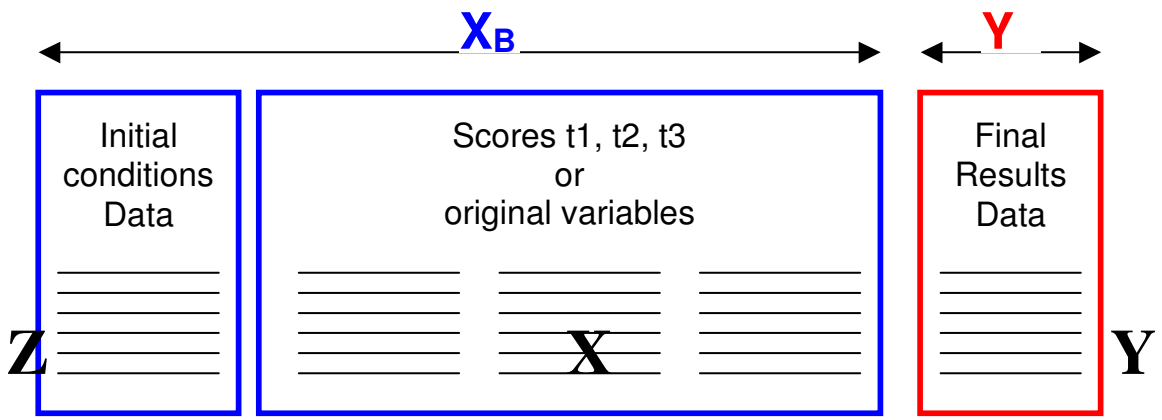






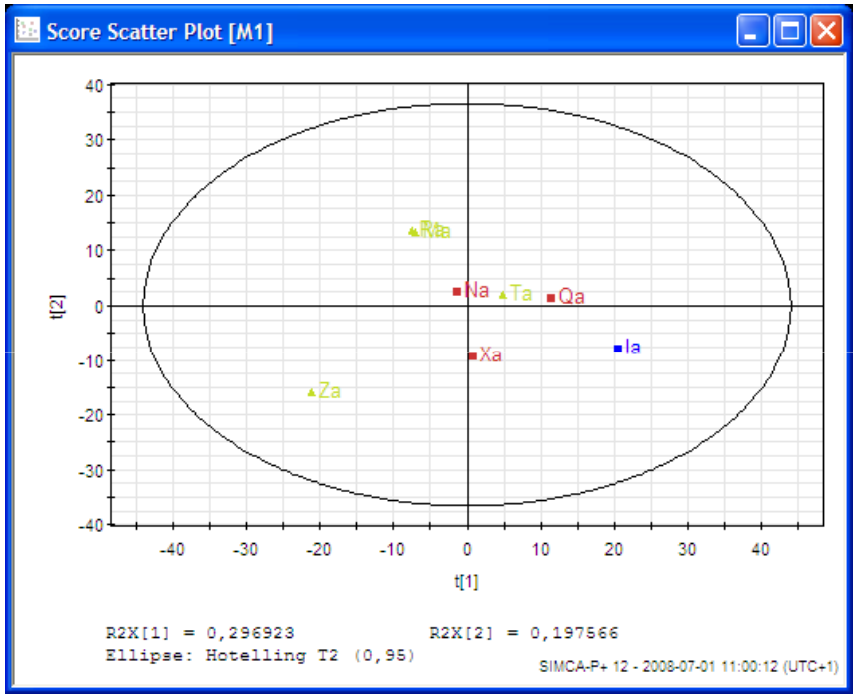
PLS - batch level: Modelling final batch results

- The initial conditions + the unfolded batch data constitute X_B
- The results constitute Y
- Sub models can be made on initial data plus data from time 1, time 1-2, time 1-3, … time 1-T
- The sub models are applied consecutively in the evolution of a batch as new data become available

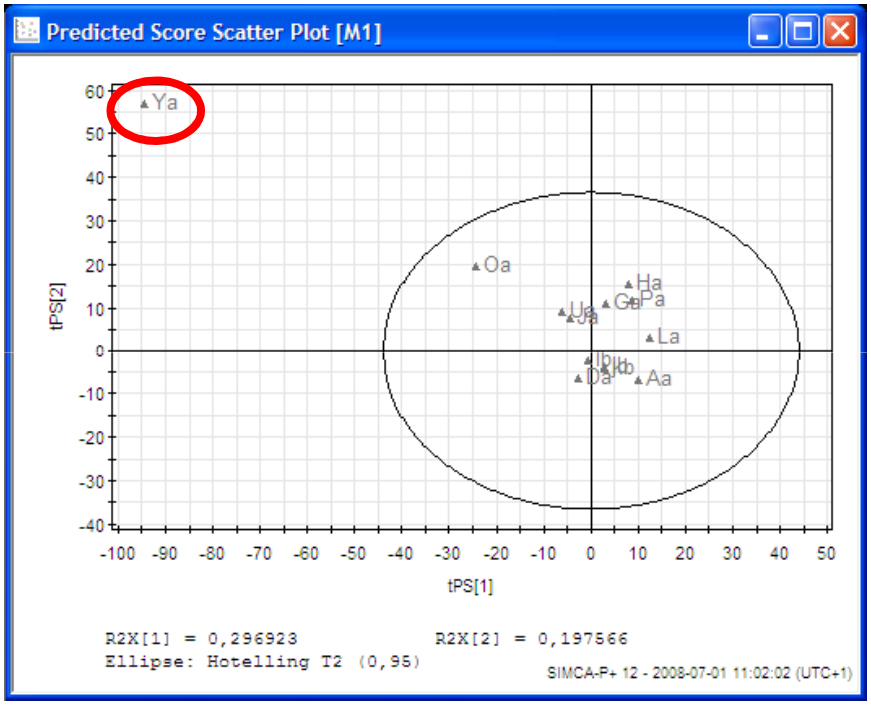


Model interpretation of batch level model

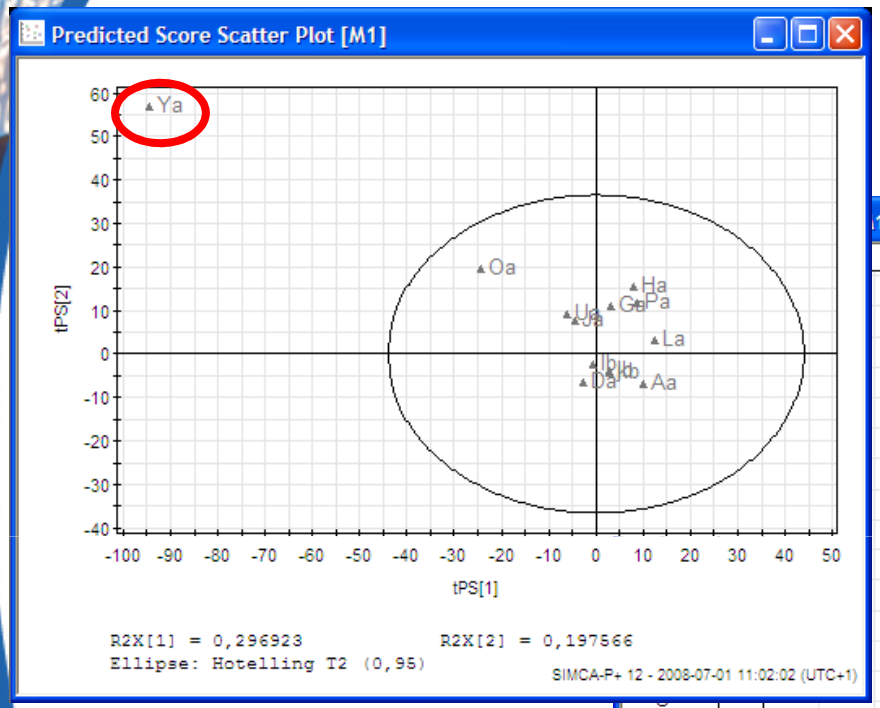
- PCA model ”good” batches

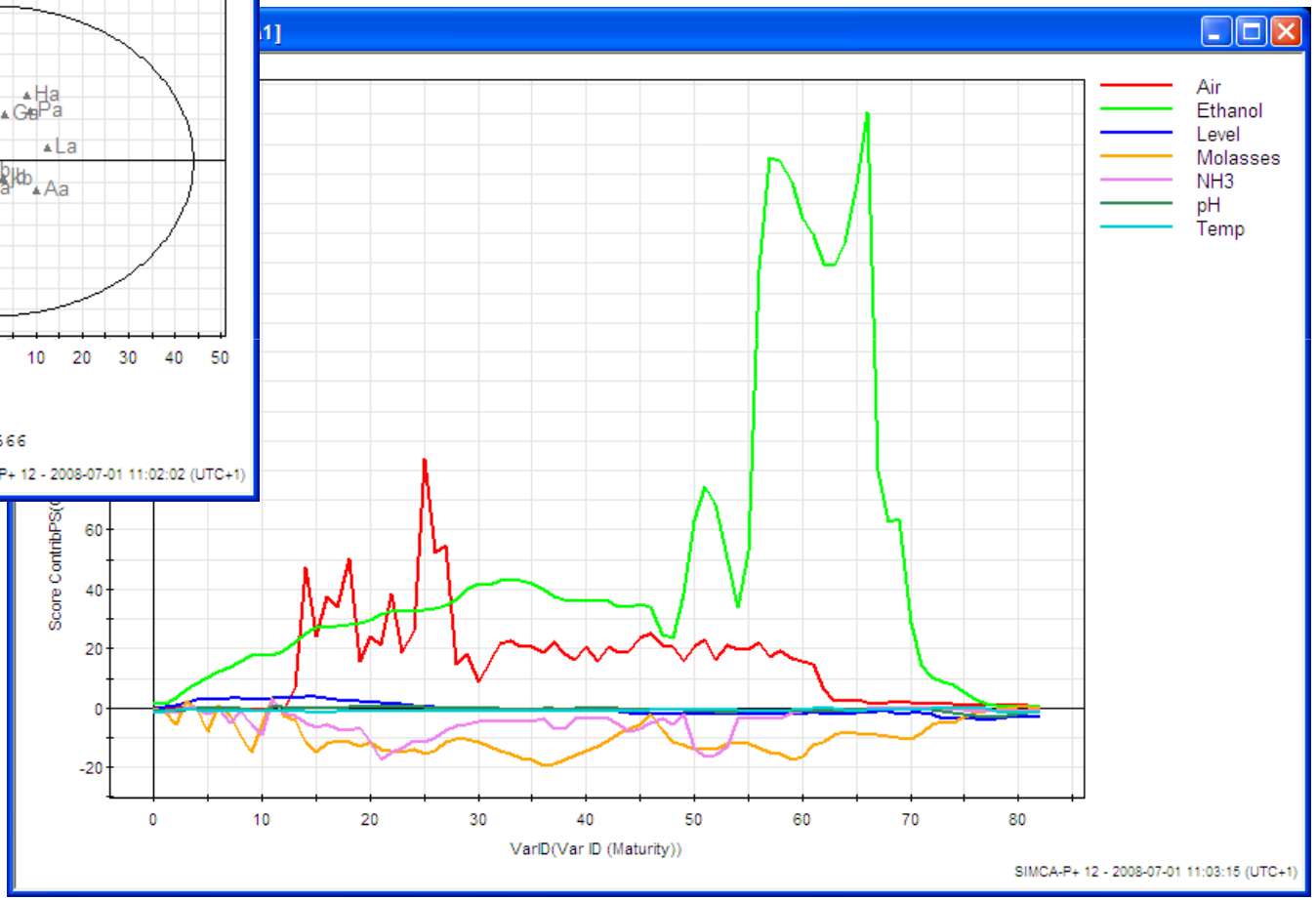


Predicted new batches



Reason for deviation





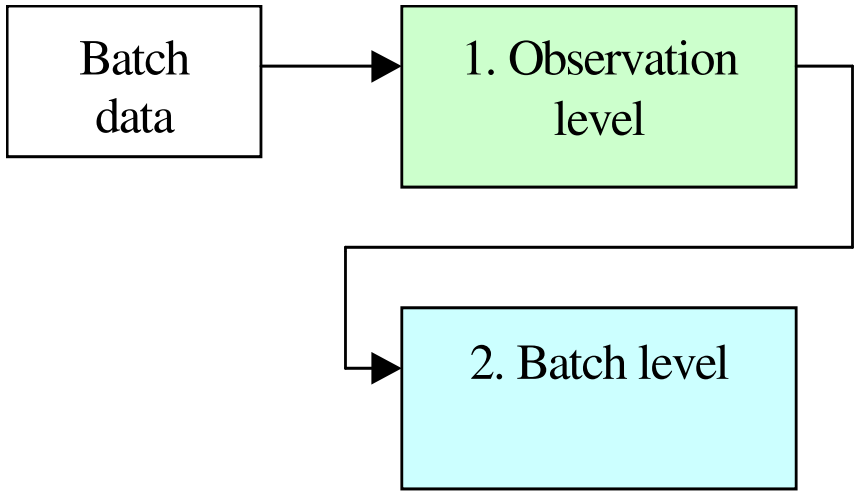
Summary - Batch modeling

- 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

UMETRICS
AN MKS COMPANY

Summary - Batch modelling in practice

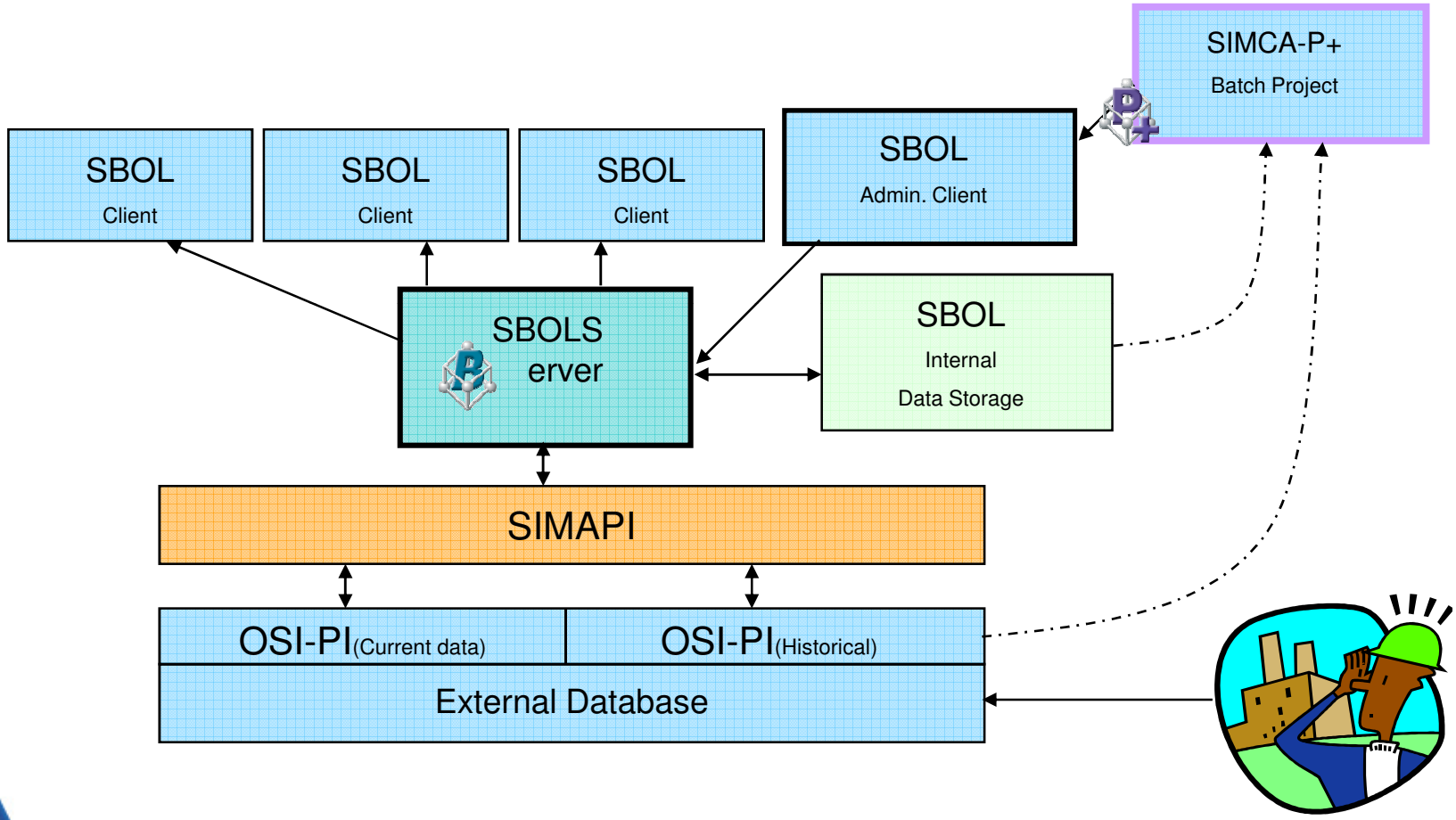
- Modelling and execution are made on two levels
  - **Observation level**
    - working with individual observations
    - monitoring the evolution of the batch
    - classifying current phase
  - **Batch level**
    - working with the whole of the batch
    - predicting the outcome of the batch



UMETRICS AN MKS COMPANY

SBOL set-up and configuration

- Windows based software



On/ In-line monitoring and prediction
SIMCA-Batch On-Line (SBOL)

- On-line Multivariate batch supervision
  - Monitoring
  - Prediction
  - Control
- Complete process supervision
- Process path
- Process cell
- Off-line
- On-Line
- All variables can be used, including spectral data etc