

Biometrics Society meeting, York, UK, March 31, 2006

---

35 Years of Chemometrics,  
from a sidekick to an obsession,  
illustrated by some case stories

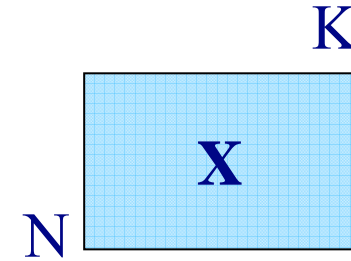
Svante Wold

Chemometrics, Chem.Inst., Umeå Univ., SE \*  
and Umetrics Inc., Kinnelon, NJ, USA

[www.umetrics.com](http://www.umetrics.com)

\* just retired

# Plan



- Some history – a number of random events
- Many variables ( $K$ ) in chromatography and spectroscopy (and their combinations), & gene arrays, ....
  - (a)  $K \gg N$  is possible, and (b) much more information
- A “Minimalist” Approach to Chemometrics
- PCA, SIMCA classification
- PLS, PLS-DA, OPLS
- Experimental design in "latent variables" (scores)
- Illustrated by examples from chemistry/biology, with some thoughts on the relationship between chemistry, chemometrics, and statistics, academia, and industry.

## How did I get into this ?

---

- **1961** Summer job at 1.st computer in home town Alwac III e at Q Chem Dept
- 1963 Grad student in phys org chem – lousy experimentalist
- **1964** Father Herman needed programmer help  
NIPALS programmed for US Jockey Club, Missing Data
- 1966 Moved to Umeå Univ. urged by old professor
- 1969 Final synthesis in org chem, data analysis preferred
- **1970** Homology between Phys Org Chem models (Hammett, ...) & PCA
- 1971 PhD, “Chemometrics” in grant proposal,
- 1972 Gordon RC in Stats in Chem & CE; Splines,
- 1973 SIMCA method
- **1973-74** Visiting scientist at U Wis Madison (GEP Box), met Bruce K
- 1976 Research Professorship in Chemometrics, SwNSRC, .....

## 1970-75 Beginning of Chemometrics (C) Driven by “data explosion”, pros and cons

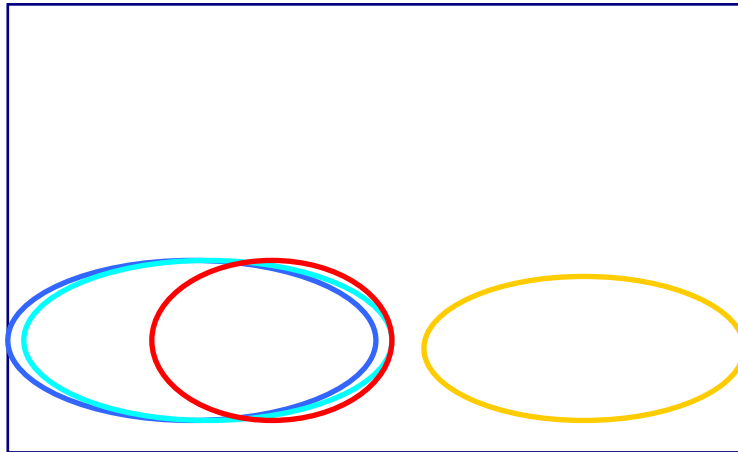
---

- B.R. Kowalski and C.F. Bender (1972-3). Pattern recognition. I. A powerful approach to interpreting chemical data. II. Linear and nonlinear methods for displaying chemical data. J. Amer. Chem. Soc., 94, 5632 – 5639 ; 95, 686 - 693.
- D.L. Massart, C. Janssens, L. Kaufman & R. Smits (1972). "Application of the theory of graphs to the optimisation of chromatographic separation schemes for multicomponent samples". Anal. Chem., 44, 2390-2399
- D.L. Massart & H. De Clercq (1974). "Application of numerical taxonomy to the choice of solvents in TLC". Anal. Chem., 46, 1988-92
- S. Wold (1971), “Chemometrics” – in grant proposal to SwNSRC
- S. Wold (1972-4). Spline-funktioner - ett nytt verktyg i data-analysen.  
Kemisk Tidskrift (3) 34., Technometrics 16, (1974) 1-11
- Chemometrics Society (Kowalski, Wold, et al, June 10, 1974)

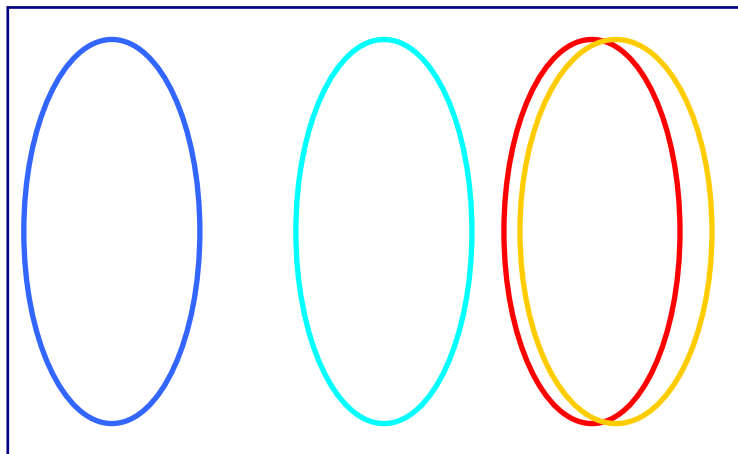
(a) The "scientific approach" is to consider one single variable (factor, ...) at a time -- ***COST***

---

6



3



### Obsidian Artifacts

10 elements (Fe, Ti, Ba, Ca, ..., Y, Zn) analyzed by X-ray fluorescence

4 classes + pred set

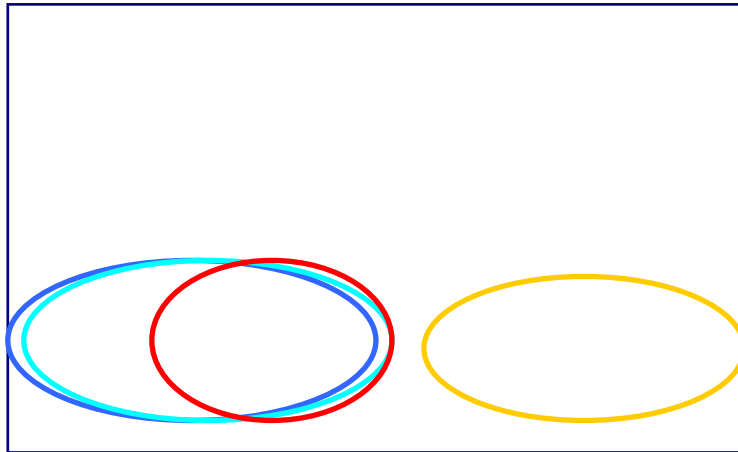
$N = 45 + 29$ ,  $K = 10$

Kowalski et al,  
Anal.Chem. Vol.44  
(1972) 2176.

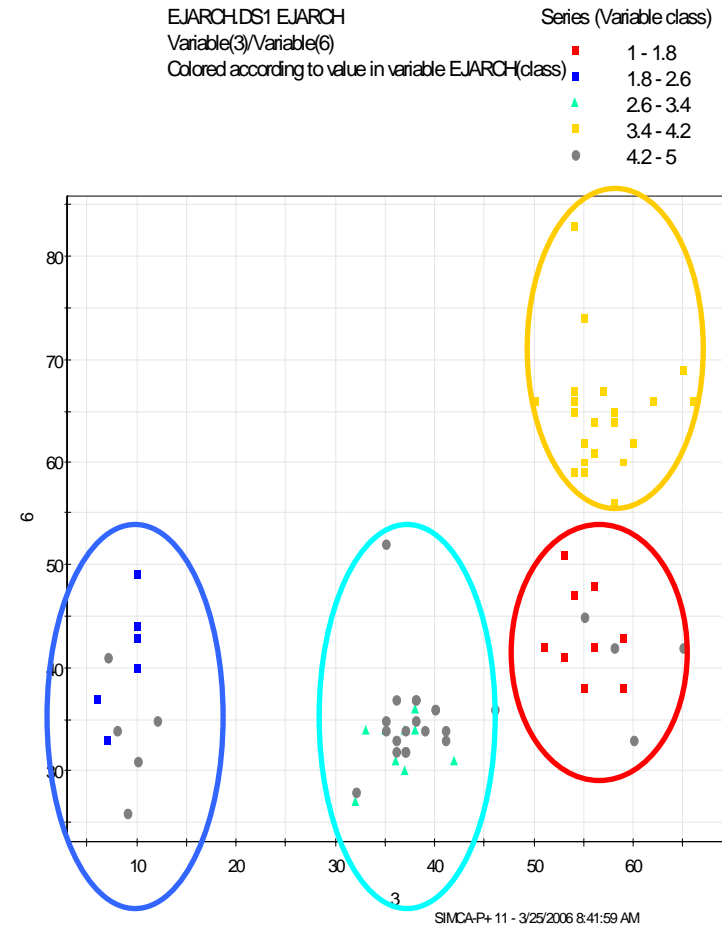
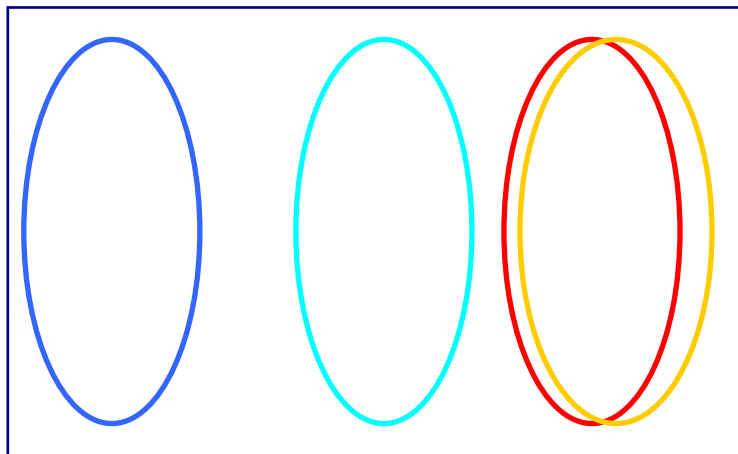
(EJ\_ARCH)

(b) Two variables *together* show much more than “2 x univariate”

6

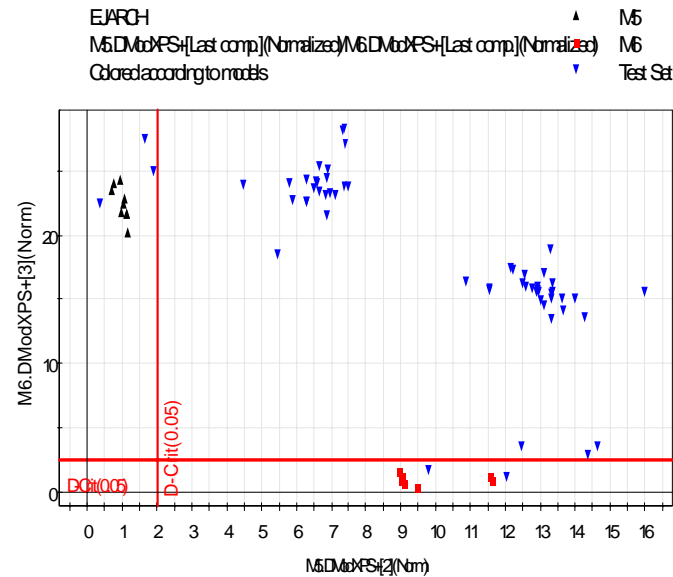


3

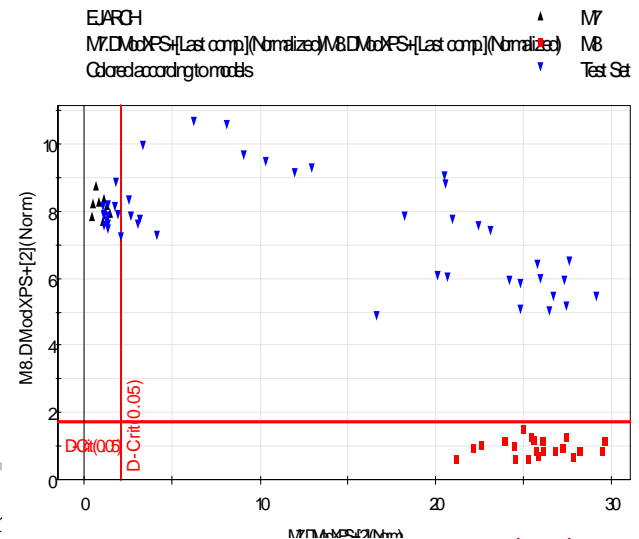


# SIMCA (disjoint PCA), $X_g$ centered & scaled

- A separate PC or PLS model is fitted to each class training set
- New observations classified to (a) the closest class, (b) provided that the probability of class belonging is large enough

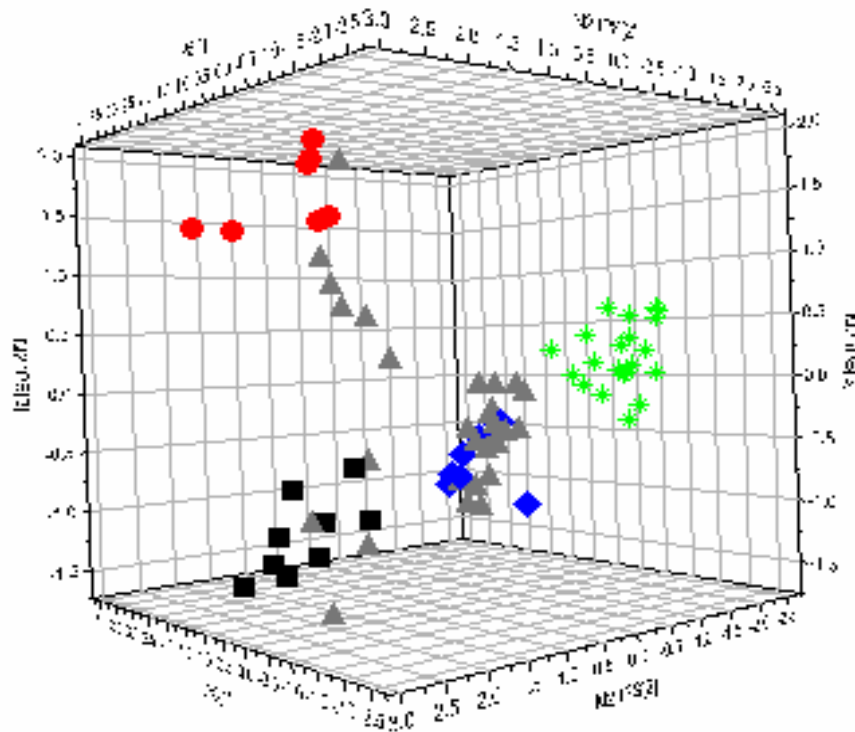


← 1 & 2  
3 & 4  
↓

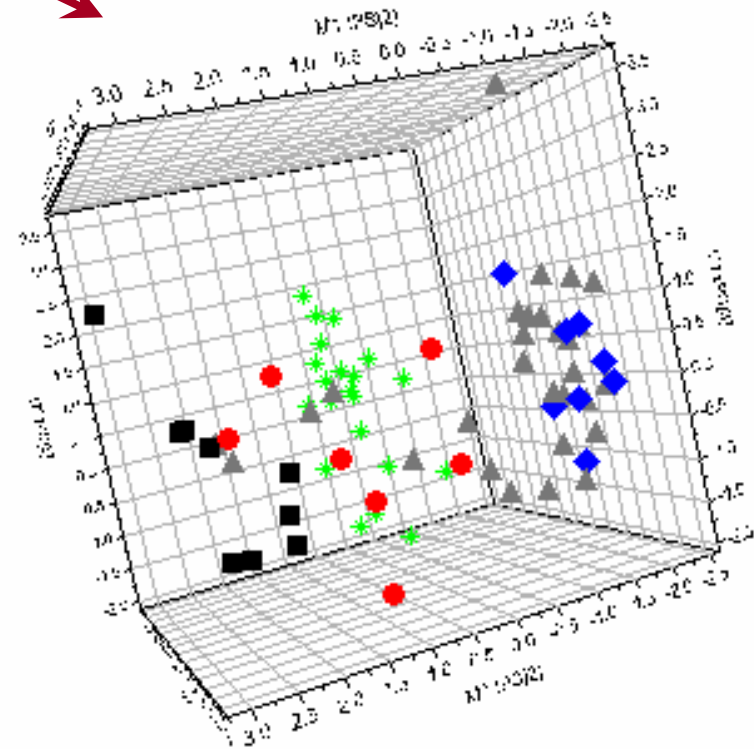


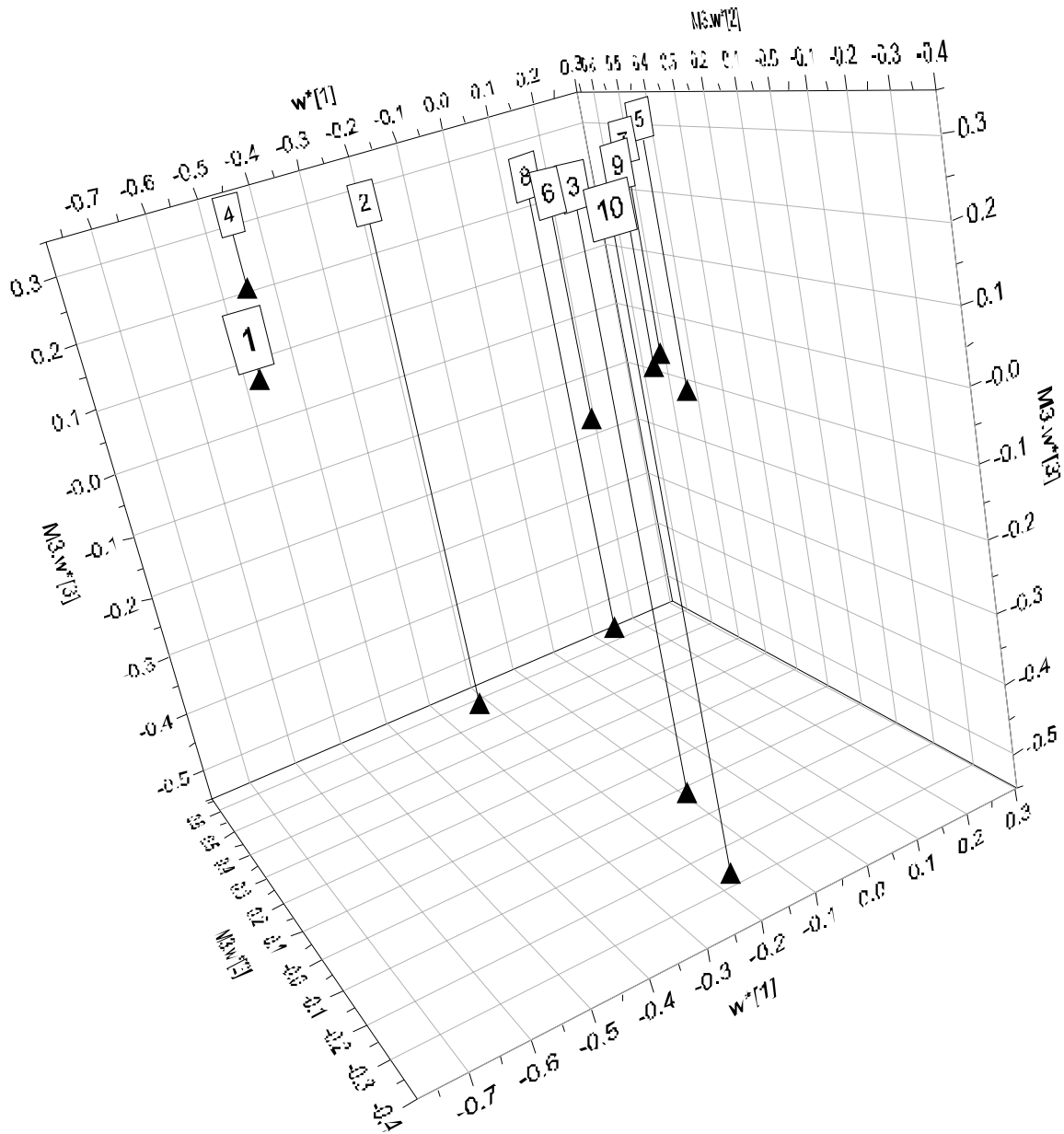
# PLS-DA & PCA, 3D scores scatter plots (3 sign. comp.s)

EJARCH.M2 (PLS-DA), PLS-DA, X=UV, PS-EJARCH  
 tPS[Comp. 1]/tPS[Comp. 2]/tPS[Comp. 3]  
 Colored according to classes in M2



EJARCH.M1 (PCA-X), PCA, X UV scaled, PS-EJARCH  
 tPS[Comp. 1]/tPS[Comp. 2]/tPS[Comp. 3]  
 Colored according to classes in M1





- 1 = Fe**
- 2 = Ti**
- 3 = Ba**
- 4 = Ca**
- 5 = K**
- 6 = Mn**
- 7 = Rb**
- 8 = Sr**
- 9 = Y**
- 10 = Zn**

Two variables show much more than one

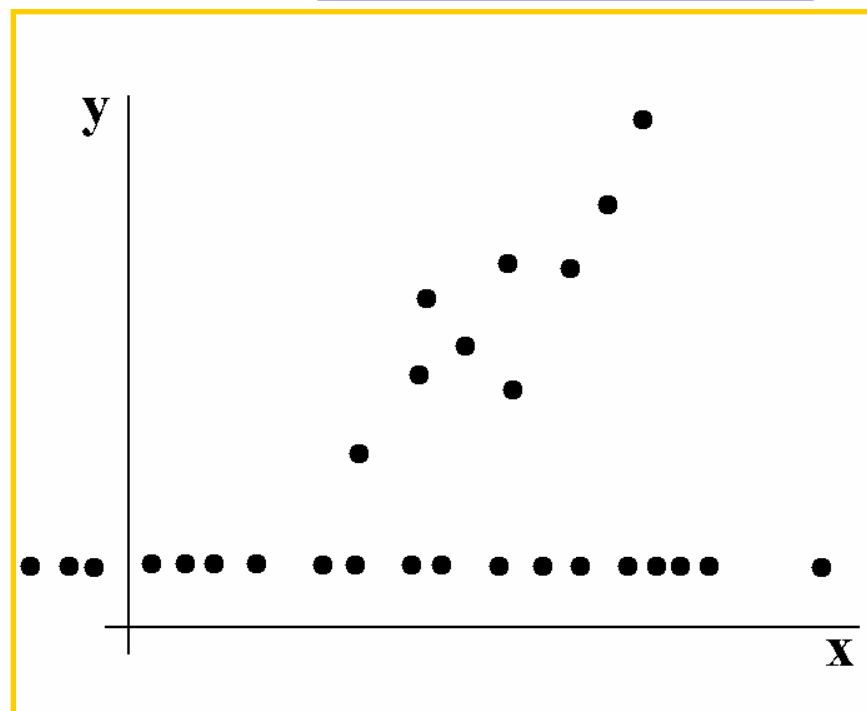
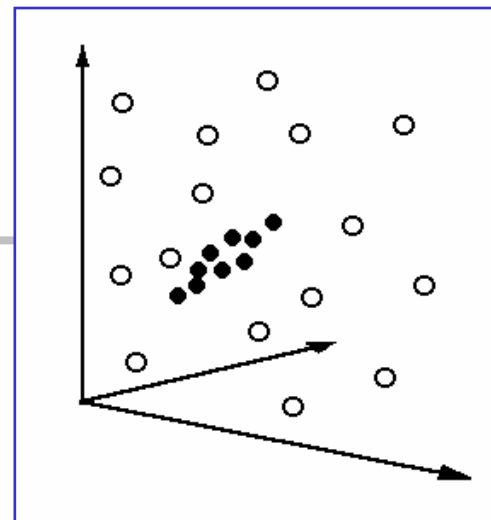
K variables show more than 2, & results are more stable

---

This still surprises us chemists

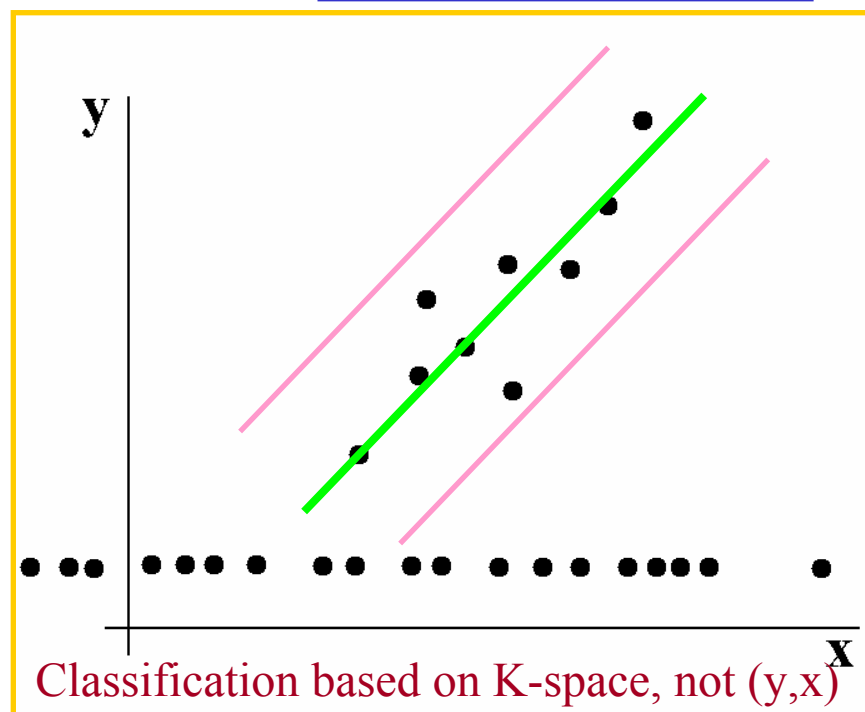
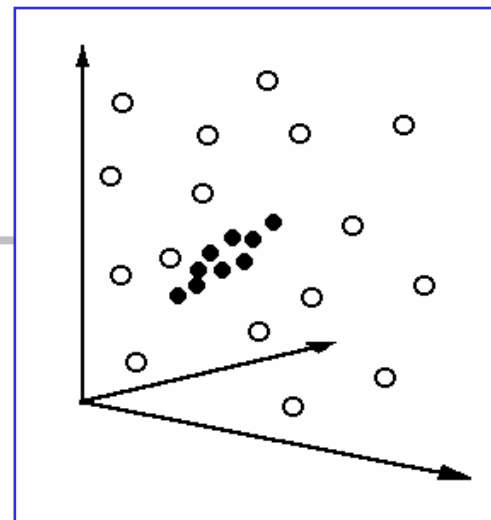
## Any modelling is a combination of classification & other modelling

- Asymmetric classification = MSPC
- Asymmetric QSAR – inactives of two types (Nordén et al., QSAR 1983)
  - Those close to the model and with low activity
  - Those far from the model (inactive = dissimilar)
- Same in process modelling & soft sensors
- Process modelling; often = classification (grades) + separate PLS models for each class (ex whisky)



## Any modelling is a combination of classification & other modelling

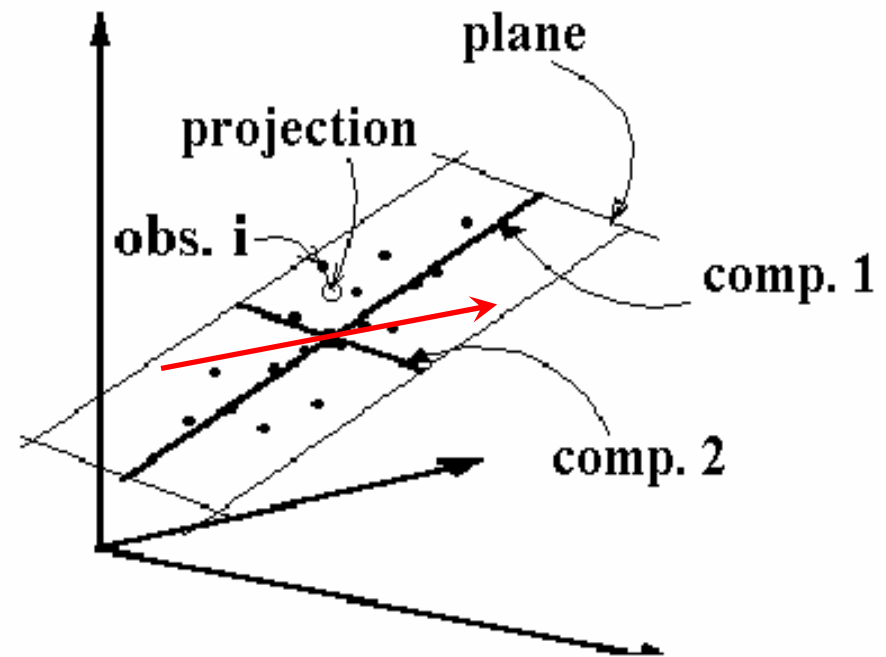
- Asymmetric classification = MSPC
- Asymmetric QSAR – inactives of two types (Nordén et al., QSAR 1983)
  - Those close to the model and with low activity
  - Those far from the model (inactive = dissimilar)
- Same in process modelling & soft sensors
- Process modelling; often = classification (grades) + separate PLS models for each class (ex whisky)



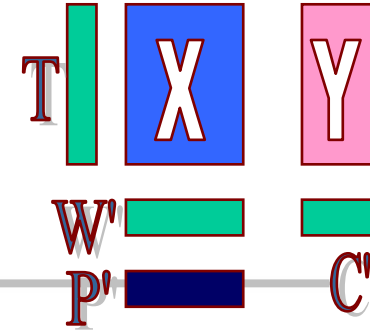
# PCA → PLS (H.Wold, et al., 1975-1982)

## Idea (starting from PCA)

- Similar objects close in K-space
- Monotonous relation to important “properties”, Y (PC regression, PCR)
- Problem with PCR: y not used in X-model development
- PLS: develop scores (**t**) like PCA, but “weights” (**w**) based on covariance between X and y  
⇒ “Regression” also with  $K \gg N$



## Herman Wold, $\approx 1975 - 1981$



- In PCA, PLS, (SVD,, ...):  $\mathbf{X} = \mathbf{T} \mathbf{P}' + \mathbf{E}$   
any model parameter, e.g.,  $t_{ia}$ , is a normalized vector product =  
= least squares slope = weighted average of K elements

$$t_{ia} = \mathbf{x}_i' \mathbf{p}_a / (\mathbf{p}_a' \mathbf{p}_a) \quad (\text{PCA}); \quad t_{ia} = \mathbf{x}_i' \mathbf{w}_a / (\mathbf{w}_a' \mathbf{w}_a) \quad (\text{PLS});$$

- scores,  $\mathbf{t}_a$ , are more precise the larger is K
- elements in  $\mathbf{t}_a$  are close to normally distributed, closer the larger is K (everything else equal)

**But** B. Nadler and R.C. Coifman. The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration.

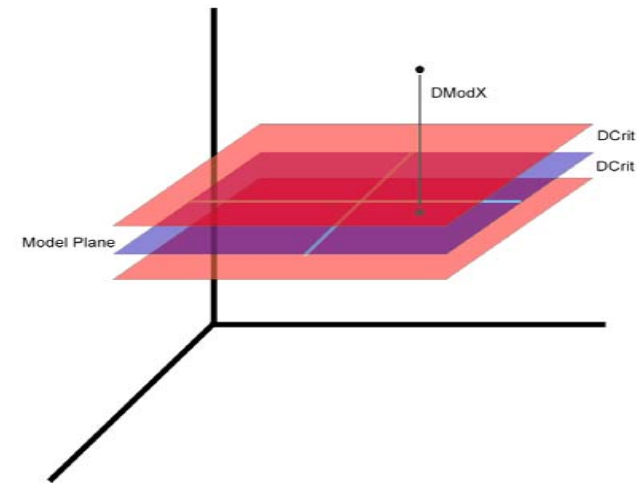
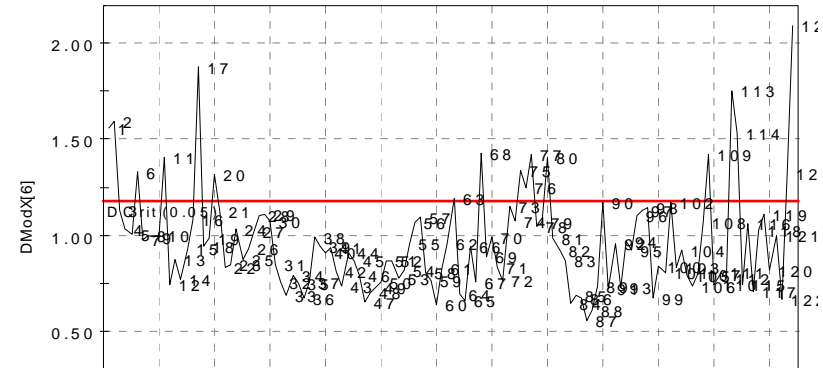
J.Chemom. 19 (2005) 107-118

Pred-error contains a term proportional to  $s^2 K^2 / n^2$  !?

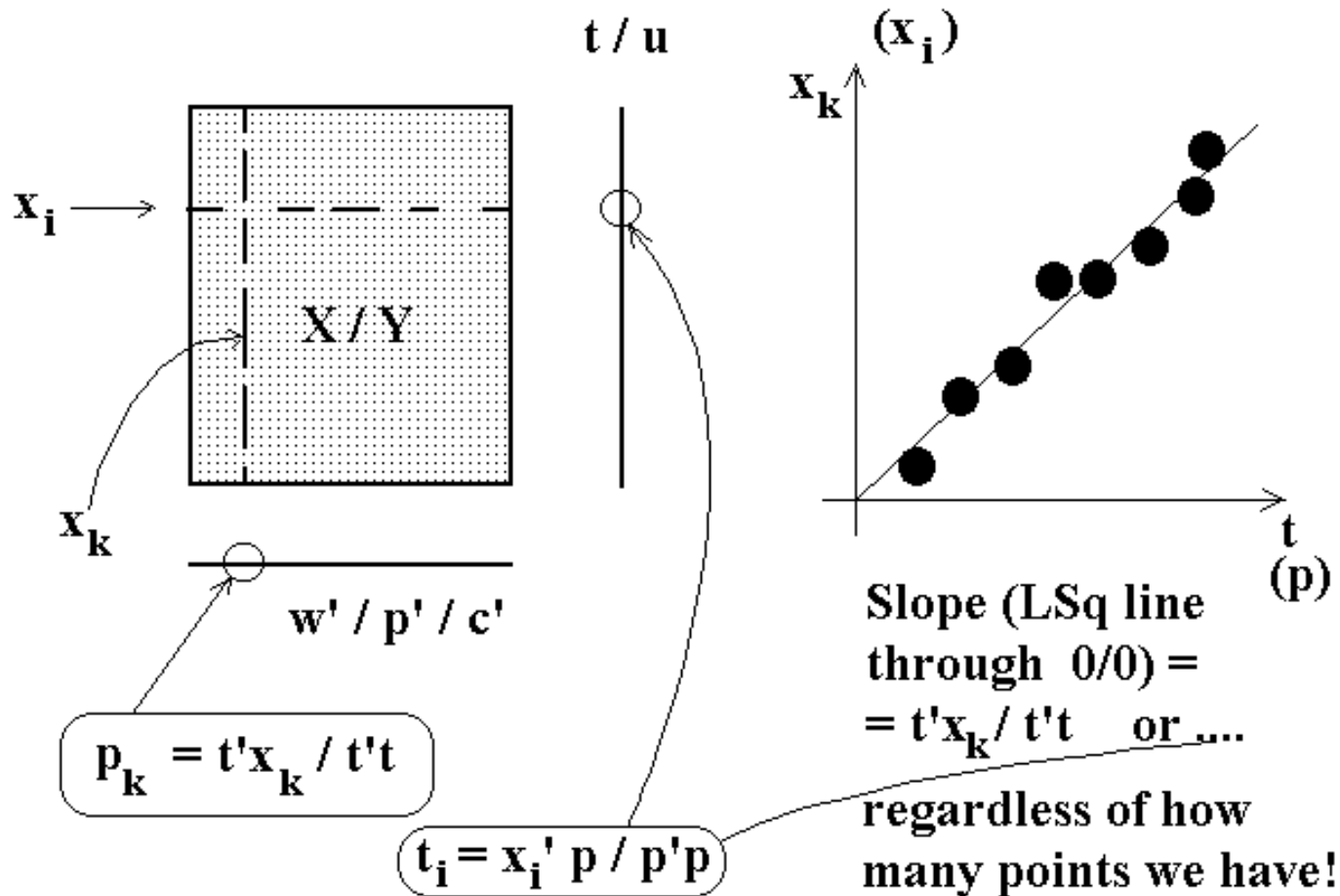
# Residuals of observations (row-wise).

Same for PLS as for PCA, but two spaces, X and Y

- X-residuals,  $E = X - TP' = [e_{ik}]$   
row SD =  $DModX_i$   
column criterion  $R_k^2$
- Y-residuals,  $F = Y - TC' = [f_{im}]$   
row SD =  $DModY_i$   
column criterion  $R_m^2$
- Critical values of  $DModX$  and  $DModY$  from F-distributions



NIPALS = sequence of vector-matrix multiplications  
 = partial least squares steps (PCA, PLS, ...) = EM per comp.



# PLS-discriminant analysis (PLS-DA) is analogous to linear discriminant analysis (LDA)

## *Classes tight*

MLR  $\Rightarrow$  LDA

Fisher, 1936-38.

PLS  $\Rightarrow$  PLS-DA

Wold et al., 1986

Also singular  $\mathbf{X}'\mathbf{X}$

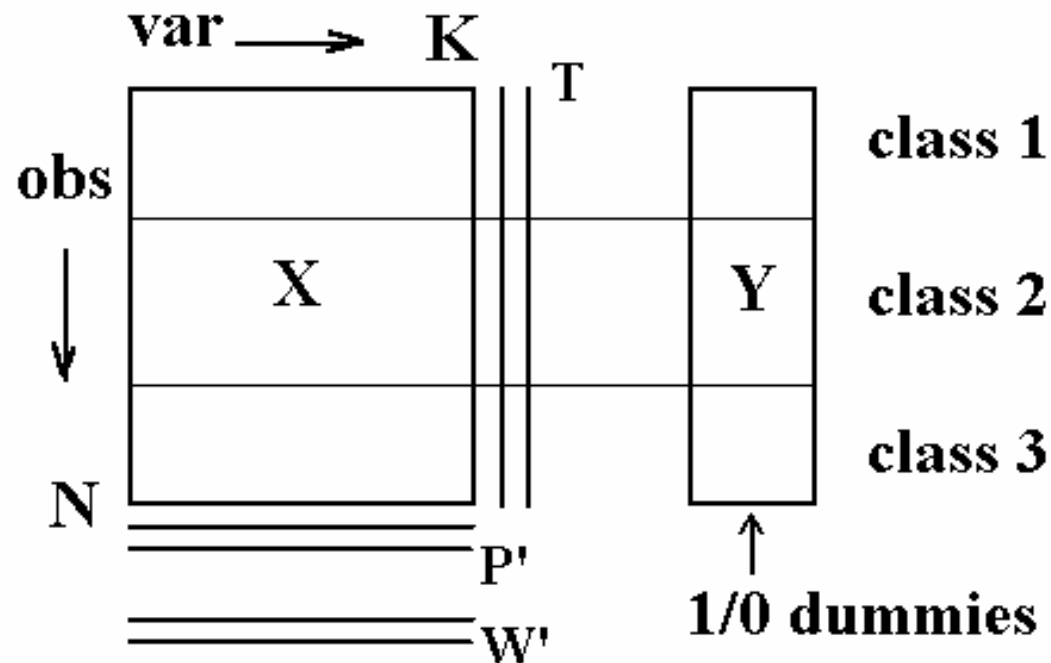
e.g.,  $K \gg N$

M Barker and W

Rayens

J. Chemom. 17 (2003)

166 - 173



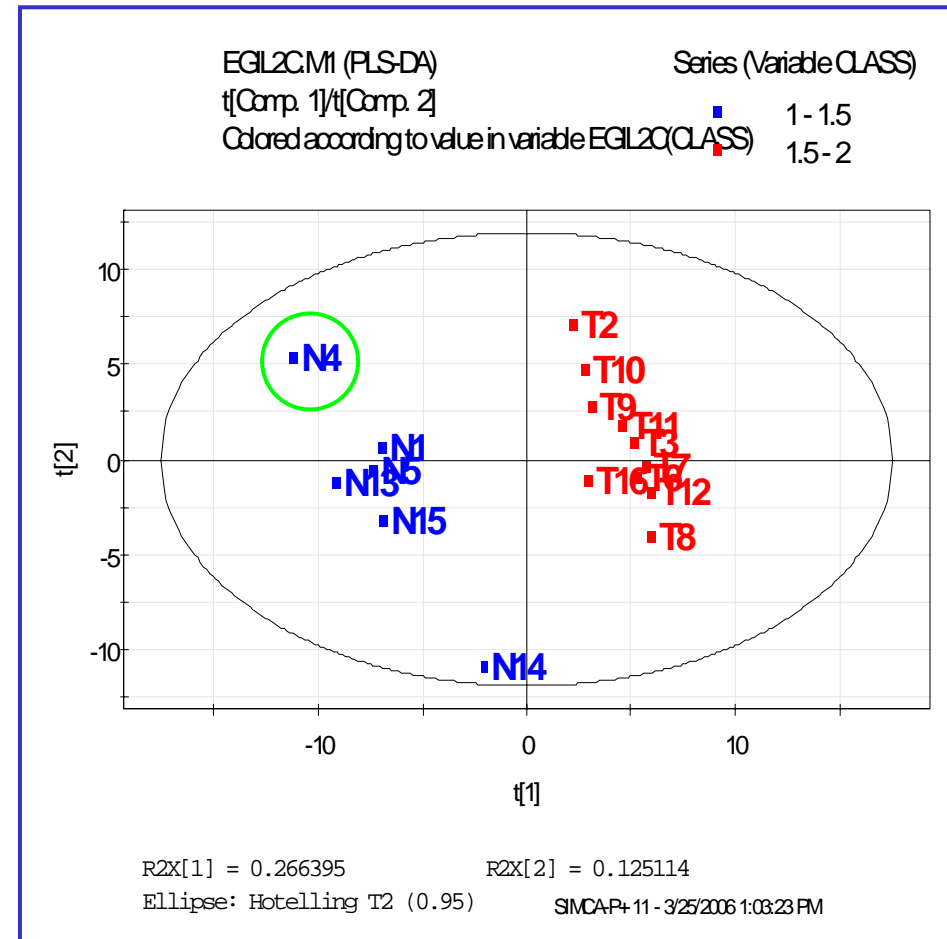
Plot of X-scores (T) shows class separation

PLS-weights ( $w_{ak}$ )  $\leftrightarrow$  discrimination power

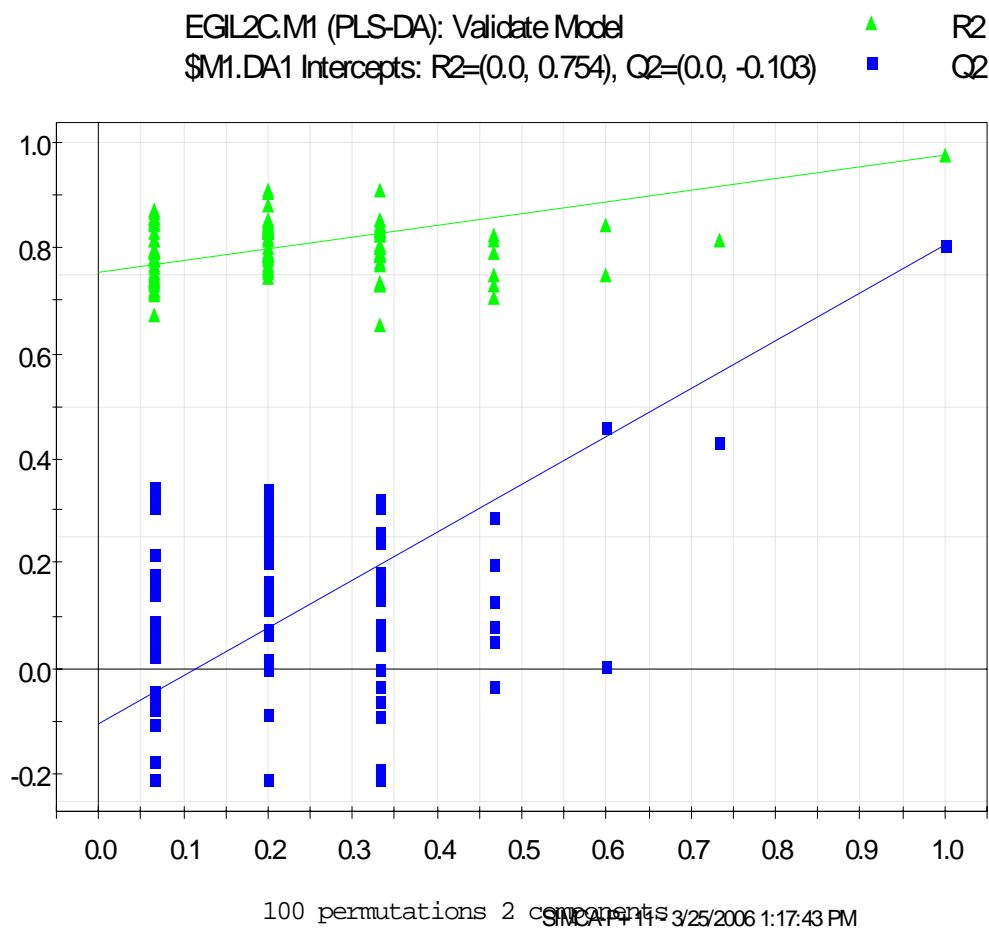
# $K \gg N$ is possible -- and good ! (ex tumors)

- Two classes of samples – tumor and normal brain tissue
- Characterized by GC (capillary)
- $N = 16, K = 156$
- E.Jellum, I.Bjornson, R.Nesbakken, E.Johansson and S.Wold (1981). Classification of human cancer cells by means of capillary gas chromatography and pattern recognition analysis. J.Chromatogr. 217, 231 - 237.

## 1.st MVA with $K \gg N$ ??



# Validation is essential, because the chemometrics models are often used for predictions & decisions

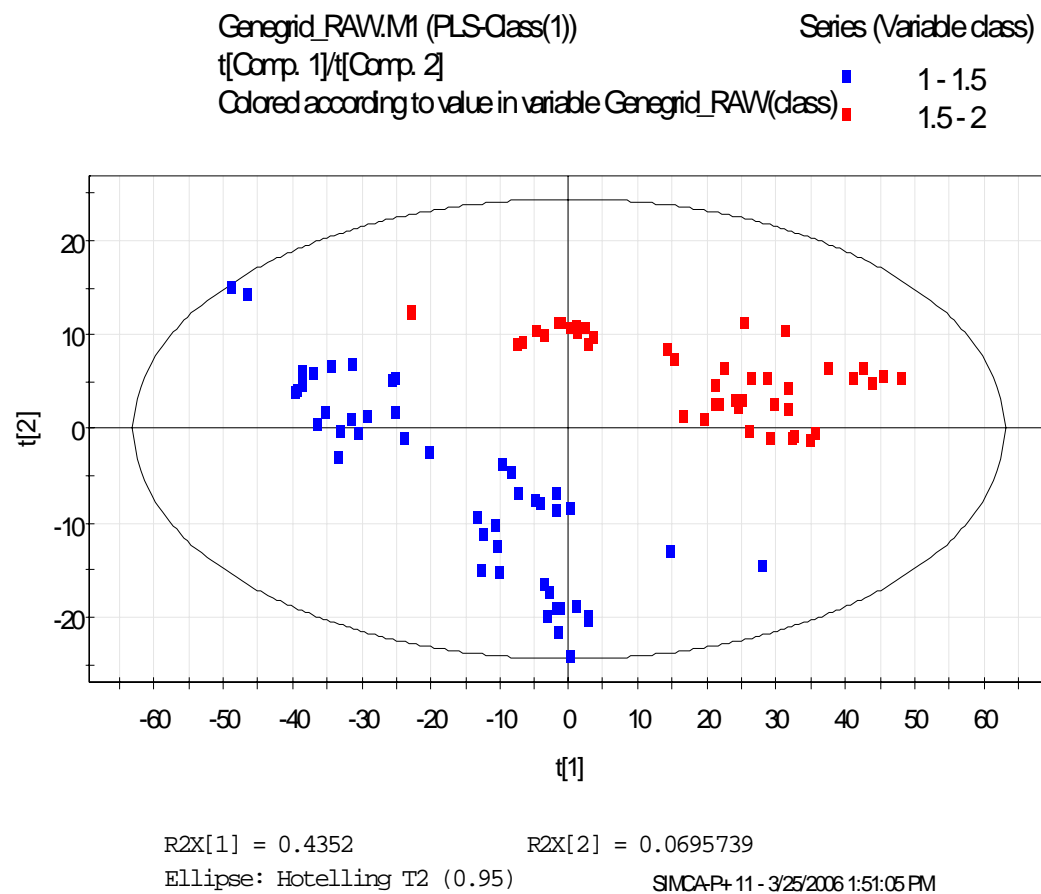


## Validation approaches

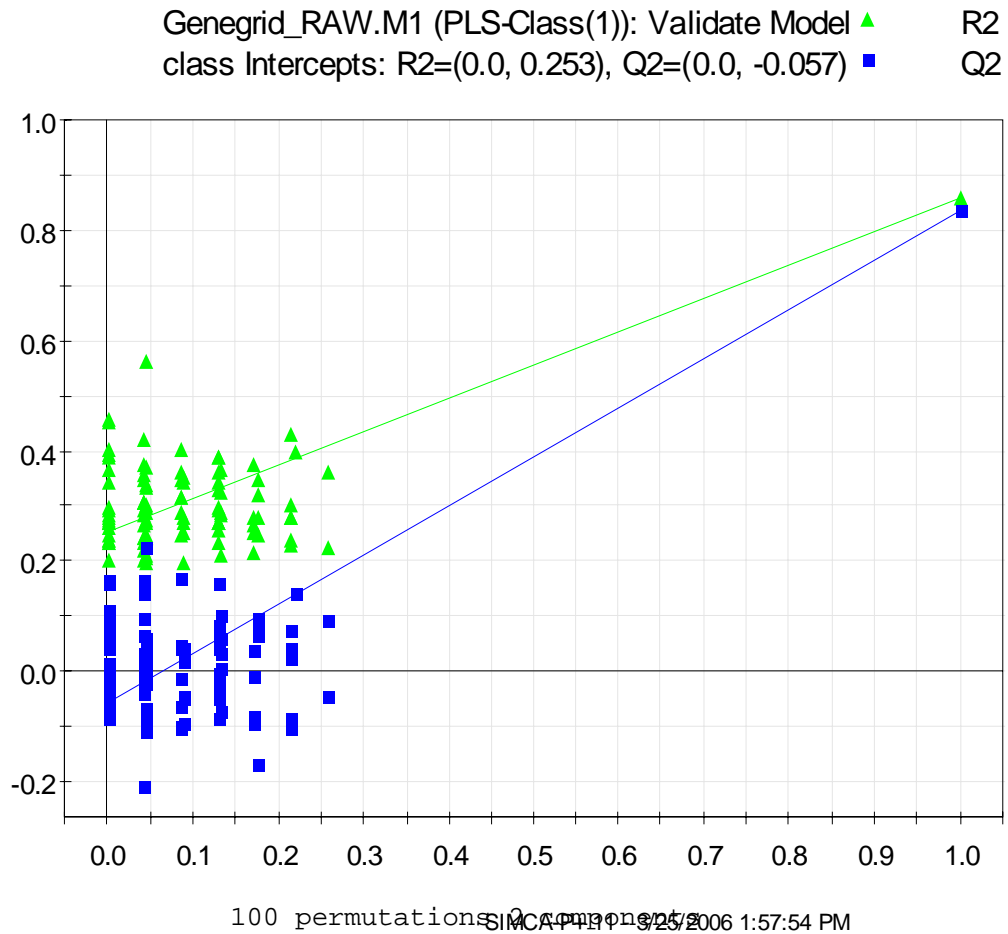
- Cross-validation
- Permutation tests
- Understanding
- Representative Prediction Sets

# And some data with very many variables (mega-variate), PLS-DA

- Gene arrays
- Controls
- Low exposure
- $N = 92$ ,  $K = 1611$
  
- $A = 3$
- $R^2 = 0.86$
- $Q^2 = 0.68$

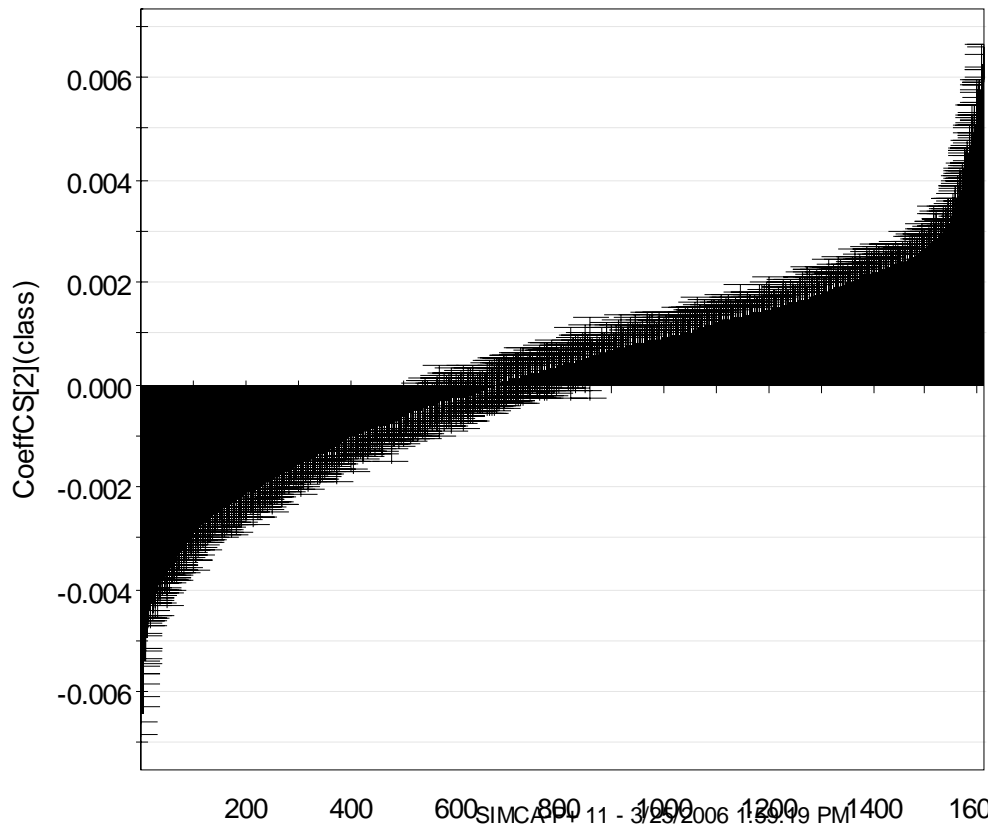


# Validation, 100 random permutations of y



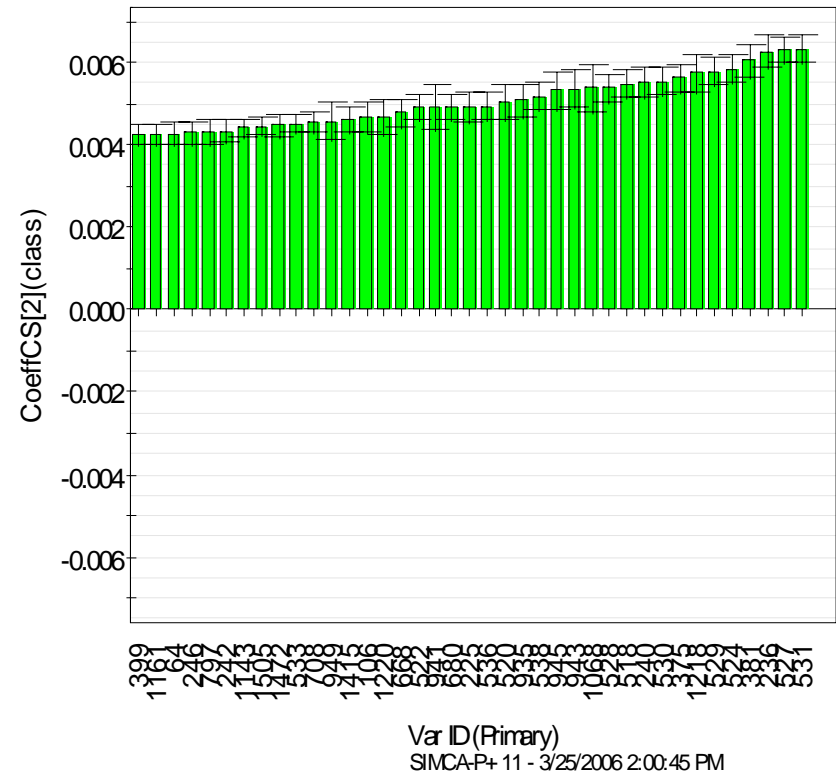
# Coefficients, sorted, with confidence intervals (JK)

Genegrid\_RAW.M2 (OPLS)  
CoeffCS[Last comp.](class)

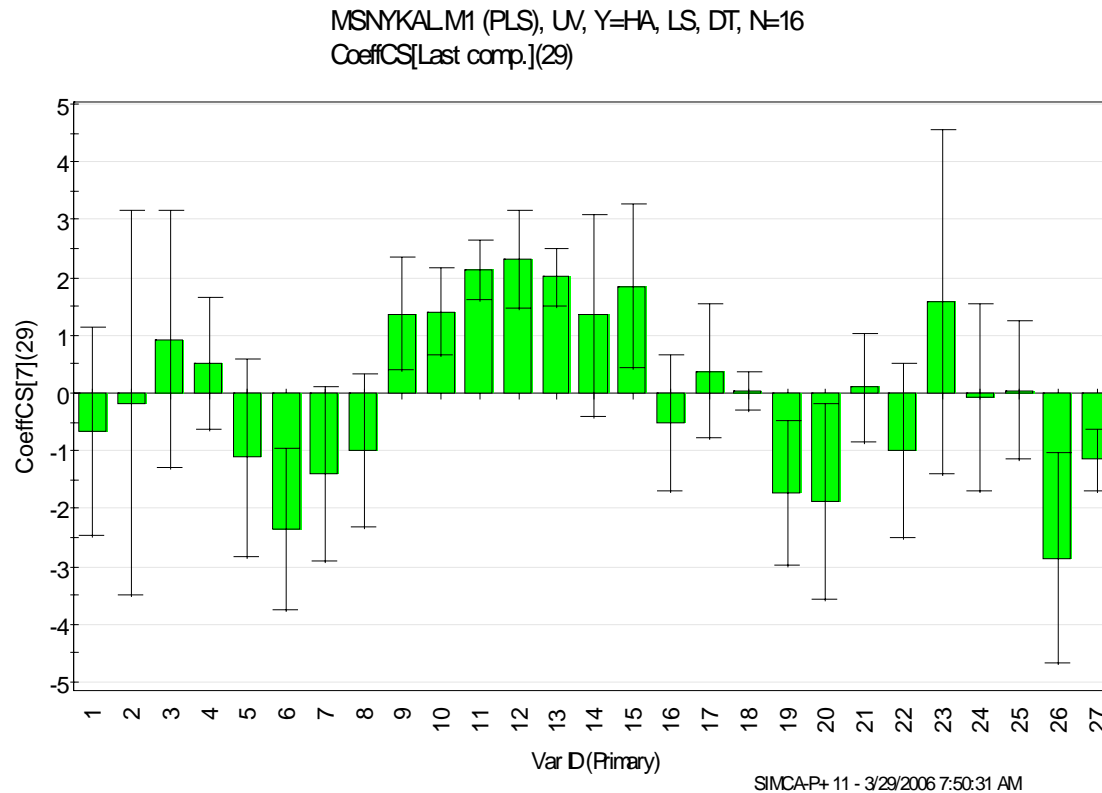


Zoomed on largest (right)

Genegrid\_RAW.M2 (OPLS)  
CoeffCS[Last comp.](class)



# PLS regression coefficients, B (one set per y-variable)



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

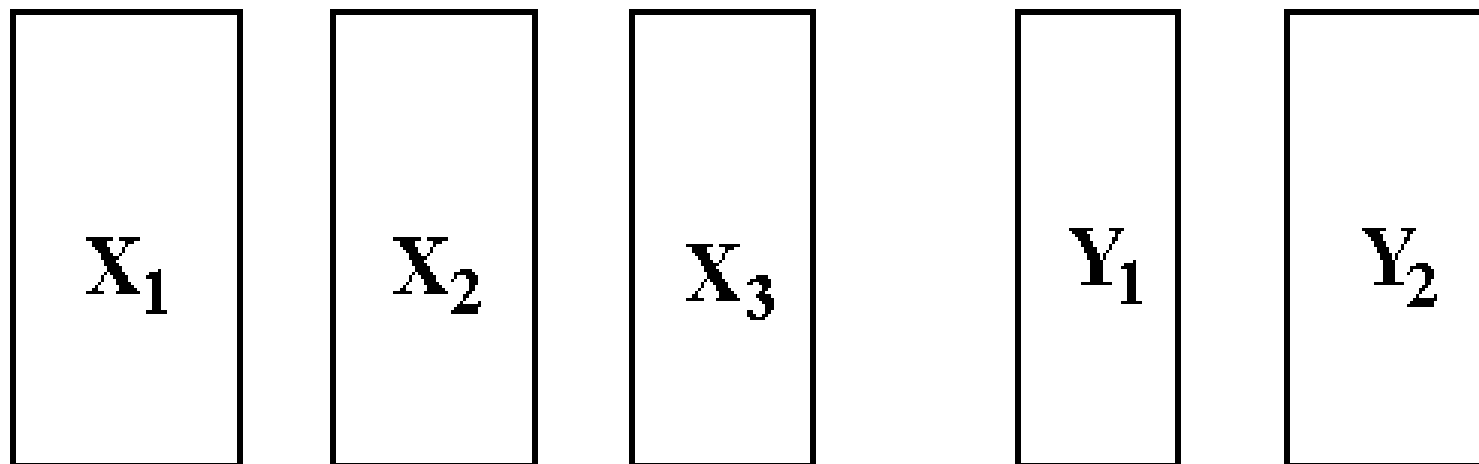
remember:  
these coefficients  
are usually **NOT**  
**independent.**

Compare mixtures

Confidence intervals (interpretation !) from Jack-knifing.

## Dividing into blocks $\Leftarrow$ objectives, knowledge

---



### Process:

$X_1$  = Raw mtrl,  $X_2$  = Cleaning,  $X_3$  = Reactor,  $X_4$  = Work Up,  
 $Y_1$  = Product quality,  $Y_2$  = Costs

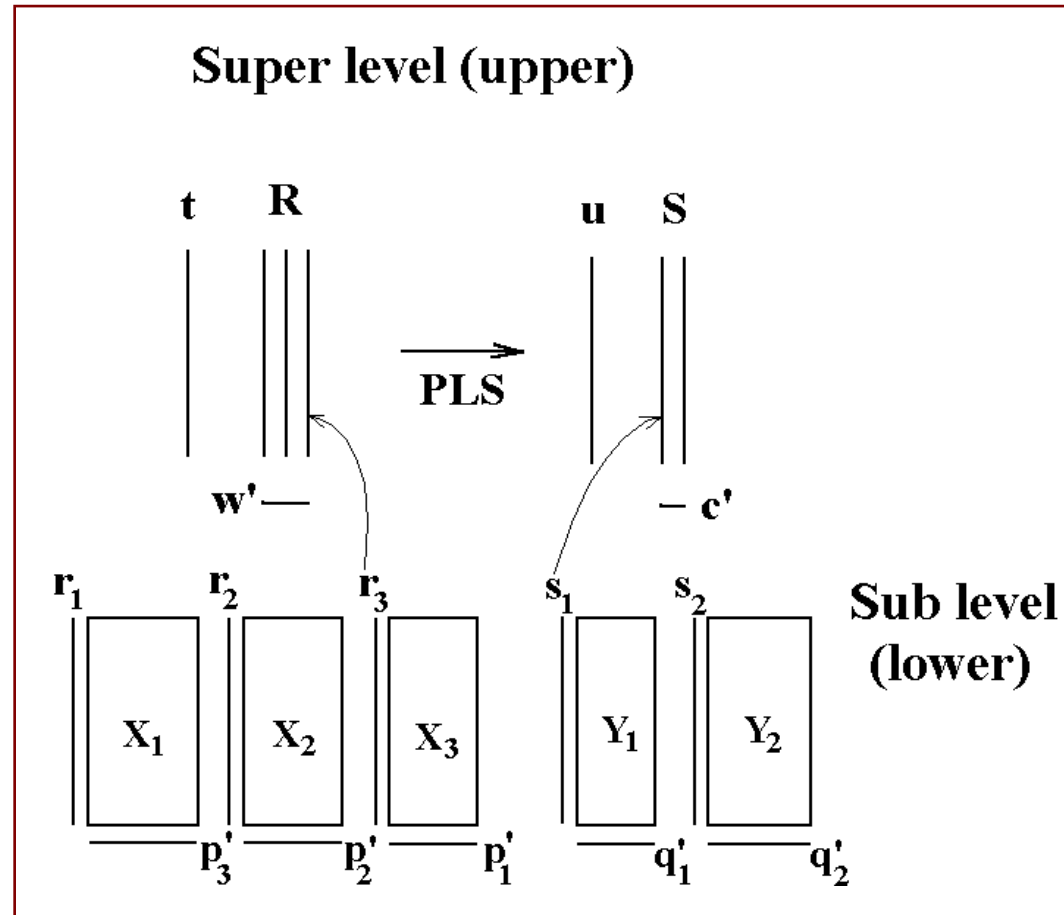
### QSAR:

$X_1$  = left side of molecule,  $X_2$  = spacer,  $X_3$  = right side.  
 $Y_1$  = *in vivo*,  $Y_2$  = *in vitro*,  $Y_3$  = Tox, etc.

# The block scores are variables in the “super” model

Many variants:

- No Y's (hier PCA)
- Few Y's; (H-PLS)  
Y unblocked
- Few X's; (H-PLS)  
X unblocked
- Many X's and Y's  
X and Y blocked  
(H-PLS)



Two variables show much more than one

K variables show more than 2, & results are more stable

K = 1200, N = 42 is possible, results are interpretable, ....

---

This still surprises us chemists very much

# Chemometrics (C) “Information aspects of chemistry”

---

- Chemistry 1970     wet chemistry, reagents, data poor  
Bottle neck was data generation.  
Strategy was hypothesis testing  
++ fostered thinking; -- slow
- Chemistry 2006     instrumental chemistry, data overflow  
Bottle neck is thinking, data interpretation.  
Strategy should be **getting good data, and  
the efficient interpretation of these data (C)**  
++ great potential, fast; -- AI temptation
- My personal view: Without thinking and enjoyment, no science

## Some philosophy

## Chemometrics $\neq$ Statistics

---

- < 1900      Science =      Search for “truth”
- 1900 - 30    Planck, Heisenberg, Dirac, ..., Gödel
- > 1930      Science =      Set of useful connected models

-----  
Now, what is then “knowledge”

Science:      ???

Technology:    working models; interpretation = connecting

Tukey, Breiman, Snee, Hoerl, Munck, ...

Two Cultures: {Hypotheses, ML} / {Data Analysis, LSq, LV}

GEP Box:    All models are wrong, but some are still useful

$\Rightarrow$  All hypothesis are wrong, but ...



# Quantitative Modelling, PLS. Relationships between sets of multivariate data, X and Y

---

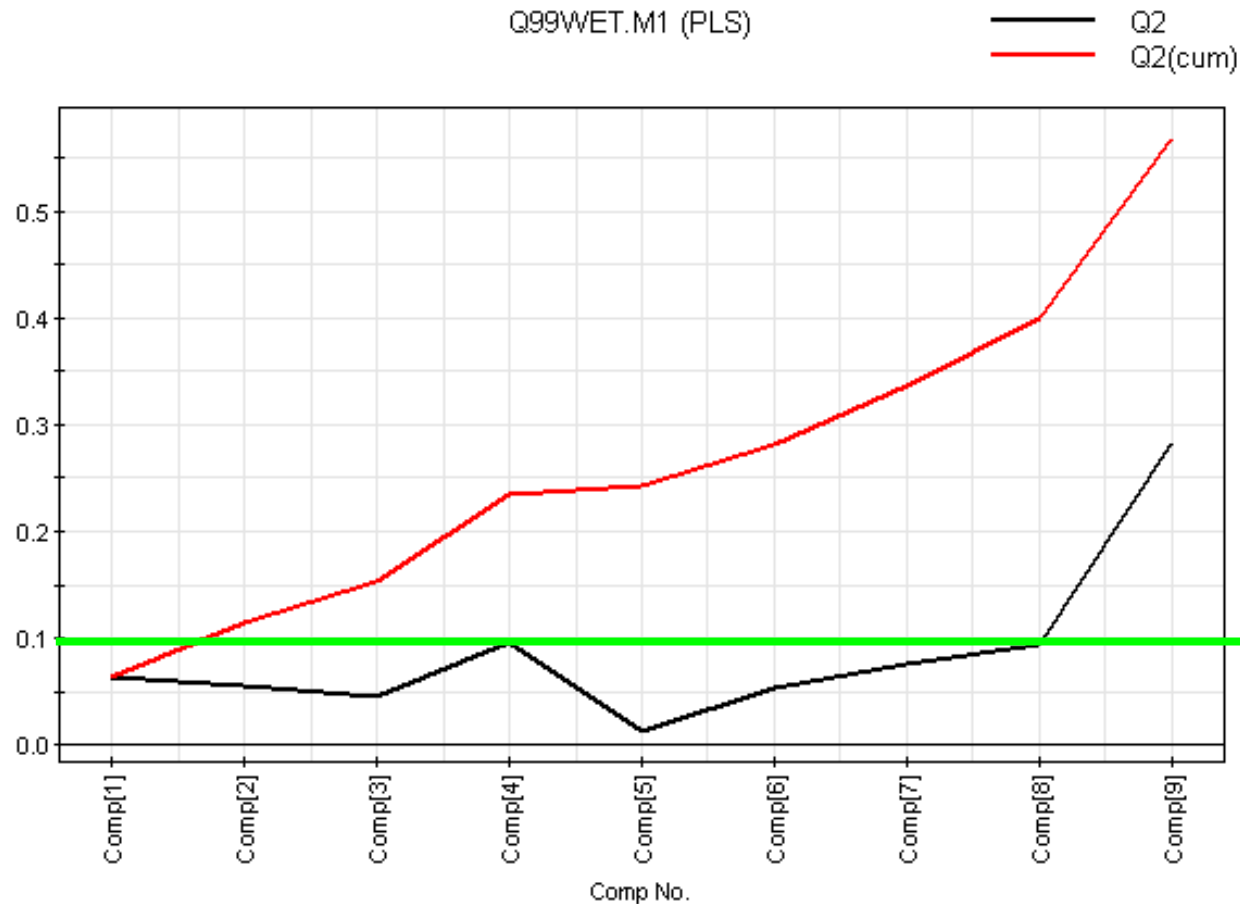
- Process modelling and optimization
- Chemical composition ↔ Quality  
physical measurements Biol. Activity  
Med. Classification, “omics”
- Chemical structure ↔ Reactivity, properties  
biol. activity
- Multivariate calibration  
Signals (spectra) ↔ Concentrations  
Energy content, Age, Taste

## Example Q99

---

- Tembec, Quebec, 1999 (Marc Champagne)
- Problems to measure a minor constituent in pulp, “Q99”, which caused some problems in paper production
- Lab method slow ( $> 6$  hours)  $\Rightarrow$  often *post mortem*
- NIR spectroscopy ( $K=2151$ ,  $N = 158$ ) & PLS tried by consultant (Bob Meglen, CO) without much success, NIR dominated by water (pulp is  $> 70$  % water)
- Marc C tried the same data with OSC + PLS, which worked well, verified by Bob M, later opponent for H. Antti
- Here shown by OPLS – simpler, and same results
- Data  $\Rightarrow$  training ( $N_1 = 113$ ) and pred sets ( $N_2 = 45$ )  
Q99  $> 7$  excluded

PLS 1999 “not significant” according to CV, *but*,  
OSC + PLS and OPLS are (also other CV lay-out)



# The O-PLS method – simple extension of the NIPALS algorithm

Trygg, J. (2001). Parsimonious multivariate models. PhD thesis, Umeå University.

Trygg, J. and S. Wold (2002). J. Chemometr. **16**(3): 119-128.

Model of **X**: 
$$\mathbf{X} = \begin{bmatrix} | \\ \hline \end{bmatrix} \begin{bmatrix} \text{---} \\ \hline \end{bmatrix} \mathbf{t}_p \mathbf{p}_p^T + \begin{bmatrix} | \\ \hline \end{bmatrix} \begin{bmatrix} \text{---} \\ \hline \end{bmatrix} \mathbf{t}_o \mathbf{p}_o^T + \mathbf{E}$$

Model of **Y**: 
$$\mathbf{Y} = \begin{bmatrix} | \\ \hline \end{bmatrix} \begin{bmatrix} \text{---} \\ \hline \end{bmatrix} \mathbf{t}_p \mathbf{c}_p^T + \mathbf{F}$$

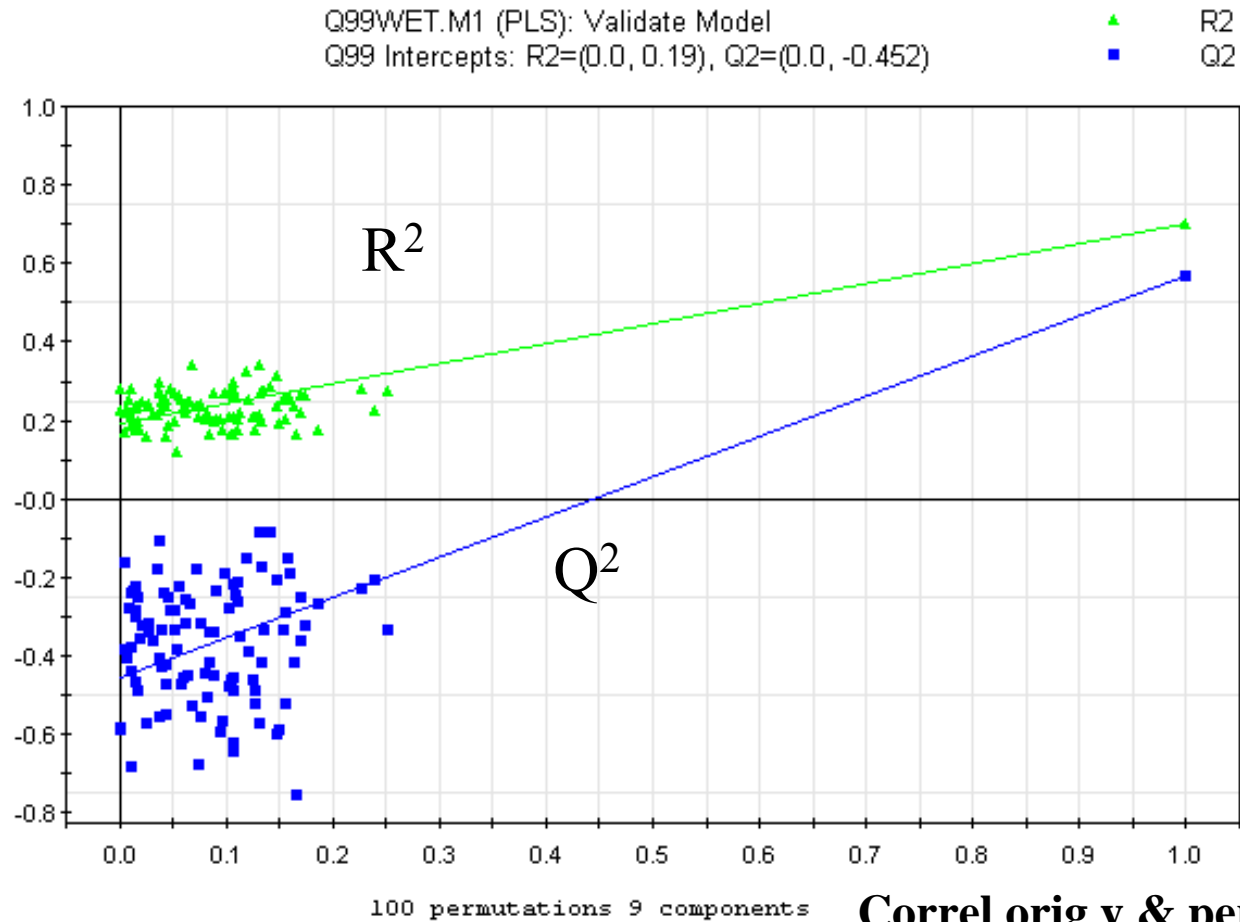
Prediction  $\mathbf{X} \leftrightarrow \mathbf{Y}$

$$\mathbf{X} = \begin{bmatrix} | \\ \hline \end{bmatrix} \begin{bmatrix} \text{---} \\ \hline \end{bmatrix} \mathbf{t} \mathbf{p}^T + \mathbf{E}$$

PLS

$$\mathbf{Y} = \begin{bmatrix} | \\ \hline \end{bmatrix} \begin{bmatrix} \text{---} \\ \hline \end{bmatrix} \mathbf{t} \mathbf{c}^T + \mathbf{F}$$

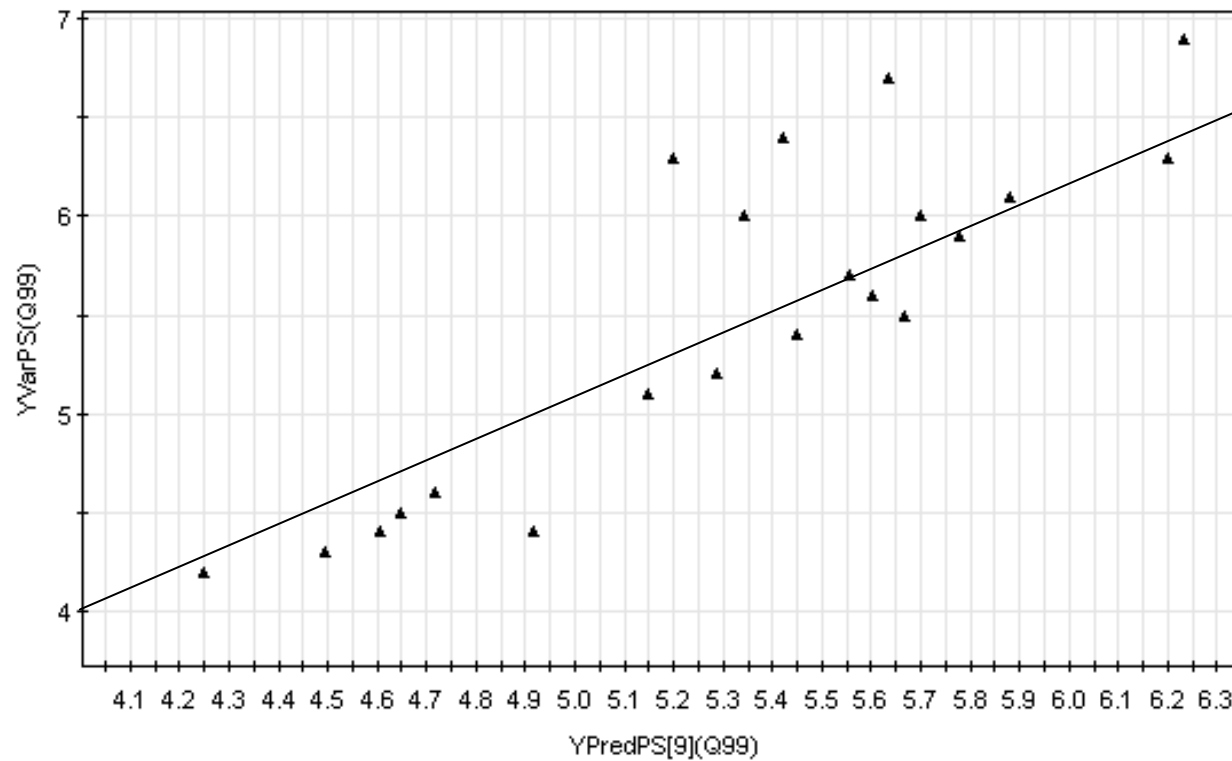
With  $A=9$ ,  $N= 113$ , and  $K = 2151$ , risk for overfit ?



100 random permutations of  $y$ , followed by full PLS / OPLS with CV (9 comp.s)

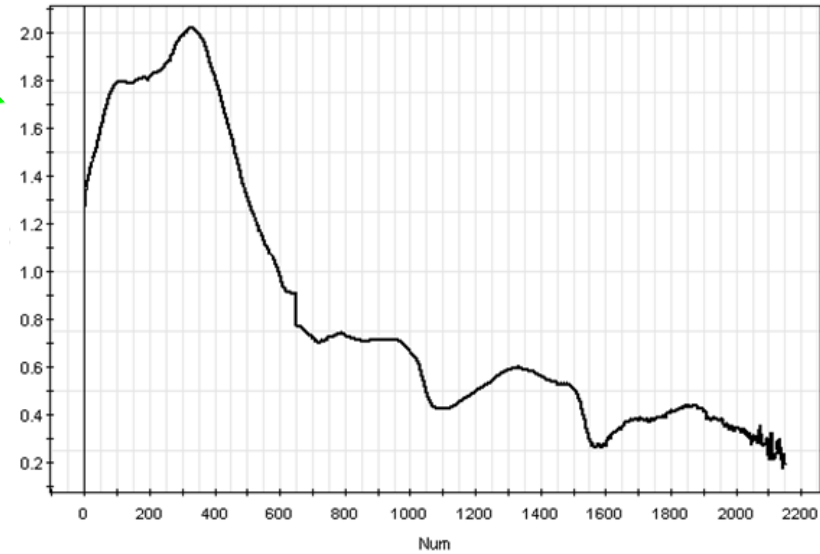
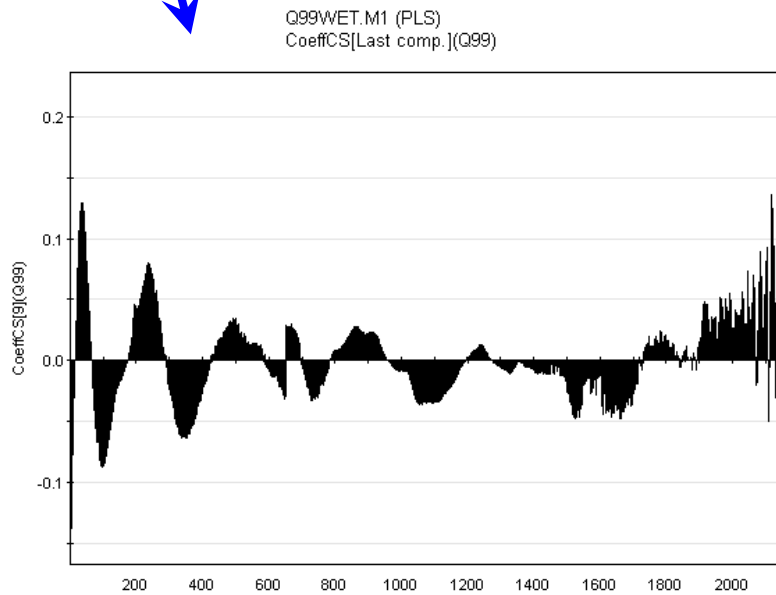
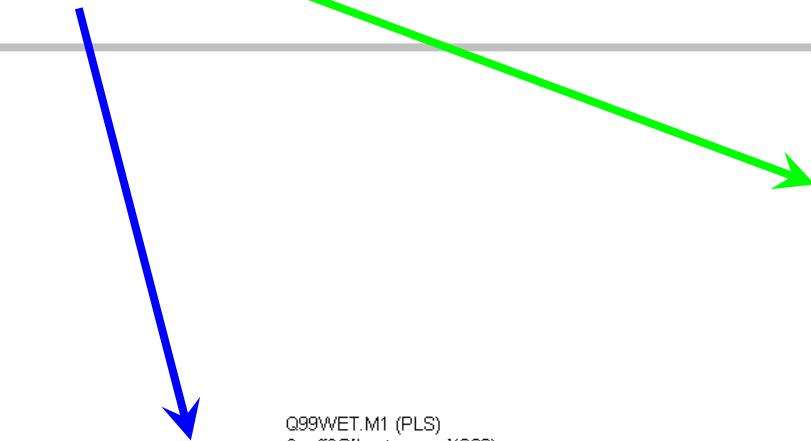
# Prediction set predicted by OPLS (identical to PLS)

Q99WET.M2 (OPLS), PS-Complement Model 2  
YPredPS[Last comp.](Q99)/YVarPS(Q99)

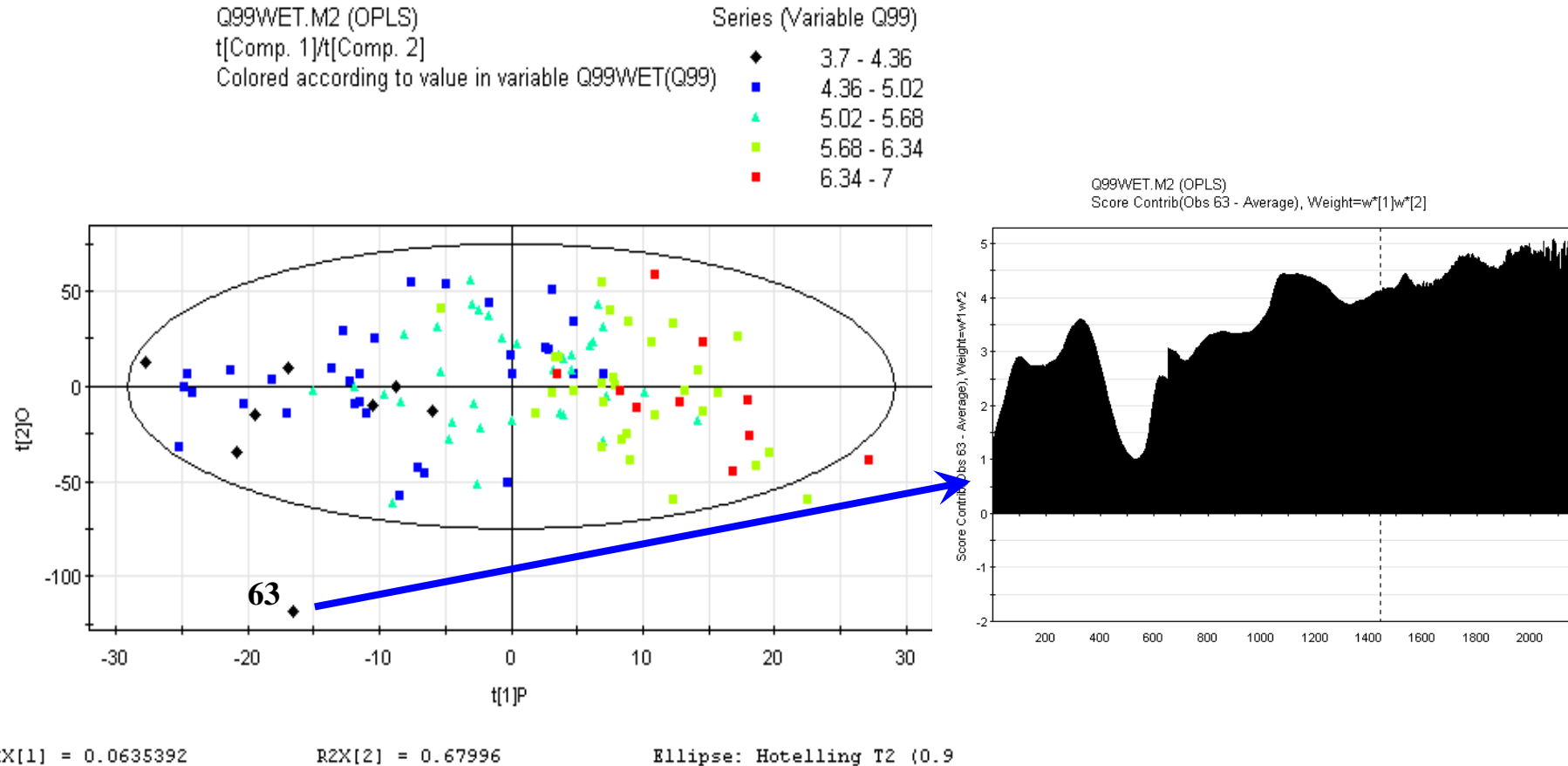


RMSEP = 2.89836

OPLS has 1 component related to y → interpretable coeff.s  
PLS in this case has 9 → ???



# OPLS scores 1 & 2, colored by Q99-conc



## The “contribution plots”: Show what has happened in the individual observation (interpreting upsets)

---

- A residual SD (DModX), e.g., point 71, is suspect

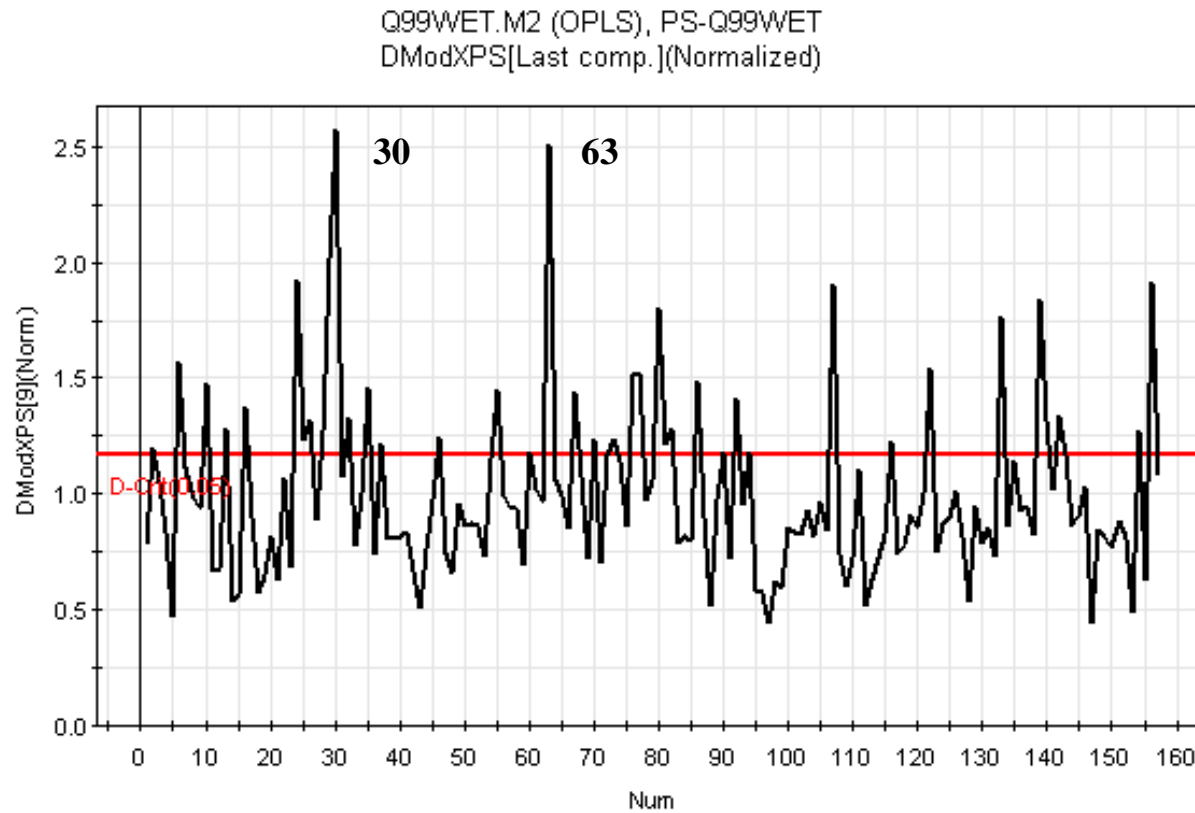
we look at the residuals  $z_k = e_{71, k}$   
times a weight (e.g.,  $v_k = \text{sqrt}(R_k^2)$  )

- A score value (e.g. point 65) is suspect

we look at the scaled data  $z_k = (x_{65, k} - \text{xavg}_k) * w_{s_k}$   
times a weight ( $v_k = p_k, f(p_a, p_b)$ , or,  $v_k = \text{sqrt}(R_k^2)$  )

- Contributions =  $z_k v_k$  ; for  $k=1,2,\dots,K$  (each variable)
- The “contribution” plots identify “culprit” variables
  - (Ron Swanson, Kodak, 1992)

# DModX (obs resid sd), OPLS after A=9

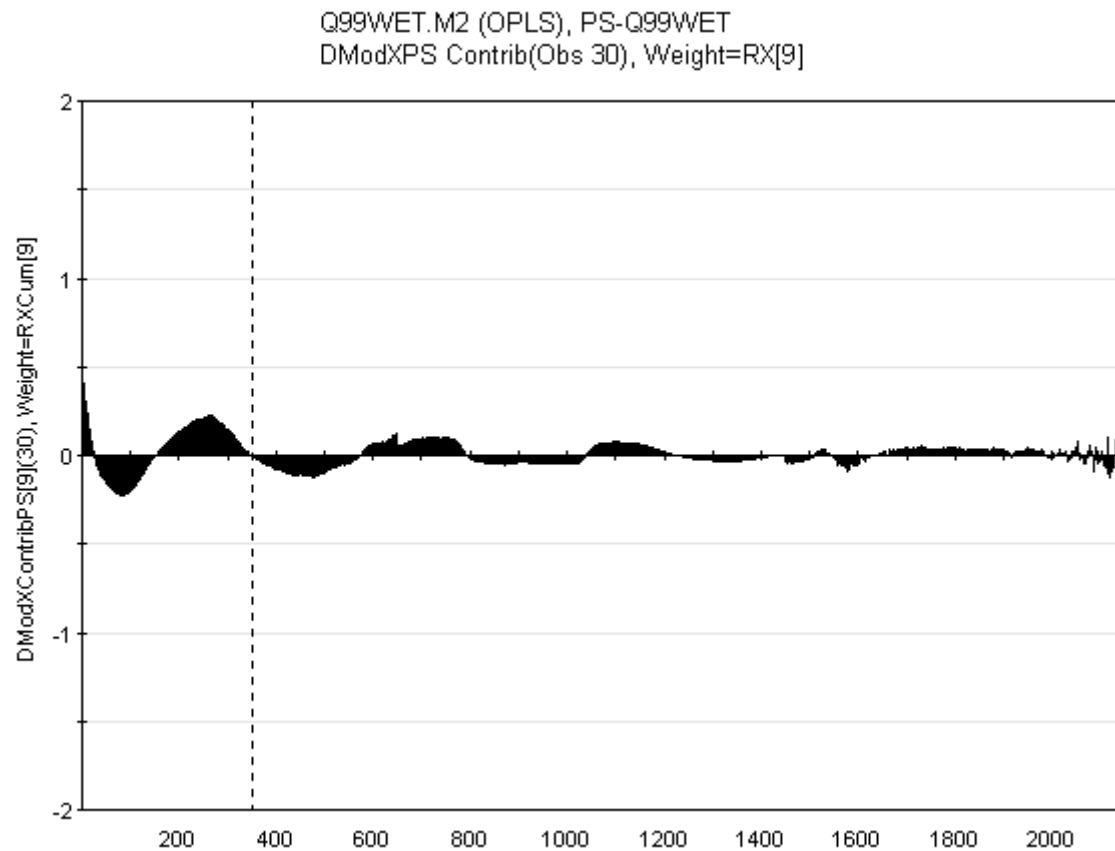


M2-D-Crit[9] = 1.175

1 - R2X(cum)[9] = 0.0009694

# Contribution plot, obs 30

---



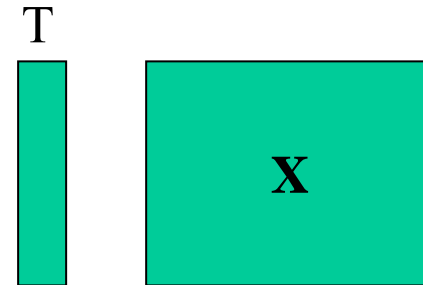
## Conclusions Q99

---

- In retrospect, always easy
- PLS implementation 1999 too high limit for CV significance
- OSC removed most of the water peak and the subsequent PLS analysis gave a good model (one component) with good predictions
- OPLS does the same, but easier
  
- In applications where PLS gives many components because of strong secondary variation in X (non-related to Y), OPLS facilitates the interpretation by rotating the solution so that the y-related X-variation sits in component 1.

## Use of scores $T = \{ t_a \}$ as design variables; MV design

- When  $X$  are numerous and collinear
- DoE in  $X$  is meaningless and undoable
- But DoE in the scores  $T$  (which summarize  $X$  in a small number of components) is meaningful and doable
- Typically D-optimal in  $T$ , where  $T$  is derived by PCA or PLS of the candidate set ( $X$ )



Solvents  
Catalysts  
Molecules  
Oils  
Lots of raw materials  
Manufacturing parts

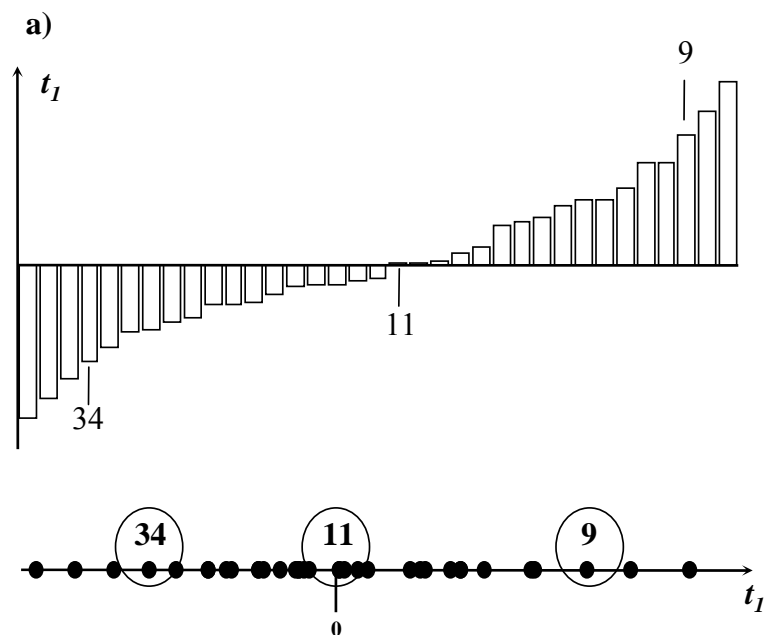
## x.2. A combinatorial chemistry application

### Motivation: Better biological testing

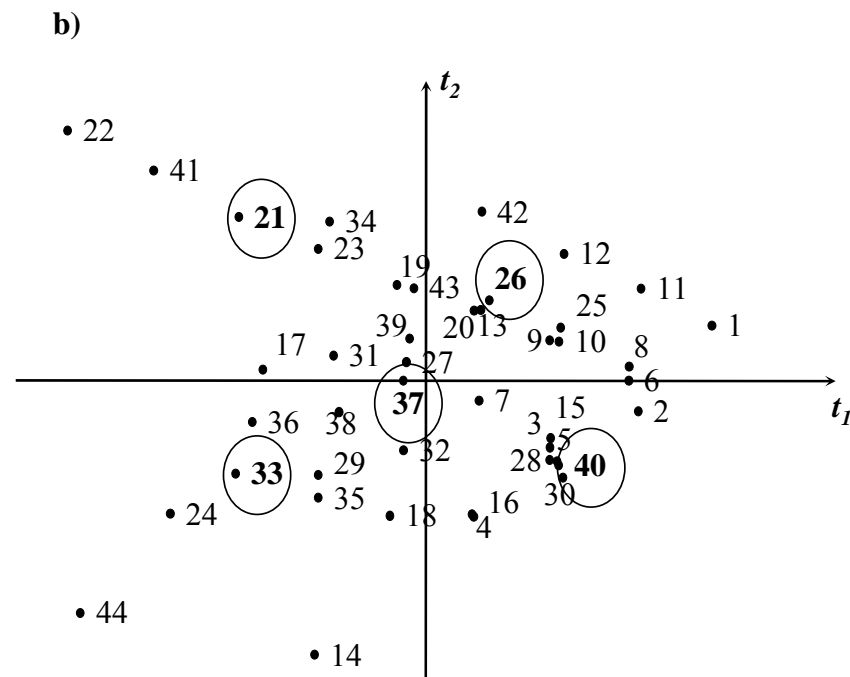
---

- Primary amine ( $n_1=35$ )+ aromatic aldehyde ( $n_2=44$ )
- Full library:  $35 \times 44 = 1540$  products
  - are all these really needed ?? *Can we test them* ????
- Selecting representative sets of building blocks
  1. Characterization of the candidate structures (K=11 & 54)
  2. Making a compact representation (PCA: A=1 & 2)
  3. Selecting spanning compounds ( $n_1=3$ ,  $n_2 = 5$ )
- Combining the sets of building blocks, a new design
  - A design with 9 compounds will do a good job

## PC scores of Amines (left) & Aldehydes (right)



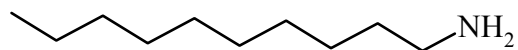
Amines, A=1 (size & lipophil)



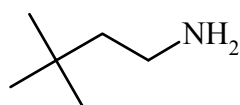
Aldehydes, A=2 (size & polariz.)

## Selected structures

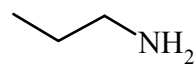
### Primary Amines



(9)

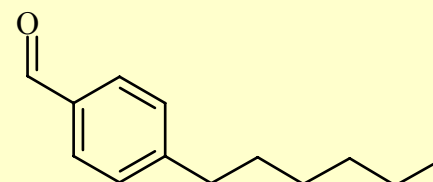


(11)

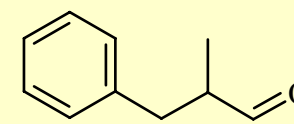


(34)

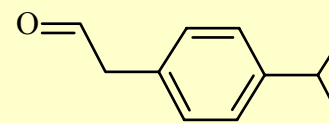
### Aromatic Aldehydes



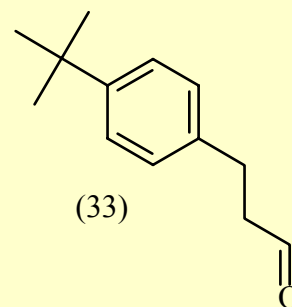
(21)



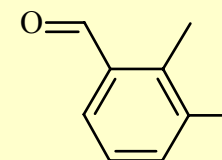
(26)



(37)



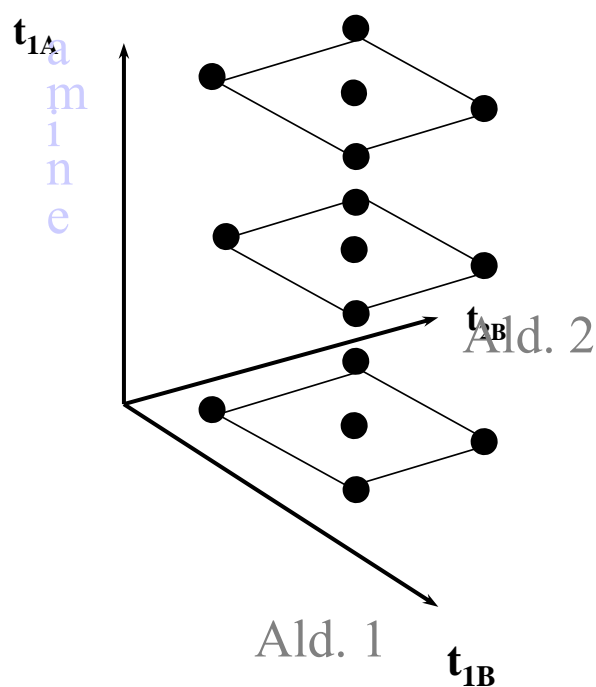
(33)



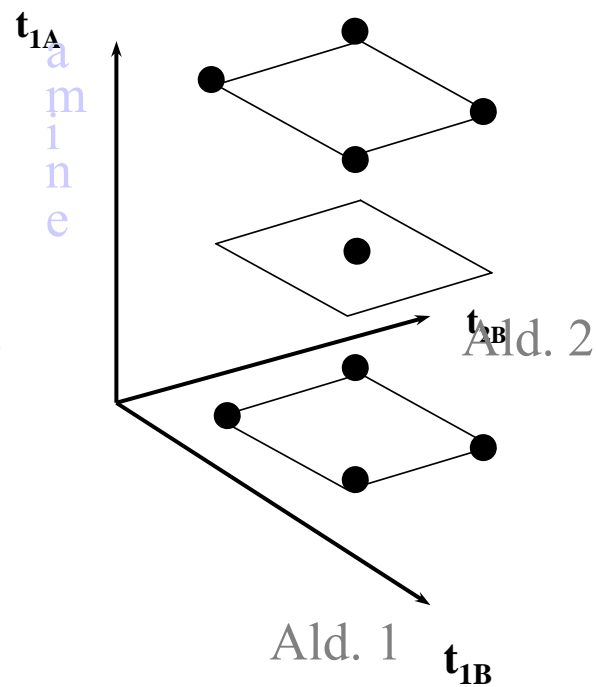
(40)

Best combination of BB1 (amines) & BB2 (aldehydes) is also a design !

---



$N = 3 * 5 = 15$  is rather much !



$N = 3 + 5 + 1 = 9$  is better !

## Conclusions – MV design

---

- Design makes data representative
- In complex systems, variables are not independent.
- Hence, design in latent variables (scores) is natural
- Approach:
  1. Multivariate characterization of “objects” (e.g., molecules)
  2. PCA (sometimes PLS) to get scores; “latent variables”
  3. Design in these scores – D-optimal or classical, e.g., factorial
  4. This give a selection of representative objects
  5. Measure / Experiment carefully
  6. PLS model based on these objects
  7. Validation, Interpretation, prediction, etc.

Final words -- Chemometrics is evolving – new application areas; “-omics”, PAT,..., simpler approaches (Hier, OPLS), ...

---

- MV projection methods (LV methods), the workhorses of chemometrics, provide a straight forward approach for direct analysis and interpretation of highly multivariate data.
- Geometrical interpretation as windows into MV data space with associated plots makes the approach “understandable” and even appealing for chemists, biologists & engineers
  - Scores ( $\mathbf{t}_a$ )  $\leftrightarrow$  coordinates in pertinent windows  $\leftrightarrow N(0, \sigma_{ta})$
  - Loadings  $\leftrightarrow$  orientation of windows & correlations [ $\mathbf{x}_k, \mathbf{t}_a$ ]
  - Obs residual SD (row of resid mx)  $\leftrightarrow$  distance betw obs & model
  - Contribution plots  $\leftrightarrow$  pattern of individual observation(s)
- Use computation based approaches such as CV and JK, consistent with data-analytical branch of statistics

---

THE END

(hopefully followed by a lively discussion)

Thanks for Your attention