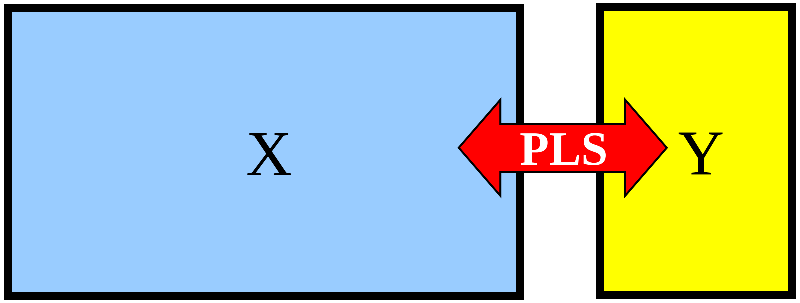
PLS - Partial Least Squares Projection to Latent Structures

- Notation
- Scaling
- Geometric interpretation
- (Algebraic solution)
- Outliers
- Residuals
- Cross Validation (CV)
- Prediction
- Summary
- Applications



Quantitative modelling, PLS

- **Find relationships between blocks of multivariate data, x and y**
- **Predict one block from the other for new observations (samples). (Predict y for known x)**

Applications:

- **Process modelling and optimisation**
- **Chemical composition ⇔ Quality**
  Physical properties — Biological activity
- **Chemical structure ⇔ Reactivity**
  Properties
  Biological activity
- **Multivariate calibration**
  Signals (spectra) ⇔ Concentrations
  Properties
  Age, Taste, ...

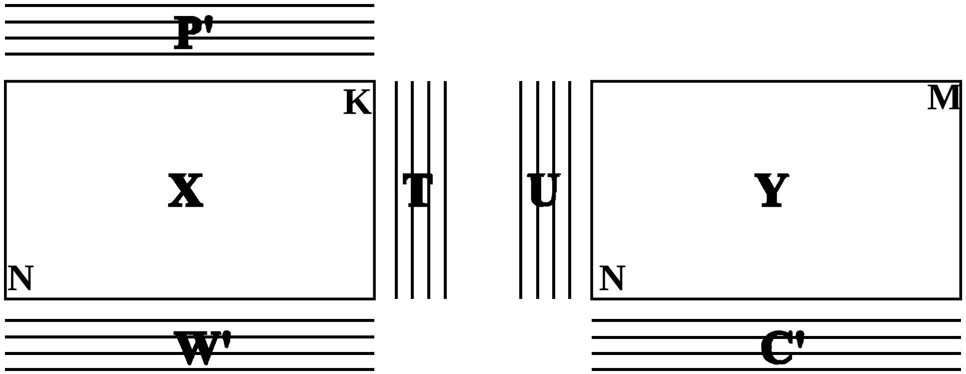
Notation - PLS

K = number of X-variables

M = number of Y-variables

N = number of observations (samples)

A = number of PLS components



T = matrix of X-scores with columns $t_1,..,t_A$ (vectors)

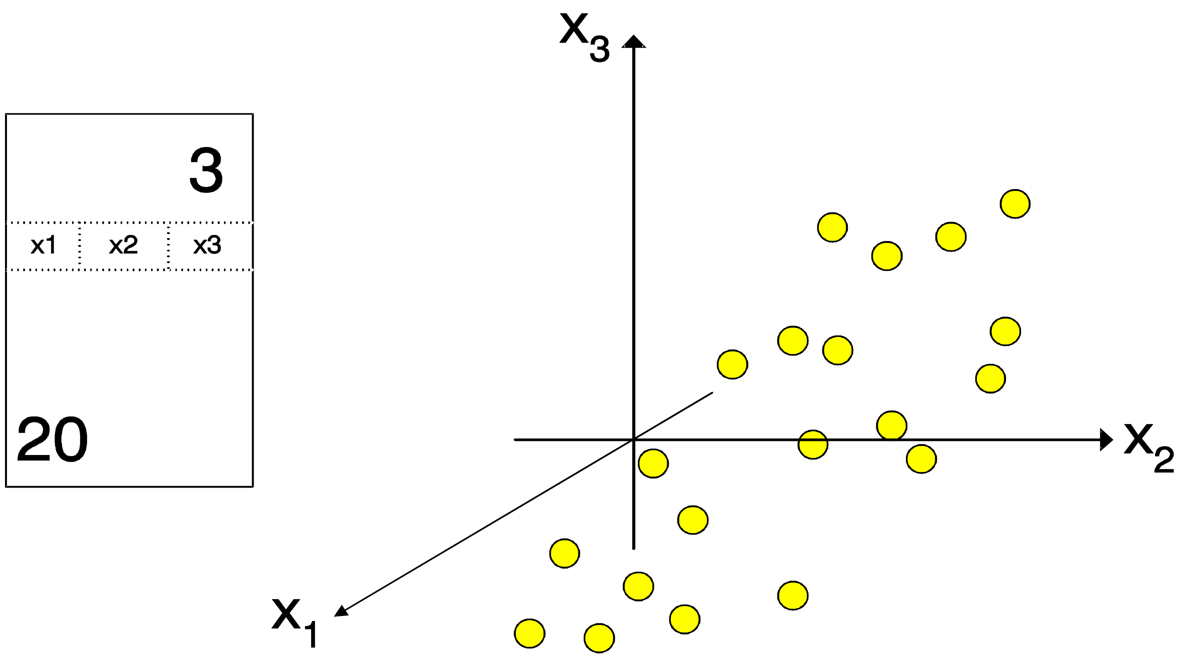
P = matrix of X-loadings with columns $p_1,..,p_A$ (vectors)

W = matrix of PLS X-weights with columns $w_1,..,w_A$ (vectors)

U = matrix of Y-scores with columns $u_1,..,u_A$ (vectors)

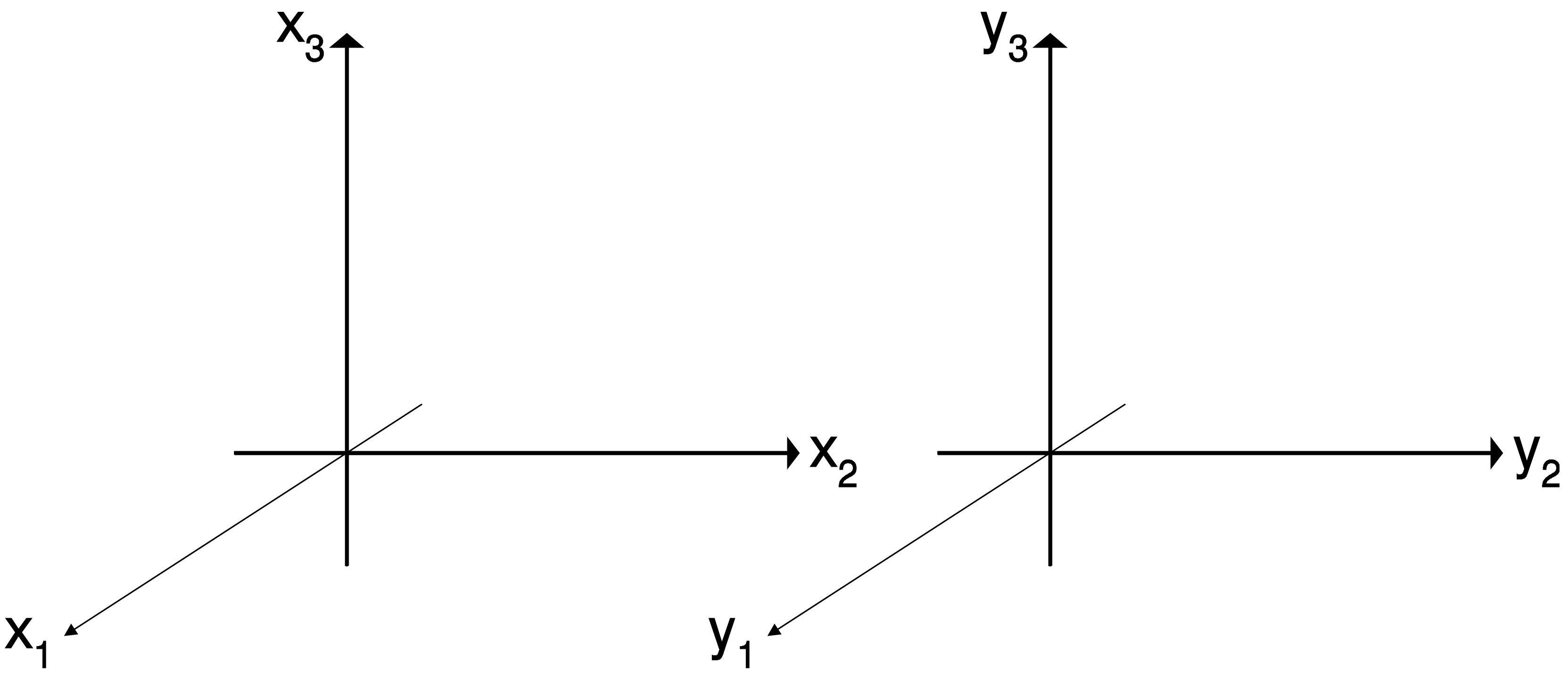
C = matrix of PLS Y-weights with columns $c_1,..,c_A$ (vectors)

Scaling of variables



- Define length of variable axis (X andY spaces)
- **Recommended:** Make all variable axis the length 1 (Auto scaling)

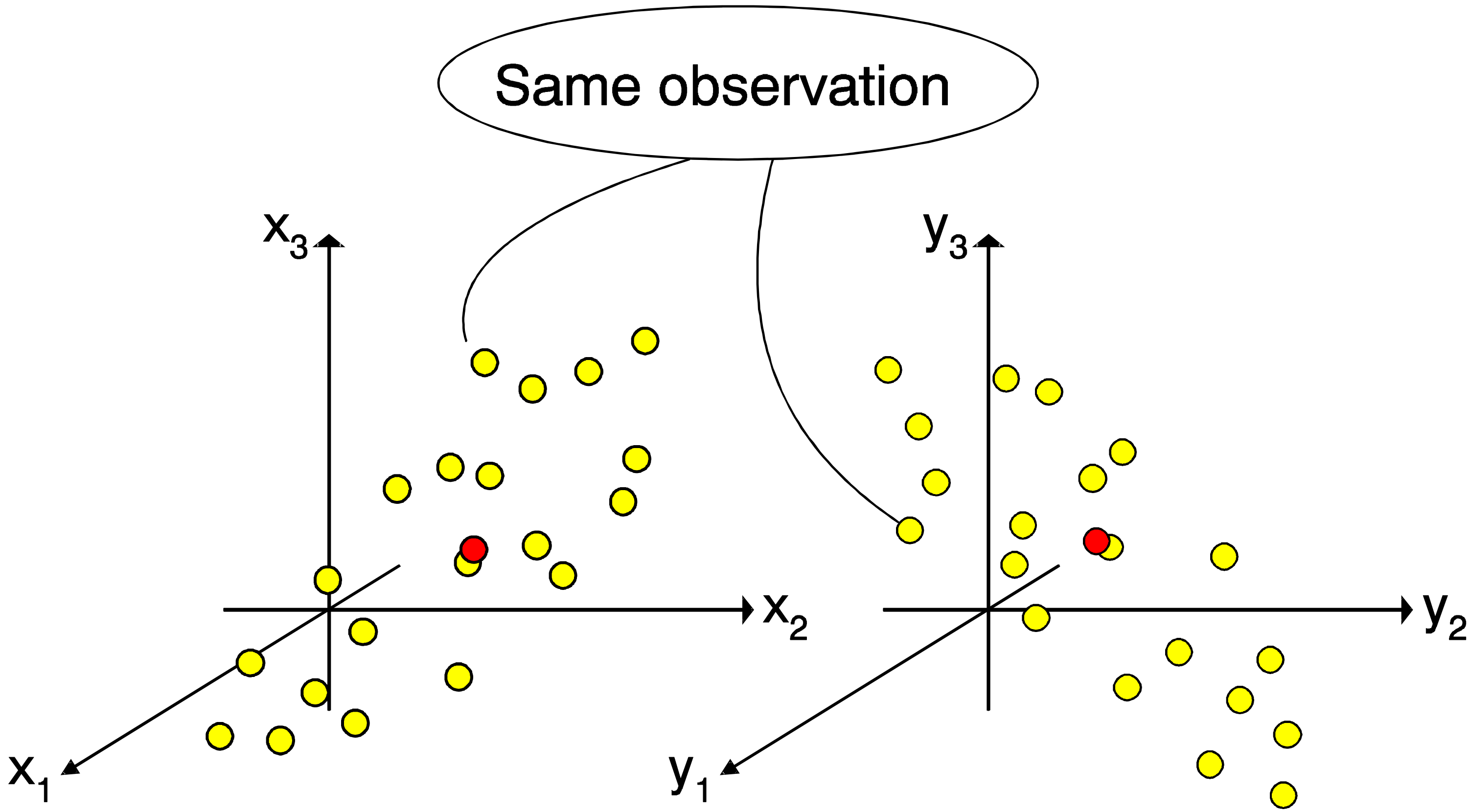
PLS - Geometric interpretation, 1



- For each matrix, X and Y, a space is defined with K respective M dimensions (here K=M=3)

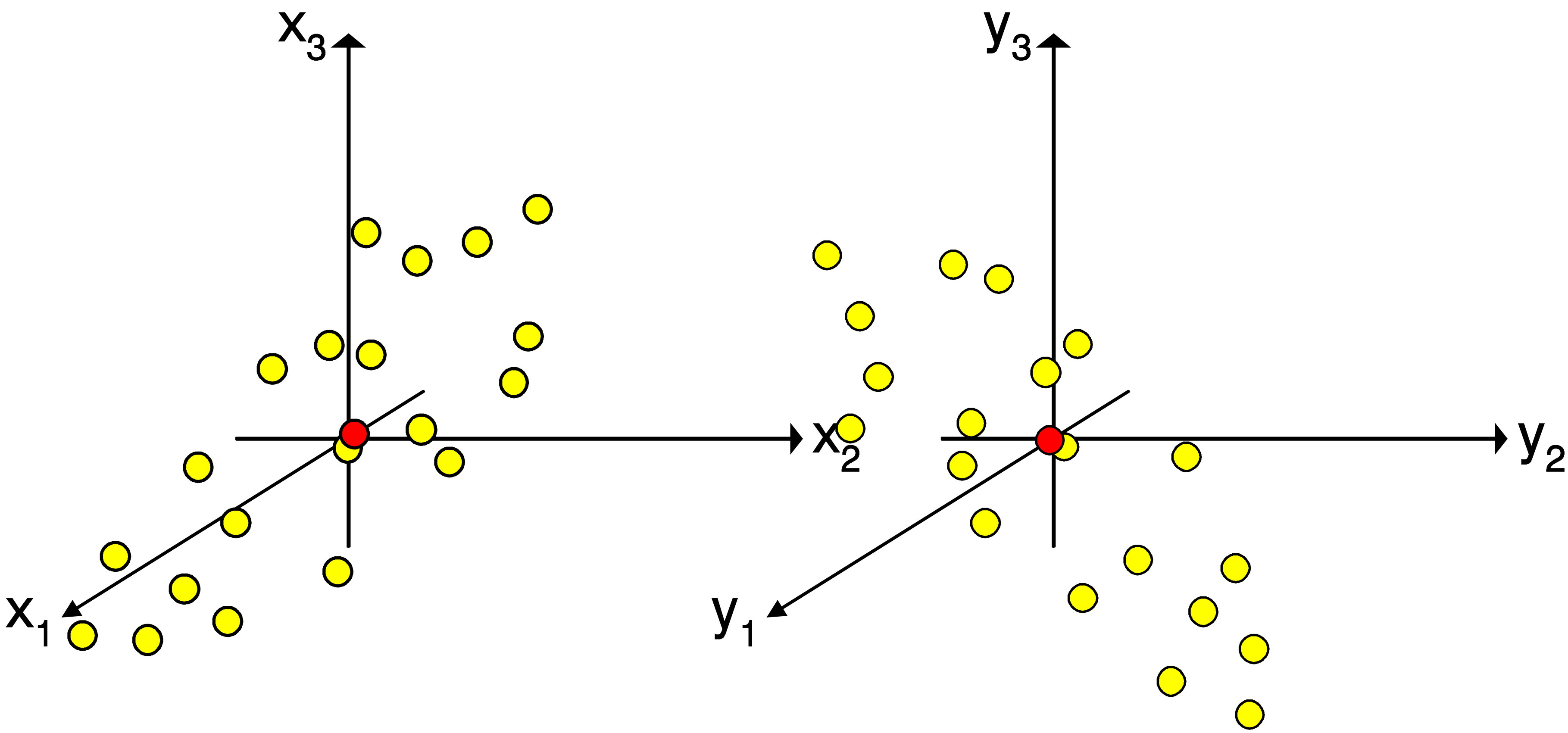
- Each variable in X and Y is described by a co-ordinate axisl with the length defined by the scaling, usually V = 1 (1/Sdev) (When variables are measured in different units)

PLS - Geometric interpretation, 2



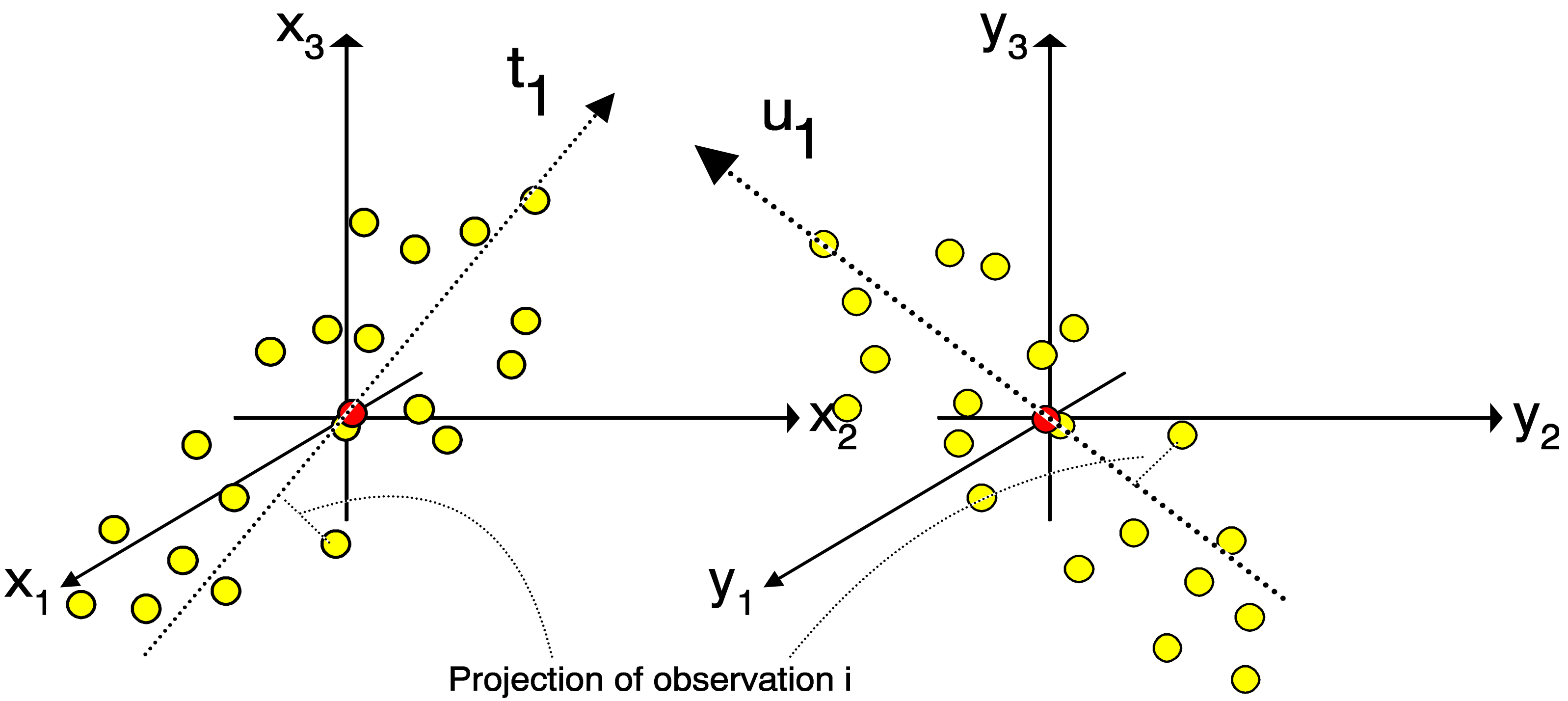
- Every observation (sample) defines a point in both the X and Y spaces.
- Similar to PCA, the first step is to mean centre the data; this means moving the centre of the point swarm to the origin of the variable space (X, Y)

PLS - Geometric interpretation, 3



- After the mean centering the points are centred in the two variable spaces.

PLS - Geometric interpretation, 4



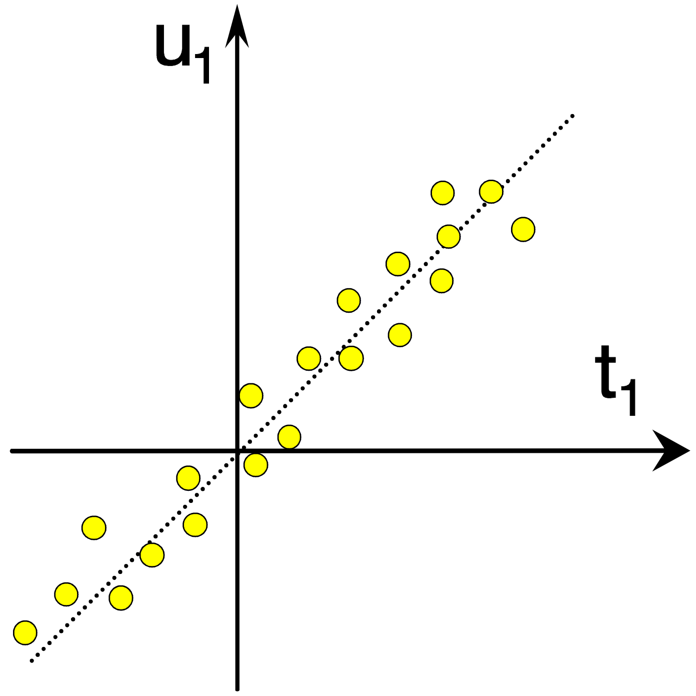
- The first PLS component is a line in the X space and a line in the Y space, fitted so that...

  a) it's a good summary of the variation in X and Y.

  b) so that the co-variation between the scores t_1 and u_1 is maximised.

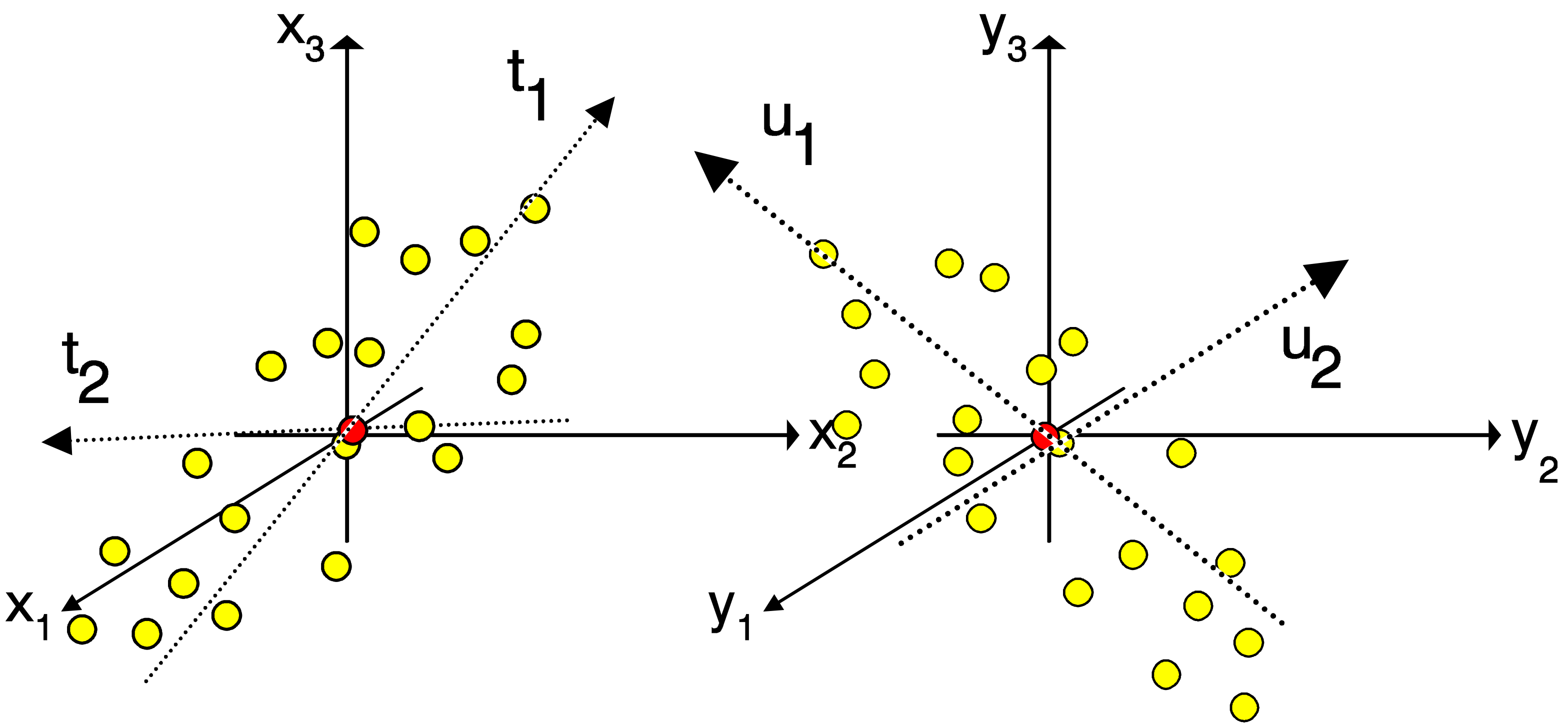
- The lines runs through the center of the point swarms (the origin of the variable spaces).

PLS - Geometric interpretation, 5



- Scores, t_1 and u_1, for the two spaces , X and Y, are connected and correlated through an **"inner relation"** $\boldsymbol{u_{i1} = t_{i1} + h_i}$ (where h_i is a residual)

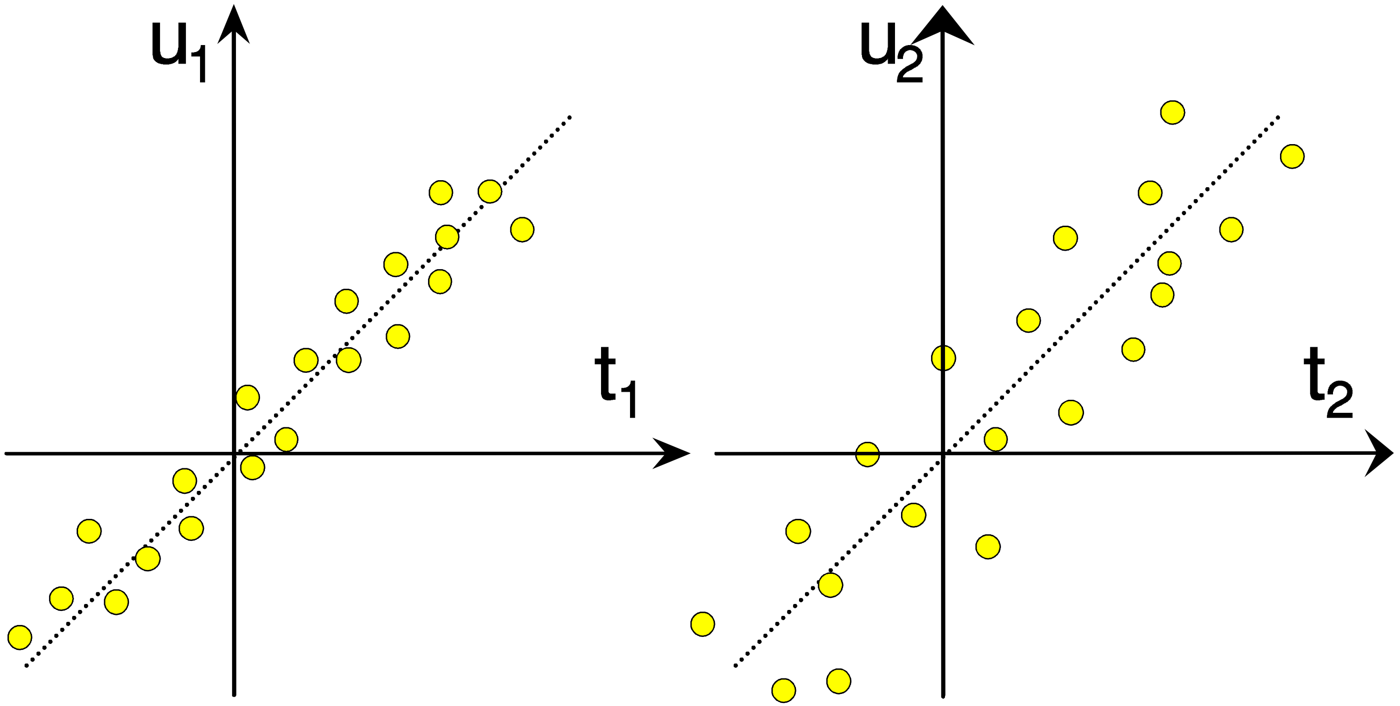
PLS - Geometric interpretation, 6



- The second PLS-component is represented by lines in the X and Y spaces **orthogonal** to the lines describing the first component, these lines also run through the centre of the point swarms.

- These lines, t_2 and u_2, enhance the variation description and correlation as much as possible.

PLS - Geometric interpretation, 7



- **Scores for the second component (t_2 and u_2) are correlated, but not as highly as in the first (t_1 and u_1)**

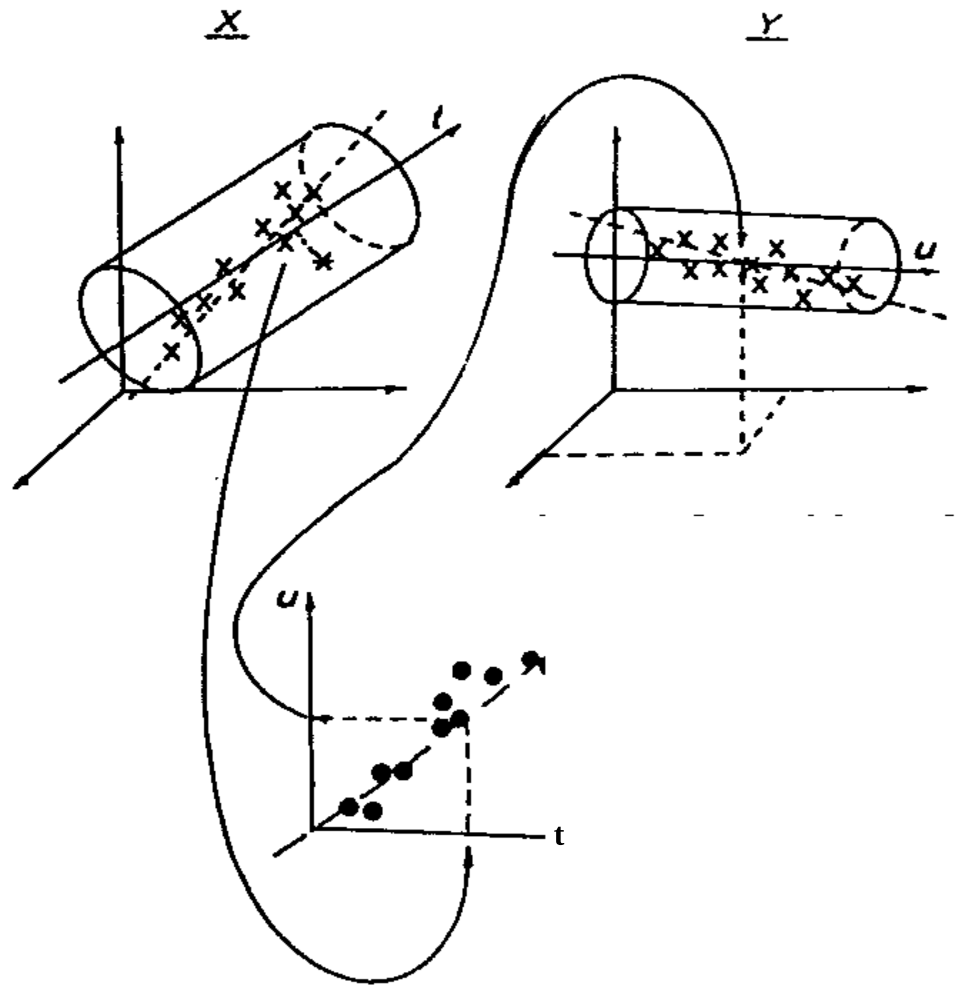
- **By introducing X values for a new observation into the model, t1 and t2 score values are obtained for this observation. Via the inner relation we we can then find the u1 and u2 score values, that gives us the possibility to estimate the Y values for the new observation *(prediction)***

Predictions, PLS

• A new observation is similar to the model samples (the training set) if it falls within the described tolerance cylinder in X space (the confidence interval) – **Check in scores and DModX**

• If the sample fits the model it can be projected onto the model in X(t). The value from the projection on t (score value) can then be put into the T-U relation, which gives the u score value for the sample.

• The u score value can then be put into Y space and predicted Y values can be obtained by deciding which Y values that correspond to the u score obtained for the new observation.



PLS – By Hand (2X, 1Y)

Two X:s and one response Y.
10 samples for model calculations with known y values.
4 samples for prediction!

	x1	x2	y1
1	-1.32	-1.24	-1.3
2	-1.08	-0.39	-1.06
3	-0.85	-0.82	-0.82
4	-0.61	-1.24	-0.82
5	-0.38	-0.39	-0.34
6	0.09	0.47	0.14
7	0.56	0.04	0.63
8	0.8	0.9	1.11
9	1.27	1.76	1.11
10	1.51	0.9	1.35
11	*4.8*	*4.76*	*?*
12	*1.98*	*1.76*	*?*
13	*0.09*	*0.04*	*?*
14	*-0.61*	*1.33*	*?*

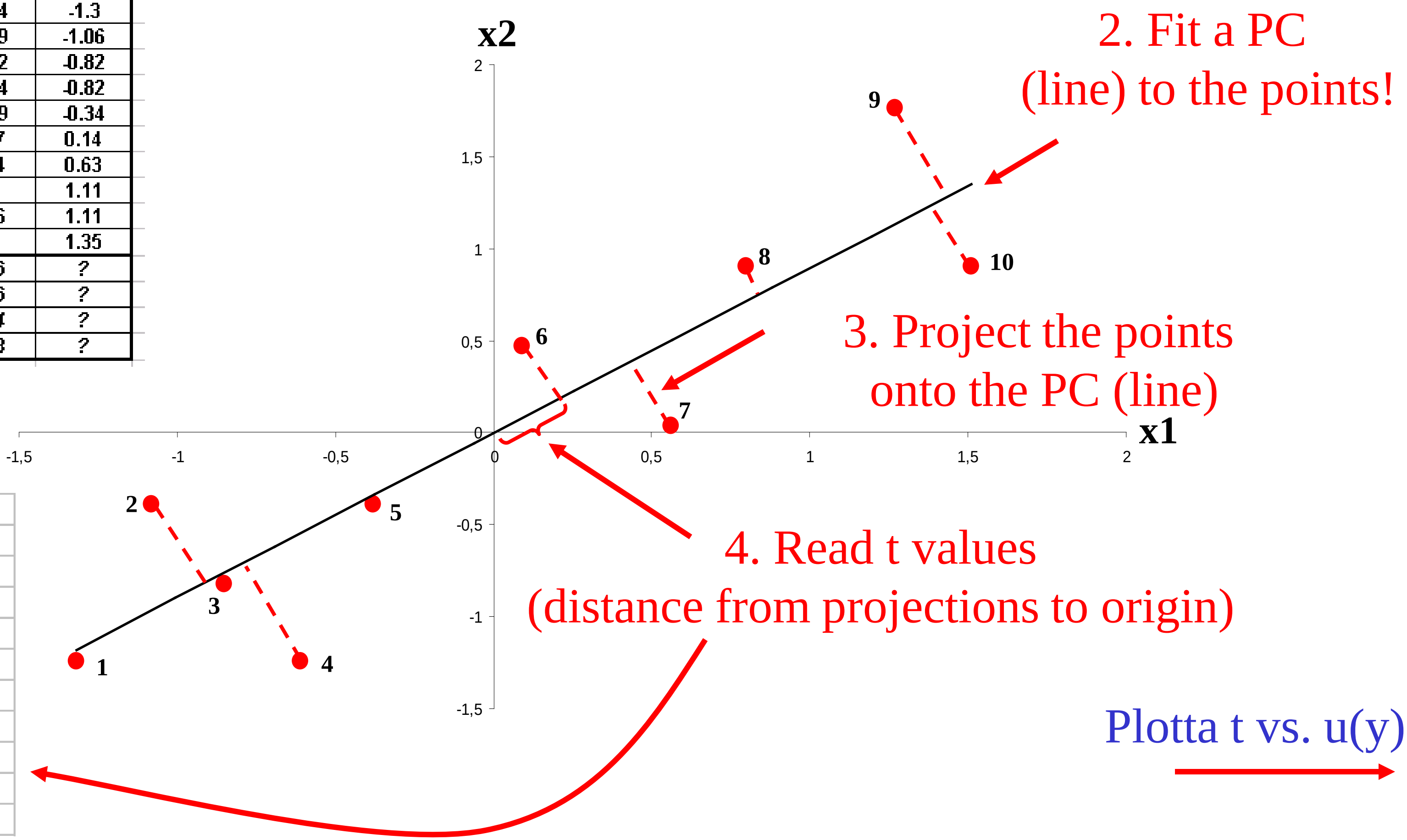
Data mean centred and scaled to V=1!

- **Plot x1 against x2 (samples 1-10)**
- **Fit a principal component**
- **Project the points down onto the PC**
- **Calculate the scores (t_i) for the observations**
- **Plot t against u(y) (u = y if only one y)**
- **Put observations 11-14 into X space**
- **Make sure the new observations fit the model**
- **Calculate t scores for observations fitting the model**
- **Go to the t/u(y) plot and predict y values.**
- **WHAT ABOUT 2 Y?**

PLS – By Hand (2X, 1Y)

1.Plot x1 against x2!

	x1	x2	y1
1	-1.32	-1.24	-1.3
2	-1.08	-0.39	-1.06
3	-0.85	-0.82	-0.82
4	-0.61	-1.24	-0.82
5	-0.38	-0.39	-0.34
6	0.09	0.47	0.14
7	0.56	0.04	0.63
8	0.8	0.9	1.11
9	1.27	1.76	1.11
10	1.51	0.9	1.35
11	*4.8*	*4.76*	*?*
12	*1.98*	*1.76*	*?*
13	*0.09*	*0.04*	*?*
14	*-0.61*	*1.33*	*?*



	t1
1	-1.37
2	-1.02
3	-0.98
4	-0.89
5	-0.43
6	0.24
7	0.48
8	0.93
9	1.59
10	1.54

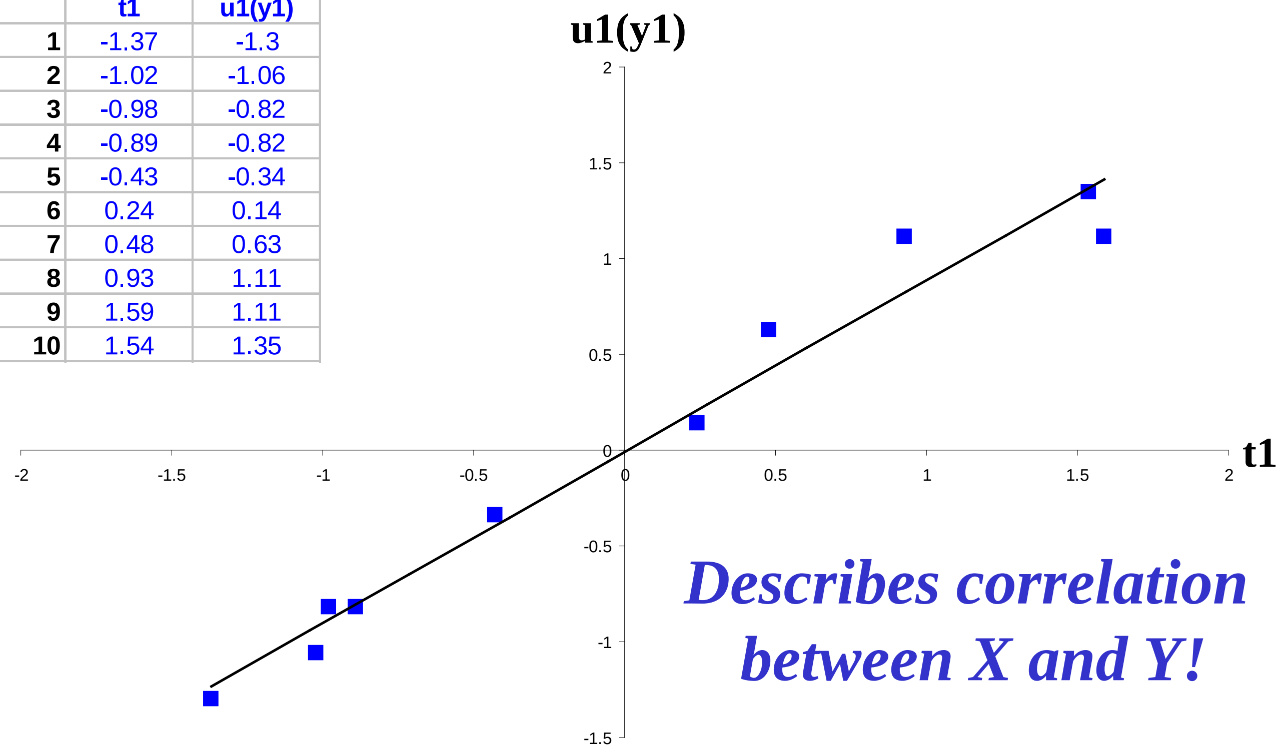
Plotta t vs. u(y)

PLS – By Hand (2X, 1Y)

1.Plotta t1 against u1(y1)!

”inner relation”

	t1	u1(y1)
1	-1.37	-1.3
2	-1.02	-1.06
3	-0.98	-0.82
4	-0.89	-0.82
5	-0.43	-0.34
6	0.24	0.14
7	0.48	0.63
8	0.93	1.11
9	1.59	1.11
10	1.54	1.35



Describes correlation between X and Y!

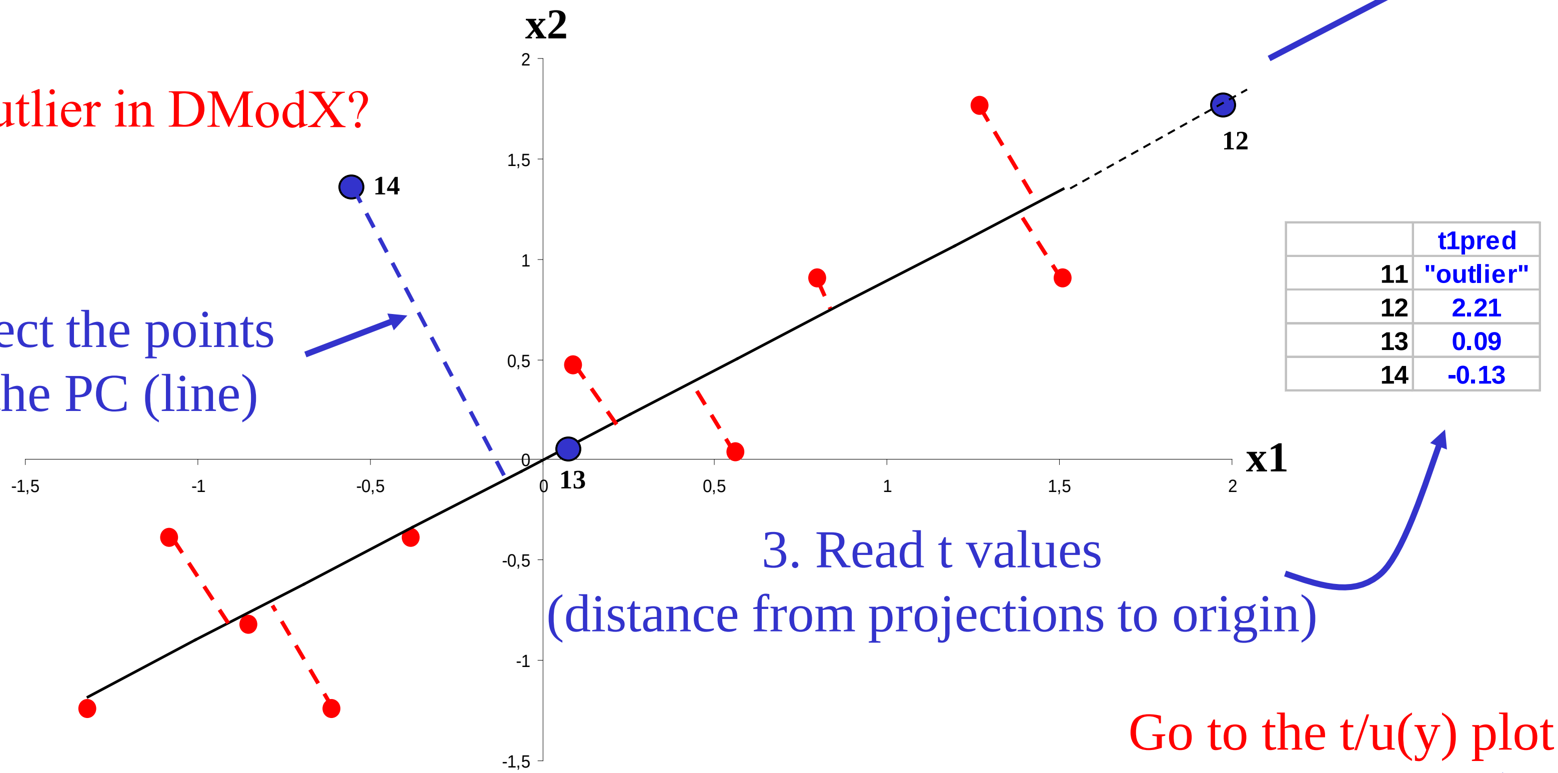
PLS – By Hand (2X, 1Y)

1.Plot samples 11-14

(4.8, 4.76)

Outlier in scores!

Outlier in DModX?



2. Project the points onto the PC (line)

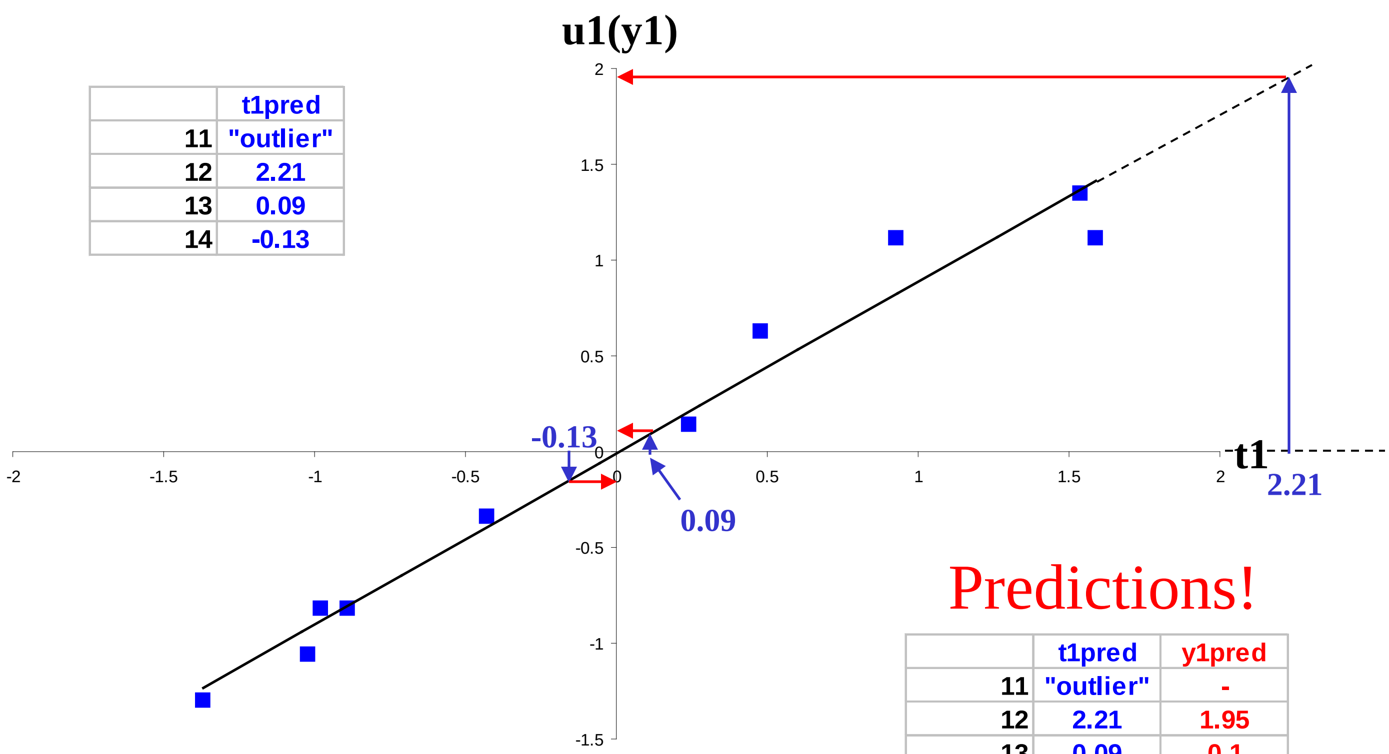
3. Read t values (distance from projections to origin)

	t1pred
11	"outlier"
12	2.21
13	0.09
14	-0.13

Go to the t/u(y) plot

PLS – By Hand (2X, 1Y)

Predict (read) y values for samples 11-14!

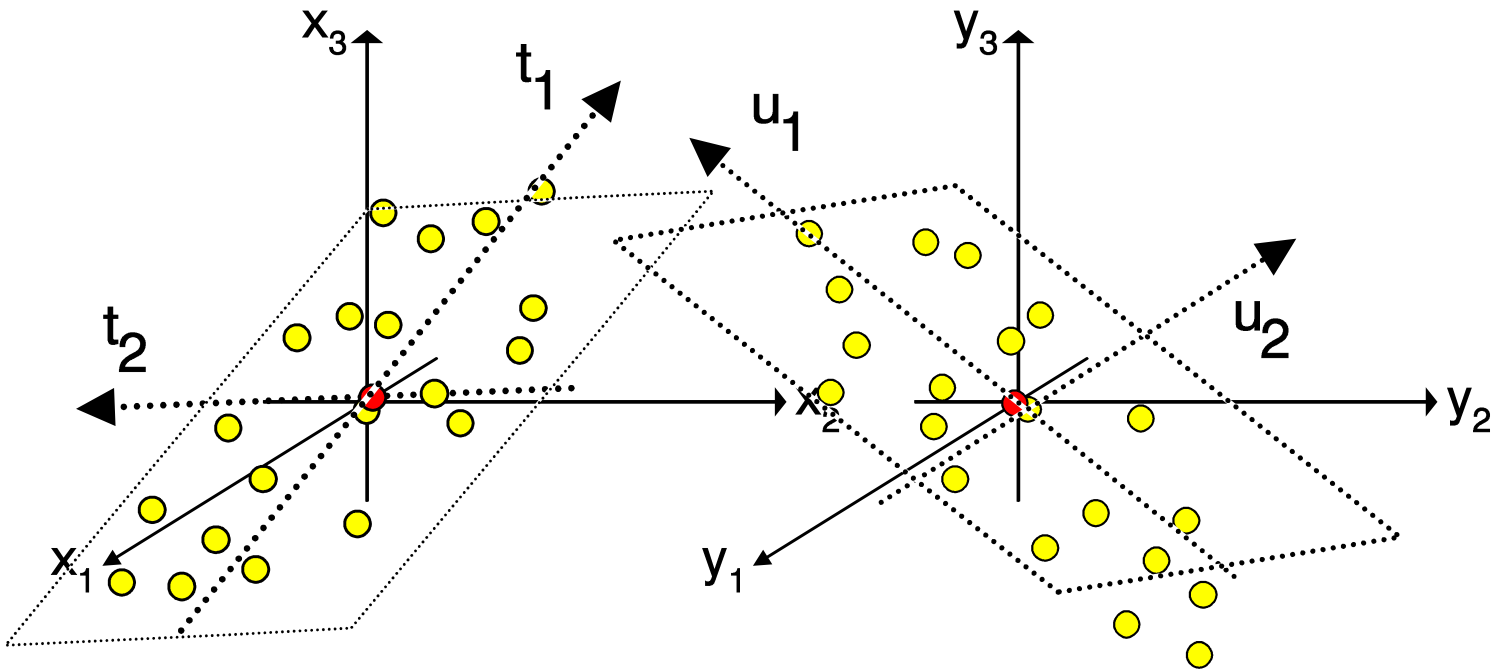


	t1pred
11	"outlier"
12	2.21
13	0.09
14	-0.13

Predictions!

	t1pred	y1pred
11	"outlier"	-
12	2.21	1.95
13	0.09	0.1
14	-0.13	-0.15

PLS - Geometric interpretation, 8

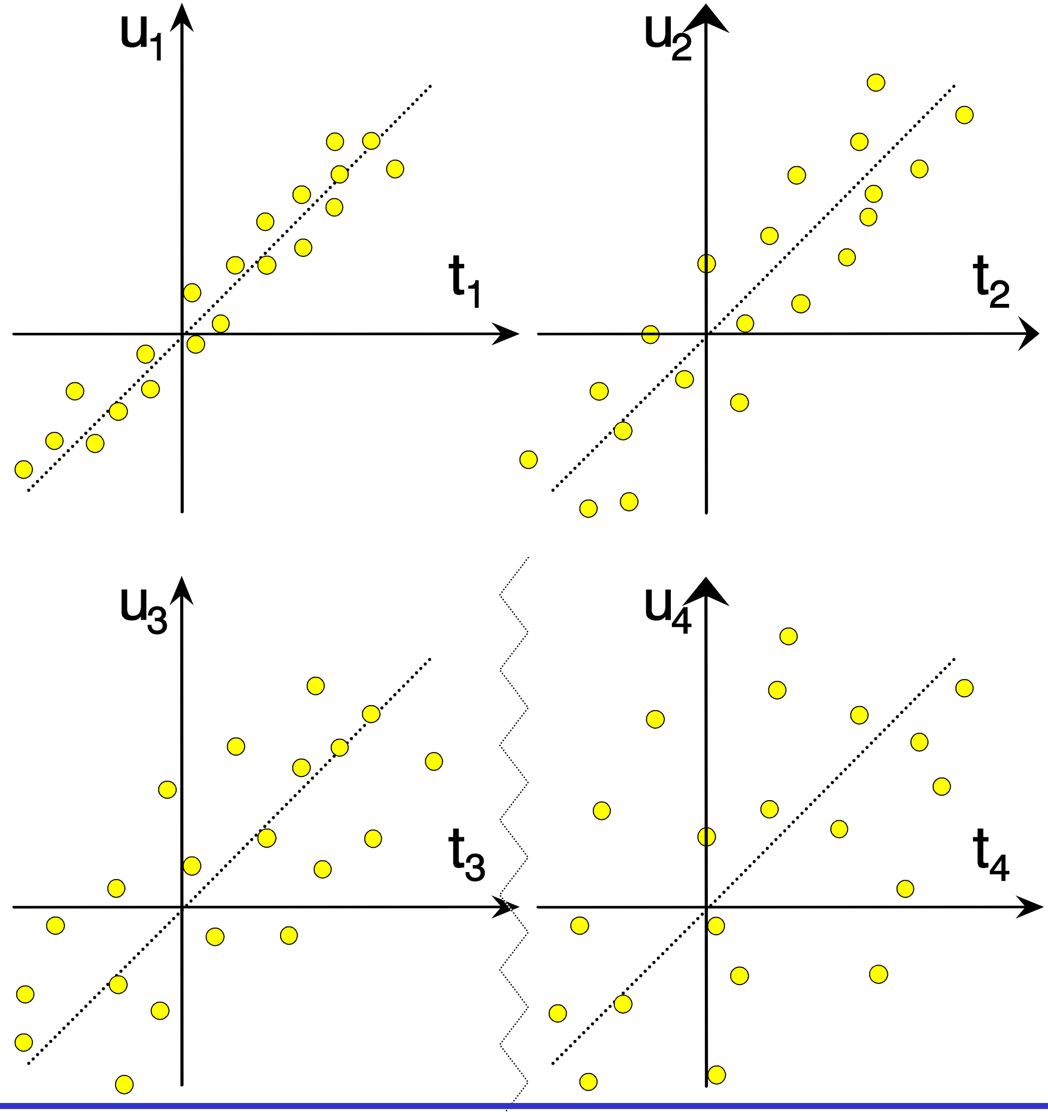


•PLS creates planes (windows) in X and Y space.

• The variability around the X plane is used to calculate **tolerance intervals** within which observations similar to the model samples (the training set) can be found. This is of interest for both classification as well as prediction.

PLS - Geometric interpretation, 9

- **Subsequent plotting of pairs of X and Y scores provides a picture of the correlation structure.**



PLS - Overwiev

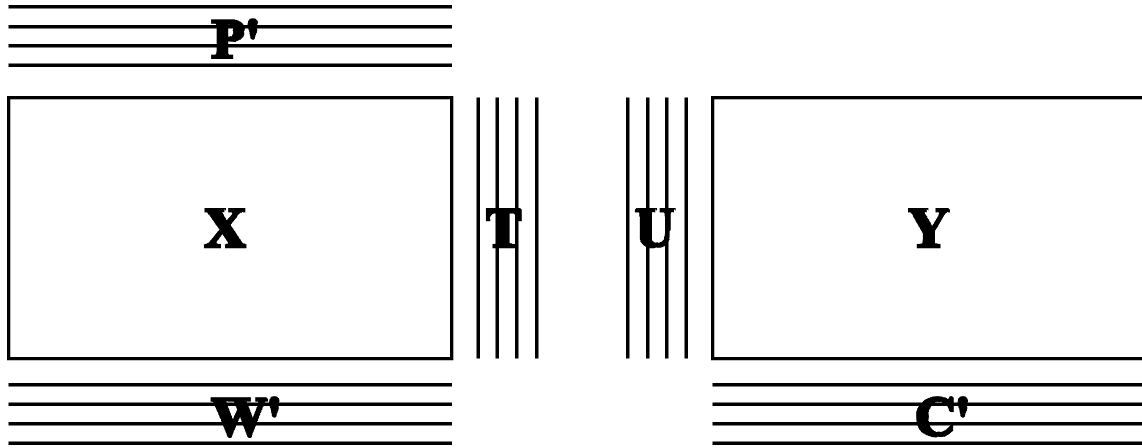
$$\mathbf{X} = \mathbf{1} * \bar{\mathbf{x}} + \mathbf{T} * \mathbf{P}' + \mathbf{E}$$

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

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

(since **U = T + H**)

(inner relation)



PLS

Projection of X that gives a good approximation of X, and correlates to Y

⇔

PCA

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

Properties of PLS parameters

• **For each component:**

1) **t** is linear combinations of **X** with weight **w**

- **t** is a **summary** of the **X variables** that are **correlated to Y**

2) **u** is linear combinations of **Y** with weight **c**

- **u** is a **summary** of the **Y variables**

3) **w** is **the correlation coefficients** the **x variables** and **u**

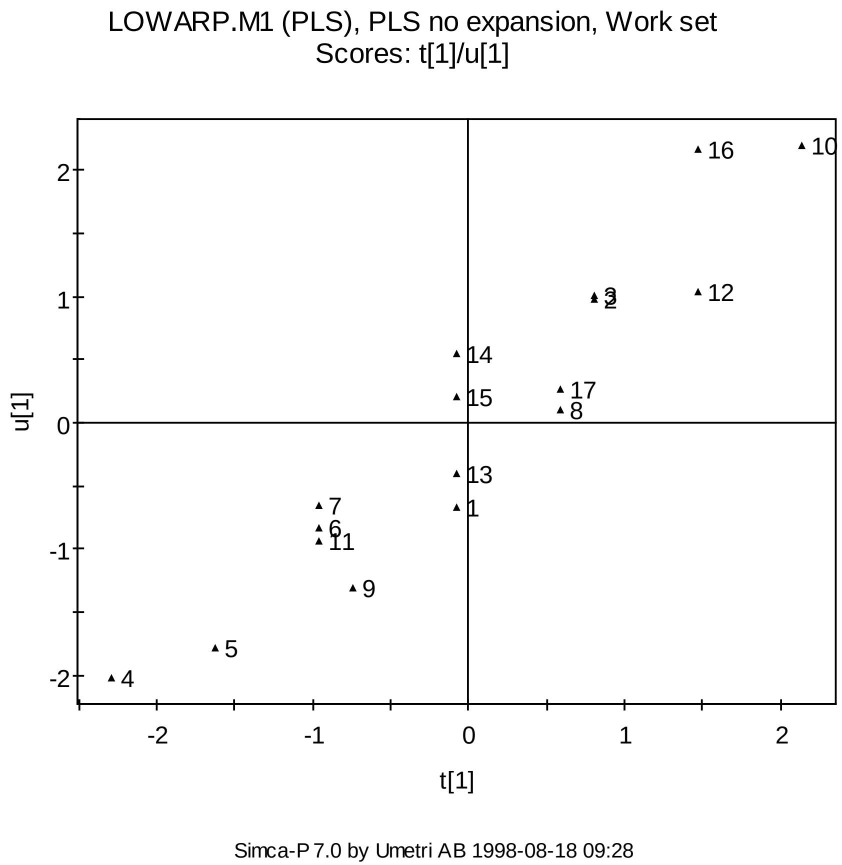
- Columns (variables) in **X** strongly correlated to **Y** get high weights**, w**

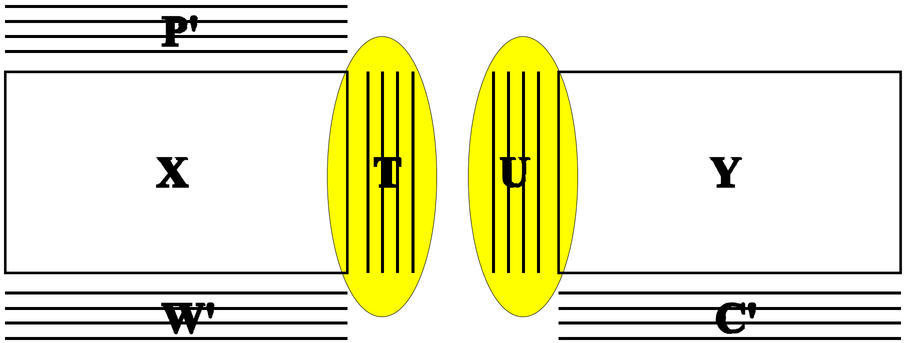
4) **After convergence, for orthogonality:**

- **p** is calculated so that **t*p'** is the best approximation of **X**
- **t*p'** is subtracted from **X** for calculation of the next component.

PLS - Application

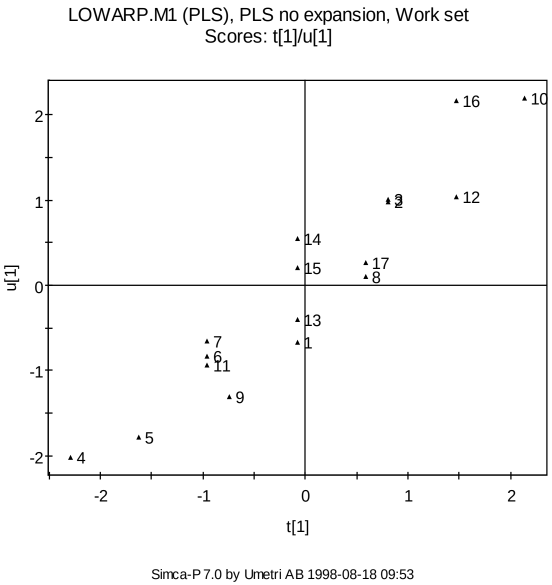
- Investigate correlation structure X/Y; LOWARP example, 17 polymers, 4 X, 14 Y



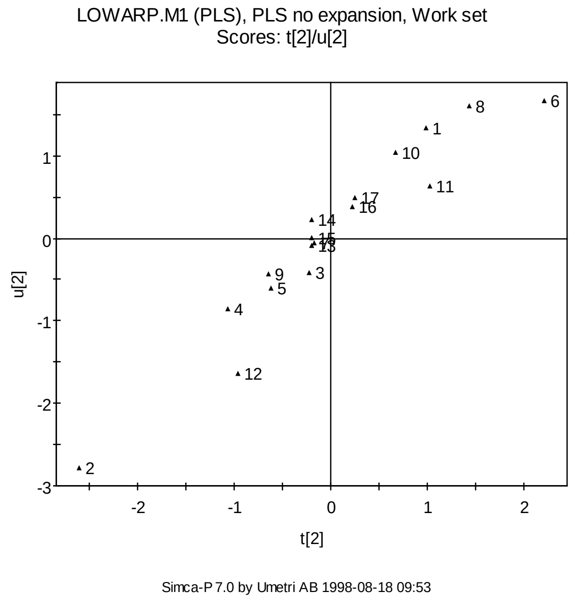


PLS – score plots

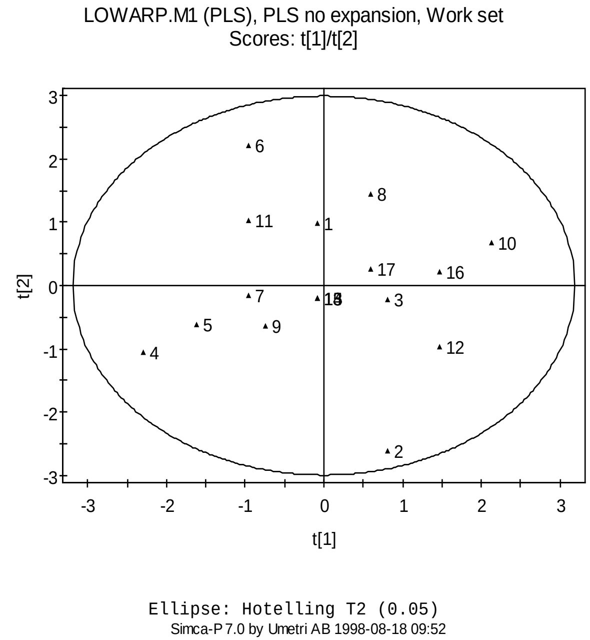
t_1/u_1 show relations between $X(t_1)$ and $Y(u_1)$ in the first dimension



t_2/u_2 show relations between $X(t_2)$ and $Y(u_2)$ in the second dimension

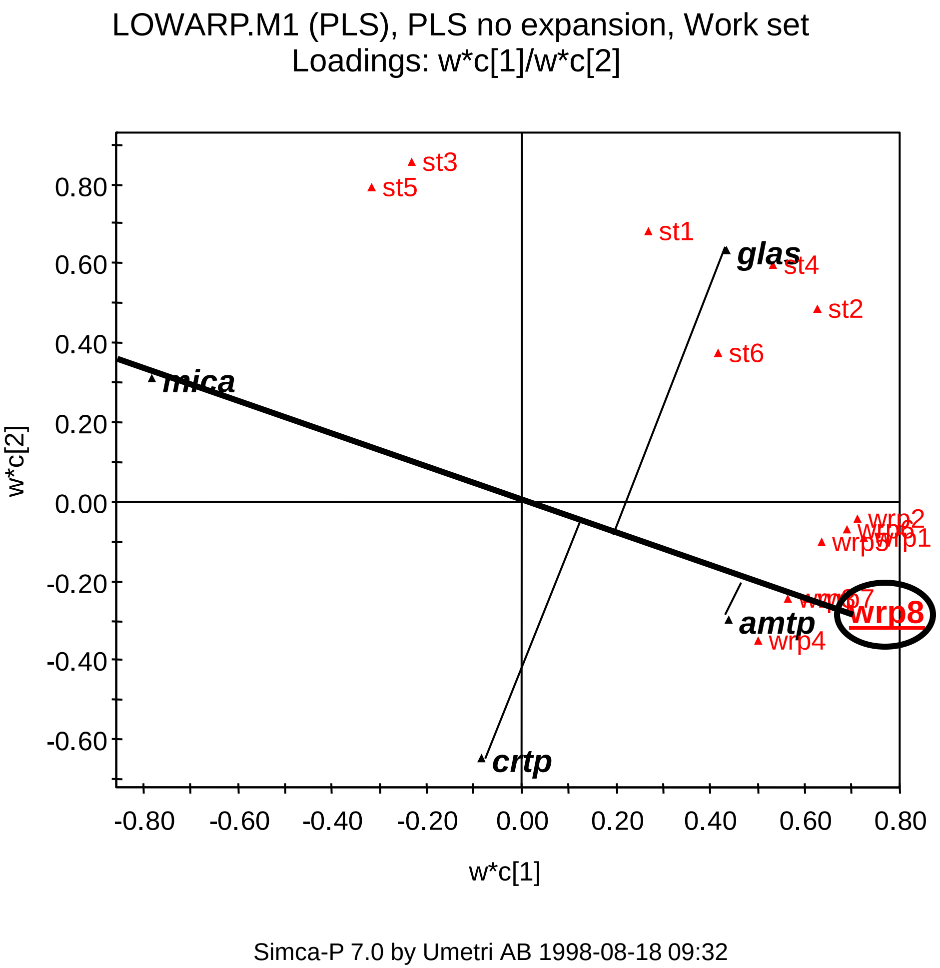


t_1/t_2 show similarities/dissimilarities between observations in two dimensions

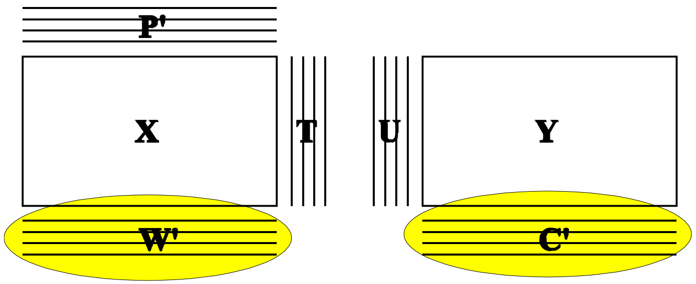


- **u_1/u_2 can also be investigated to find similarities/dissimilarities between observatons (samples) in Y.**

PLS – Interpretation of variable correlations

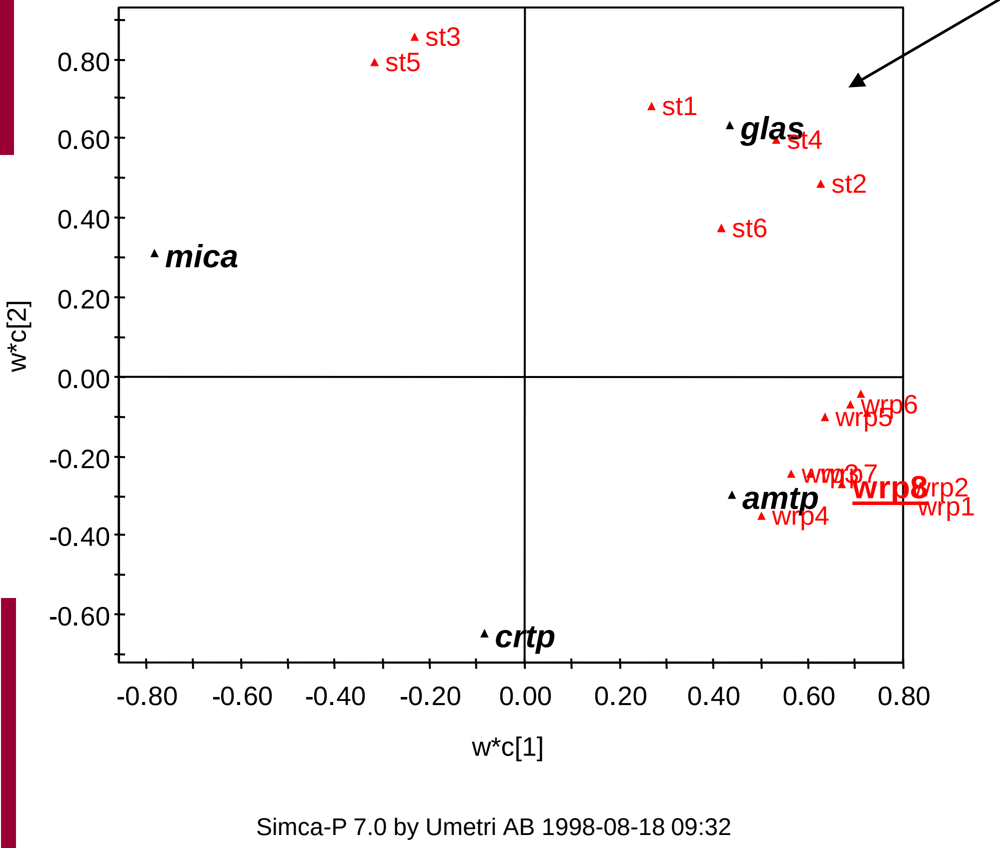


Find an important **y-variable** (e.g. **wrp8**).
Draw a line from **wrp8** through the origin (0,0).
Project all **x-variables** onto the line.
X-variables showing a large distance from the projection to the origin (0,0) are important for explaining **wrp8**.
X-variabler on the same side of (0,0) as **y (wrp8)** have got positive influence (pos. korrelation)
X-variabler on opposite side of (0,0) compared to **y (wrp8)** has got negative influence (neg. korrelation).



PLS – Interpretation of models

Variable correlation between X and Y



a) **loadings**, wc, w*c

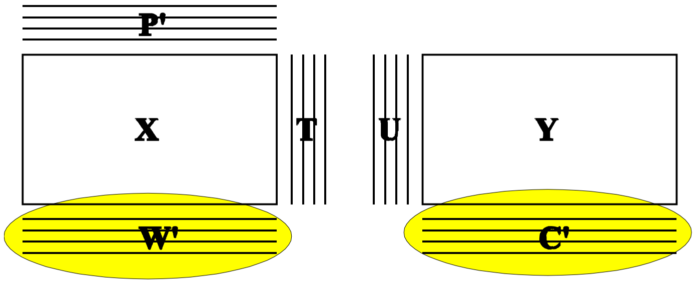
b) **regression coefficients**

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

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

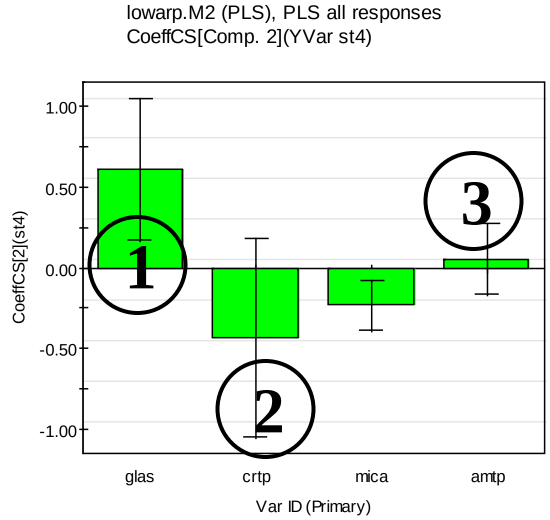
(Manne 1987)

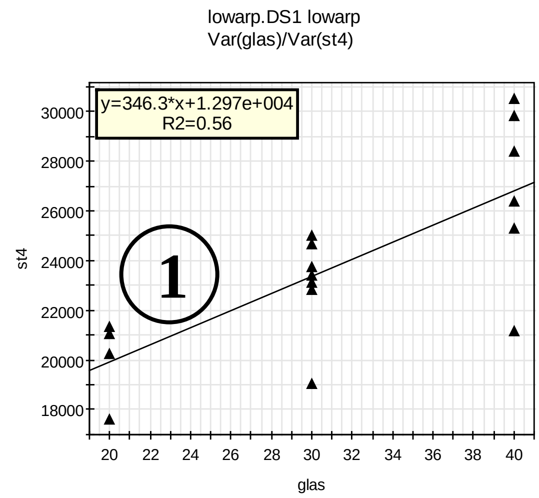
Size and sign of the regressions coefficient (b) defines influence of a x-variable.

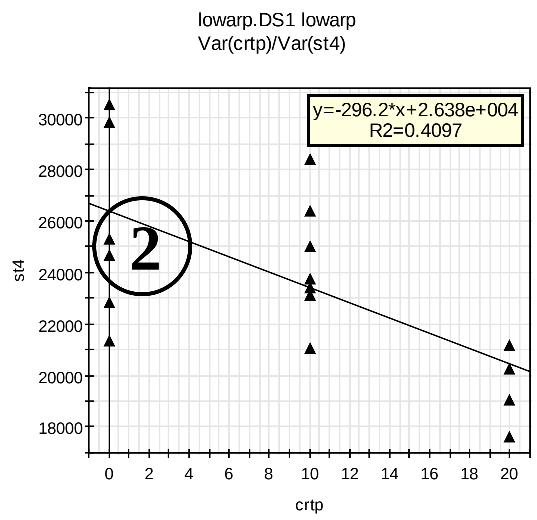


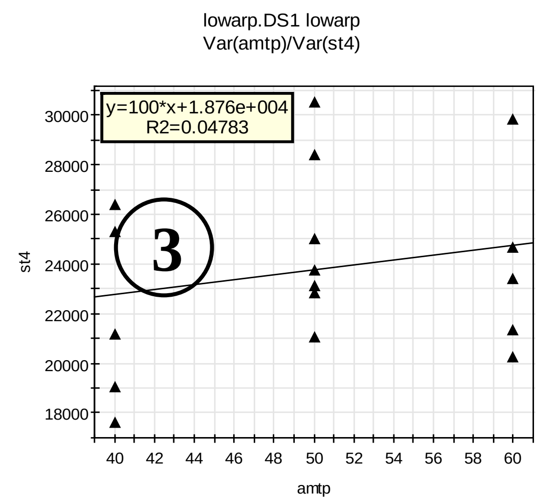
PLS – regression coefficients

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



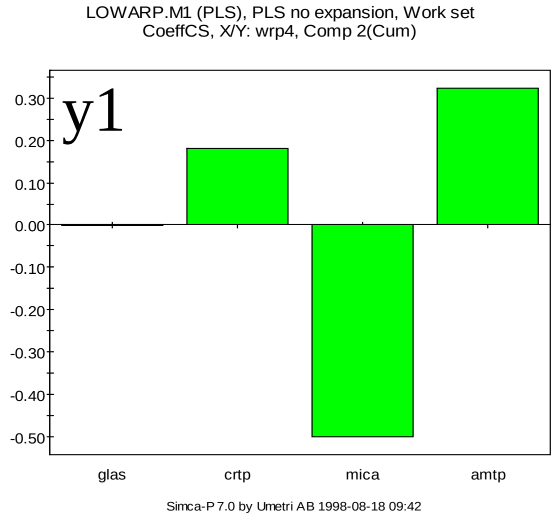


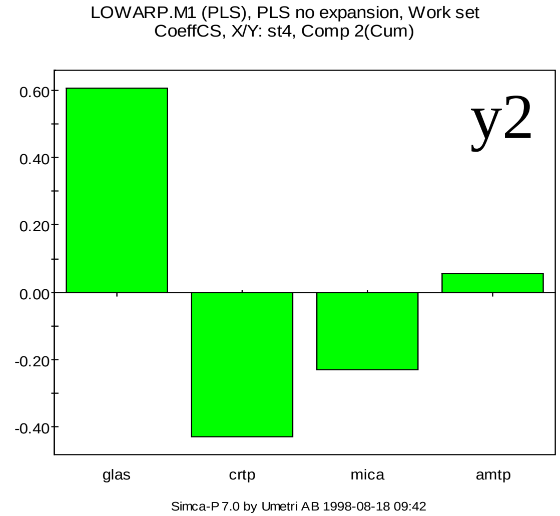


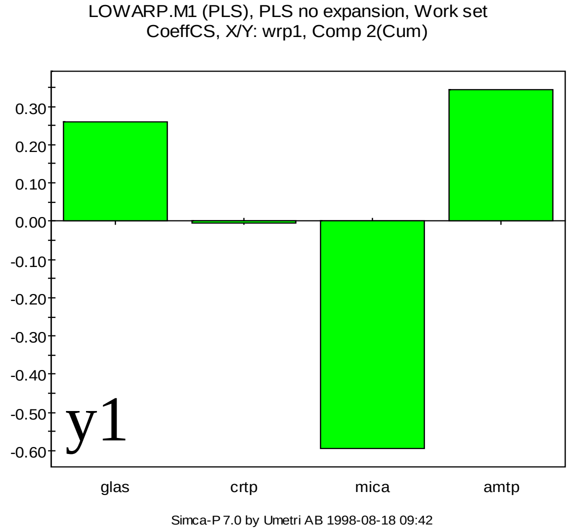


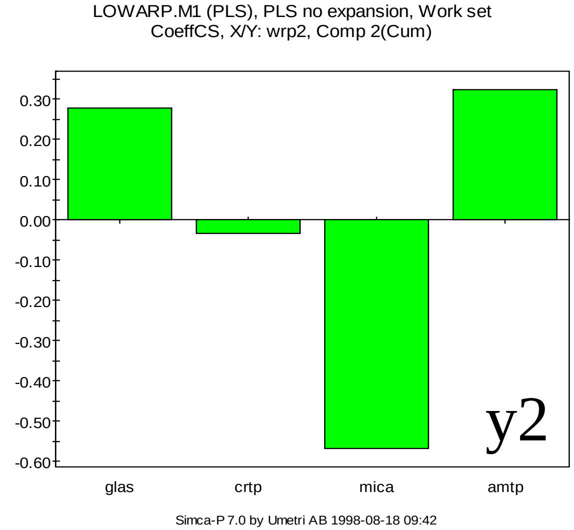
PLS – regression coefficients

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









PLS - Diagnostics

- **Observations** - outliers (strong, moderate)

- **Variables** – which varaibles are well explained?

- **Models** – Cross Validation (CV)

PLS - Diagnostics (Observations)

- **Strong outliers, groups, inhomogeneities,...**

  PLS plots:

  1) X space (t_1,t_2..)

  2) Y space(u_1,u_2,..)

  3) X,Y space(t_1,u_1,..)

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

  Plot DModX against observation number :

  Object X $\rightarrow$ RSD: Distance to Model (DModX).

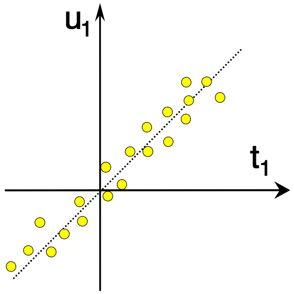
  Check that no observation has got high DModX.

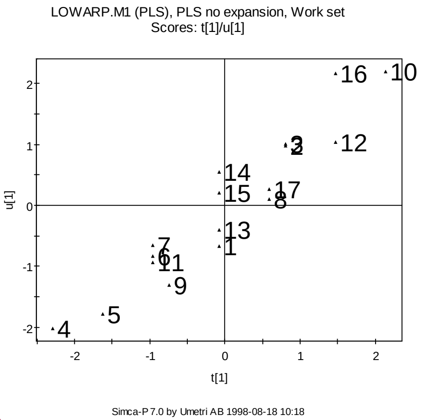
  Plot DModY mot observation number:

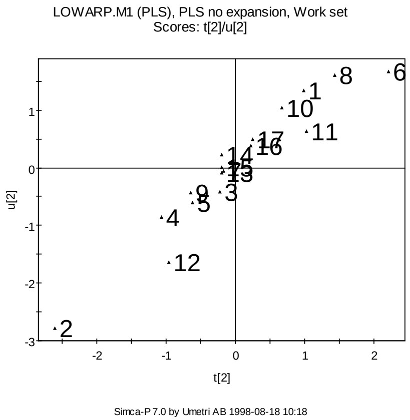
  Check that no observation has got high DModY.

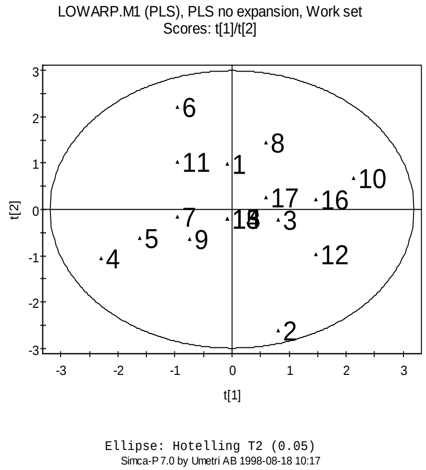
PLS - Strong outliers

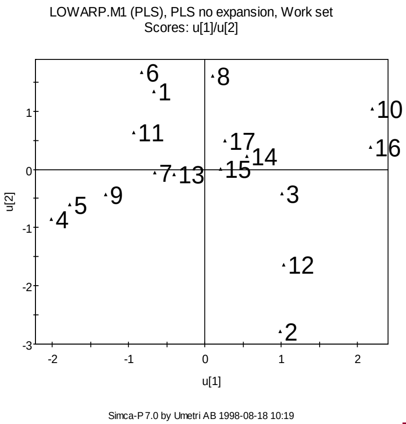
An observation can be an outlier in:











t/u space **t/u space** **t space** **u space**

deviation from correlation structure **dev. in measurements** **dev. in properties**

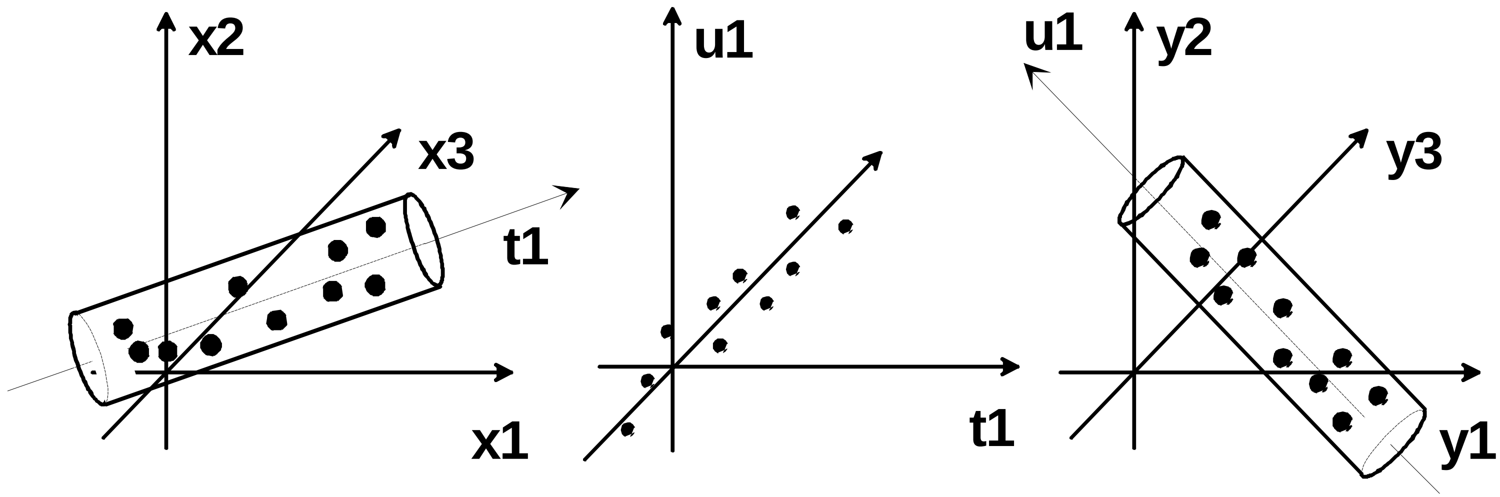
PLS - Moderate outliers

- Moderate outliers can be detected by investigating the PLS residuals, E and F:

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

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

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



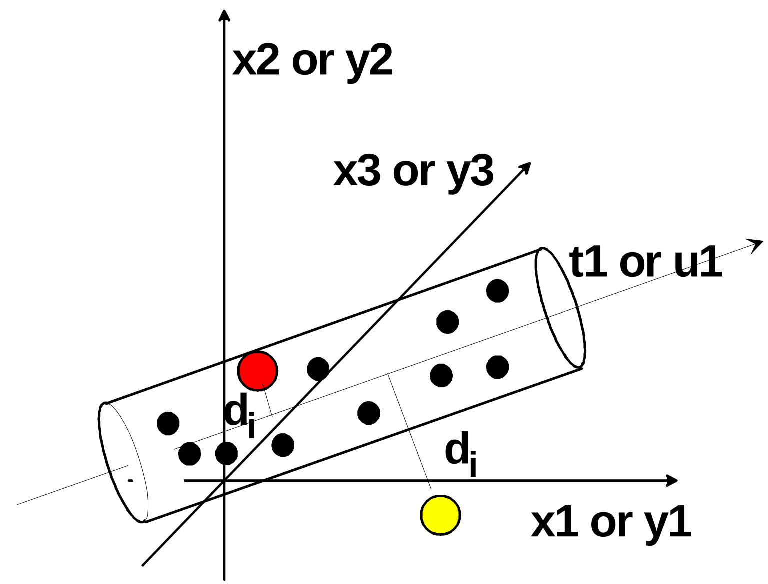
• The residual matrices E and F are used to calculate the diameter for the ”beer cans“ surrounding the data points in X respective Y. (“beer cans” = tolerance limits).

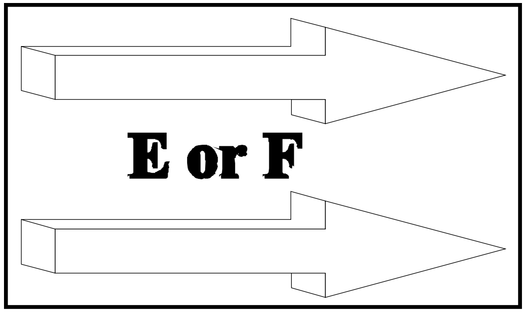
PLS - Moderate outliers

Moderate outliers can be detected by investigating the residual for each sindividual observation (DModX, Y)

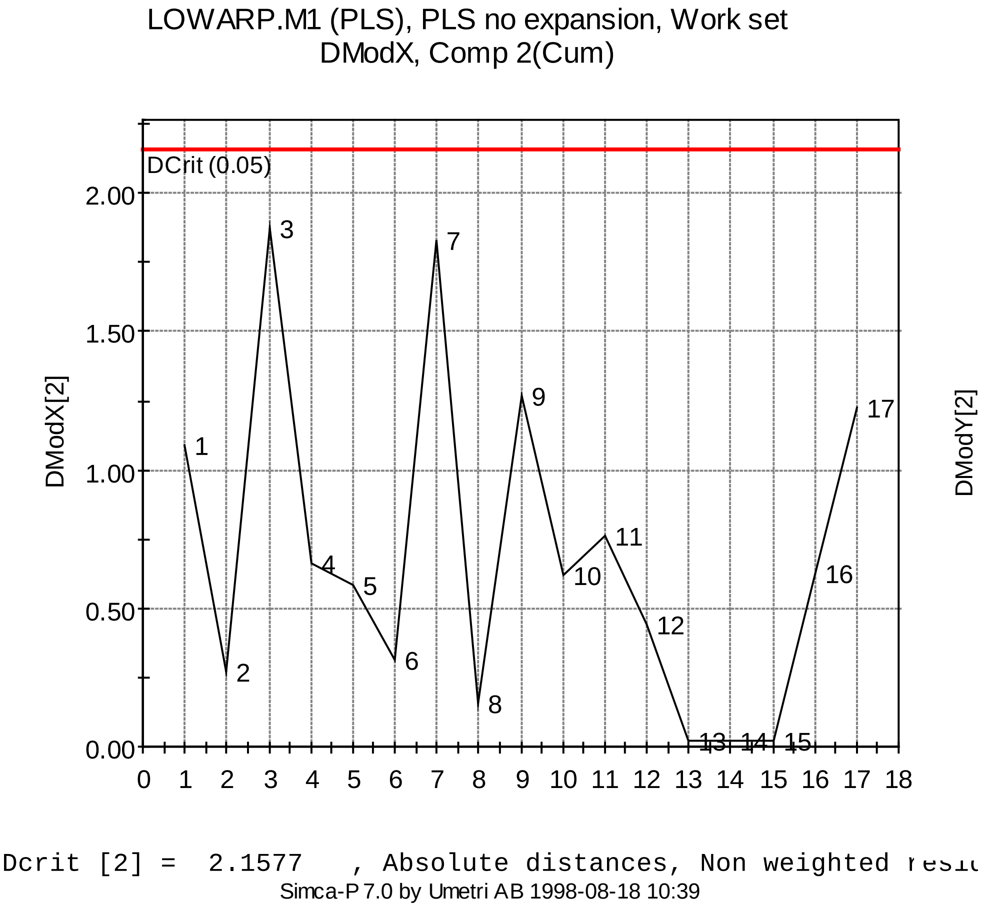
An observation with extremely high DModX, Y compared to the other observations should be discarded as an outlier.

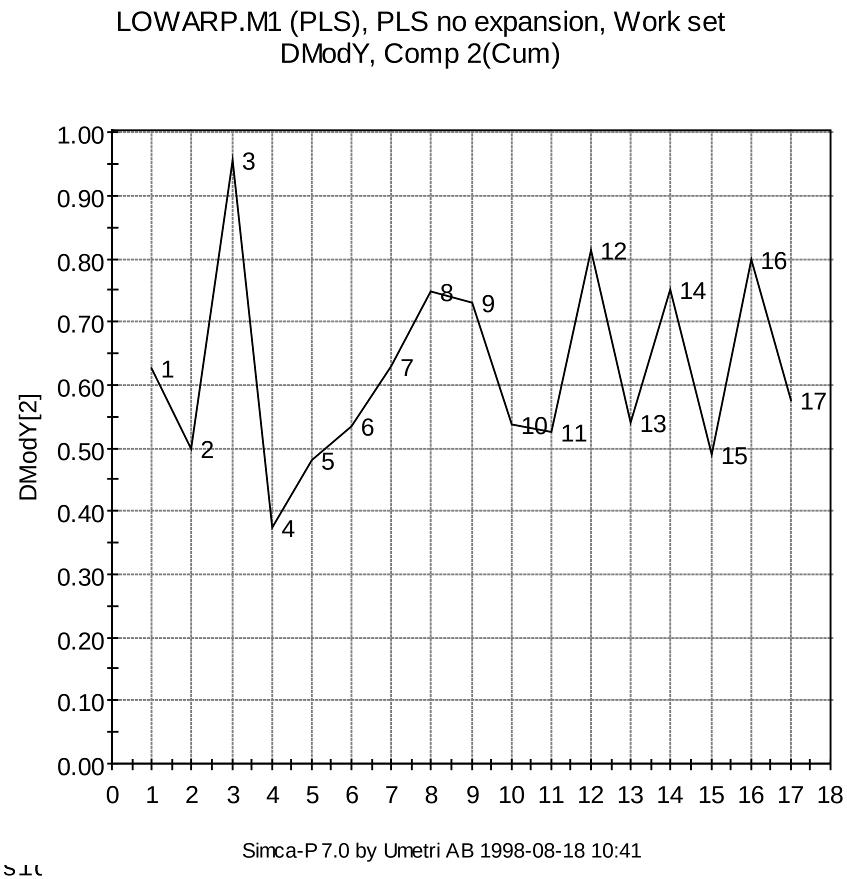
An observation with marginally higher DModX, Y doesn't have to be discarded unless the model is affected in a negative sense.





PLS - Moderate outliers





- There are no moderate outliers in the example above !

PLS - Diagnostics (Models)

• Validation is used to investigate if the existing model is the best alternative from a predictive point of view and to estimate over fit.

SIMCA includes two alternative "internal" validation methods.

1. Cross Validation (CV)

   To define optimal model complexity (number of PLS components).

2. (Permutation of responses)

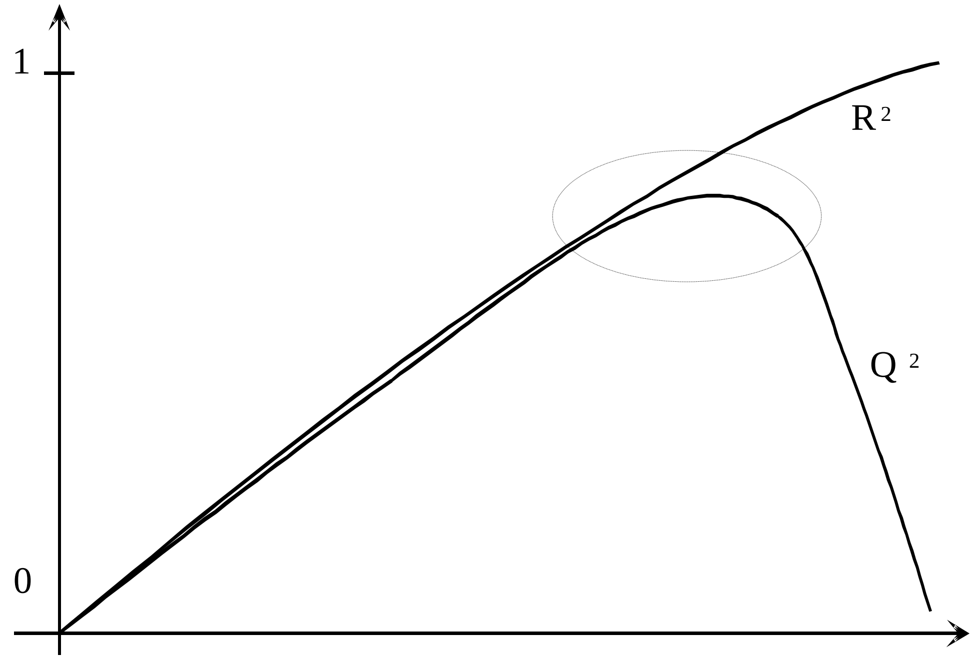
PLS - R2/Q2

- **Question:** How can we decide the optimal number of PLS components for the model?

- **Method:** Cross Validation (CV); CV simulates the predictive ability for a PLS-model.

- A model must not be over fitted, i.e. model noise in it's components.

- Trade of between fit and predictive ability.



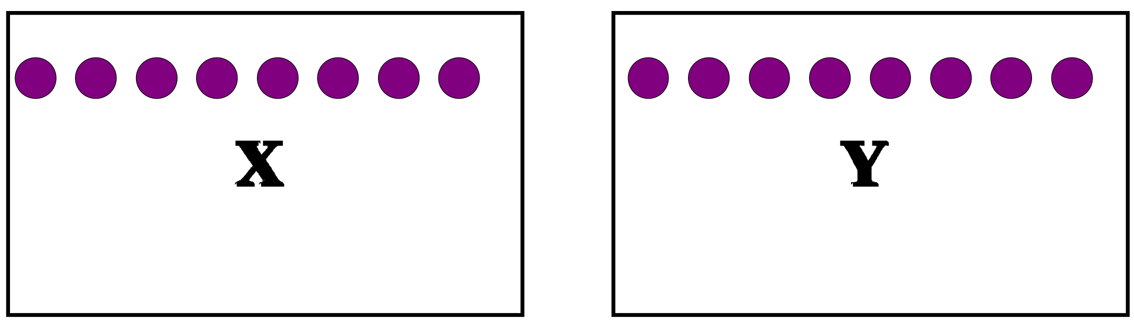
R^2 estimates the fit to the data "goodness of fit"
How much of the variation in **Y** that is described by the model.

Q^2 estimates the predictive ability.
How much of the variation in **Y** that can be predicted by the model.

PLS – Cross Validation (CV)

- Data is divided into G groups, usually 5-10 (default in SIMCA 7 groups).
- A model is fitted with one group excluded.
- The excluded group is predicted by the model ⇒ part-PRESS (Predictive Residual Sum of Squares or Prediction Error SS)
- This is repeated G timesr; and then all part-PRESS are summed to get PRESS.
- If another PLS component (a) enhances the predictive ability compared to (a-1) PLS components, the new component is included in the model.
- **OBS!**: In PLS data is removed row wise. PCA calculates CV Q^2 based on X-data, in PLS based on Y-data.

Data removed row wise!



PLS – R^2 and Q^2

- **PRESS** is the sum of the squared differences between predicted and observed y-values. (based on CV)

$$\mathrm{PRESS} = \sum (y_{im} - \hat{y}_{im})^2$$

- **PRESS can be translated to Q^2, which is without unit as is R^2**

R^2, Q^2 varies between 0 and 1

$Q^2 = 1 - \mathrm{PRESS}/SS_{total}$

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

Q^2 > .5 Good (Depending on appl.)

Q^2 > .9 **Brilliant** (Depending on appl.)

IMPORTANT!

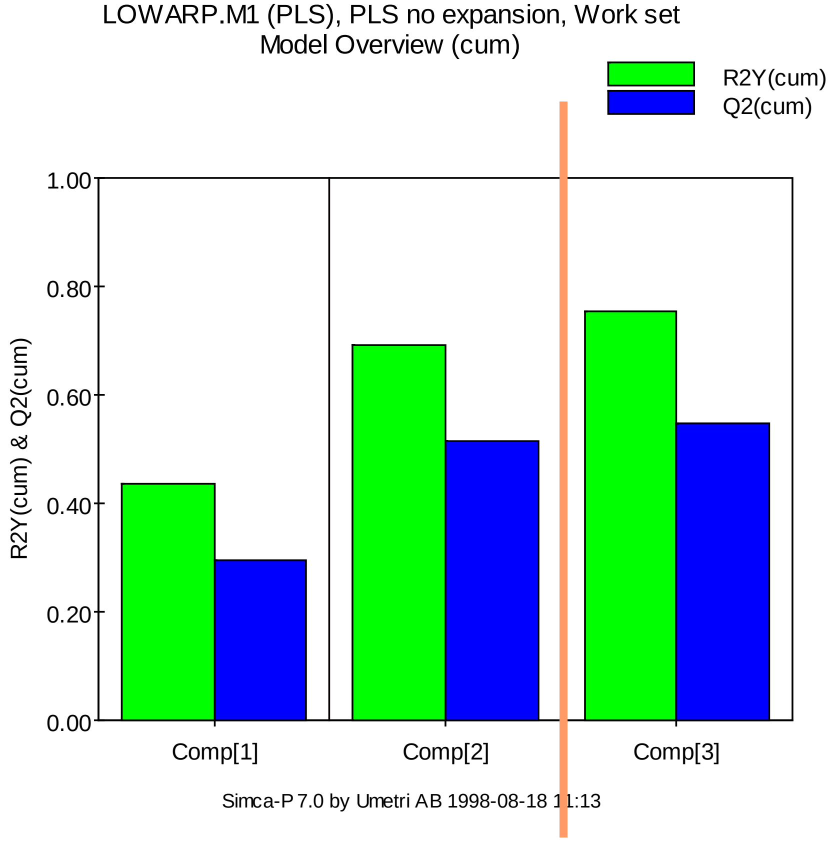
1. A high R^2 is a prerequisite for a high Q^2.

2. High R^2 and Q^2 is wanted.

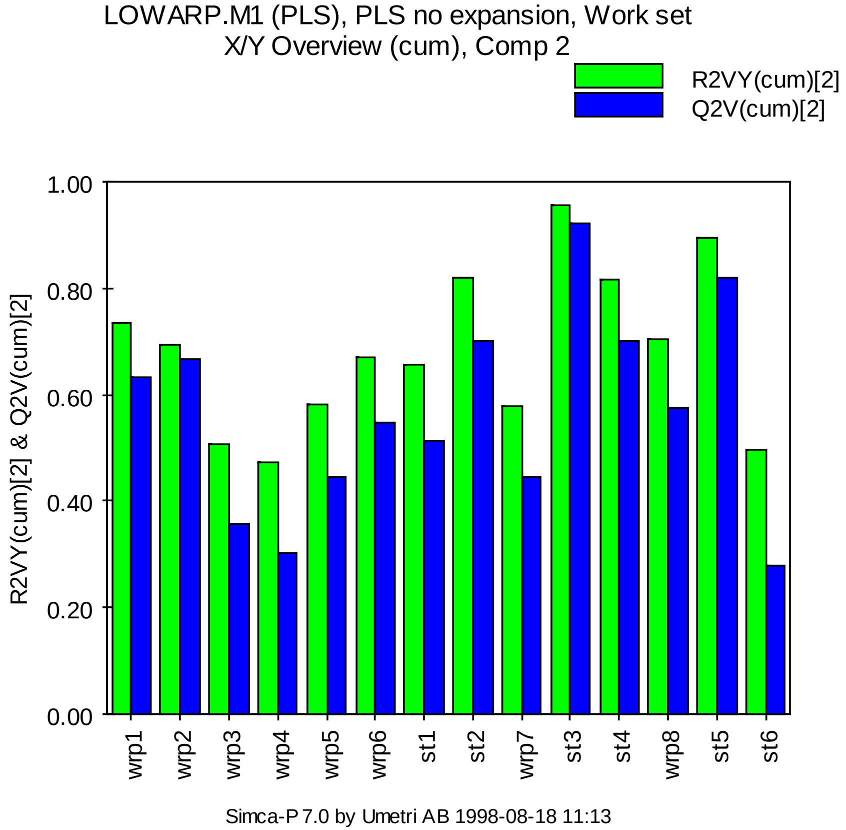
3. The difference between R^2 and Q^2 should not be too large.

PLS – Model complexity - Example

For each PLS component



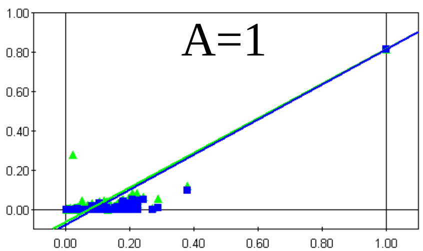
For each Y variable

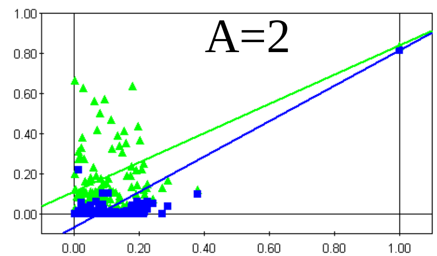


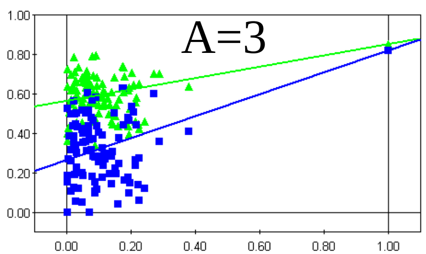
Response permutation - "Validate"

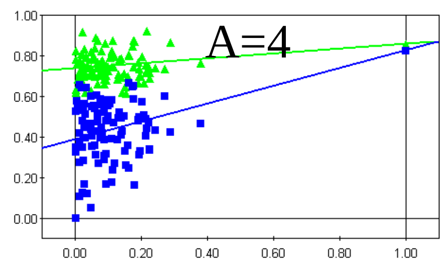
- To check if the existing model is the best predictive alternative and to decide the degree of overfit.
- Rules: Y-axis intercept
$R^2 < 0.3$, and $Q^2 < 0.05$
- If the R^2-line is close to horisontal, this is an indication of overfit

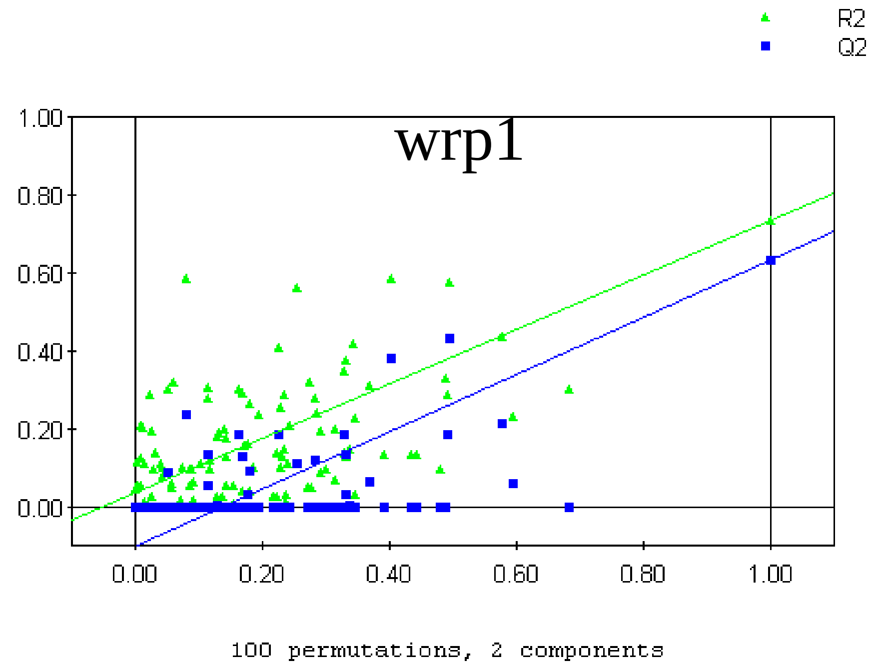
/	Factors					Responses			
/	1	2	3	4		Randomise wrp1 in new columns			
ONum	glas	crtp	mica	amtp	/	wrp1	Wrp1:1	Wrp1:2	Wrp1:3
1	40	10	10	40		0.9	3.7	0.6	0.3
2	20	20	0	60		3.7	0.6	3.6	0.6
3	40	20	0	40		3.6	0.3	1.2	1.2
4	20	20	20	40		0.6	1.2	0.3	3.7
5	20	10	20	50		0.3	0.9	0.9	3.6
6	40	0	20	40		1.2	3.6	3.7	0.9



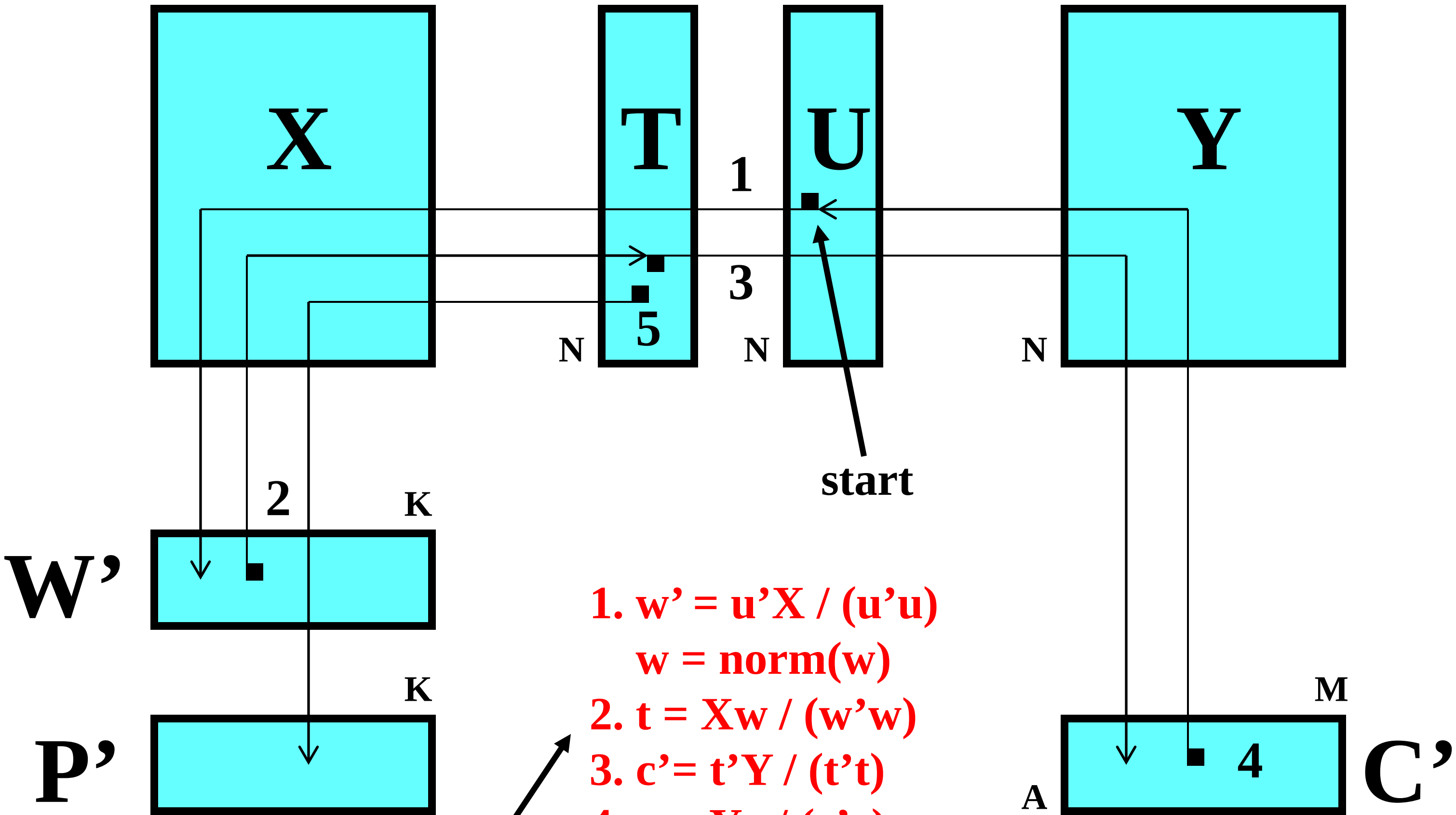








PLS - *NIPALS*



1. $w' = u'X / (u'u)$
   $w = norm(w)$
2. $t = Xw / (w'w)$
3. $c' = t'Y / (t't)$
4. $u = Yc / (c'c)$

---

5. $p' = t'X / (t't)$
6. $E = X - tp'$
7. $F = Y - tc'$

Itererate and check for convergence for $(t_{new}-t_{old}) / t_{old}$

PLS - Summary (predictions)

X→(W,P) T → U → (C) Y

- **Coefficients**
- **Confidence intervals**
- **Residuals**

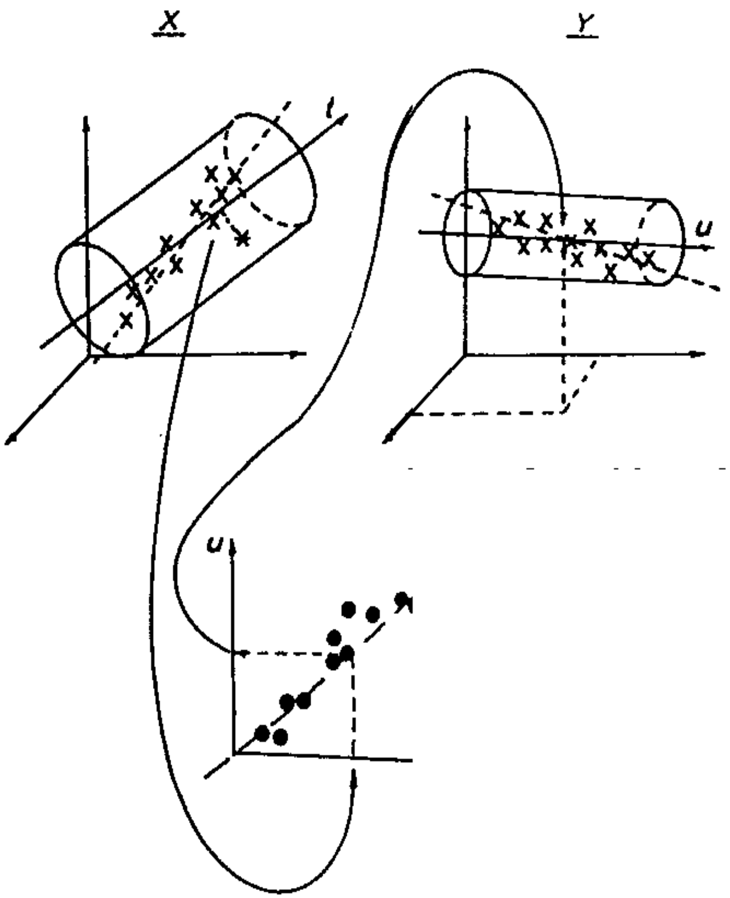
- New observation ⇒ x_i ⇒ PLS Model ⇒

  1) distance to model (DModX) in X-space

  2) y = predicted y

A new observation fits the model (tränings set) if it falls within the tolerance cylinder in X space.

If that's the case the PLS model can be used to predict y values for the new observation.



PLS - Summary

- **Modelling:** The variation in the data tables **X** and **Y** is described by (hyper)-planes + residuals **(E, F)** and an "inner relation" between **U** and **T**.

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

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

$$\mathbf{U} = \mathbf{T} + \mathbf{H}$$

- **Number of components:** Cross Validation, Q^2 compass
- **Residuals:** DModX, Y (distance to model) - Moderate outliers
- **Strong outliers:** X (t-scores), Y (u-scores), Correlation XY (t/u)
- **Similarities/Dissimilarities:** Observations (t-, u-scores)
  Variables (loadings p, weights w och c)

PLS - Applications

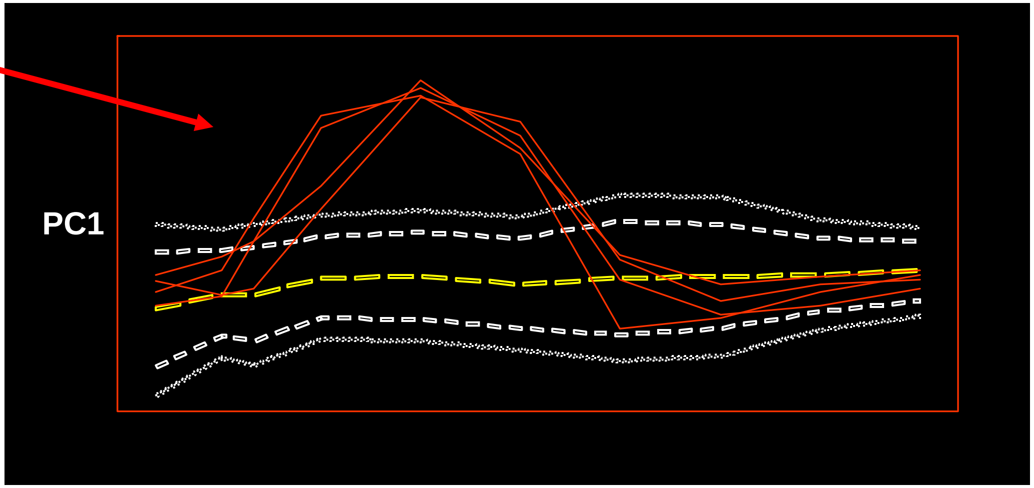
	X	Y
• Process modelling	Process variables	Results
• Structure-activity	Structure descript.	Biol. activ.
– Pharmaceutical optimisation		
– Pesticides		
– Toxicity, ...		
Composition-property	Composition	Outcome
– Polymer blends		
– Cosmetics		
– Food, drink		

PLS - Applications

	X	**Y**
• **Multivariate calibration**	Signals	Conc.
– Protein, Fat	Spectra	Amounts
– Wood, Pulp, Fibres		
– Alcohol in wine		
– Energy		
– ….		
• **Optimisation**	Factors	Responses
• **Discriminant analysis**	Variables	Dummy var.

PLS – Special Cases

- **PLS time series (Batch)**



- Non-linear PLS
- **Multiblock PLS**
  - Hierarchical models
  - Consensus PCA
  - Hierarchical block structures in X and/or Y
- Multi-way PLS
- **Orthogonal Signal Correction (OSC) – O- and O2-PLS better versions**