Multivariate statistical methods for the analysis, monitoring and optimization of processes

Jay Liu
Dept. of Chemical Engineering
Pukyong National University

Some properties of PLS models

- At convergence, **w**, **u**, **t**, **c** don’t change.

$$\mathbf{w} = \mathbf{X}^T\mathbf{u}\big/\mathbf{u}^T\mathbf{u}$$

Substitute for **u** $$= \mathbf{X}^T\mathbf{Y}\mathbf{c}\Big/\Big(\big(\mathbf{c}^T\mathbf{c}\big)\big(\mathbf{u}^T\mathbf{u}\big)\Big)$$

Substitute for **q** $$= \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t}\Big/\Big(\big(\mathbf{t}^T\mathbf{t}\big)\big(\mathbf{c}^T\mathbf{c}\big)\big(\mathbf{u}^T\mathbf{u}\big)\Big)$$

Substitute for **t** $$= \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}\Big/\Big(\underbrace{\big(\mathbf{w}^T\mathbf{w}\big)\big(\mathbf{t}^T\mathbf{t}\big)\big(\mathbf{c}^T\mathbf{c}\big)\big(\mathbf{u}^T\mathbf{u}\big)}_{\text{Constant, denote as }\lambda}\Big)$$

$$\therefore \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w} = \lambda\mathbf{w}$$

➔ **w** is eigenvector of $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$

Some properties of PLS models

- Also,
  - **t** is eigenvector of $\mathbf{X}^T\mathbf{X}\mathbf{Y}\mathbf{Y}^T$.
  - **u** is eigenvector of $\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{X}^T$.
  - **q** is eigenvector of $\mathbf{Y}^T\mathbf{X}\mathbf{X}^T\mathbf{Y}$.

- Orthogonal properties

$$\mathbf{w}_i^T\mathbf{w}_j = 0 \quad (i \neq j)$$

$$\mathbf{t}_i^T\mathbf{t}_j = 0 \quad (i \neq j)$$

$$\mathbf{w}_i^T\mathbf{p}_j = 0 \quad (i < j)$$

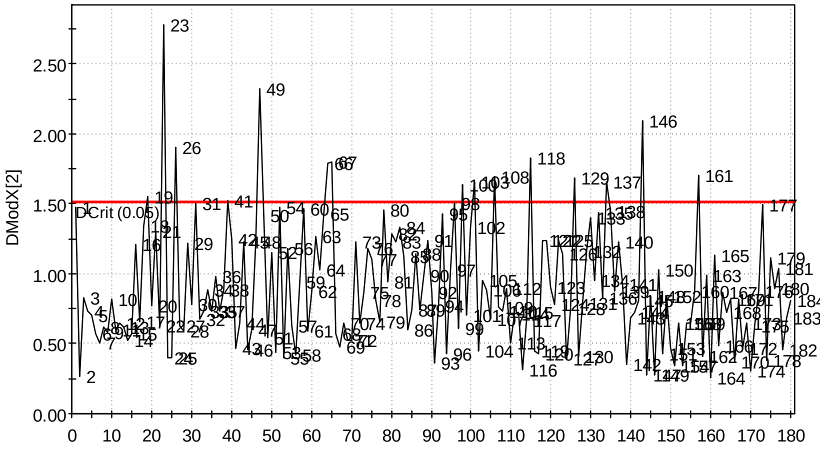
Residuals

- Measure of size of residuals (same as in PCA)
  - $R^2_{X,k}$ measures how well the model describe the variable (x_k)
  - $R^2_{Y,m}$ measures how well the model describe the variable (y_m)
    - RV2X and RV2Y in Simca-p
  - $Q^2_{X,k}$ measures how well the model predict the variable (x_k)
  - $Q^2_{X,k}$ measures how well the model predict the variable (y_m)

  - $R^2 = 1 – [SS_{residuals}/SS_{data}]$ SS = sum of squares
  - $Q^2 = 1 – [SS_{predictive\ resid.}/SS_{data}] = 1 - [PRESS/SS_{data}]$

Residuals

- Residuals of observations (row-wise)
  - Same as PCA, but two spaces, X and Y
  - X-residuals, $\mathbf{E} = \mathbf{X} - \mathbf{TP}^{T}$
    row SD = DModX_i
    column criterion $R^2_{X,k}$



  - Y-residuals, $\mathbf{F} = \mathbf{Y} - \mathbf{TC}^{T}$
    row SD = DModY_i
    column criterion $R^2_{Y,m}$

  - Critical values of DMOdX/DModY from F-distribution

Cross-validation

- Analogous to PCA, PLS model dimensionality can be chosen by CV
  - Data (rows of X and Y) divided into G groups (~7)
  - Model estimated for data minus one group (G rounds)
  - Y of deleted group predicted by model
  - PRESS (prediction error sum of squares) = $\Sigma(y_i - y_{ip})^2$
  - y_{ip} = predicted by model estimated from data after deleting the ith observation

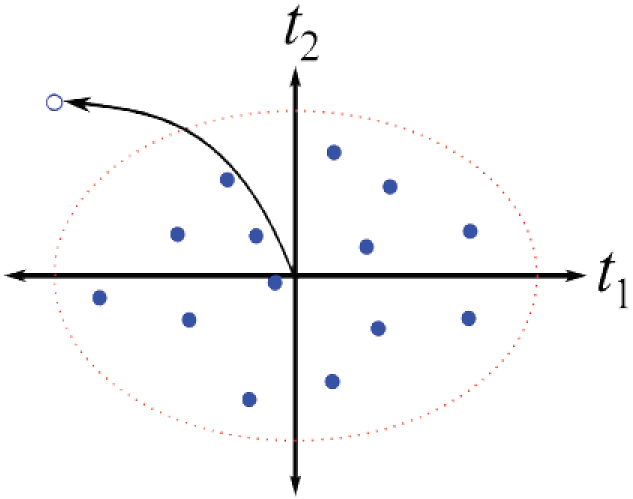
VIP (Variable importance for the projection)

- VIP is a weighted combination over all components of the squared PLS weights w_{ak}.
- SSY_a/SSY_{tot} is amount of Y variance explained by component a.
- Suggestions for usage
  - "Normal" VIP value is 1.0.
  - VIP < 0.5 indicates unimportant X's in explaining Y & the projection in X

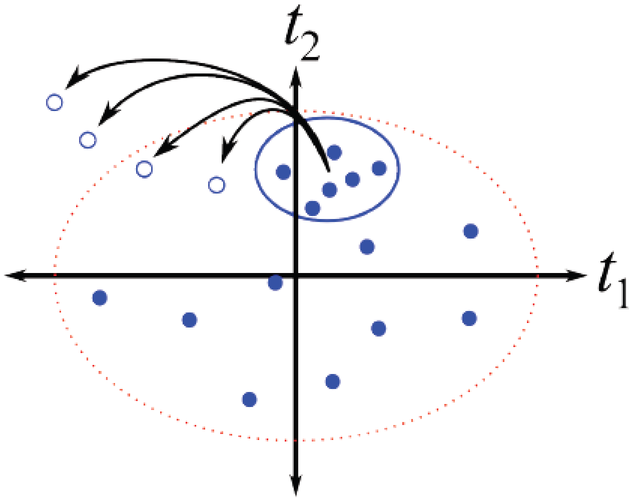
$$VIP_k^2 = K\sum\nolimits_a \left(w_{ak}^2 SSY_a \right) \Big/ SSY_{tot}$$

Contribution plot

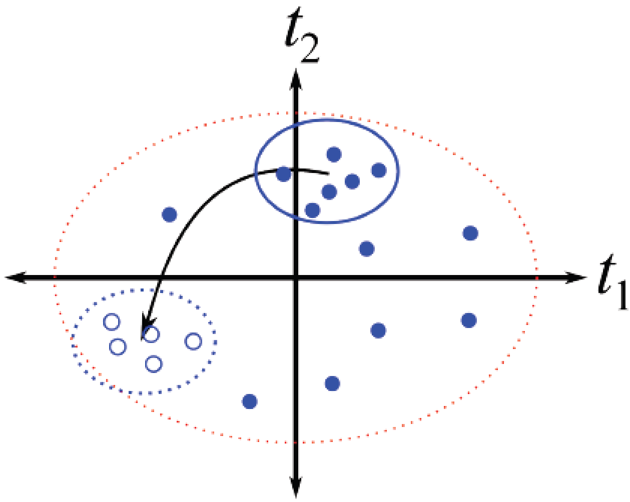
- Same as PCA
    - Also have for Y variables



From the model center to a point



Four seperate contribution plots to learn why the sequence of deviations occurred



From one group to another group

PLS regression coefficients

- PLS model

$$\begin{aligned}\mathbf{Y} &= \mathbf{T}\mathbf{C}^T + \mathbf{F} \\ &= \mathbf{t}_1\mathbf{c}_1^T + \mathbf{t}_2\mathbf{c}_2^T + \cdots + \mathrm{F} \\ &= \mathbf{X}\mathbf{w}_1\mathbf{c}_1^T + \left(\mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T\right)\mathbf{w}_2\mathbf{c}_2^T + \mathbf{F}\end{aligned}$$

Making all substitutions for **t**’s

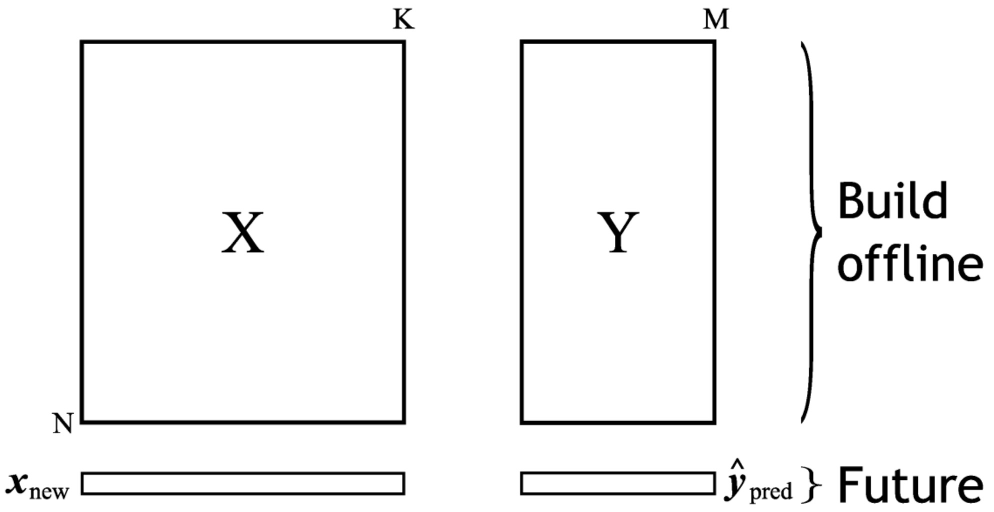
$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{F} \quad \text{where} \quad \mathbf{B} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{C}^T$$

i.e.,

$$y_m \cong b_{1m}x_1 + b_{2m}x_2 + \cdots + b_{Km}x_K$$

Size and sign of scaled and centered regression coefficients (b_{km}) indicates influence of x_k term on model for y_m.

Prediction via PLS model



Build offline

Future

$$
\begin{aligned}
t_{1,\text{new}} &= \mathbf{x}_{\text{new}}^{T} \mathbf{w}_1 \\
\mathbf{x}_{\text{new}}^{T} &= \mathbf{x}_{\text{new}}^{T} - t_{1,\text{new}} \mathbf{p}_1^{T} \\
t_{2,\text{new}} &= \mathbf{x}_{\text{new}}^{T} \mathbf{w}_2 \\
\mathbf{x}_{\text{new}}^{T} &= \mathbf{x}_{\text{new}}^{T} - t_{2,\text{new}} \mathbf{p}_2^{T} \\
\textit{etc} &
\end{aligned}
$$

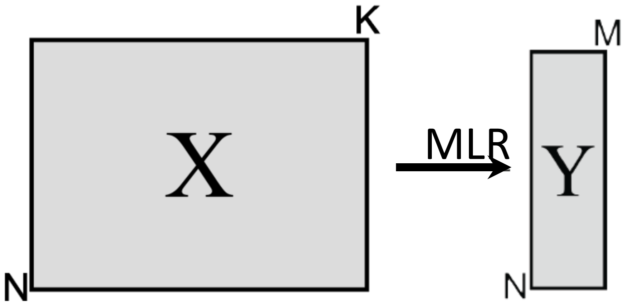
Using this approach, we not only get prediction of y but also get t_a's and DmodX.

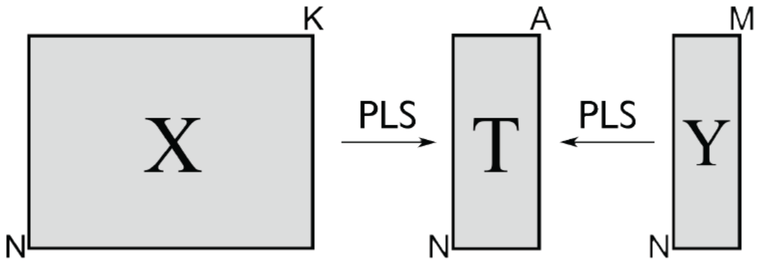
Collect all the $t_{a,\text{new}}$ score values in $\mathbf{t}_{\text{new}}$

Then, $\hat{\mathbf{y}}_{\text{new}}^{T} = \mathbf{t}_{\text{new}}^{T} \mathbf{C}^{T}$

Relation to MLR

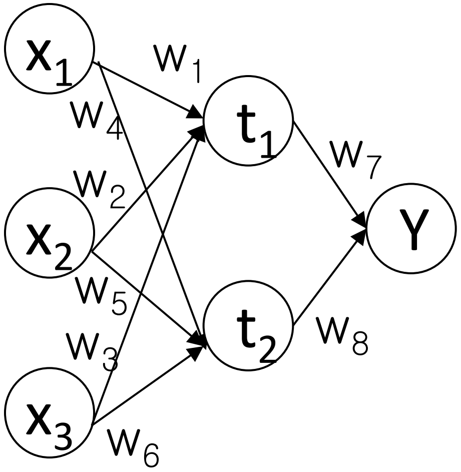
- PLS contains MLR as a special case
  - When the X variables are few and fairly independent
  - And A → K (A is the number of PLS component)
  - Then T → reformulation of X
  - PLS → MLR





Relation to Neural Networks

- In the linear case:
  - $Y = \Sigma t_a c_a$
  - $t_a = \Sigma x_k w_{ak}$ *



Identical to PLS, but PLS gives a unique solution (the t_a's are orthogonal and anchored due to modeling of the x-space)

(McAvoy & Qin, Computers and Chemical Engineering, 16(1992) 379-391)

Tutorials

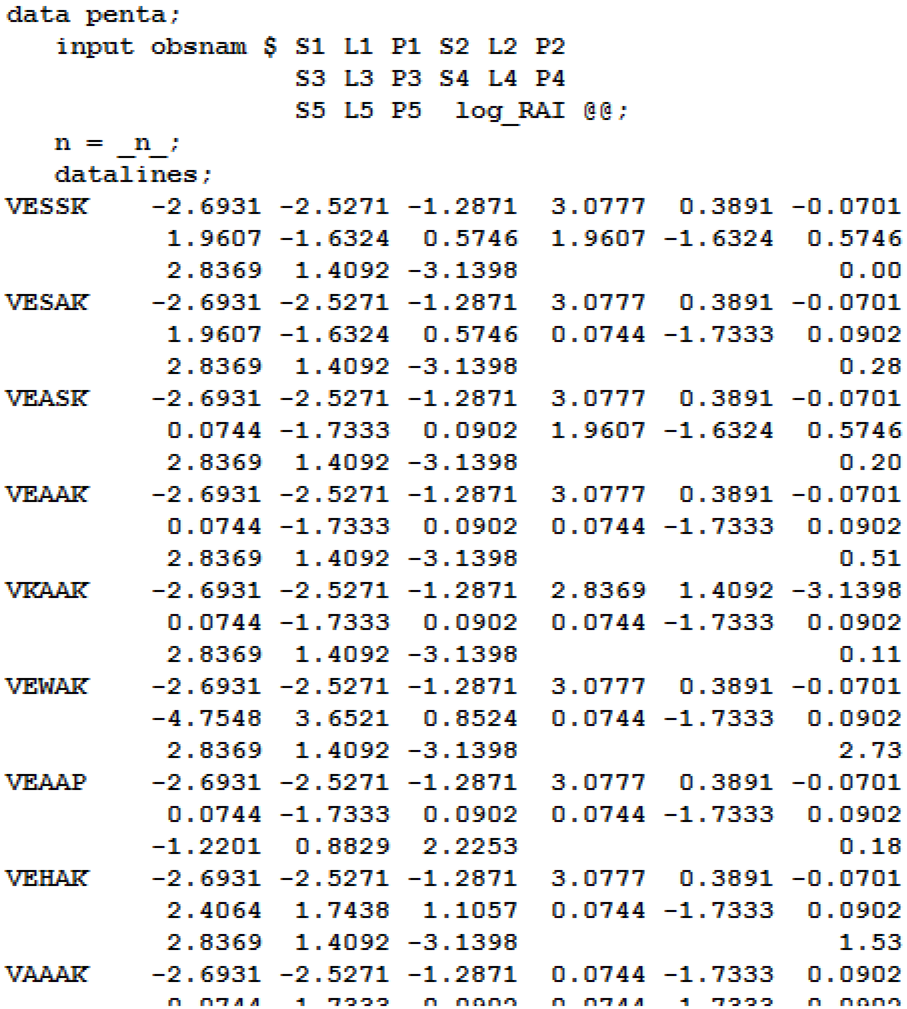
- Drug discovery
  - New drugs: chemicals that are biologically active.
  - Testing chemicals for biological activity is very expensive.
  - Prediction of biological activity from cheaper chemical measurements is desirable
  - Measurements: size, lipophilicity, and polarity at various sites on the molecule
- Dataset
  - 30 chemical compounds
  - 16 measurements including the activity (represented by the logarithm of relative activity)

Originally from

Ufkes *et.al*. (1978), "Structure-Activity Relationships of Bradykinin-Potentiating Peptides," European Journal of Pharmacology, 50, 119.

Ufkes *et al*. (1982), "Further Studies on the Structure-Activity Relationships of Bradykinin-Potentiating Peptides," European Journal of Pharmacology, 79, 155.

Tutorials



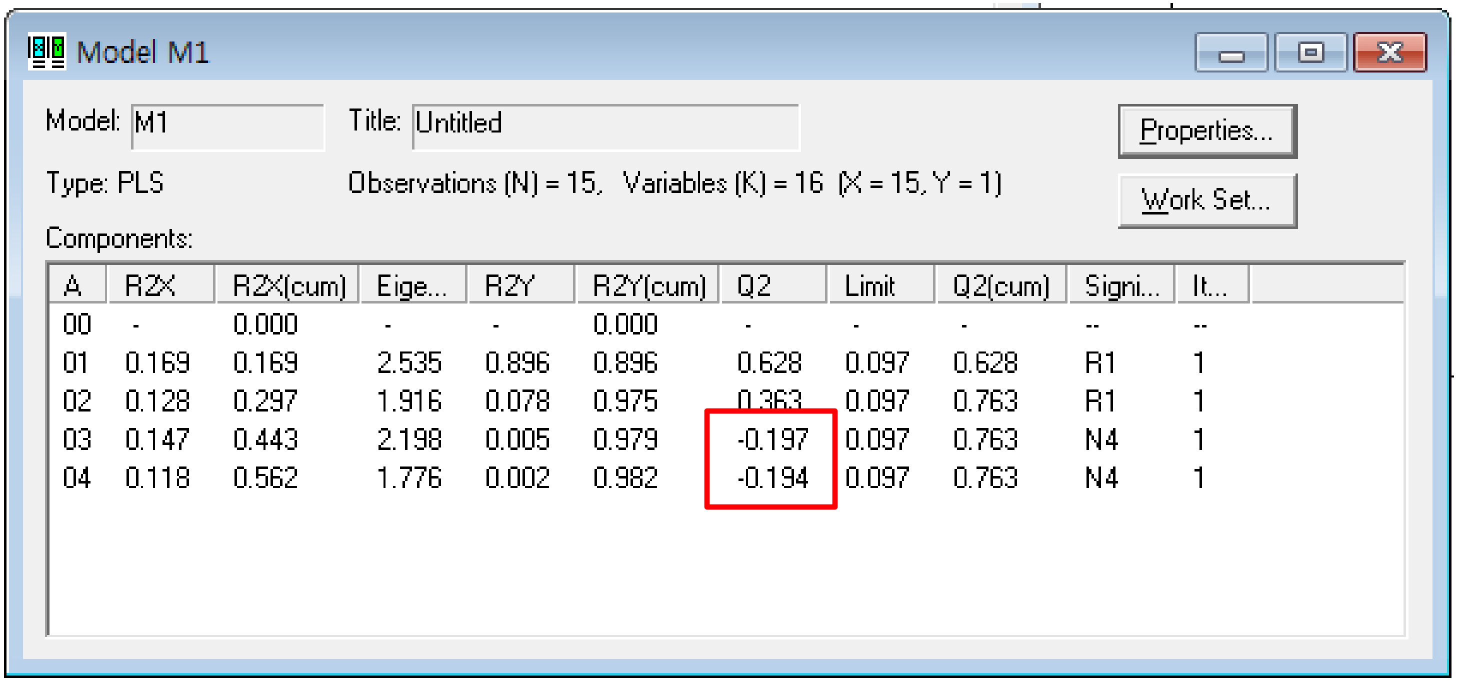
```
data penta;
   input obsnam $ S1 L1 P1 S2 L2 P2
                  S3 L3 P3 S4 L4 P4
                  S5 L5 P5  log_RAI @@;
   n = _n_;
   datalines;
VESSK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
          1.9607 -1.6324  0.5746  1.9607 -1.6324  0.5746
          2.8369  1.4092 -3.1398                    0.00
VESAK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
          1.9607 -1.6324  0.5746  0.0744 -1.7333  0.0902
          2.8369  1.4092 -3.1398                    0.28
VEASK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
          0.0744 -1.7333  0.0902  1.9607 -1.6324  0.5746
          2.8369  1.4092 -3.1398                    0.20
VEAAK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
          0.0744 -1.7333  0.0902  0.0744 -1.7333  0.0902
          2.8369  1.4092 -3.1398                    0.51
VKAAK    -2.6931 -2.5271 -1.2871  2.8369  1.4092 -3.1398
          0.0744 -1.7333  0.0902  0.0744 -1.7333  0.0902
          2.8369  1.4092 -3.1398                    0.11
VEWAK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
         -4.7548  3.6521  0.8524  0.0744 -1.7333  0.0902
          2.8369  1.4092 -3.1398                    2.73
VEAAP    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
          0.0744 -1.7333  0.0902  0.0744 -1.7333  0.0902
         -1.2201  0.8829  2.2253                    0.18
VEHAK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
          2.4064  1.7438  1.1057  0.0744 -1.7333  0.0902
          2.8369  1.4092 -3.1398                    1.53
VAAAK    -2.6931 -2.5271 -1.2871  0.0744 -1.7333  0.0902
```

Tutorials

- Goal
  - To predict biological activity with chemical measurements (that are easily available)
  - To understand latent structure of chemical measurements
  - To find which measurements are more important in predicting biological activity

Tutorials

- Two components seems adequate.



Model: M1 Title: Untitled

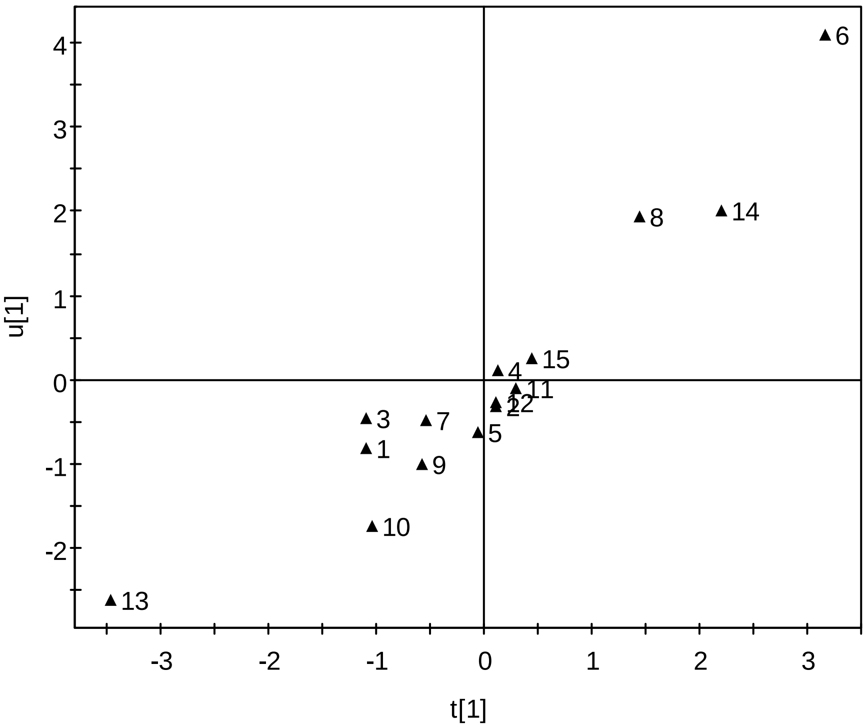
Type: PLS Observations (N) = 15, Variables (K) = 16 (X = 15, Y = 1)

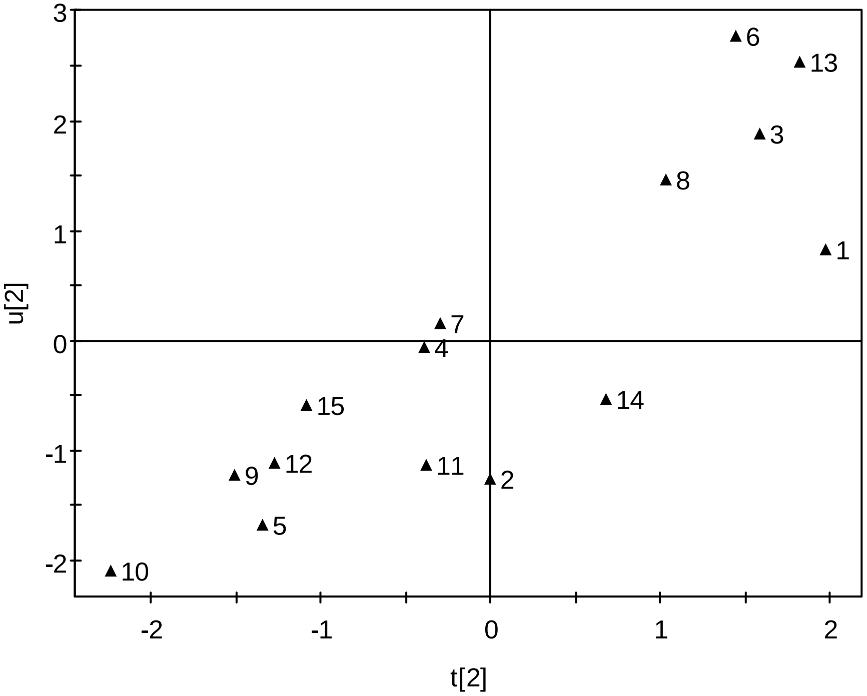
Components:

A	R2X	R2X(cum)	Eige...	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)	Signi...	It...
00	-	0.000	-	-	0.000	-	-	-	--	--
01	0.169	0.169	2.535	0.896	0.896	0.628	0.097	0.628	R1	1
02	0.128	0.297	1.916	0.078	0.975	0.363	0.097	0.763	R1	1
03	0.147	0.443	2.198	0.005	0.979	-0.197	0.097	0.763	N4	1
04	0.118	0.562	1.776	0.002	0.982	-0.194	0.097	0.763	N4	1

Tutorials

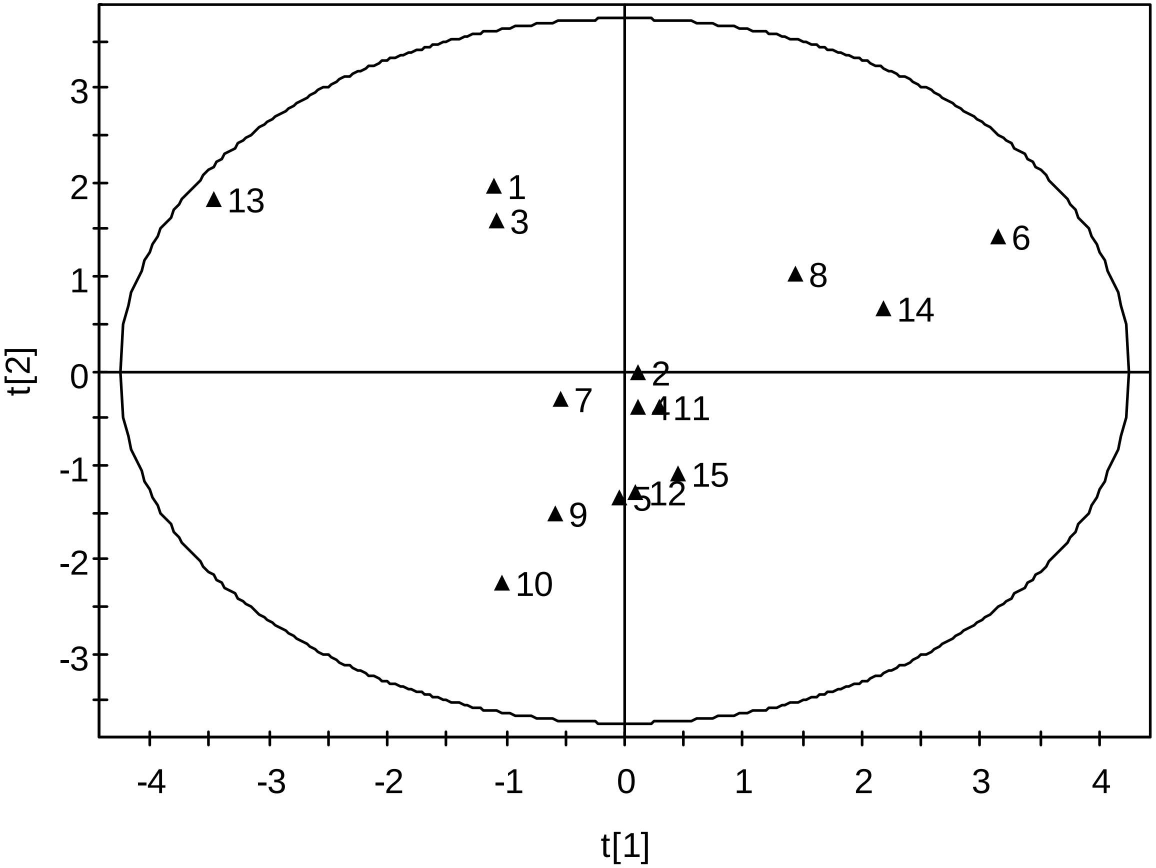
- t vs. u plots verify linear relationships ($t_i = u_i + e$)





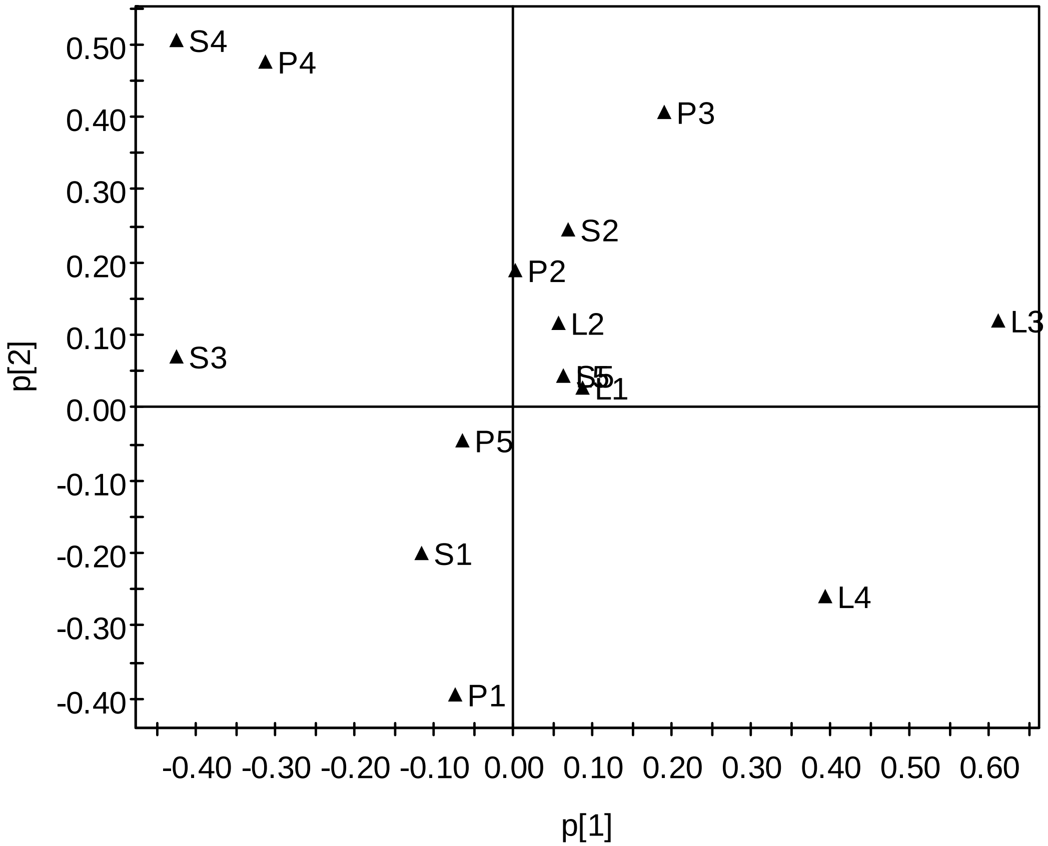
Tutorials

- Groups/clusters or outliers can be found in x-score plots.



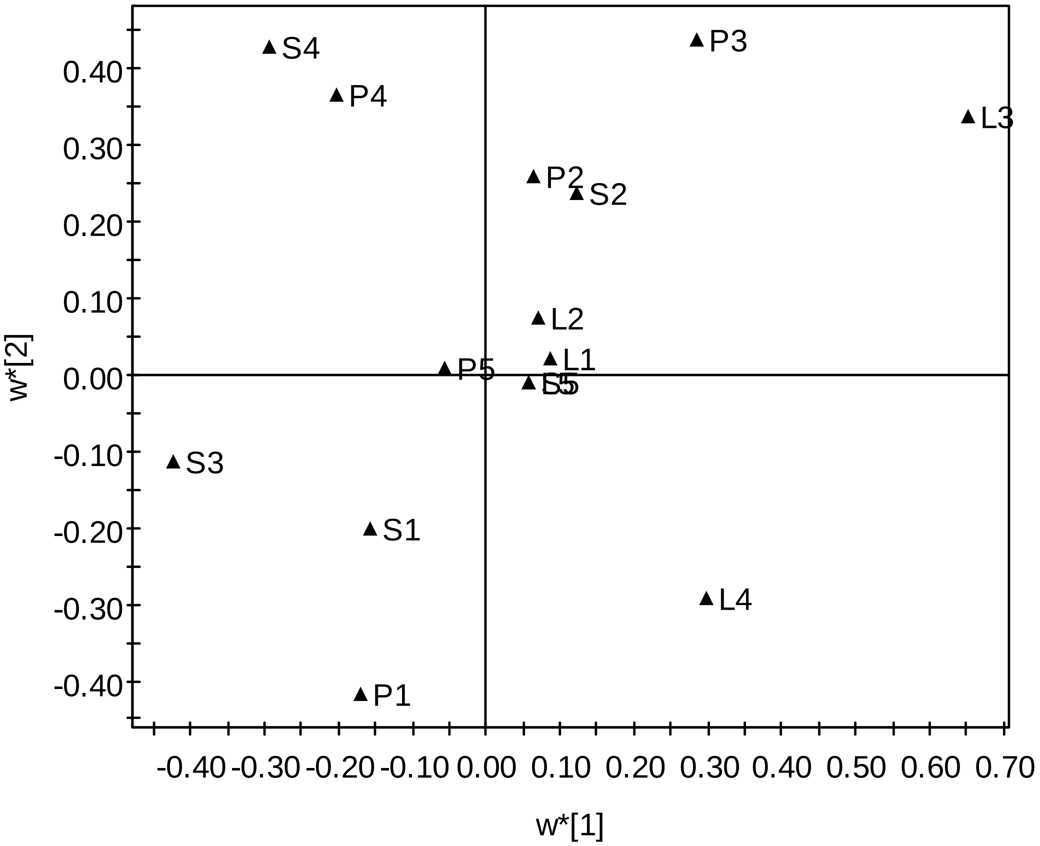
Tutorials

- And which measurements may be responsible for that clusters and/or outliers.



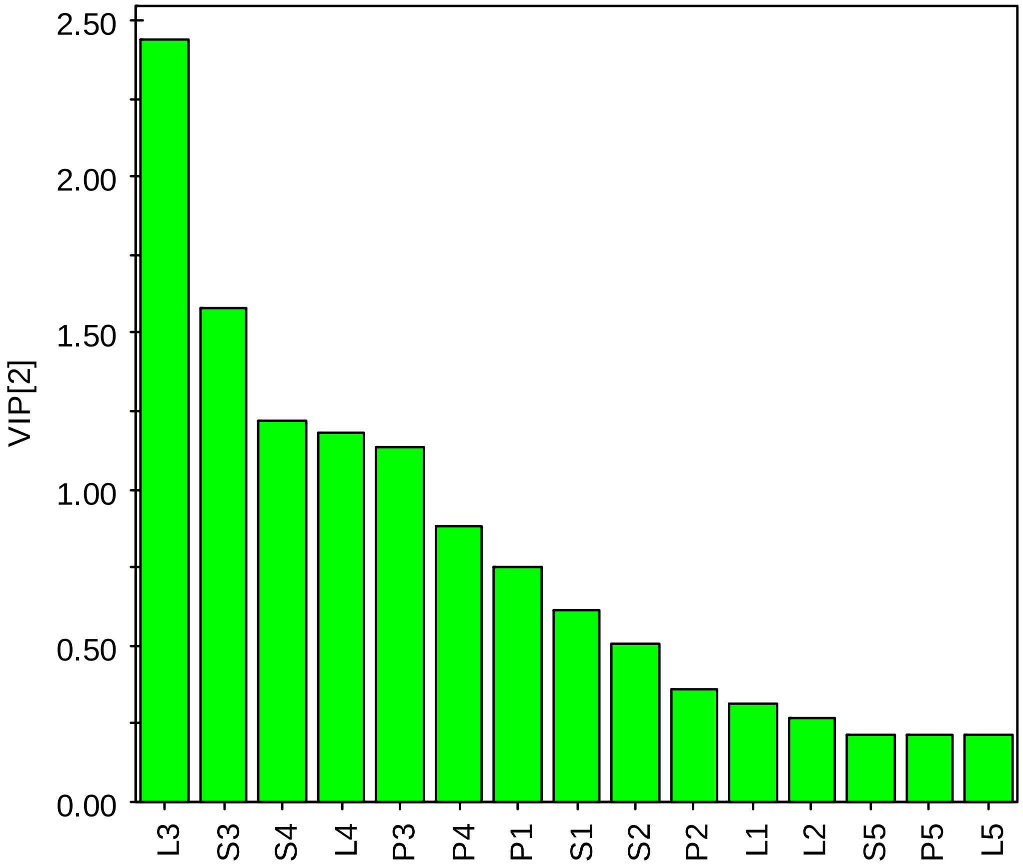
Tutorials

- Weight (w*) plots can tell relative importance of chemical measurements in predicting biological activity..



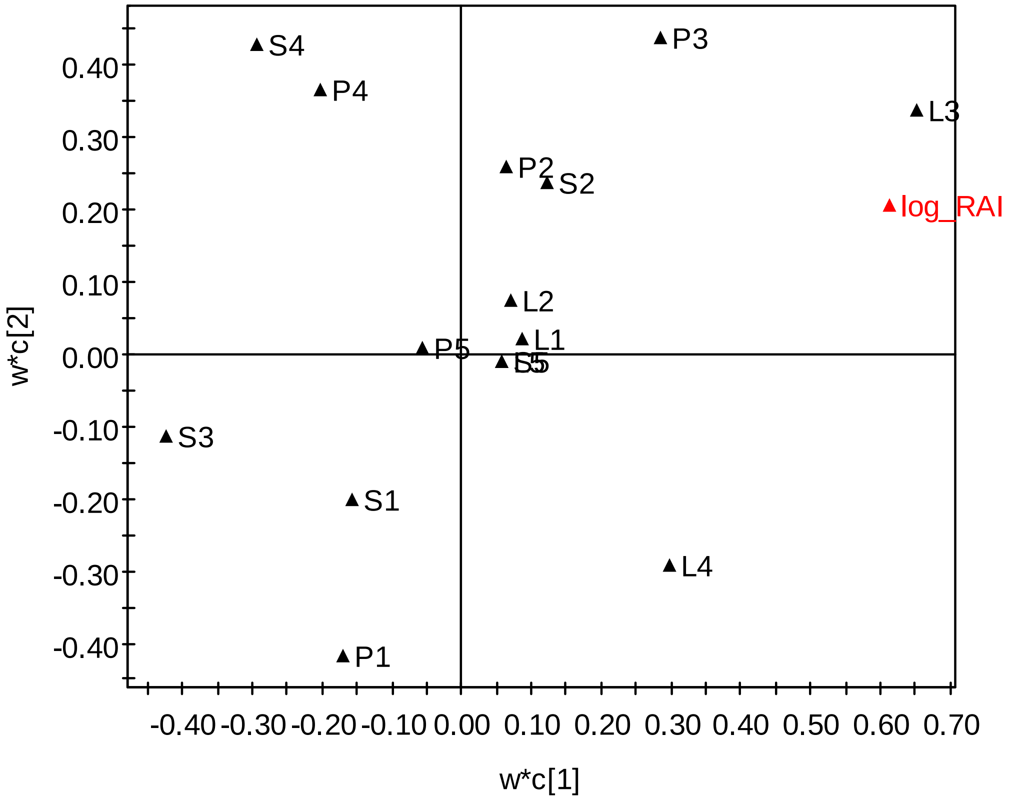
Tutorials

- A VIP plot can reveal this more easily.



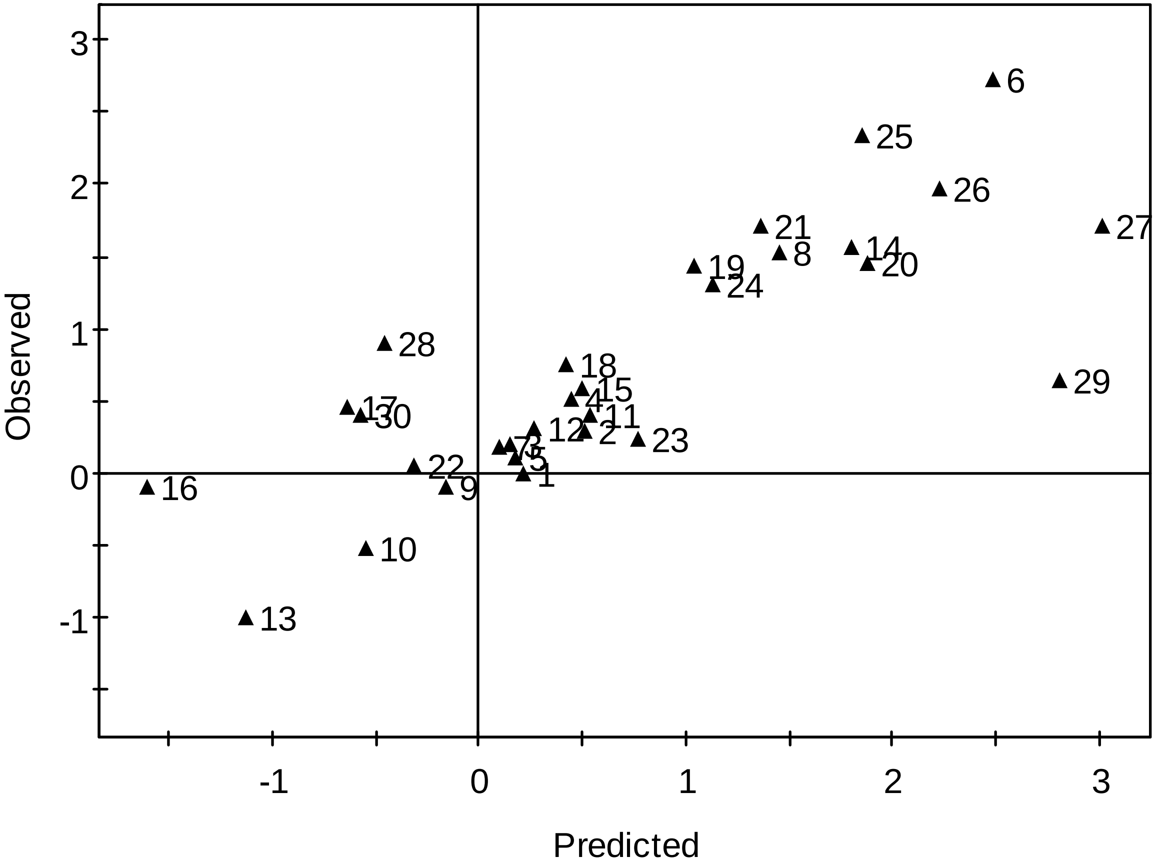
Tutorials

- w*c plots show the correlation structure between X and Y. One sees how the X and Y variables combine in the projections, and how the X variables relate to the Y variables.



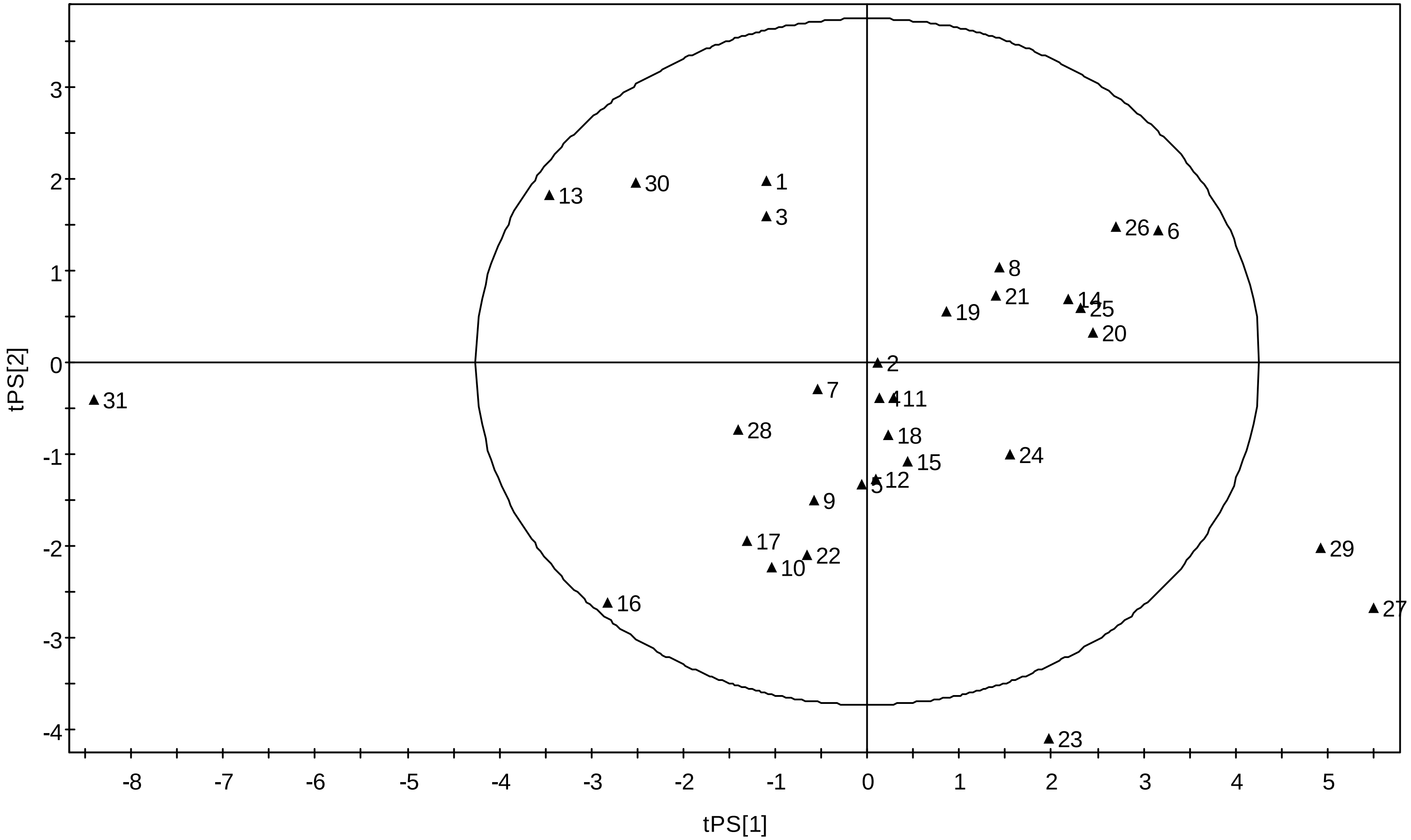
Tutorials

- Prediction for remaining 15 compounds



Tutorials

- Prediction for remaining 15 compounds



Confidence interval in PCA & PLS

- Approximate confidence interval/regions based on distribution assumption
  - Since $t_a\left(=\mathbf{p}_a^T\mathbf{x}\right)$ is a linear function of many x's, by the **Central Limit Theorem**, $t_a \sim N\left(0, \sigma_t^2\right)$ even if the individual x's are not normally distributed.

  → Use normal theory (or t-distribution if # observation is not large) to obtain confidence intervals/regions for t_a's

※Confidence interval: for single variable
※ Joint confidence interval for more than two variable? → confidence region

Confidence interval in PCA & PLS

1. 100(1-α)% confidence interval of t_a

$$\pm t_{\alpha/2}(df)\cdot s_{t_a} \qquad statistic \pm A \times \sigma_{statistic}$$

Value that depends on P.D.F of the statistic & confidence level α

Standard error of the statistic

2. Joint 100(1-α)% confidence region of t's

$$T^2 = (\mathbf{x}-\bar{\mathbf{x}})^T \mathbf{S}_{\mathbf{x}}^{-1} (\mathbf{x}-\bar{\mathbf{x}})$$

$$\mathbf{S}_{\mathbf{x}} = \hat{\Sigma}_{\mathbf{x}} = \frac{1}{N}\mathbf{X}^T\mathbf{X} = \frac{1}{N}\mathbf{P}\left(\mathbf{T}^T\mathbf{T}\right)\mathbf{P}^T = \mathbf{P}\mathbf{S}_t\mathbf{P}^T$$

$$\begin{aligned} T^2 &= (\mathbf{x}-\bar{\mathbf{x}})^T \mathbf{P}\mathbf{S}_t^{-1}\mathbf{P}^T (\mathbf{x}-\bar{\mathbf{x}}) \\ &= \mathbf{t}^T S_t^{-1}\mathbf{t} \\ &= \sum_{a=1}^{K} \frac{t_a^2}{s_{t_a}^2} \end{aligned}$$

$$\mathbf{S}_t = \begin{bmatrix} s_{t_1}^2 & 0 & \cdots & 0 \\ 0 & s_{t_2}^2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & s_{t_K}^2 \end{bmatrix}$$

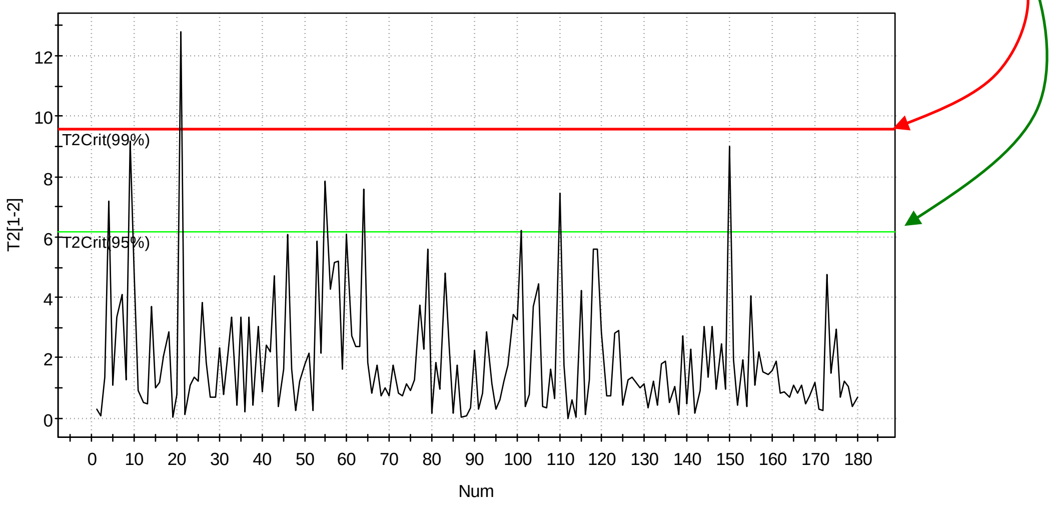
Confidence interval in PCA & PLS

- Upper 100(1-α)% confidence limit on T^2 is given by

$$T_\alpha^2 = \frac{(N-1)(N+1)K}{N(N-K)} F_\alpha(K, N-K)$$

- If only A component are used,

$$T_A^2 = \sum_{a=1}^{A} \frac{t_a^2}{s_{t_a}^2} \sim \frac{(N-1)(N+1)A}{N(N-A)} F_\alpha(A, N-A)$$



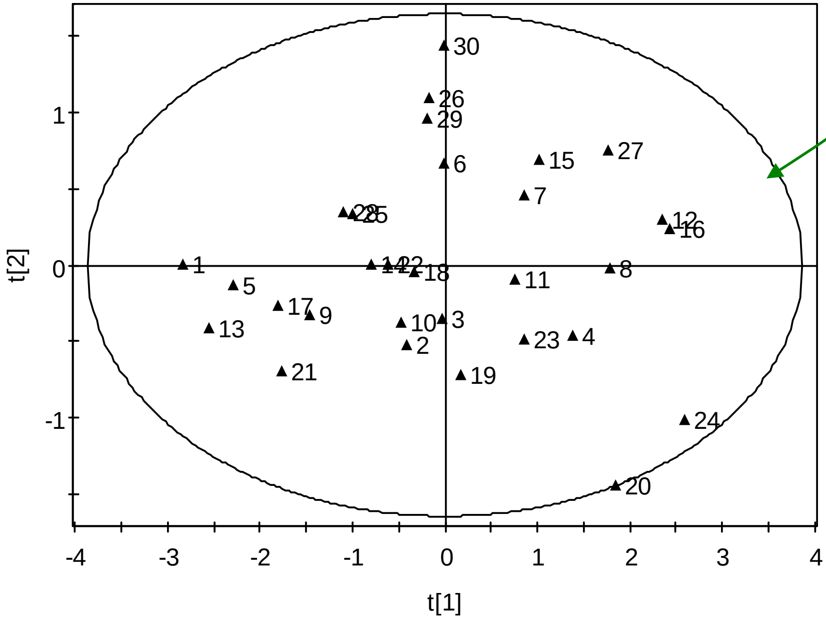
Confidence interval in PCA & PLS

- Or in space of t_1, t_2, …

$$\frac{t_1^2}{s_{t_1}^2}+\frac{t_2^2}{s_{t_2}^2}=\frac{(N-1)(N+1)2}{N(N-2)}F_\alpha(2,N-2)$$

constant

This is the equation of an ellipse in space of t_1 and t_2.



Confidence interval in PCA & PLS

3. SPE confidence interval (by Jackson, 1991)

$$Q = \left(\mathbf{x} - \hat{\mathbf{x}}\right)^T \left(\mathbf{x} - \hat{\mathbf{x}}\right) \left(\equiv SPE\right)$$

Critical upper 100(1-α)% confidence limit on Q is give by

$$Q_\alpha = \theta_1 \left[\frac{Z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 \left(h_0 - 1\right)}{\theta_1^2} + 1 \right]^{1/h_0}$$

Where

$$\theta_1 = \sum_{a=A+1}^{K} \lambda_a = Tr\left(\mathbf{E}\right) \qquad \theta_3 = \sum_{a=A+1}^{K} \lambda_a^3 = Tr\left(\mathbf{E}^3\right)$$

$$\theta_2 = \sum_{a=A+1}^{K} \lambda_a^2 = Tr\left(\mathbf{E}^2\right) \qquad h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2}$$

※ Some S/W's use resampling methods (bootstrap, jackknife) to calculate C.I.

Readings

- Theory
  - Burnham, A.J., Viveros, R., MacGregor, J.F., “Frameworks for latent variable multivariate regression,” J. Chemometrics **10**, 31-45, (1996)
  - Hoskuldsson, A., “PLS regression methods” J. Chemometrics **2**, 211-228, (1988)
- General
  - Wold, S., Sjöström, M., and Eriksson, L., “PLS regression: A basic tool of chemometrics,” Chemometrics and Intelligent Laboratory Systems, 58, 109-130, (2001)
- Applications
  - Gossen, P.D., MacGregor, J.F., Pelton, R.H., “composition and particle diameter for styren/methyl methacrylate copolymer latex using UV and NIR spectroscopy,” applied spectroscopy, **47**(11), 1852-1870
  - MacGregor, J.F., Jaeckle, D., Kiparissides, C. and Koutoudi, M., “process monitoring and diagnosis by multi-block PLS methods,” AIChE J., 40(5) 826-838, (1994)