Institute of Applied Physics “Nello Carrara”

Introduction to

Partial Least Square Regression

Dr. Leonardo Ciaccheri PhD

Regression analysis

This lesson is focused on two popular regression tools **Principal Component Regression** (PCR) and **Partial Least Square Regression** (PLSR or simply PLS).

Principal Component Regression

- PCR simply combines PCA with Multivariate Linear Regression (MLR) for predicting a quantitative target variable.
- PCs are good regressors. **High variance** reduce noise in the model. **Orthogonality** avoids collinearity problems. Probability of overfitting is reduced.
- The drawback of PCR is that it weights **predictor variables** (X) according to variance, and not correlation with **target variable** (Y). If there is strong interference, irrelevant PCs must be kept in the model in order to get a good prediction.

Partial Least Square

- PLS is a more sophisticated regression tool, which overcome these drawbacks.
- PLS looks for **factors** showing **good covariance with Y**. This favors both accuracy and robustness.

How PLS works

PLS factors are chosen imposing the following properties:

1. They are orthogonal
2. Factor-1 has the maximum covariance with target variable.
3. Factor-n has the highest covariance with target variable in the sub-space orthogonal to Factor-1 ... Factor-(n-1)

PLS uses information from both X and Y variables for determining factorial axes. This requires a more complex mathematic than PCA.

$$\underset{(N \times K)}{\mathbf{T}} = \underset{(N \times M)}{\mathbf{X}} \underset{(M \times K)}{\mathbf{W}}$$

T = X-score matrix
W = weight matrix

$$\underset{(N \times M)}{\mathbf{X}} = \underset{(N \times K)}{\mathbf{T}} \underset{(K \times M)}{\mathbf{P}} + \underset{(N \times M)}{\mathbf{R_x}}$$

P = **X**-loading matrix
$\mathbf{R_x}$ = X-residual matrix

Y-scores

A fundamental difference between PLS and PCR is that the former models both X and Y matrices. Therefore PLS produces scores and loadings also for Y matrix.

$$\underset{(N \times 1)}{\mathbf{Y}} = \underset{(N \times K)}{\mathbf{U}}\ \underset{(K \times 1)}{\mathbf{C}} + \underset{(N \times M)}{\mathbf{E}_y}$$

U = Y-score matrix
C = Y-loadings matrix
$\mathbf{E}_y$ = Residual matrix

X-scores and Y-scores are correlated. Therefore, Y can also be written as function of T.

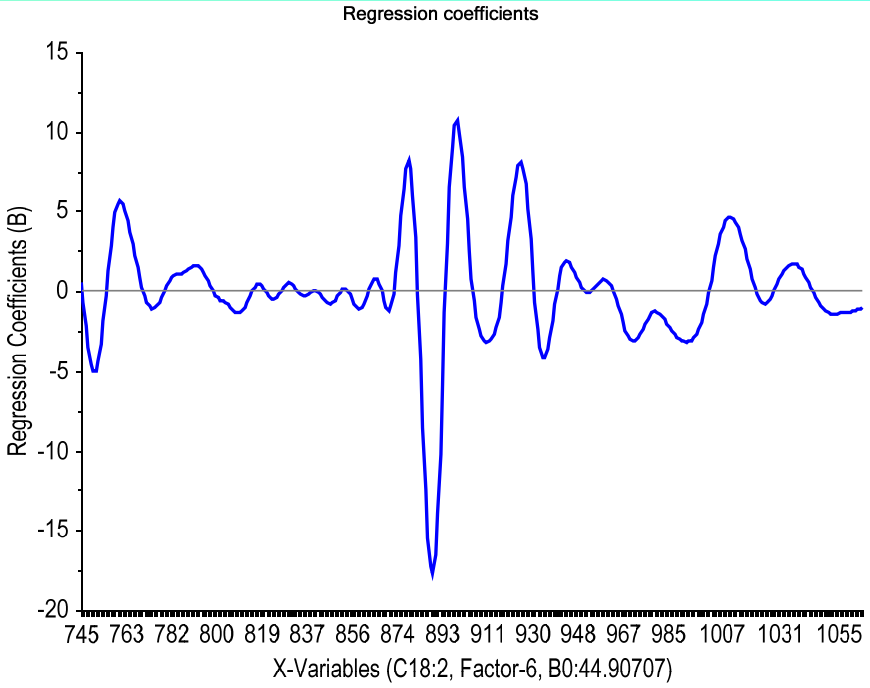
$$\underset{(N \times 1)}{\mathbf{Y}} = \underset{(N \times M)}{\mathbf{T}}\ \underset{(M \times 1)}{\mathbf{C}} + \underset{(N \times M)}{\mathbf{R}_y}$$

$\mathbf{R}_y$ = Y-Residual matrix (regression residuals)

$\mathbf{R}_y$ is different from $\mathbf{E}_y$, because X-scores (T) only approximate Y-scores (U). From regression point of view, **$\mathbf{R}_y$ is the important matrix**.

Regression Coefficients

- By expressing T as function of X and W, the regression coefficient, B, can be calculated. B is a linear combination of W-columns with coefficients given by C.
- Vector B allows predicting Y directly from the X matrix. It also reveals important variables.
- Interpretation of B is similar to that of loadings. Important variables have coefficients far from zero, either positive or negative.



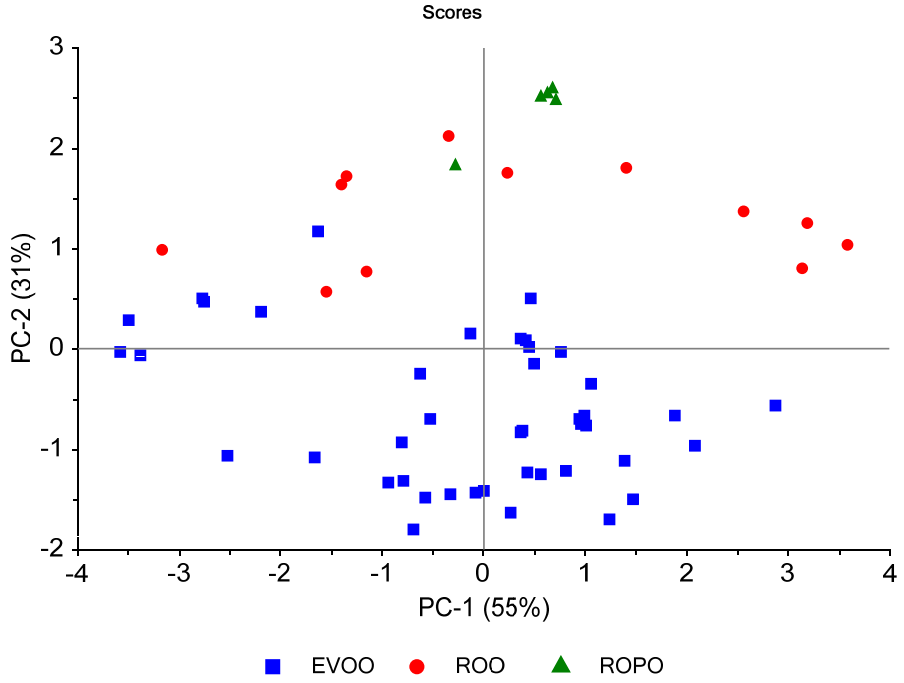
$$\underset{(M \times 1)}{\mathbf{B}} = \underset{(M \times K)}{\mathbf{W}} \; \underset{(K \times 1)}{\mathbf{C}}$$

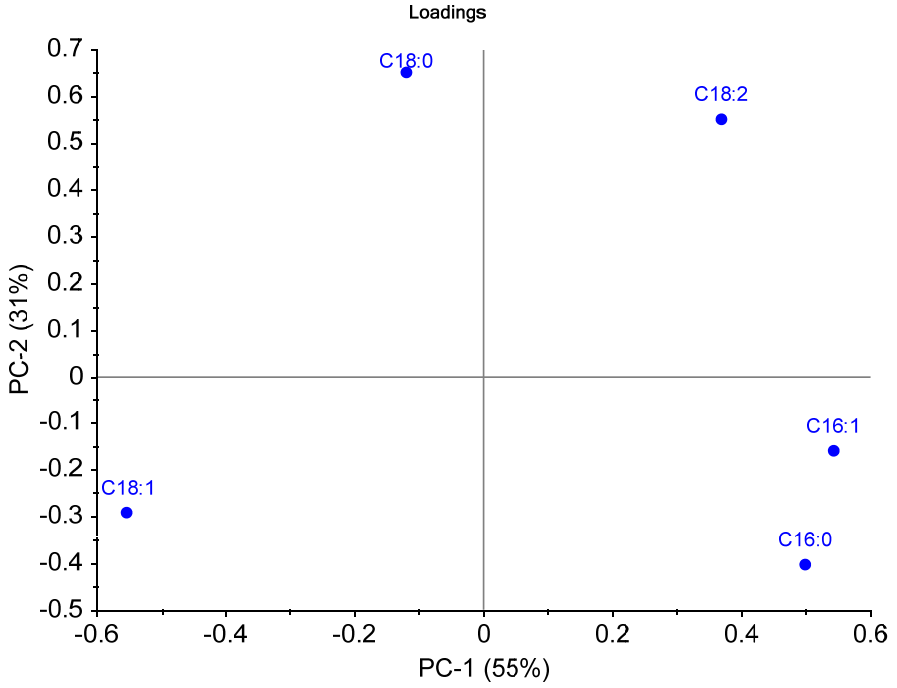
B = regression coefficients

$$\underset{(N \times 1)}{\mathbf{Y}} = \underset{(N \times M)}{\mathbf{X}} \; \underset{(M \times 1)}{\mathbf{B}} + \underset{(N \times M)}{\mathbf{R_y}}$$

PCA of fatty acids

- **Why are NIR spectra able to split oils of different categories?**
- Most of olive oil is made by fatty acids; above all: Oleic, Palmitic, Linoleic, Stearic and Palmitoleic.
- **PC2** of acidic content easily split virgin and low-quality oils. **Linoleic** and **Stearic** acids have the strongest loadings along PC2.
- **Linoleic** has **higher concentration** than Stearic, thus it is the **more probable cause** of spectra grouping.
- **Let us test PCR and PLS on predicting Linoleic acid in olive oil.**

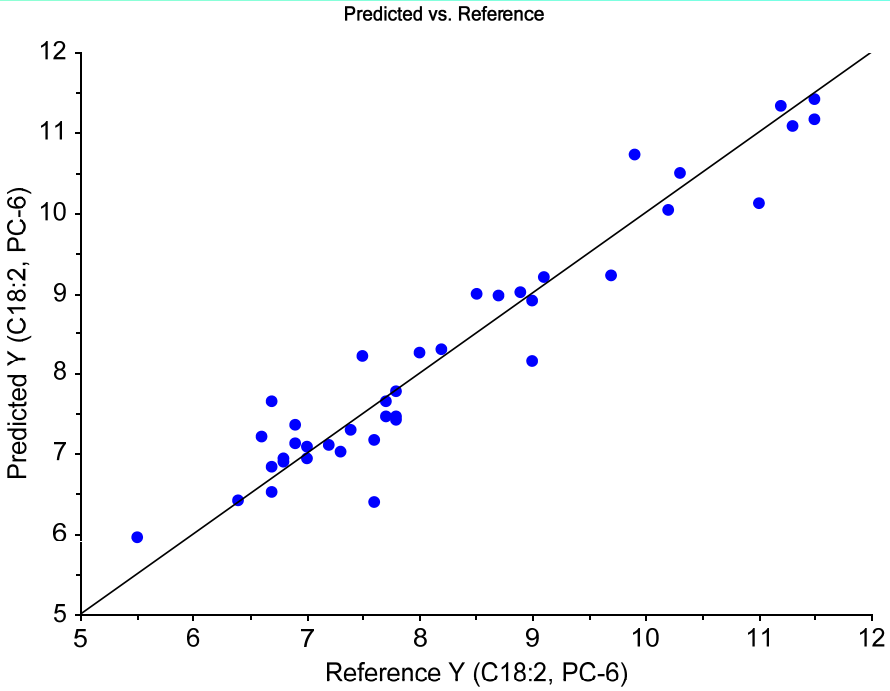


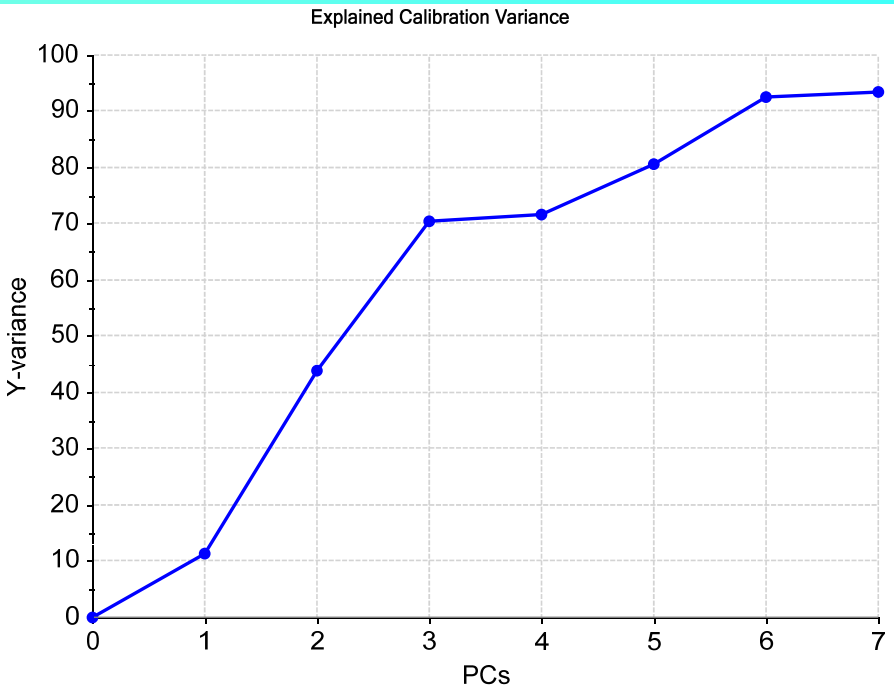


PCR

- **RMSEC** is the root mean square value of calibration residuals.
- **R^2** is the fraction of Y-variance explained by the model.
- Calibration is good, but 6 PCs are required.
- Only PC2 and PC3 capture more than 20% of Y-variance. PC4 is nearly useless.

Method	PCR
Components	6
RMSEC	0.4%
R^2	0.93

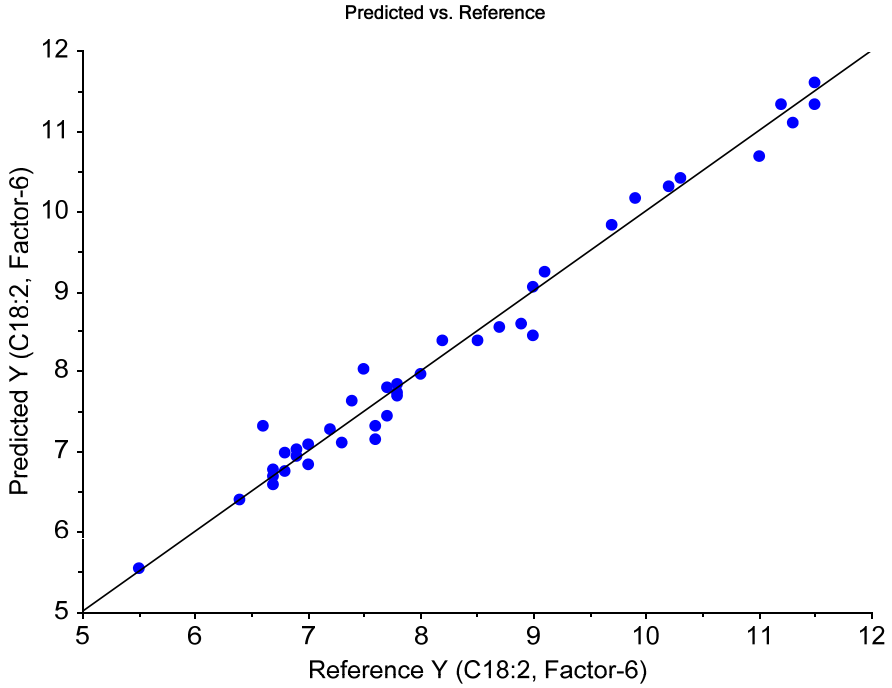


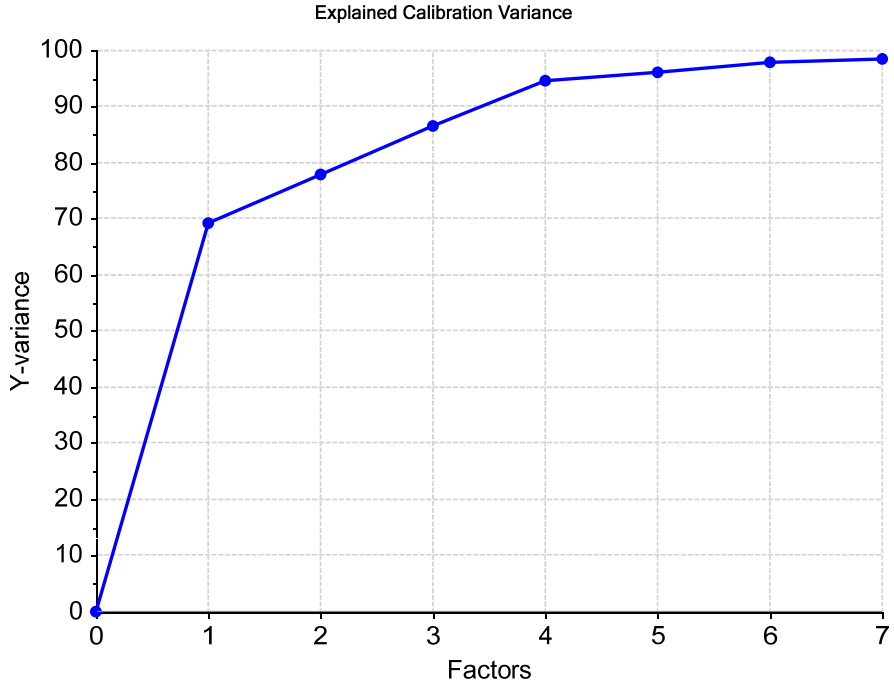


PLS

- PLS achieves lower RMSEC with the same number of factors.
- The curve of explained Y-variance raises more quickly. It explains 69% of variance with 1 factor and 95% with only 4 factors.
- Slope of the curve decrease monotonically.

Method	PLS
Factors	6
RMSEC	0.2%
R^2	0.98

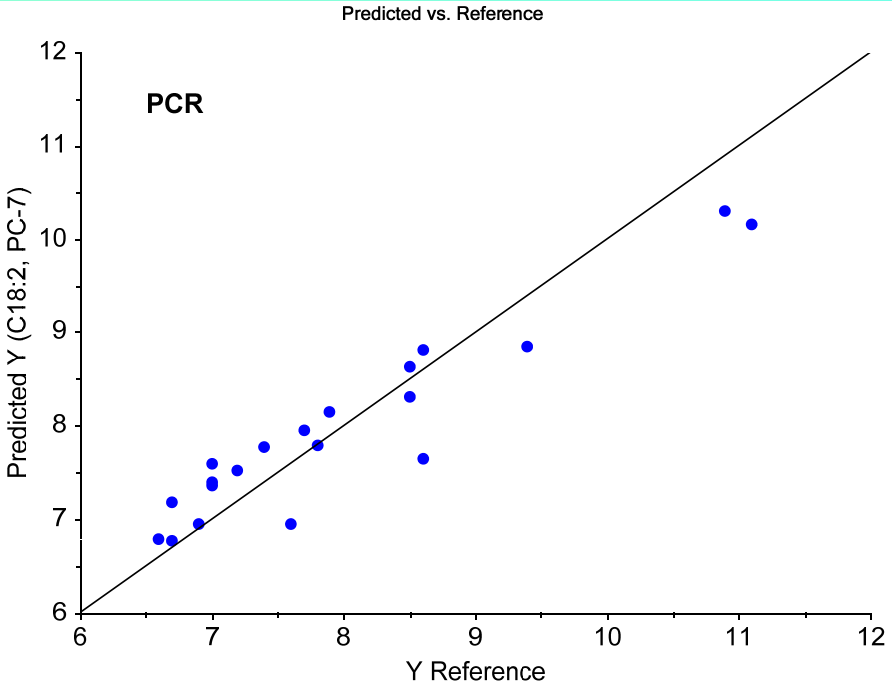


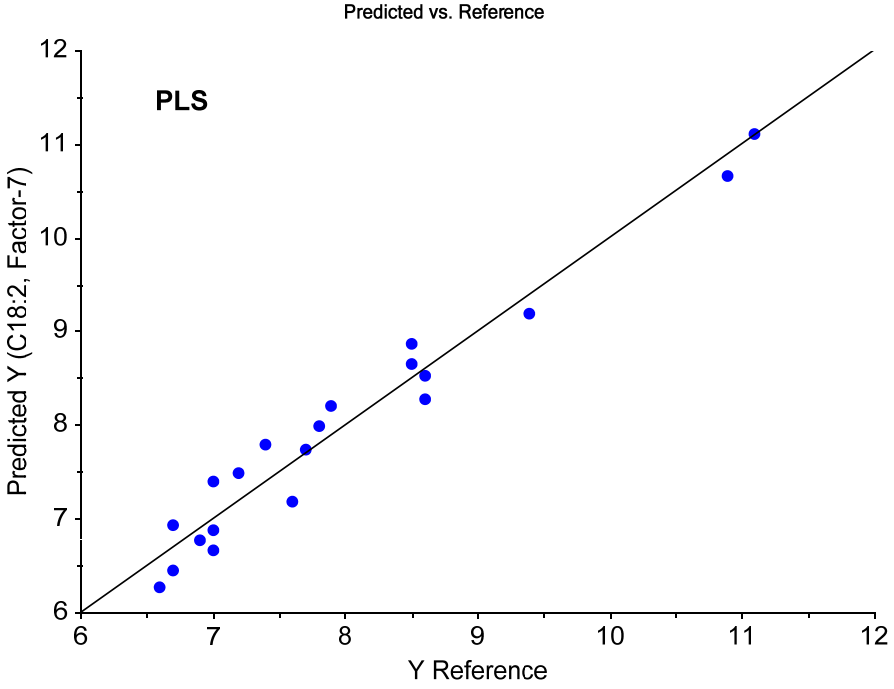


Validation of PLS and PCR

- **RMSEP** is the analogue of **RMSEC** for test set. It is usually higher than RMSEC.
- Both RMSEPs are acceptable, but **PLS** is more accurate than **PCR**.
- A new sample is required for fully validate the models.

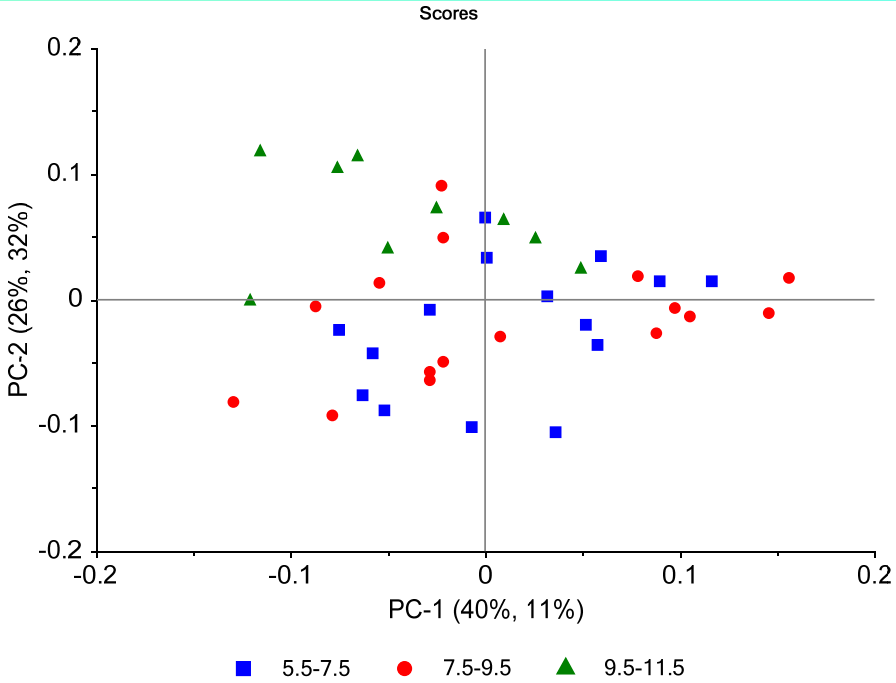
Factors	RMSE	PCR	PLS
6	RMSEC	0.4%	0.2%
	RMSEP	0.5%	0.3%

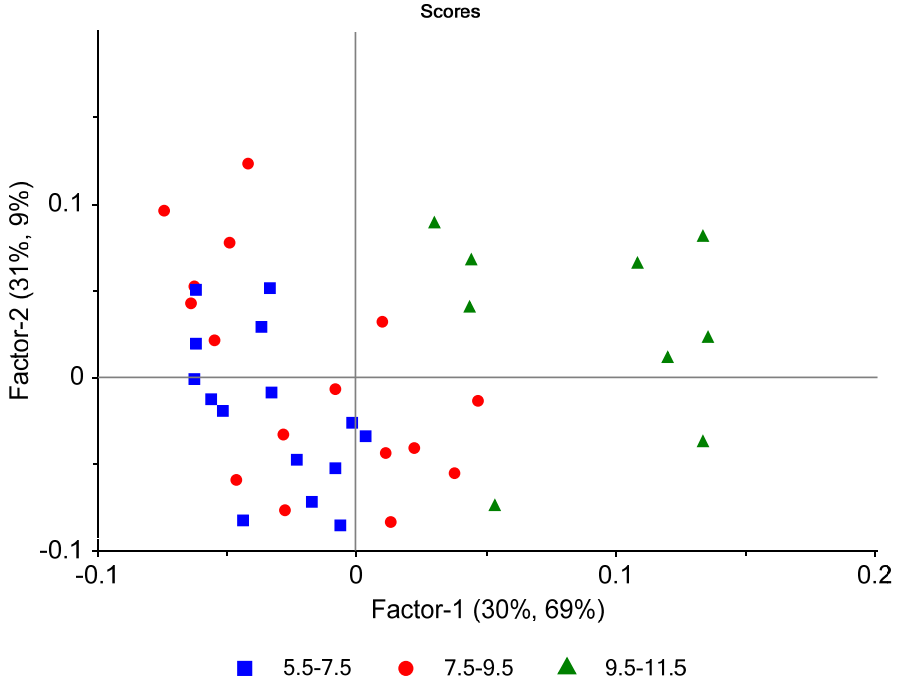




PCA scores vs. PLS scores

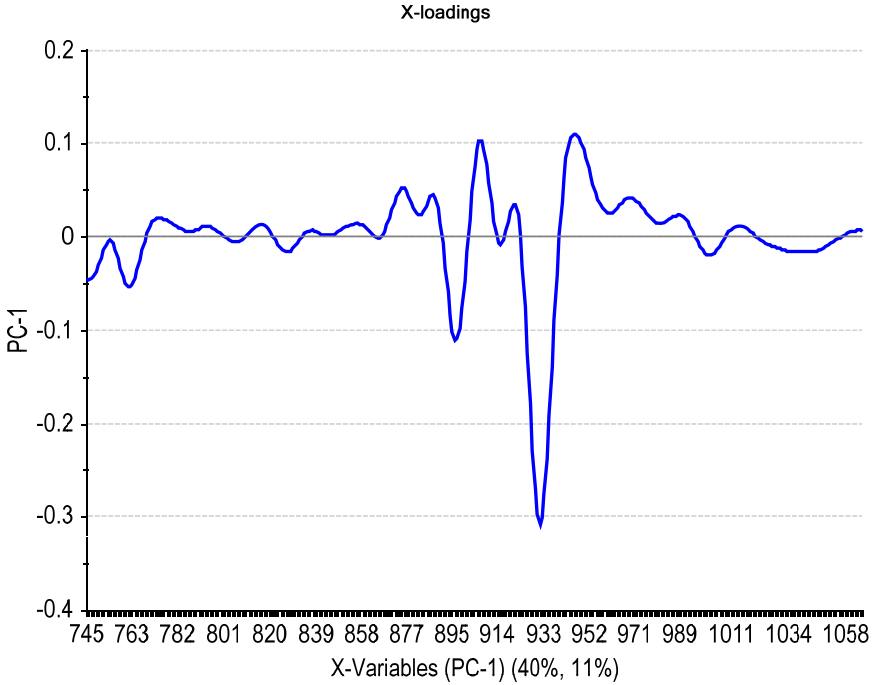
- Like PCA, PLS produces score plots, but they can be sensibly different.
- Plots below came from **PCR** (left) and **PLS** (right) models. Points are colored according to their Linoleic content, dividend into three bands.
- **PC1**, which has no predicting power. There is no separation of groups.
- **Factor 1** alone explains **69% of Y-variance**. It clearly split high-linoleic group.

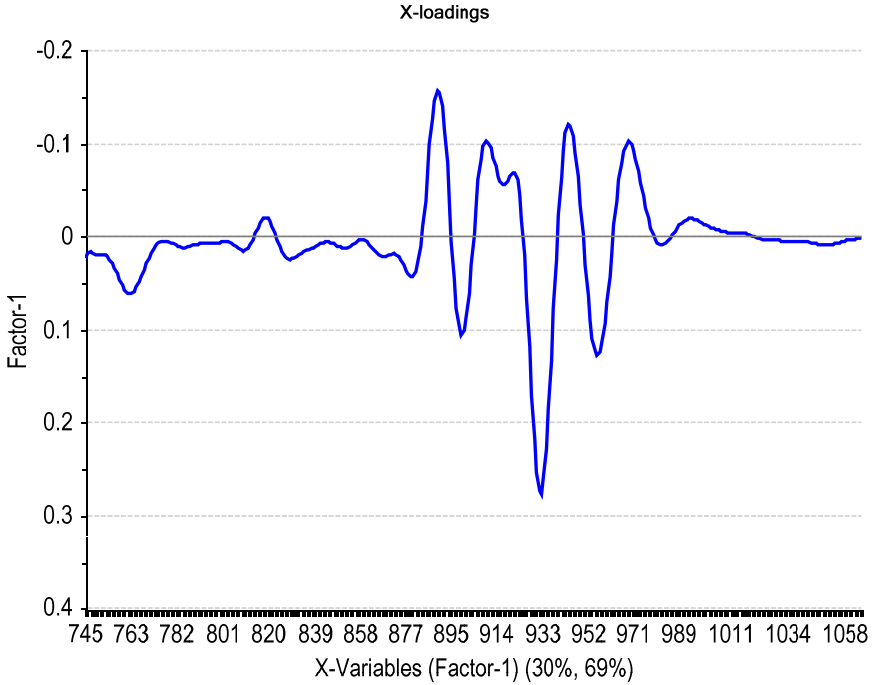




PCA loadings vs. PLS loadings (1)

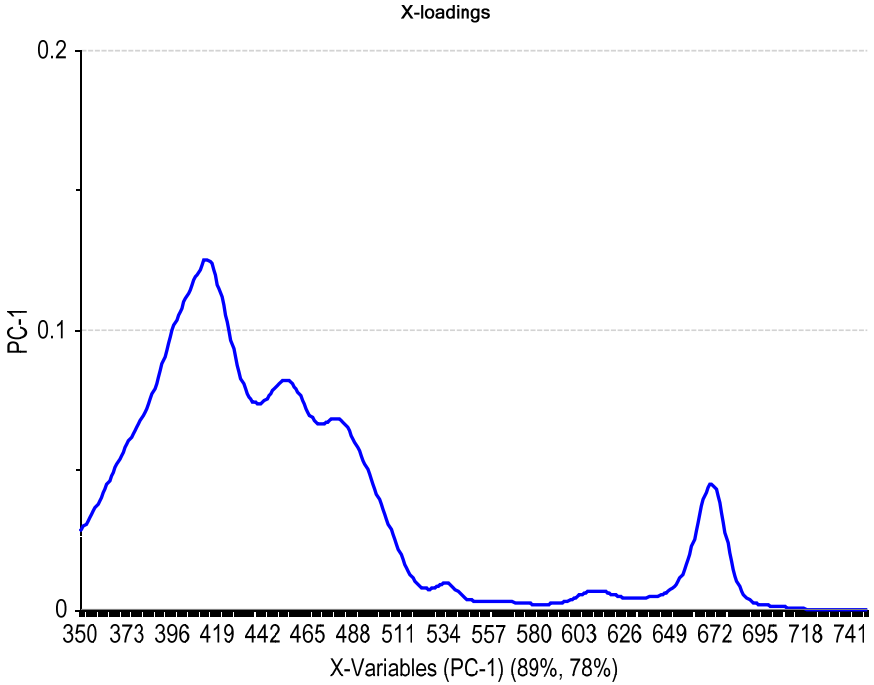
- Comparing loadings of **PC-1** (left) with those of **PLS Factor-1** (right) evident differences are observable.
- Some wavelengths are important for PLS, but not for PCR. Some wavelengths are important for both but are weighted differently.
- The axis of PLS loading has been reversed for better comparison.
- **Axis orientation is indeterminate in either PLS or PCA.**

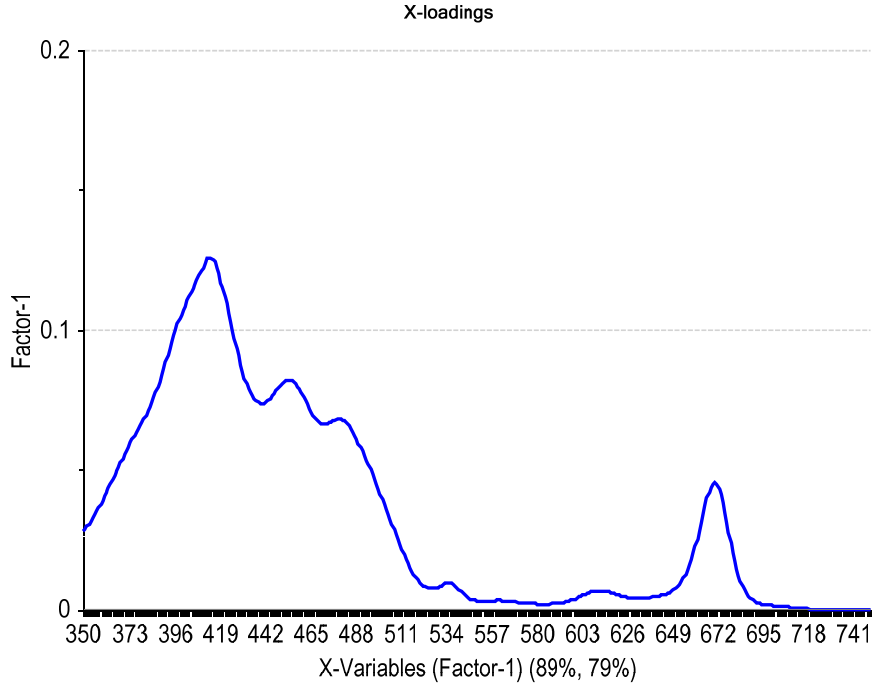




PCA loadings vs. PLS loadings (2)

- Difference between PCR and PLS is evident if Y has a weak influence on spectrum. If Y is the main absorber instead, difference between using PCR or PLS is much smaller.
- These loading plots come from models for predicting **chlorophyll** in olive oils from **visible absorption** spectra.
- The loadings of **PC-1** (left) and those of **PLS Factor-1** (right) show no evident differences.





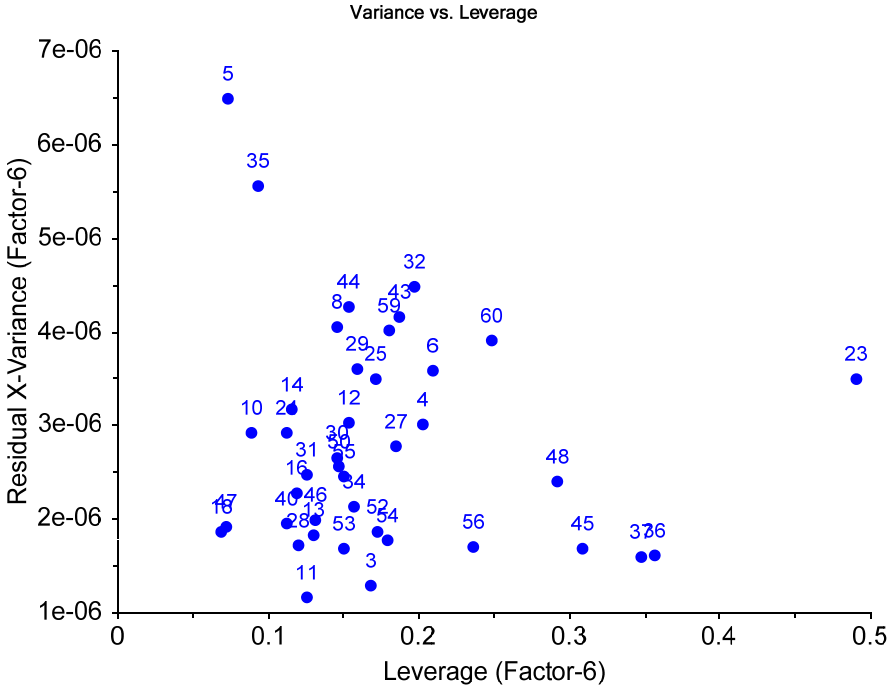
Different kind of outliers

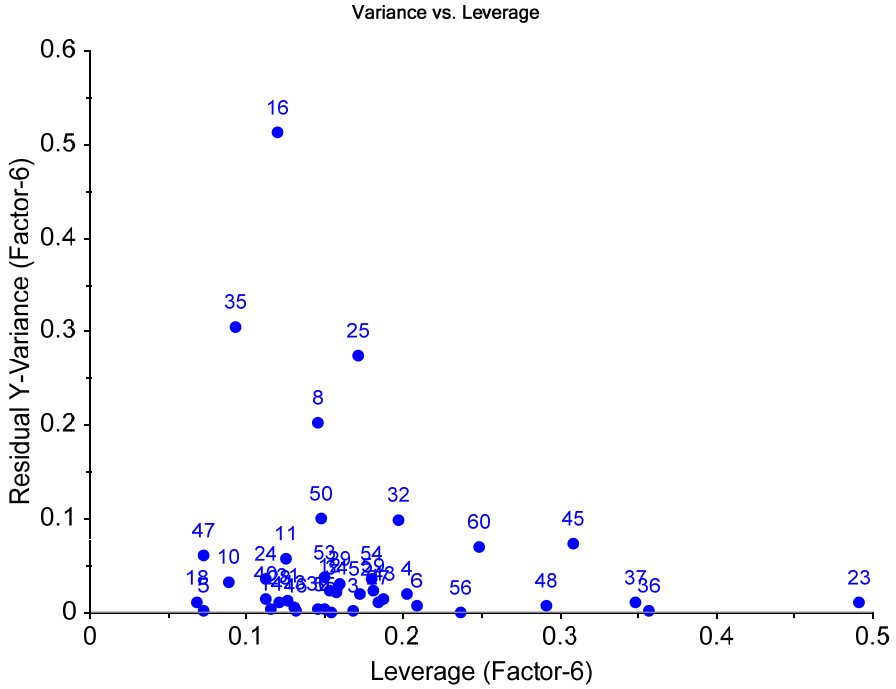
Both PCR and PLS produce **two residual matrices**.

- $\mathbf{R_x}$ says how well X matrix is represented by the model.
- $\mathbf{R_y}$ says how well target variable is predicted.

There are three reason for considering outlier an object: **high X-residuals**, **high Y-residuals** and **high influence**.

Influent objects are more critical, because they can negatively affect predictions of other samples.

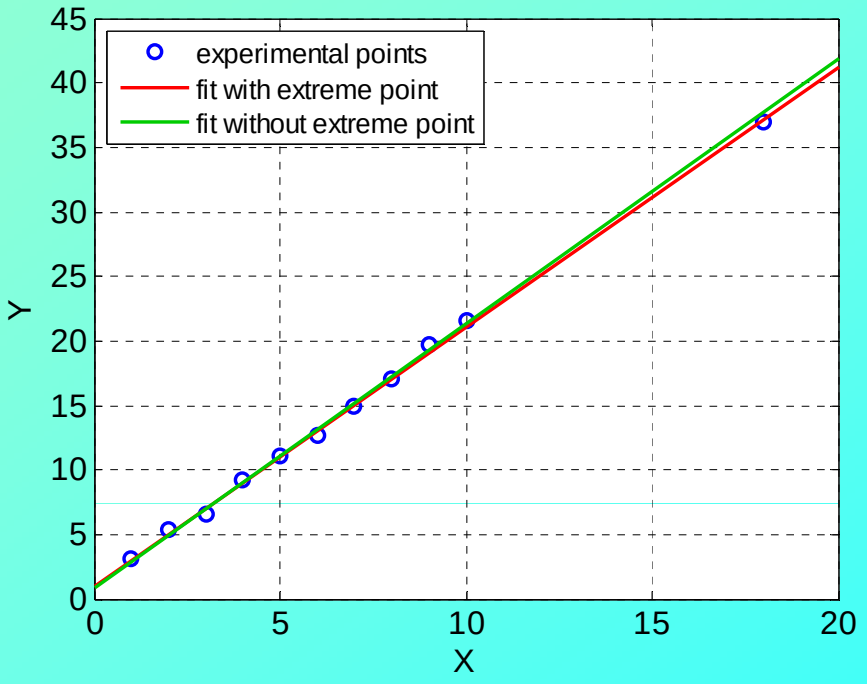


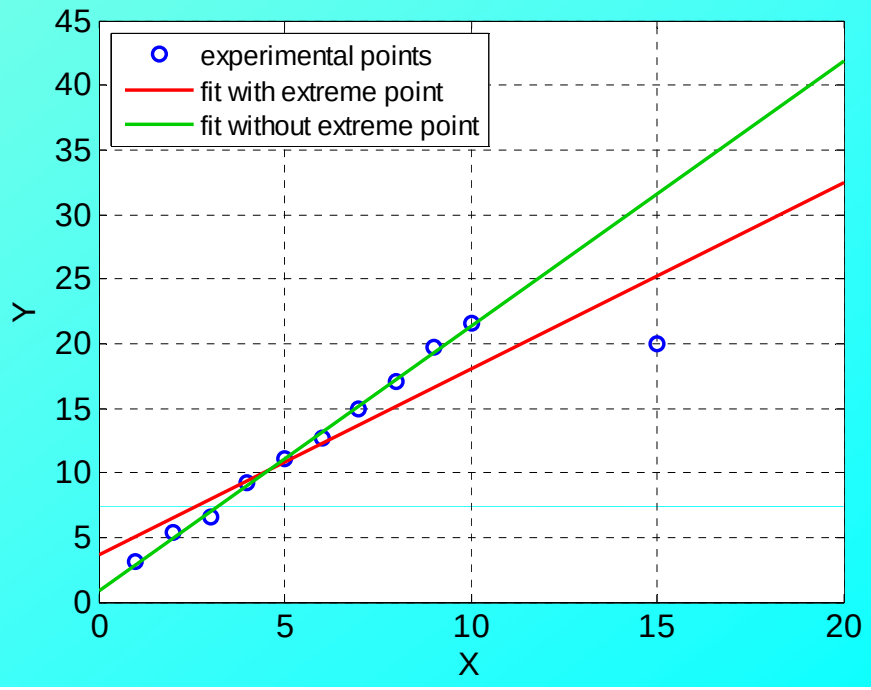


Extreme or Outliers?

These plots are examples of simple **bi-variate linear regression**.

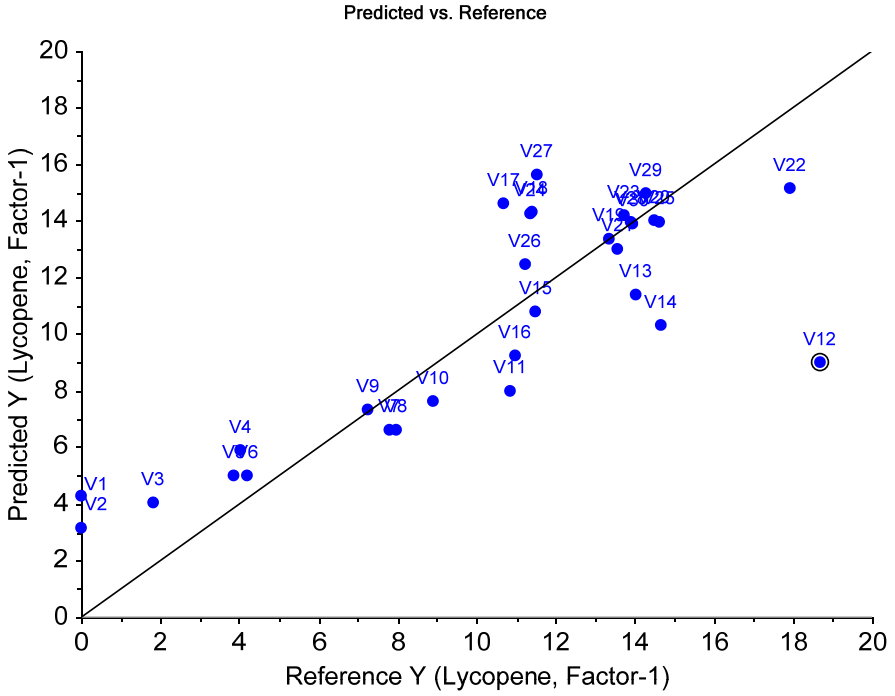
- On the **Left** is shown an **extreme sample**. It is far from others, but it obeys to the same linear relationship. Removing it minimally changes the regression line.
- On the **Right** is show an **outlier**. Not only it is influential, but it also obeys to a different X-Y relationship. Removing it sensibly changes the regression line.

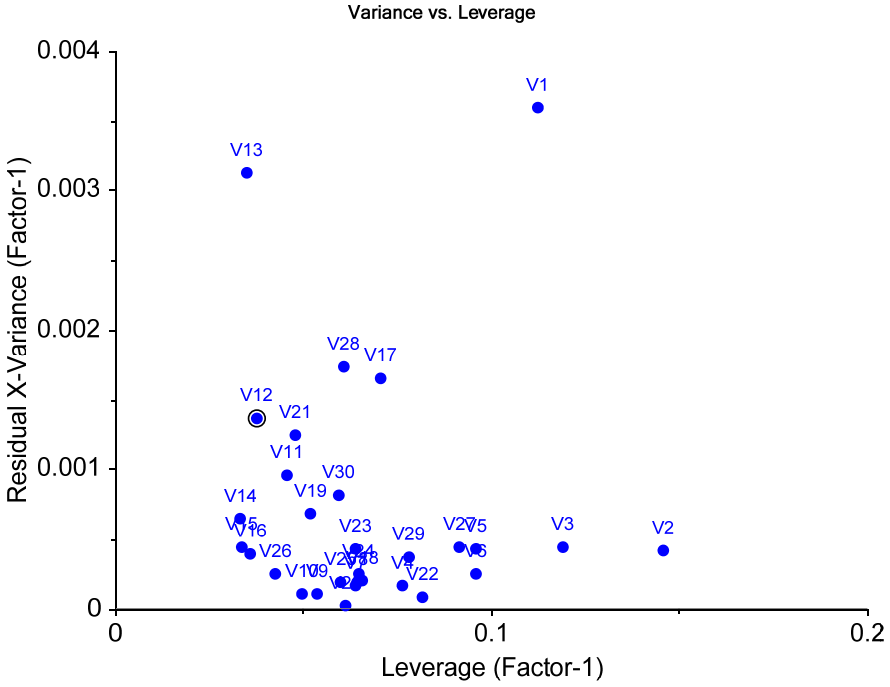


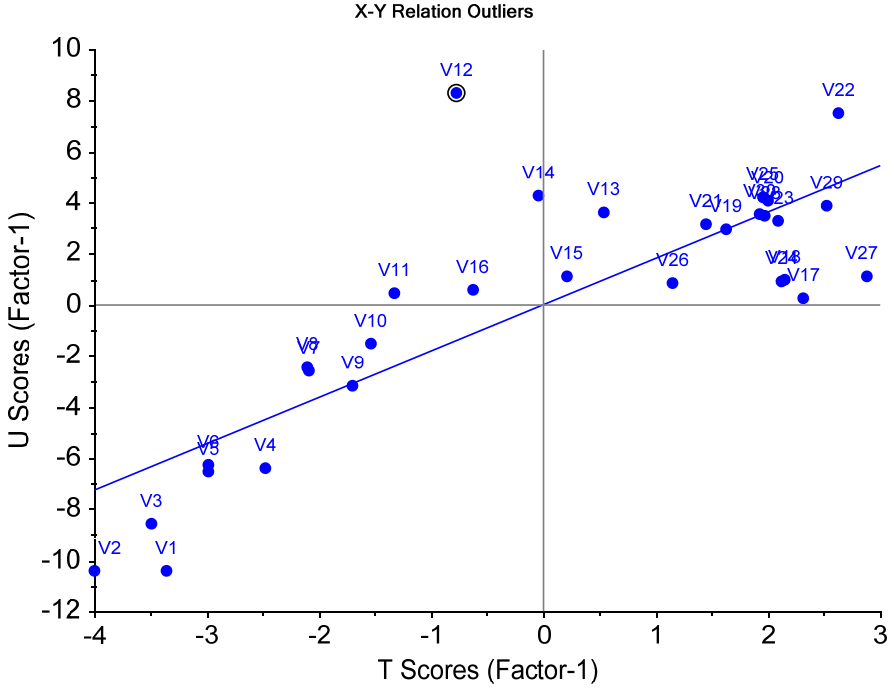


X-Y outliers

- **X-Y outliers** do not show exceptional X or Y values, but do not follow the same X-Y relationship of other samples.
- Sample **V12** is badly predicted. However its Y (right) is not exceptional, and its spectrum fit well in the model (below).
- Plotting **U vs. T**, reveals X-Y outliers, and also **non linearity** in X-Y relationship.







Conclusions

- PLS is a more efficient regression method than PCR, because it discard more irrelevant information.
- PLS is particularly useful when influence of target variable on predictor matrix is weak.
- Unlike PCR, PLS is a supervised method. It uses knowledge of target variable to determine factorial axes.
- A new, independent, sampling is necessary for validating prediction models.

Bibliography

Vandeginste, Massart,Buydens, De Jong, Lewi, Smeyers-Verbeke

Handbook of Chemometric and Qualimetric

Chapters 35, 36

Elsevier Science BV, Amsterdam, 1998

M. J. Adams

Chemometric in Analytical Spectroscopy

Chapter 6

Royal Society of Chemistry, Cambridge, 1995

S. Vold. M. Sjostrom, L. Eriksson

PLS-regression, a basic tool for chemometric

Chemometric and Intelligent laboratory Systems

vol. 58, pp. 109-130, Elsevier, 2001