LINEAR REGRESSION:

MULTIVARIATE REGRESSION

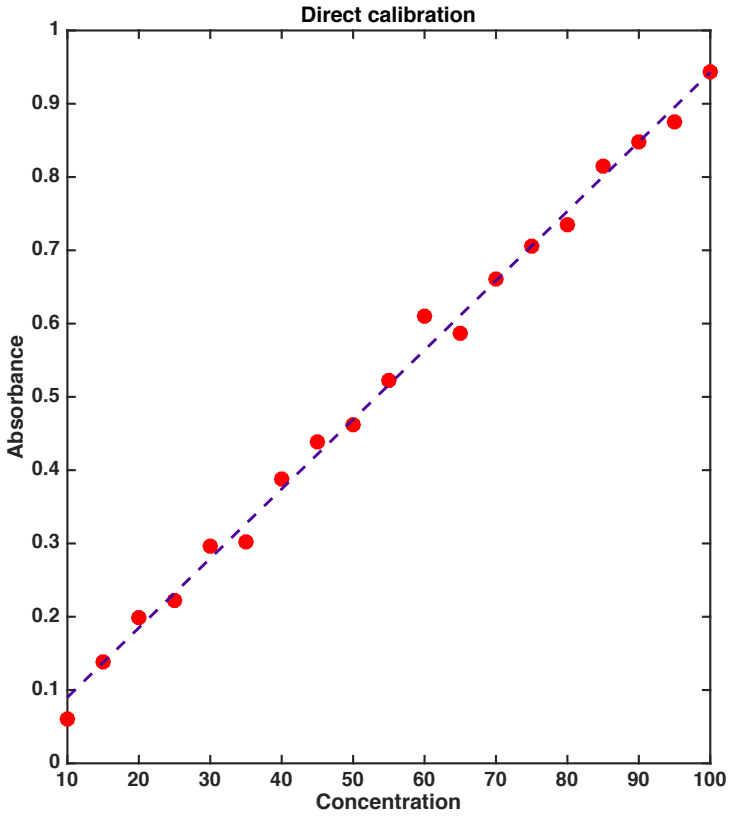
Federico Marini

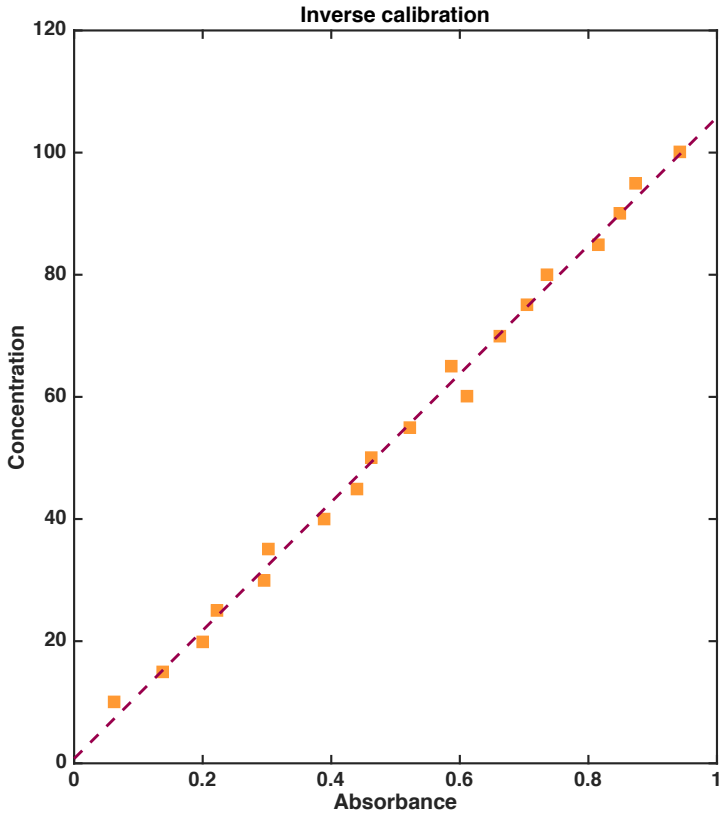
Dept. Chemistry, University of Rome "La Sapienza", Rome, Italy

SAPIENZA
UNIVERSITÀ DI ROMA

Direct vs inverse calibration

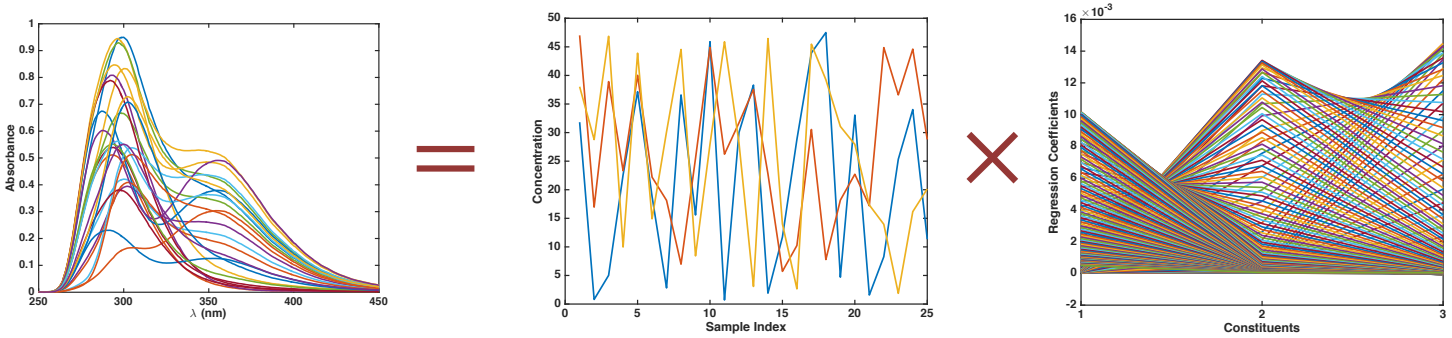
- The distinction refers to the way the relationship between signal(s) and concentration of chemical constituents is established.
  - DIRECT: signal is considered to be directly proportional to the concentration
  - INVERSE: concentration is considered to be directly proportional to the signal.





Direct vs inverse calibration - 2

- **UNIVARIATE CASE:** no substantial difference appears to exist in calibrating a regression line in a direct or in an inverse way.
  - Inverse models more efficient with small data sets and high noise level
- **MULTIVARIATE CASE:** The difference is relevant.
  - Direct models show some advantages, but present unsolvable problems regarding the variety of analytical platform/problems they can be applied to
  - Inverse models are almost always used.

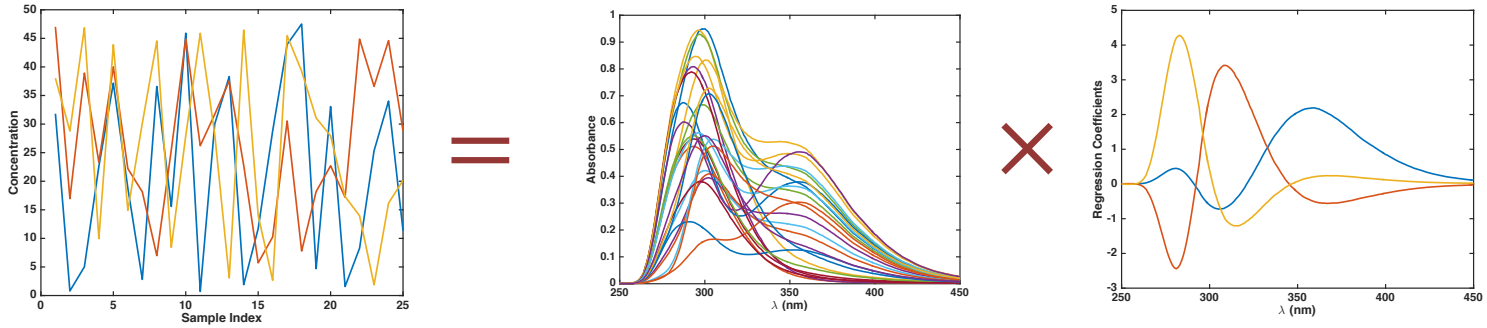


Direct Calibration

- **Requirements:**
  - Prepare mixtures of standards of all pure chemical constituents to be used as calibration samples.
  - Number of mixtures ≥ number of constituents (the larger they are the more precise the results)
  - Calibration mixtures should be as representative as possible of the combination of concentrations in future, unknown samples
- **Applicability:**
  - The concentrations of ALL the constituents in all calibration samples should be known.
  - The constituents should be the same as in future test samples.
  - Relates signals to constituents' concentrations, but not to global properties (octane number, sensorial attributes, iodine value, etc.)
- **Limitations:**
  - Sensitive to spectral correlations (constituents with severely overlapped spectra cannot be reliably determined)
  - The presence of non-modeled interferents may lead to serious errors in the determination.

Inverse Calibration

- **Applicability:**
  - Multicomponent samples where only one or a few analytes are of interest
  - Concentrations, spectra, and chemical identities of the other constituents may be unknown.
  - First proposed in the 1960's (connected to the development of NIR for the non-invasive analysis of intact material*)
- **Advantages:**
  - Allows studying of complex mixtures knowing only the concentrations of a limited number of constituents.
  - Quantitation of an analyte in the presence of interferences (if they are properly represented in the calibration samples, even if their concentrations or chemical identities are not known).



Why going multivariate?

- Instead of just using one of the variables, it makes sense to use all the measured information
- There are many significant advantages in doing so:
  - Noise reduction: More (redundant) measurements of the same phenomenon
  - Possibility of interferents: Non-selective signals can be made selective by mathematics (provided that the signal profiles of the interferents are not completely identical to that from the analyte).
  - Exploratory aspects: models provide a number of informative parameters + residuals
  - Outlier detection: detection of outlier is enhanced having multivariate data.

A first multivariate approach: MLR

Statement of the problem:

- Linear relationship between a response y_i and all the variables measured on a sample x_{ij}

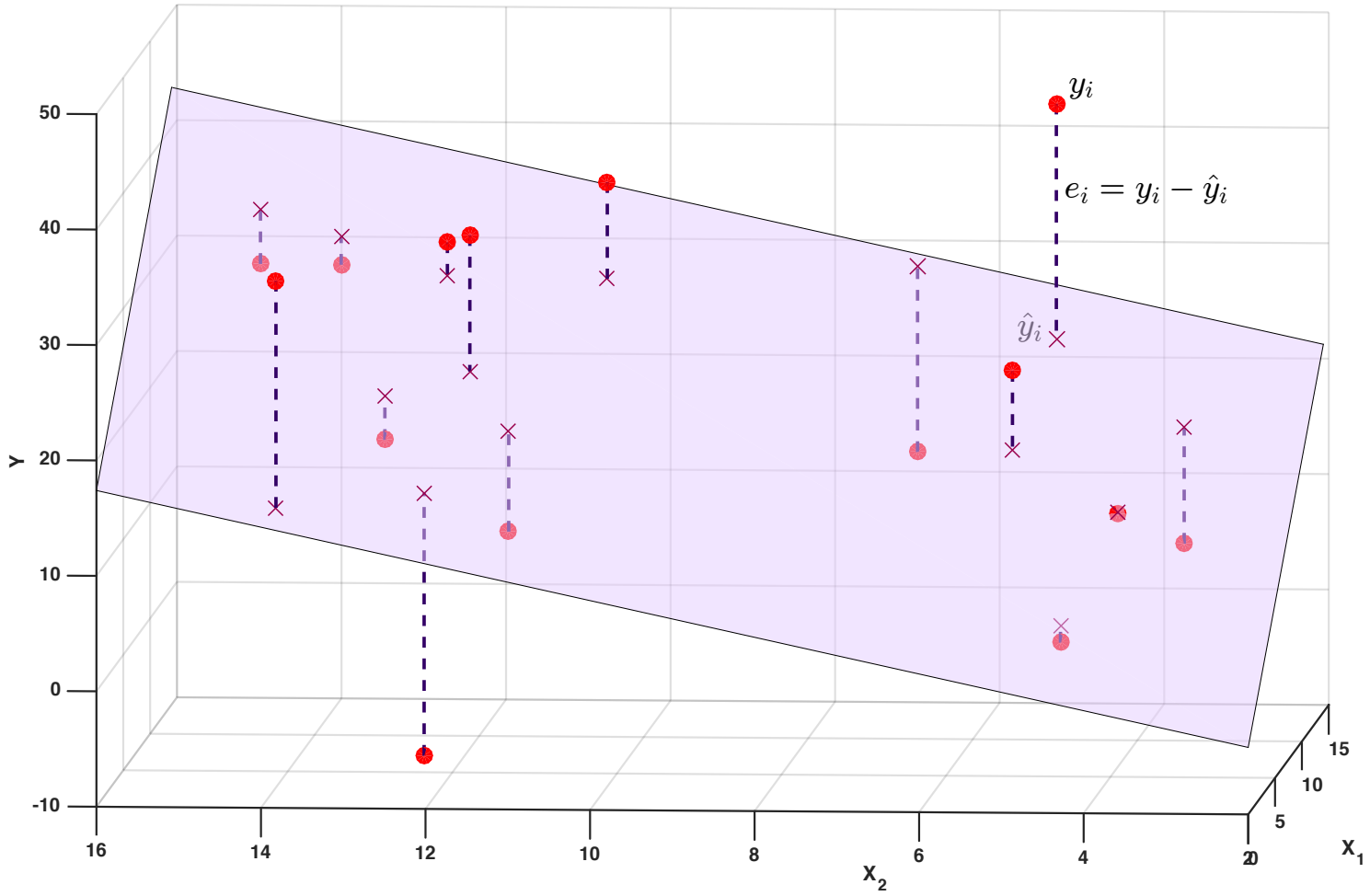
$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \cdots + b_p x_{ip} + e_i$$

- When more than a sample is analyzed:

$$y_1 = b_0 + b_1 x_{11} + b_2 x_{12} + b_3 x_{13} + \cdots + b_p x_{1p} + e_1$$
$$y_2 = b_0 + b_1 x_{21} + b_2 x_{22} + b_3 x_{23} + \cdots + b_p x_{2p} + e_2$$
$$\ldots.$$
$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \cdots + b_p x_{ip} + e_i$$
$$\ldots.$$
$$y_n = b_0 + b_1 x_{n1} + b_2 x_{n2} + b_3 x_{n3} + \cdots + b_p x_{np} + e_n$$

A first multivariate approach: MLR

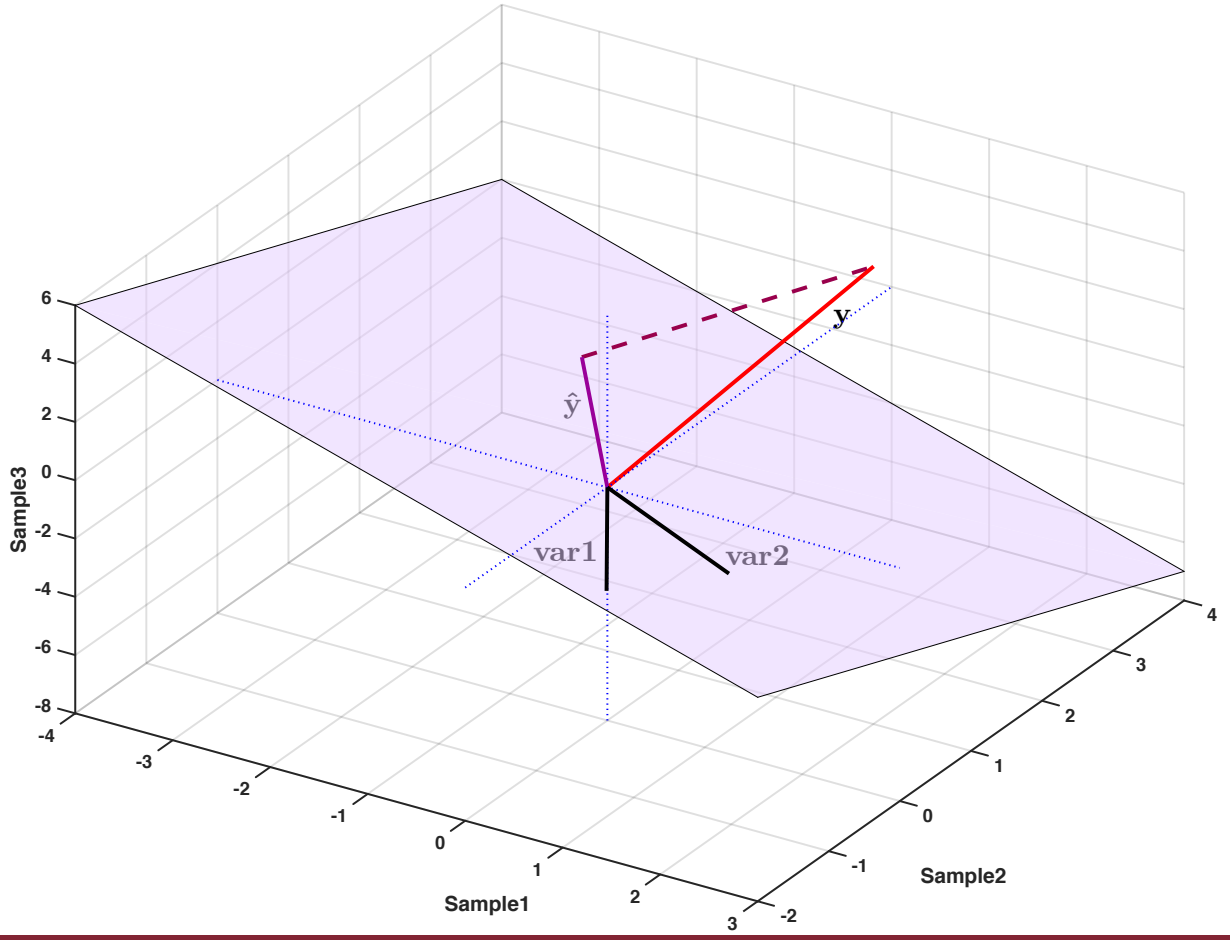
- The response estimates $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$ lie on a p-dimensional hyperplane



Least squares as projection

- Column space of a matrix: space spanned by its columns (samples are the axes).
- Regression is a projection of **y** onto the column space of **X**:

$$\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$$



Model predictions and confidence intervals

- The predicted response for any of the observations used for model building is :

$$\hat{y}_i = \mathbf{x}_i^T \mathbf{b}$$

with

$$\mathbf{x}_i^T = [x_{i1} \quad x_{i2} \quad \dots \quad \cdots \quad x_{ip} \quad 1]$$

- The corresponding prediction uncertainty is:

$$s_{\hat{y}_i}^2 = s_y^2 \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i = s_y^2 h_{ii}$$

where h_{ii} is the leverage of the i^{th} observation, which corresponds to the (i,i) element of the Hat matrix.

- Accordingly, the (1-α)% confidence interval for $\hat{y}_i$ is:

$$\hat{y}_i \pm t_{1-\alpha,n-p-1} s_{\hat{y}_i}$$

Model building and data pretreatment

- We wrote the model as: $\hat{y}_i = \mathbf{x}_i^T \mathbf{b}$

  with $\mathbf{x}_i^T = [x_{i1} \quad x_{i2} \quad \dots \quad \dots \quad x_{ip} \quad 1]$

  and $\mathbf{b} = [b_1 \quad b_2 \quad \dots \quad \dots \quad b_p \quad b_0]$
- The hyperplane fitting the data is bound to contain the point $(\bar{\mathbf{x}}, \bar{y})$, which is also the point with the smallest prediction uncertainty.

- However, in the case of mean centering for both the $\mathbf{X}$ and the $\mathbf{y}$:
  - $\boldsymbol{X}_{mc} = \boldsymbol{X} - \mathbf{1}\bar{\mathbf{x}}^T$
  - $\boldsymbol{y}_{mc} = \boldsymbol{y} - \bar{\mathrm{y}}$

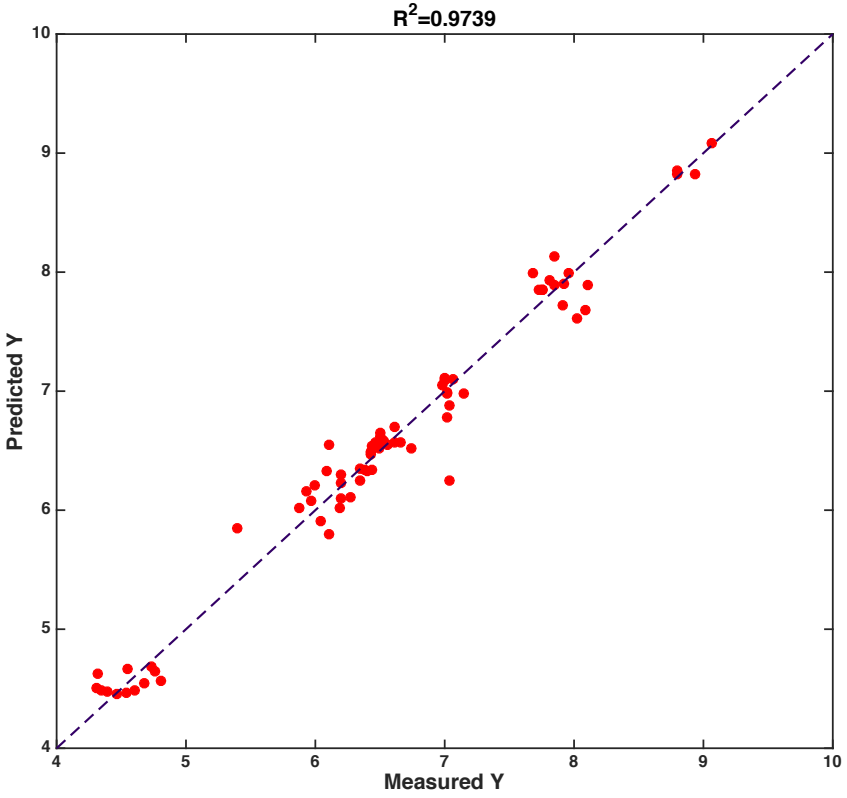
  the hyperplane fitting the data passes through the origin $(\mathbf{0}, 0)$, so the model doesn't contain the term b_0:
- Accordingly, with centered data, the model is: $\hat{y}_{i,mc} = \mathbf{x}_{i,mc}^T \mathbf{b}_{mc}$

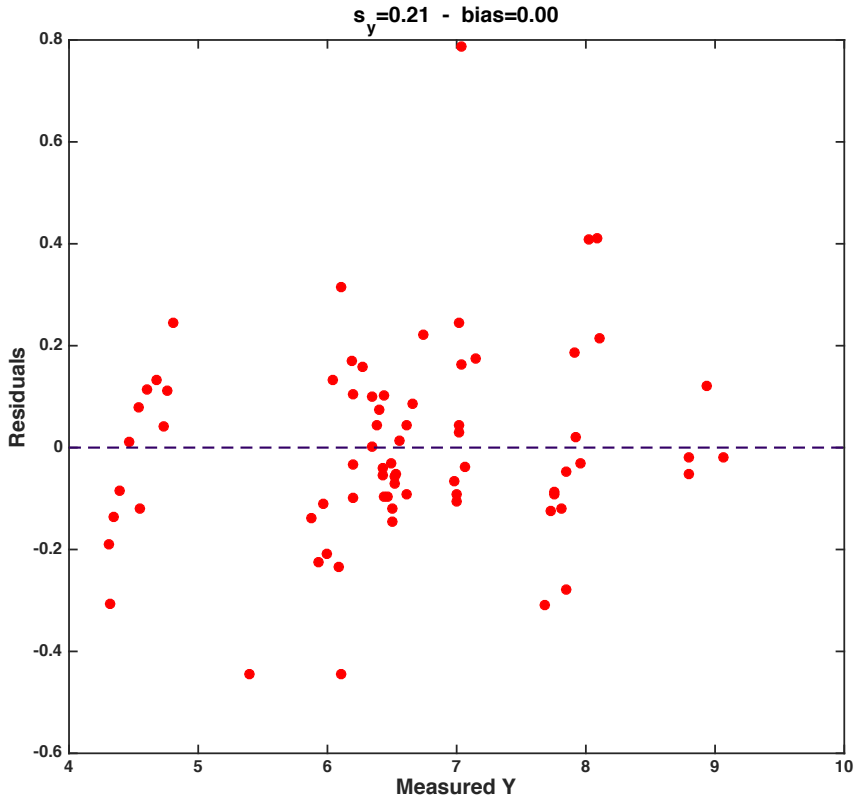
  with $\mathbf{x}_{i,mc}^T = [x_{i1} \quad x_{i2} \quad \dots \quad \dots \quad x_{ip-1} \quad x_{ip}]$

  and $\mathbf{b}_{mc} = [b_1 \quad b_2 \quad \dots \quad \dots \quad b_{p-1} \quad b_p]$

Evaluating the regression

- The quality of a multivariate regression model can be evalulated using the same figures of merit already discussed for the univariate case (bias, R^2, residuals…)





Evaluating the regression

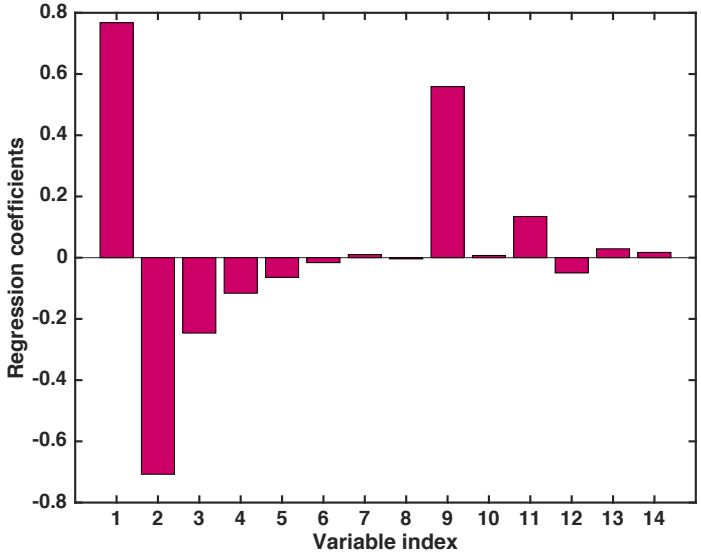
- The quality of a multivariate regression model can be evalulated using the same figures of merit already discussed for the univariate case (bias, R^2, residuals…)

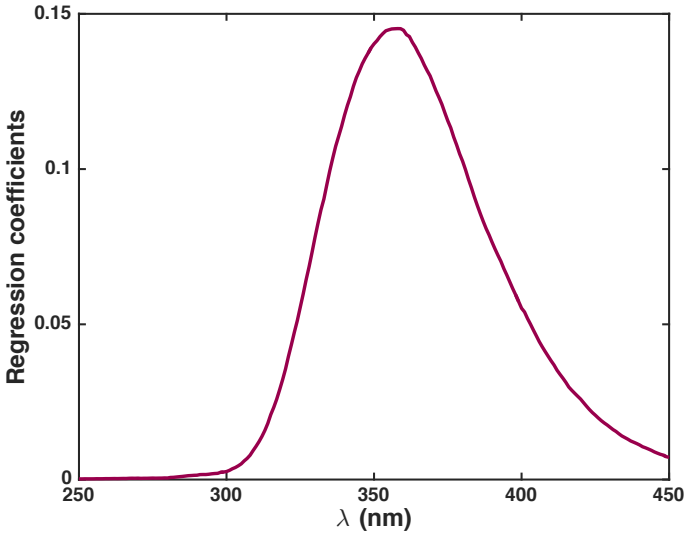
- Sometimes, to be consistent with cases where exact estimation of the degrees of freedom is not possible, the uncertainty of predictions may be expressed as:
  - Root mean square error: $RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}$

- Moreover, to compare models with a different number of variables, it is possible to define a so-called adjusted R^2 (the usual R^2 never decreases with the addition of a new variable):
  - $R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$

Interpreting the model

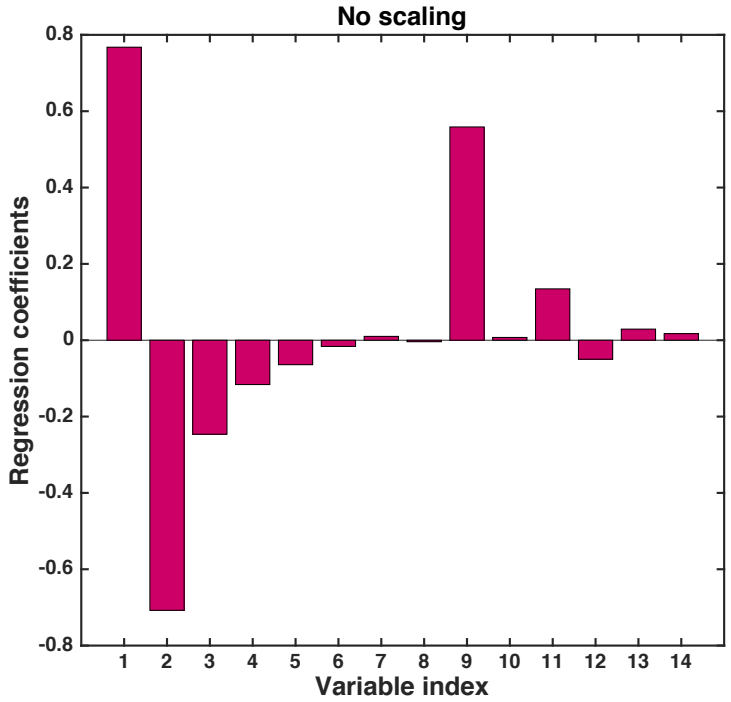
- MLR model is defined by: $\hat{\boldsymbol{y}} = \boldsymbol{Xb}$
- The relation between the $\mathbf{X}$ and the $\mathbf{y}$ is encoded in the regression coefficients:
  - Their magnitude and sign reflect the contribution of the individual X-variables in determining the value of the predicted response.
  - In the absence of interferents, they correspond to the profile of the pure constituent of interest

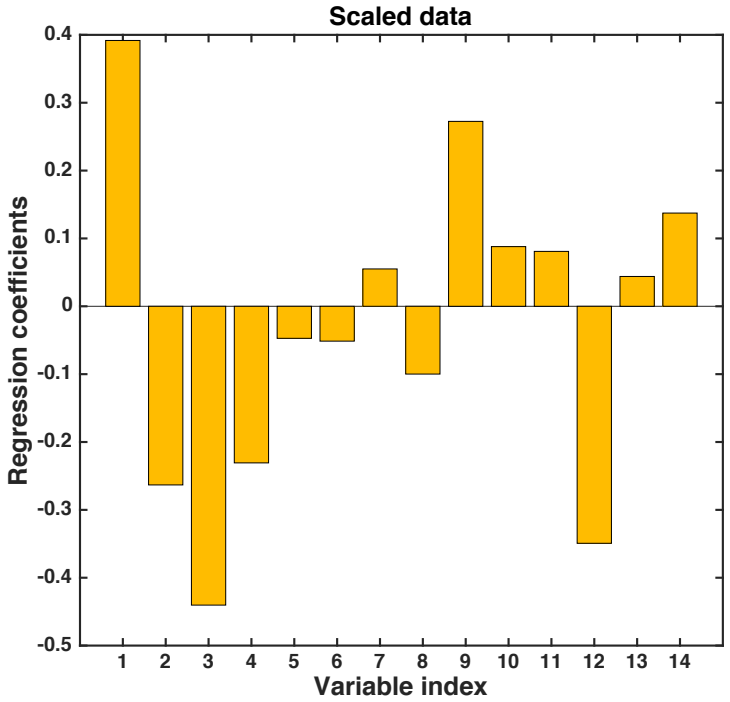




Interpreting the model: Things to be aware of

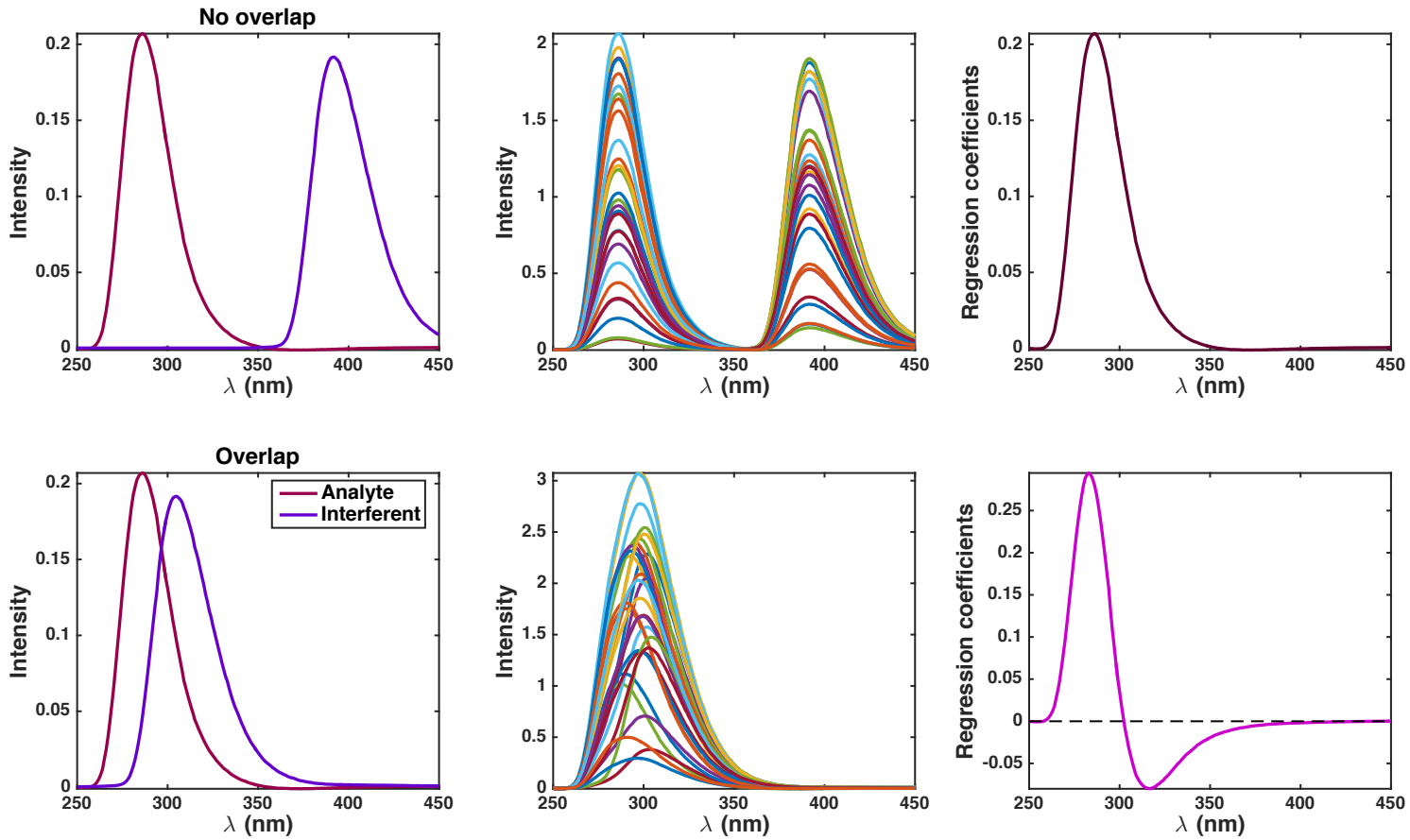
- The absolute value of the regression coefficients is influenced by the scales of the $\mathbf{X}$ and the $\mathbf{y}$ variables:
  - They can have a high magnitude just due to the relative scales of that particular X variable and the y.
  - This problem can be solved by variable scaling (as already discussed).





Interpreting the model: Things to be aware of - 2

- The presence of interferents makes interpretation of the regression coefficients less straightforward:
  - Signal of the pure constituent orthogonalized with respect to the contributions of all the interferents



What if multiple ys?

- Each y is separately regressed on **X**

$$y_{i1} = b_{01} + b_{11}x_{i1} + b_{21}x_{i2} + b_{31}x_{i3} + \cdots + b_{p1}x_{ip} + e_{i1}$$
$$y_{i2} = b_{02} + b_{12}x_{i1} + b_{22}x_{i2} + b_{32}x_{i3} + \cdots + b_{p2}x_{ip} + e_{i2}$$
$$\vdots$$
$$y_{il} = b_{0l} + b_{1l}x_{i1} + b_{2l}x_{i2} + b_{3l}x_{i3} + \cdots + b_{pl}x_{ip} + e_{il}$$

- Problem can be stated in matrix form:

$$\begin{bmatrix} y_{11} & \cdots & y_{1l} \\ y_{21} & \cdots & y_{2l} \\ \vdots & \ddots & \vdots \\ y_{i1} & \cdots & y_{il} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nl} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} & 1 \\ x_{21} & \cdots & x_{2p} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ip} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} & 1 \end{bmatrix} \begin{bmatrix} b_{11} & \cdots & b_{1l} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pl} \\ b_{01} & \cdots & b_{0l} \end{bmatrix} + \begin{bmatrix} e_{11} & \cdots & e_{1l} \\ e_{21} & \cdots & e_{2l} \\ \vdots & \ddots & \vdots \\ e_{i1} & \cdots & e_{il} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nl} \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} = \hat{\mathbf{Y}} + \mathbf{E}$$

- As in the case of a single y, the column of ones (and the terms b_{0l}) are absent for centered data

What if multiple ys? - 2

- Calculation of the model parameters by least squares gives the solution:

$$\mathbf{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{Y}$$

- This corresponds to each column of the regression coefficient matrix being given by:

$$\boldsymbol{b}_l = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}_l$$

- Since each response $\boldsymbol{y}_l$ is predicted according to:

$$\hat{\boldsymbol{y}}_l = \boldsymbol{X}\boldsymbol{b}_l$$

building a single model on multiple responses gives identical results as building l regression models on individual responses.

- This is not necessarily the case with other regression models (e.g., Partial Least Squares Regression).

Problems with MLR

- MLR is conceptually simple and generalizes univariate LS regression.

BUT

- The core of MLR is the calculation of the model parameters by the LS approach, according to:

$$\mathbf{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{Y}$$

- Due to the term $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, it can't be applied to ill-conditioned matrices:
  - Correlated variables
  - More variables than samples

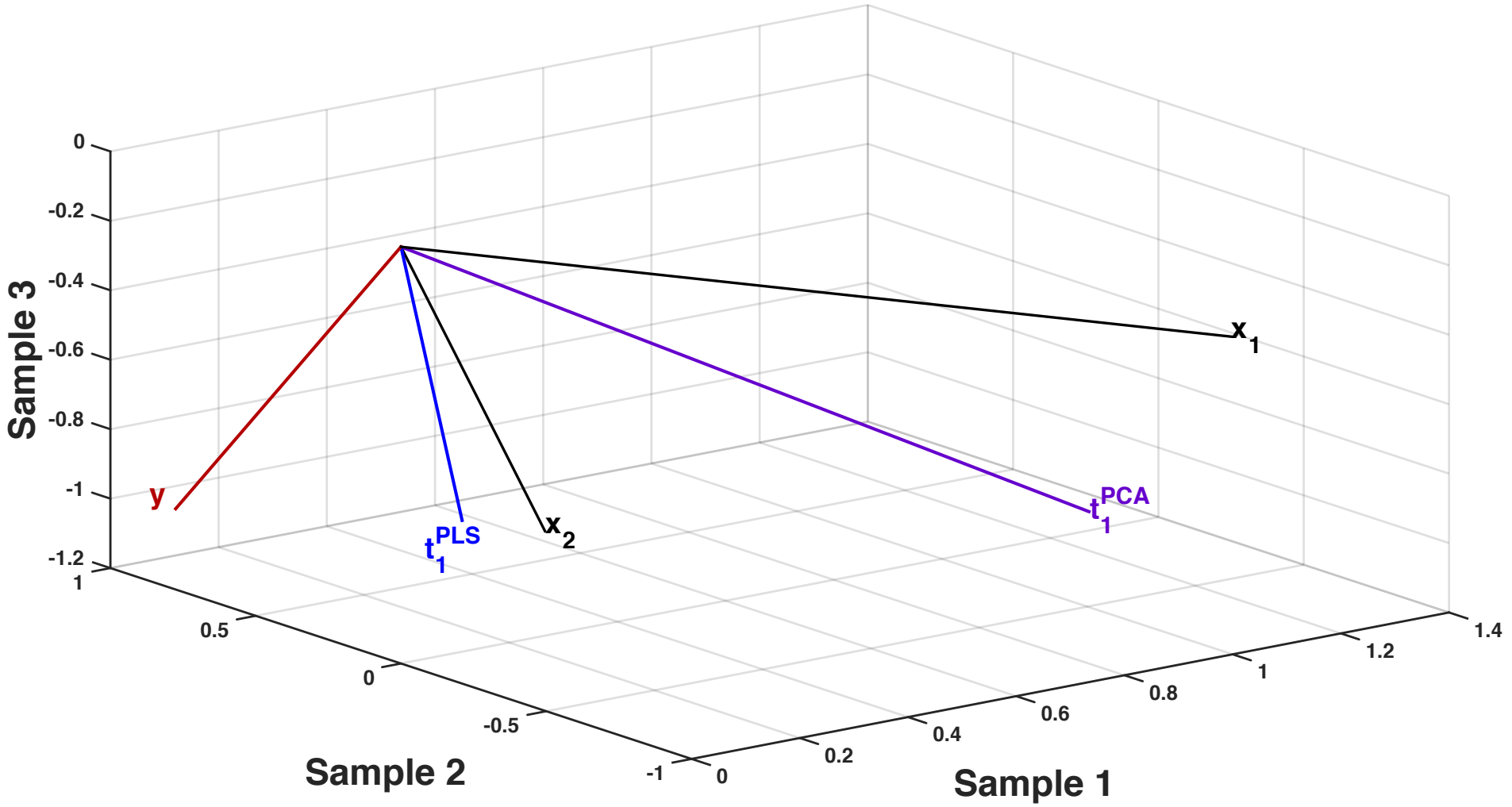
- Use of Latent variables (BILINEAR MODELING):
  - Few
  - Orthogonal

Partial Least Squares Regression (PLS)

- Scores are extracted from the **X** block so to be at the same time:
  - Explanatory of a large amount of the variance in **X**
  - Predictive for **y** (explanatory of a large amount of the variance in **y**).
- These conditions can be mathematically summarized by the following conditions/characteristics:
  - The scores $\boldsymbol{T} = [\boldsymbol{t}_1 \quad \boldsymbol{t}_2 \quad \cdots \quad \boldsymbol{t}_F] = \boldsymbol{XR}$ are extracted so to have maximum covariance with the reponse **y**: $\underset{\boldsymbol{r}_f}{\operatorname{argmax}}\, \boldsymbol{t}_f^T \boldsymbol{y} \Longrightarrow \underset{\boldsymbol{r}_f}{\operatorname{argmax}}\, \boldsymbol{r}_f^T \boldsymbol{X}^T \boldsymbol{y}$
  - $\boldsymbol{R} = [\boldsymbol{r}_1 \quad \boldsymbol{r}_2 \quad \cdots \quad \boldsymbol{r}_F]$ is a matrix collecting the weights to obtain scores from the $\boldsymbol{X}$ block
  - The scores lie in the joint column space of **X** and **y** and are orthogonal.
  - The variance in **y** is approximated by the same set of scores:

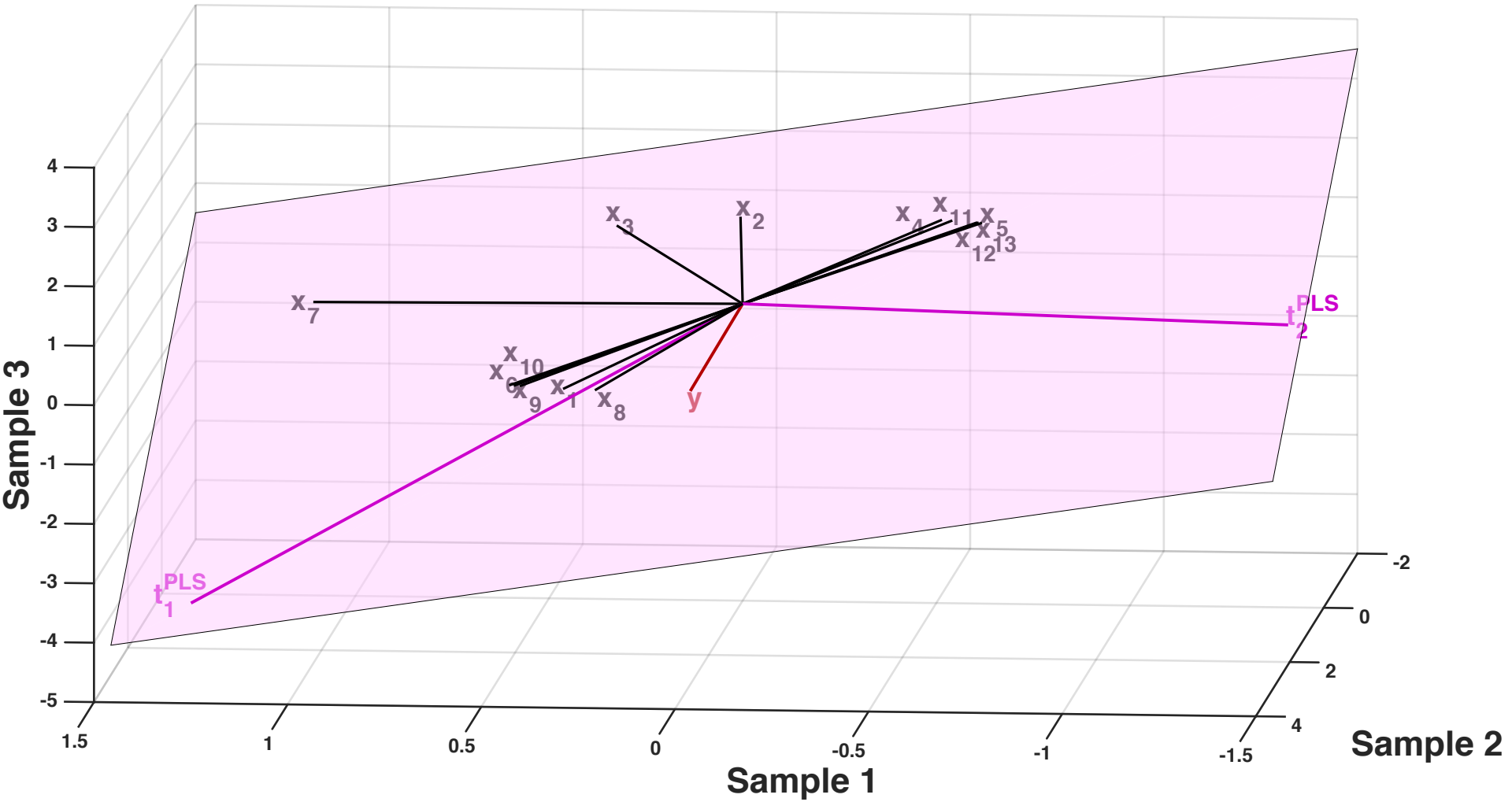
$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{q}^T + \boldsymbol{e}_y \Longrightarrow \hat{\boldsymbol{y}} = \boldsymbol{T}\boldsymbol{q}^T = q_1\boldsymbol{t}_1 + q_2\boldsymbol{t}_2 + \cdots + q_F\boldsymbol{t}_F$$

The PLS criterion graphically explained



- With respect to PCA, scores are «rotated» towards **y**

With more than 2 X-variables: The projection



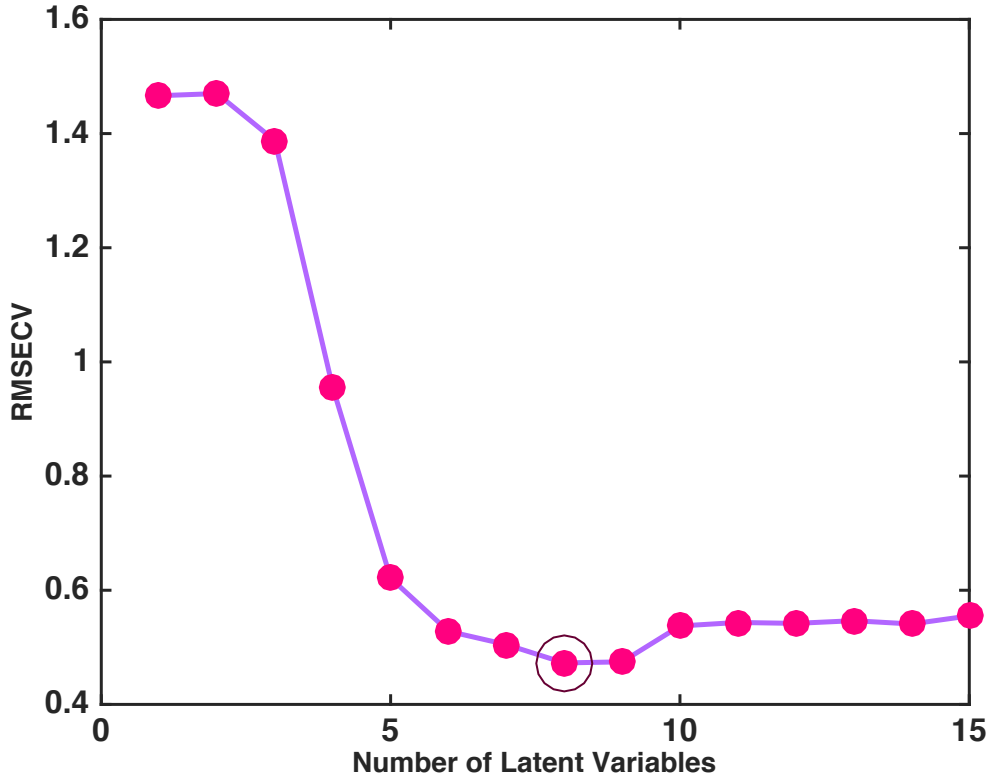
PLS: Predicting the response

- The peculiarity of PLS is to extract scores which are directly explanatory also of the **y** variance:
  - They can be used to approximate (predict) the response: $\hat{\boldsymbol{y}} = \boldsymbol{T}\boldsymbol{q}^T$
- Scores are extracted from **X** as linear combination of the variables, through the weight matrix:
  - $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{R}$
- It is then possible to directly relate the predicted response to the predictor matrix **X** as:
  - $\hat{\boldsymbol{y}} = \boldsymbol{T}\boldsymbol{q}^T = \boldsymbol{X}\boldsymbol{R}\boldsymbol{q}^T = \boldsymbol{X}\boldsymbol{b}_{PLS}$
  - $\boldsymbol{b}_{PLS}$ is a vector of regression coefficients; $\boldsymbol{b}_{PLS} = \boldsymbol{R}\boldsymbol{q}^T$
  - The possibility of expressing the PLS model in the form: $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}_{PLS}$ allows to extend to the method all the considerations already done for MLR and PCR.
  - Prediction of the response for a new sample is carried out according to:

$$\hat{\boldsymbol{y}}_{new} = \boldsymbol{X}_{new}\boldsymbol{b}_{PLS}$$

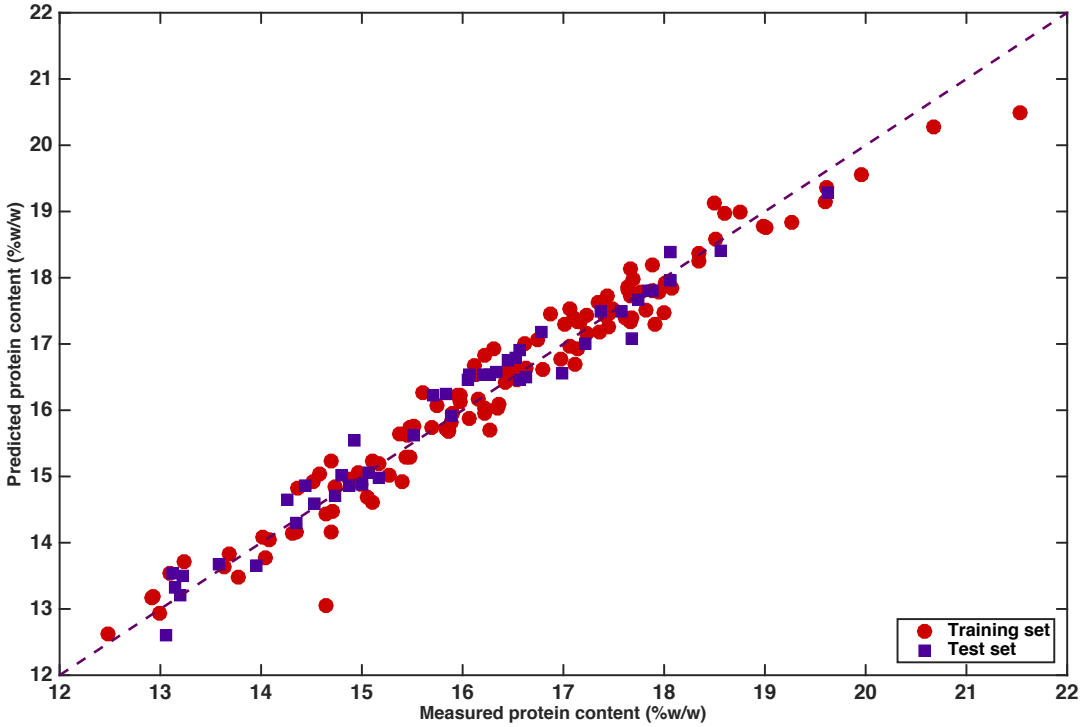
PLS: Model building and predictions

- Building a PLS model requires choosing the number of PLS components (Latent variables, LVs):
  - Selection is generally carried out through a cross-validation procedure
  - In the example, 8 components (corresponding to the minimum of RMSECV) are retained in the final model



PLS: Model building and predictions - 2

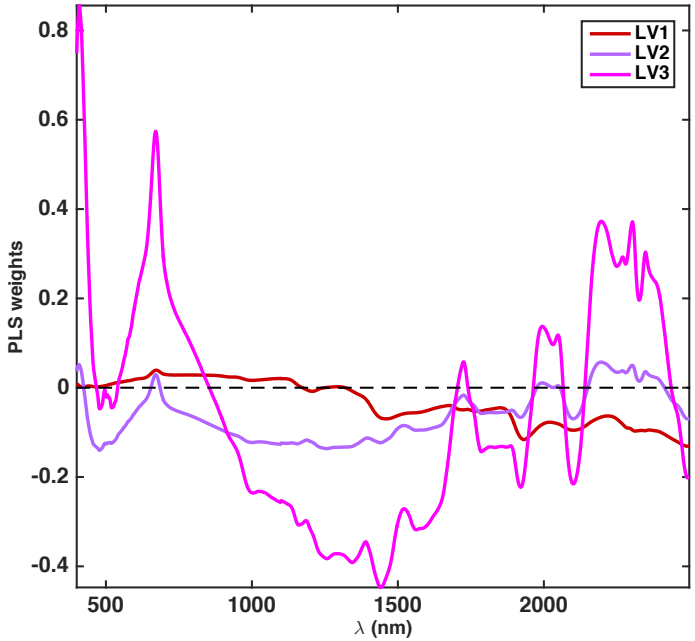
- The model is calculated on the training set and validated on the test set:
  - The «usual» figures of merit for regression can be used to evaluate the model quality

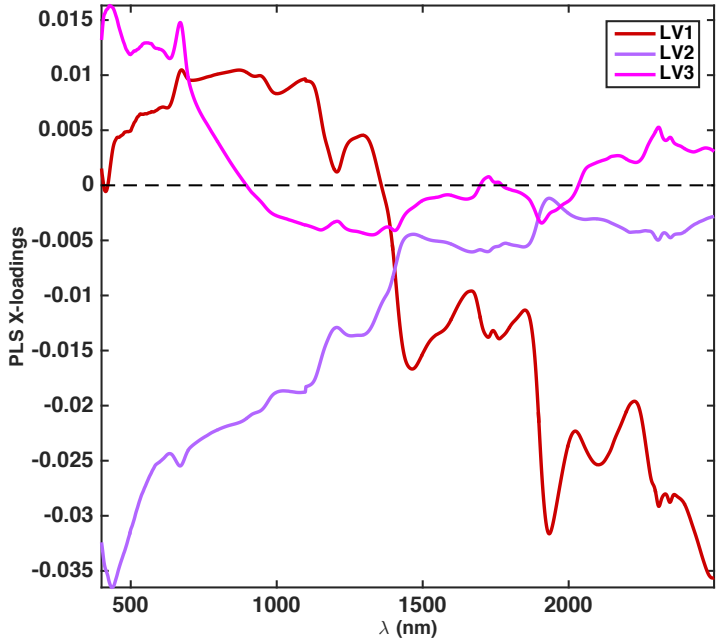


	Calibration	Validation
RMSE	0.343	0.297
bias	0.00	-0.09
R^2	0.959	0.966

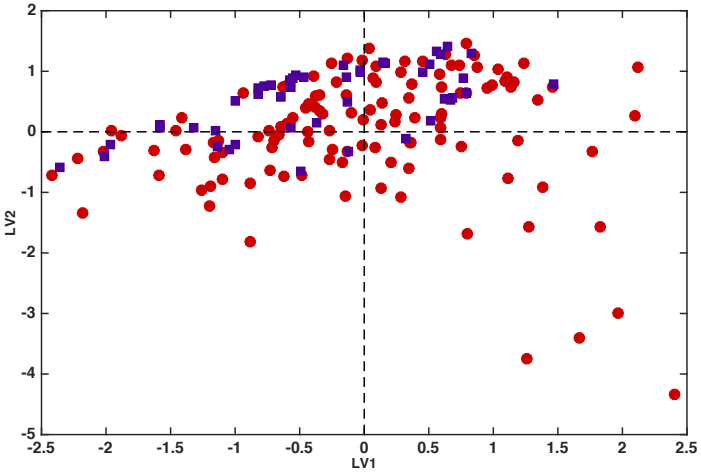
PLS: A bit more on the model

- The PLS scores are build so to capture the maximum covariance between **X** and **y** and to be orthogonal.
- These conditions are accomplished through the introduction of the weights **R**.
- Once the scores are calculated, a further set of loadings **P** is needed:
  - These loadings produce the best approximation of the variability in **X** given the set of scores **T**: $\widehat{\boldsymbol{X}} = \boldsymbol{TP}^T$
  - In PLS, these loadings are not orthogonal; they describe only the variance in **X**.

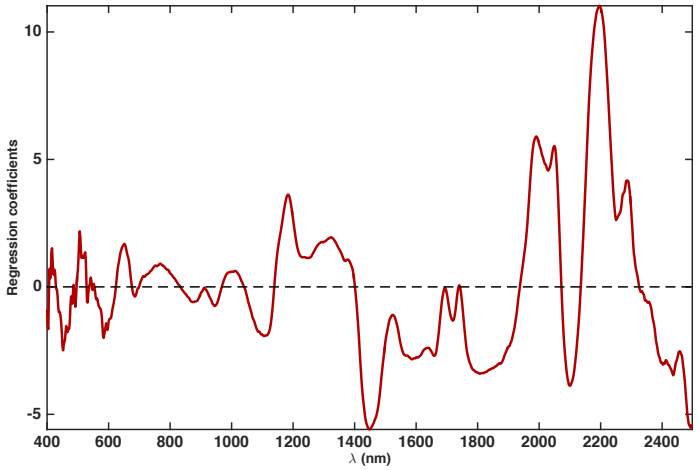




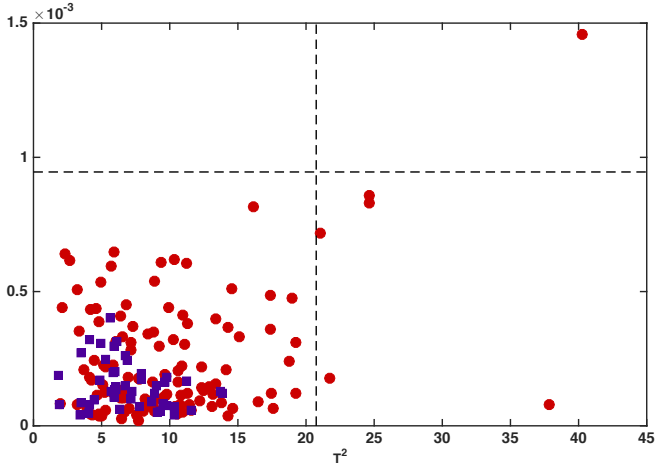
PLS for a single response: What else do we get?



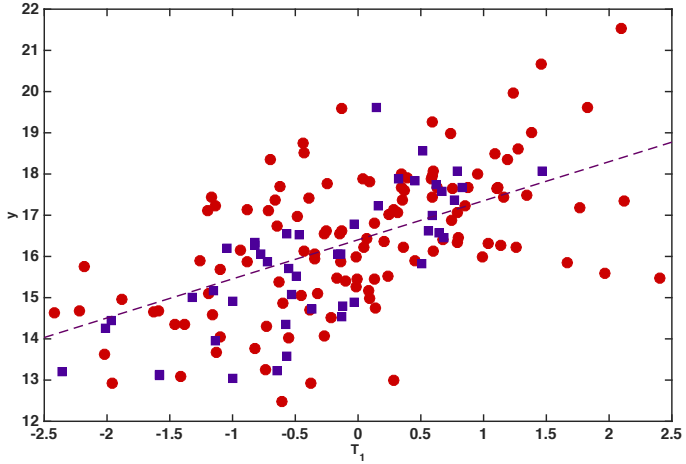
Scores plot



Regression coefficients

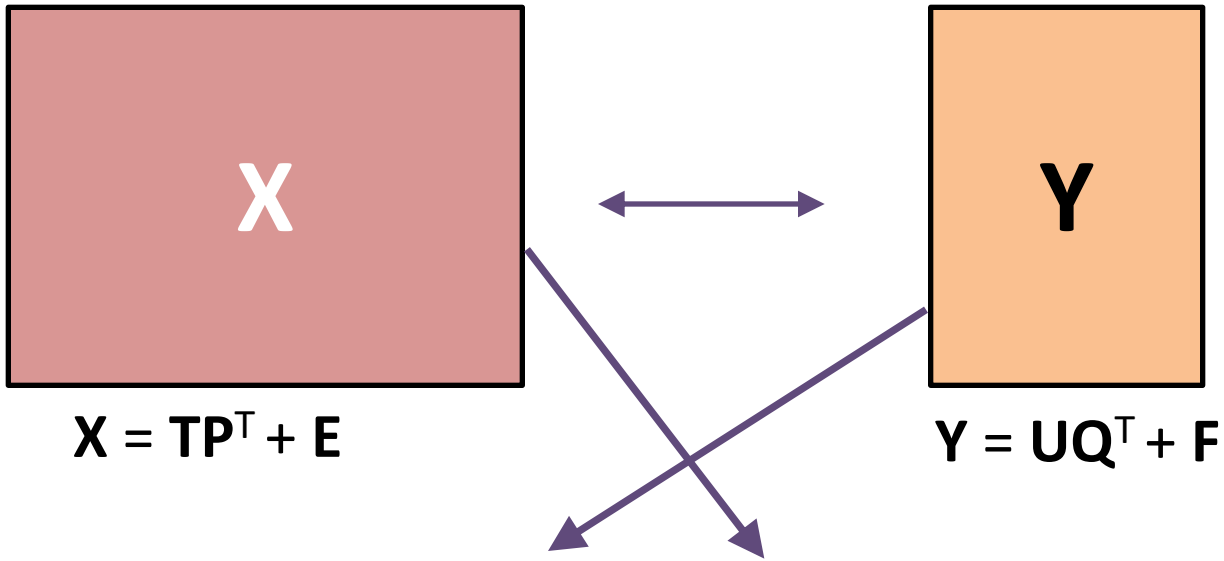


Influence plot



Inner relation

Partial Least Squares for multiple responses: The idea



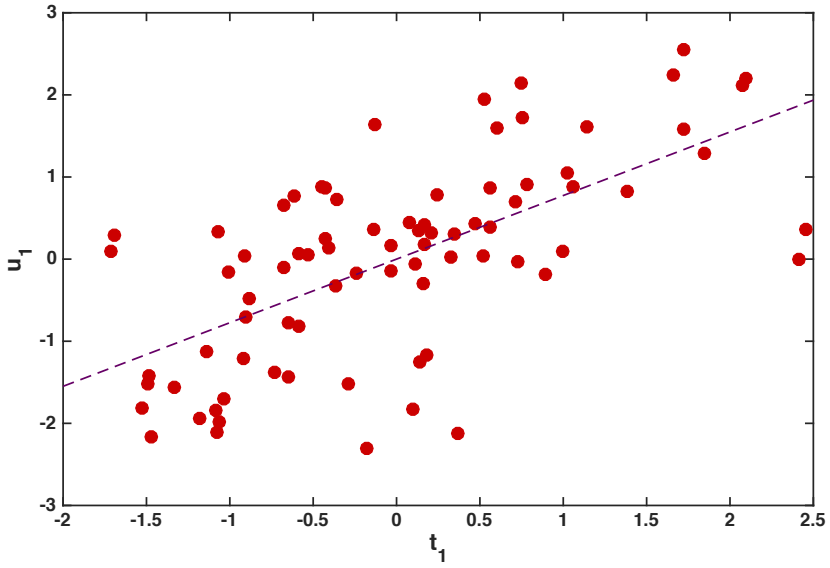
PLS for multiple Ys: A few concepts

- The responses are assumed to possess an underlying latent structure.
- Scores are built so to have maximum covariance:

$$\underset{\boldsymbol{r}_f,\boldsymbol{q}_f}{\operatorname{argmax}}\, \boldsymbol{t}_f^T \boldsymbol{u}_f \Longrightarrow \underset{\boldsymbol{r}_f,\boldsymbol{q}_f}{\operatorname{argmax}}\, \boldsymbol{r}_f^T \boldsymbol{X}^T \boldsymbol{y}\boldsymbol{q}_f$$

- Regression is accomplished between the scores (univariate linear regression):

$$\widehat{\boldsymbol{u}}_f = \boldsymbol{t}_f c_f \Rightarrow \widehat{\boldsymbol{U}} = \boldsymbol{T}\boldsymbol{C}$$



PLS for multiple Ys: Summarizing the model

- The responses are approximated by the bilinear model of the block as:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{Q}^{T}.$$

- The values of the Y-scores are predicted from the X-scores (inner relation):

$$\widehat{\boldsymbol{U}} = \boldsymbol{T}\boldsymbol{C}$$

- X-scores are calculated from the X-variables through the X-weights:

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{R}$$

- Also in the case of PLS regression for multiple Y, prediction of the responses can be expressed directly in terms of the original variables, by combining the information above into a single equation:

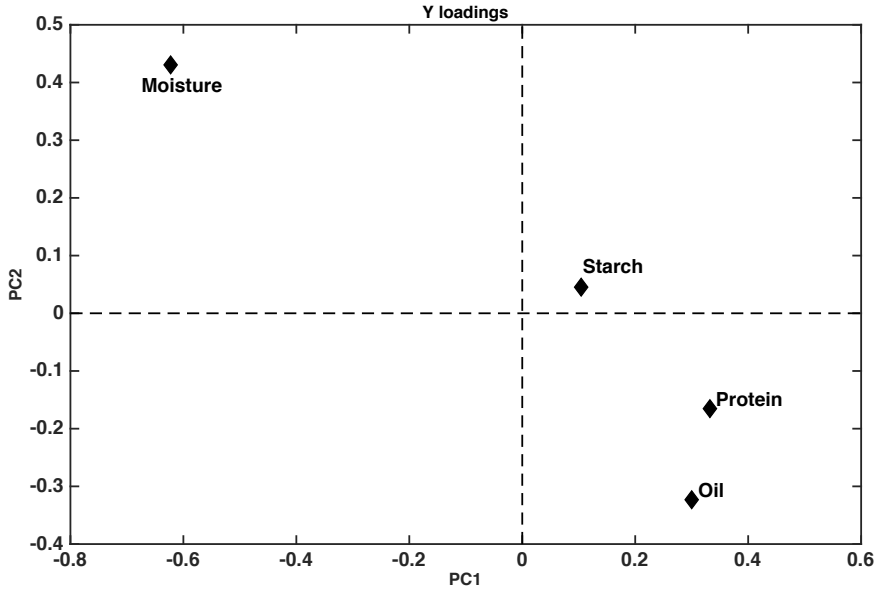
$$\widehat{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{Q}^{T} = \boldsymbol{T}\boldsymbol{C}\boldsymbol{Q}^{T} = \boldsymbol{X}\boldsymbol{R}\boldsymbol{C}\boldsymbol{Q}^{T} = \boldsymbol{X}\boldsymbol{B}_{PLS}$$

- Where the matrix of regression coefficients is equal to:

$$\boldsymbol{B}_{PLS} = \boldsymbol{R}\boldsymbol{C}\boldsymbol{Q}^{T}$$

PLS for multiple Ys: Considerations

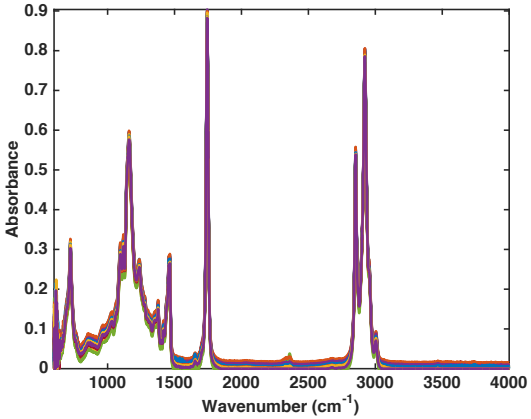
- Contrarily to the case of MLR and PCR, in PLS there is a substantial difference in building a single model to calibrate all the responses or L individual models, one for any response.
- A single PLS-2 (multiple-y) model is recommended when the Y variables are correlated and share an underlying structure.
- PCA of the Y matrix can help choosing the most appropriate approach

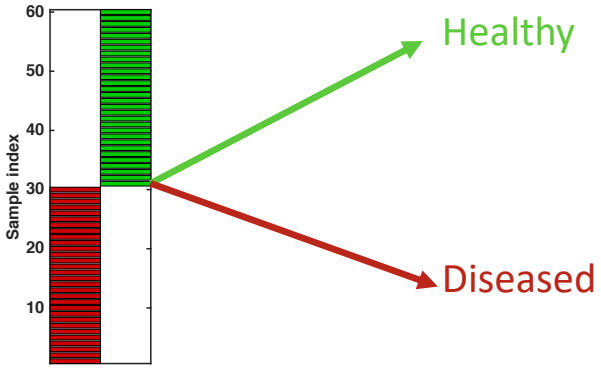


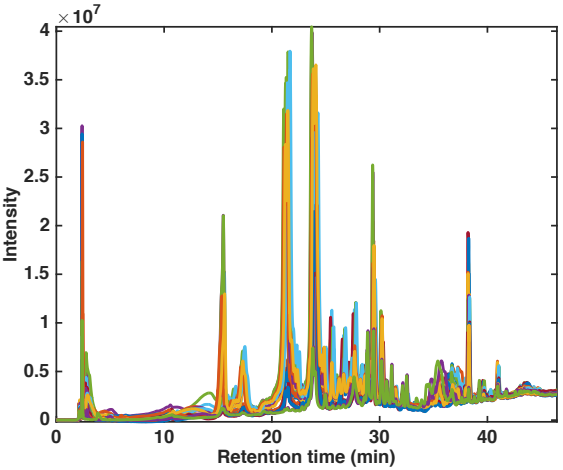
Partial Least Squares Discriminant Analysis (PLS-DA)

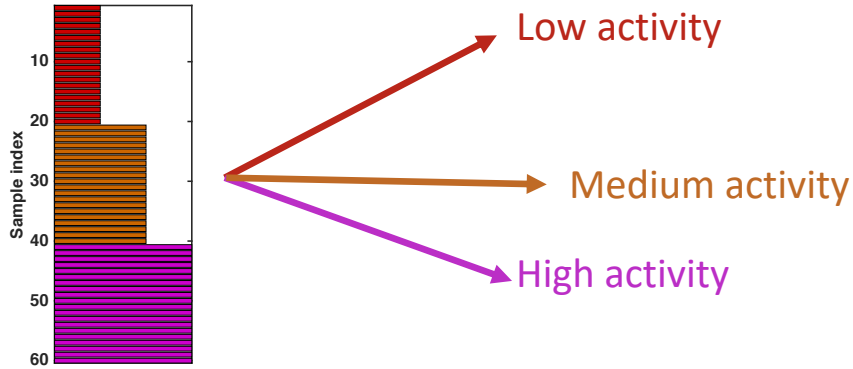
Classification: Intro

- Sometimes, data are collected in order to predict a qualitative property, i.e., a response that can take only discrete values



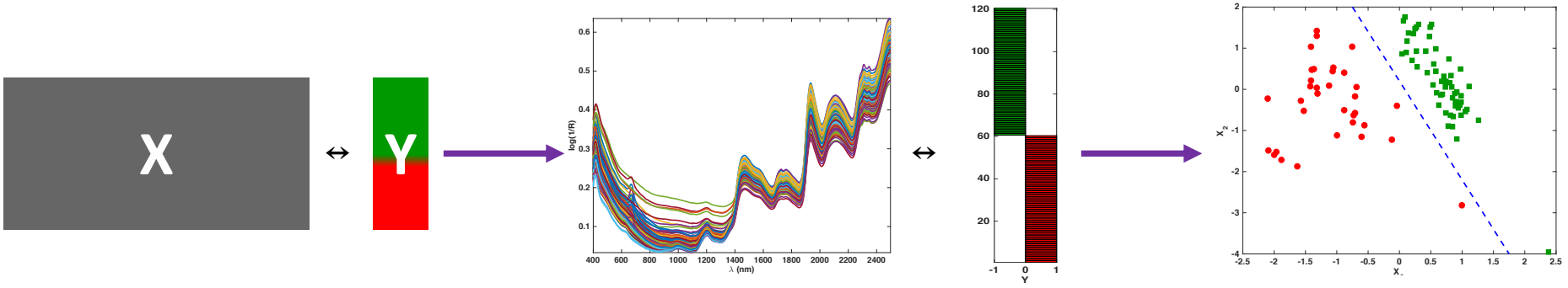






Classification

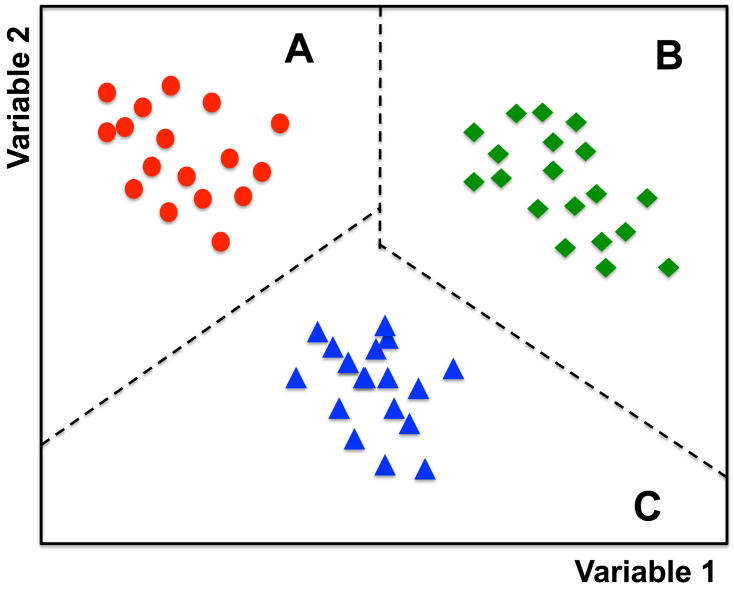
- Each value that the discrete response can take is called a class or category.
- Category or class is a (ideal) group of objects sharing similar characteristics.

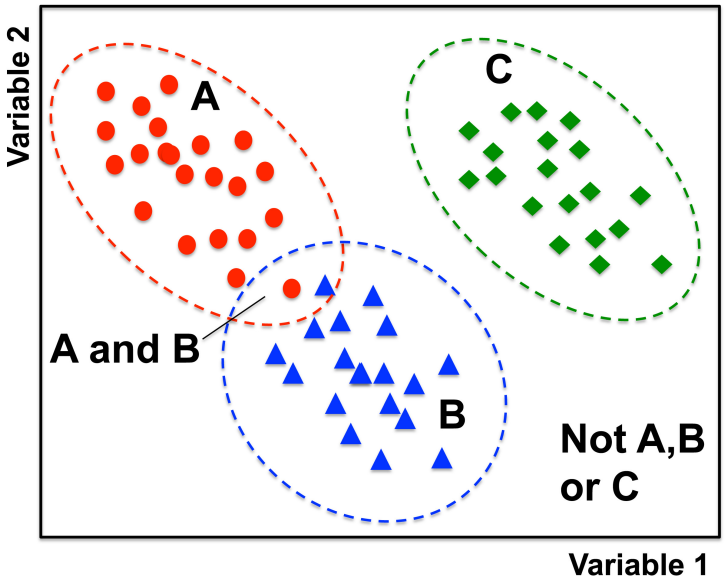


CLASSIFICATION:

- "To find a criterion to assign an object (sample) to one category (class) based on a set of measurements performed on the object itself"
- In classification, categories are defined a priori (≠ cluster analysis)

Classification approaches





DISCRIMINANT TECHNIQUES:

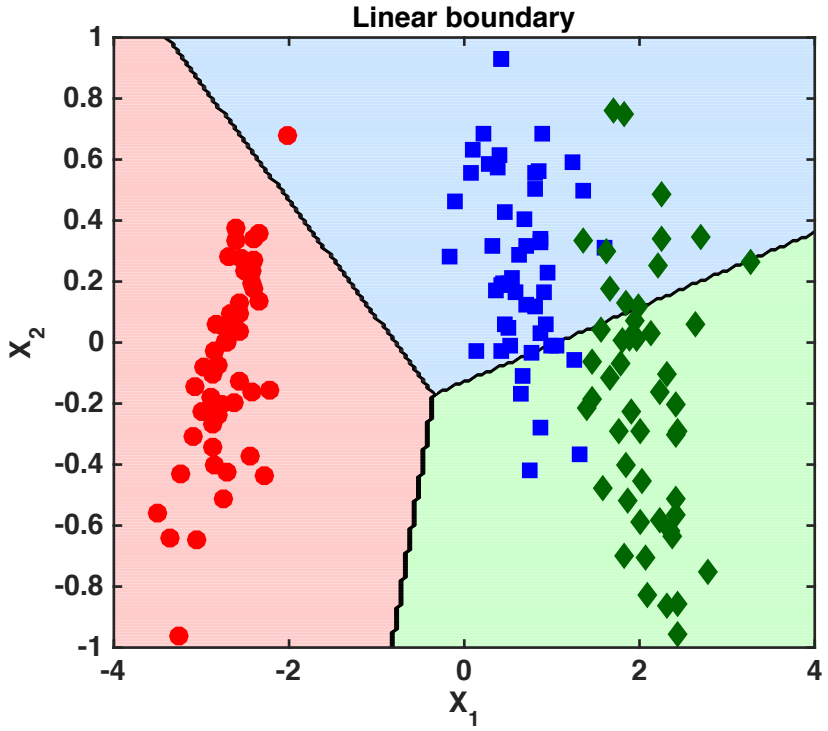
- Estimate the optimal boundaries (decision surfaces) which separate the different classes in the multidimensional space.
- Samples are always classified to one and only one of the categories in the training set

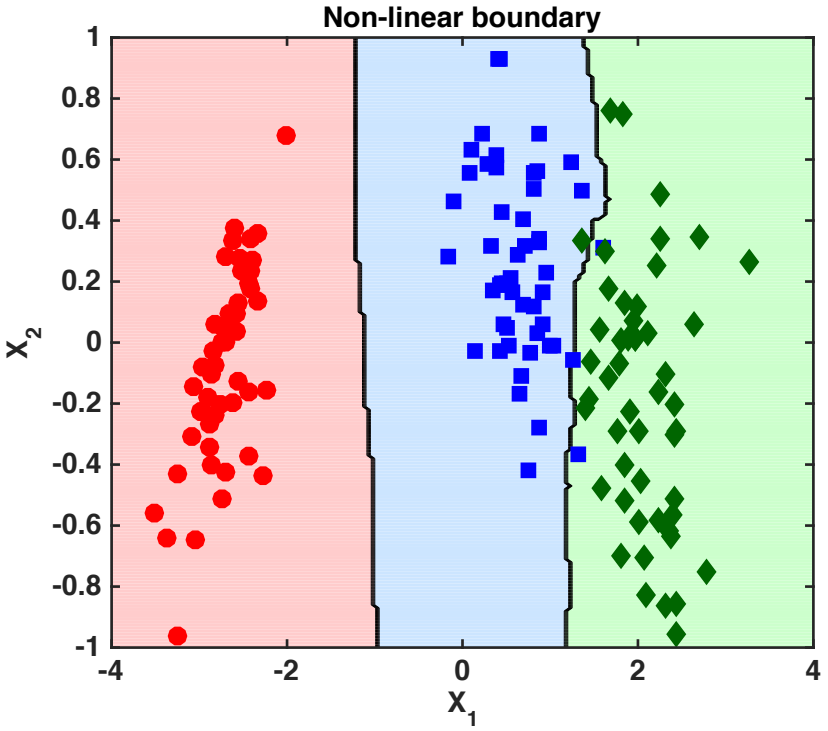
MODELING TECHNIQUES:

- Focus on looking for similarities among samples belonging to the same class.
- Each category is modeled individually.
- A sample can be assigned to one class, to more than one class or to no class at all.

Discriminant techniques

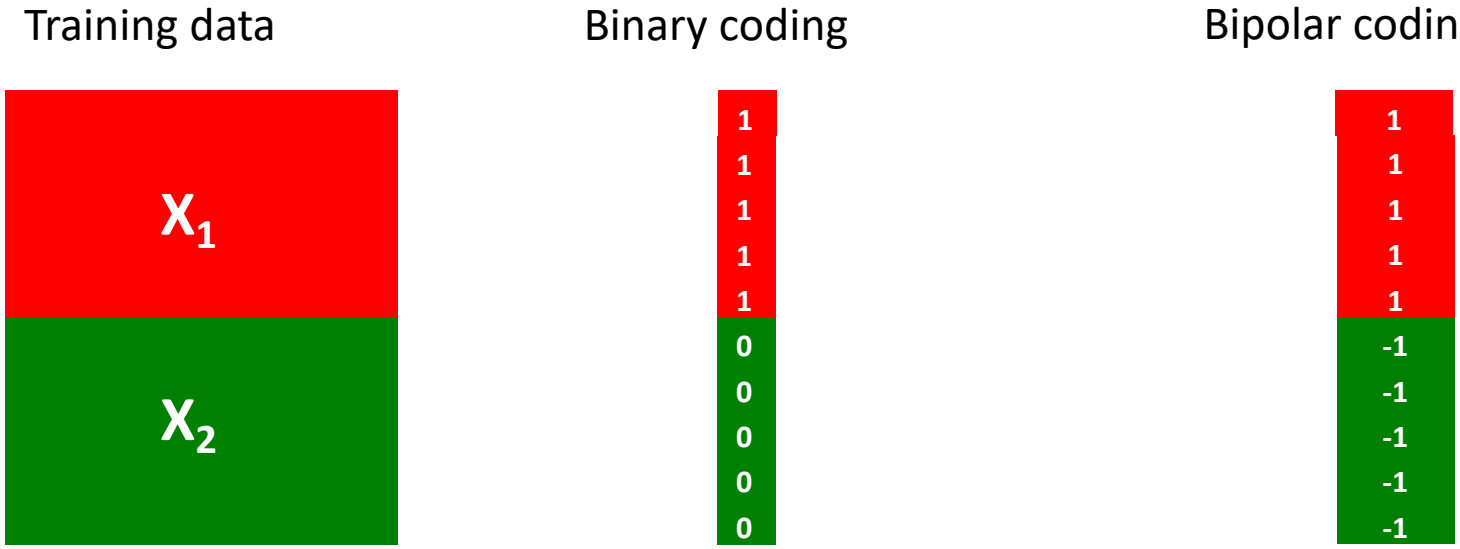
- Look for surfaces (decision boundaries) dividing the multivariate space into as many regions as the number of classes in the training set.
- Associate to any vector of measurements $\boldsymbol{x}_i$ a predicted class $\hat{c}_i$ which can be one and only one of the categories represented in the training set.
- Classification boundaries can take any arbitrary mathematical form



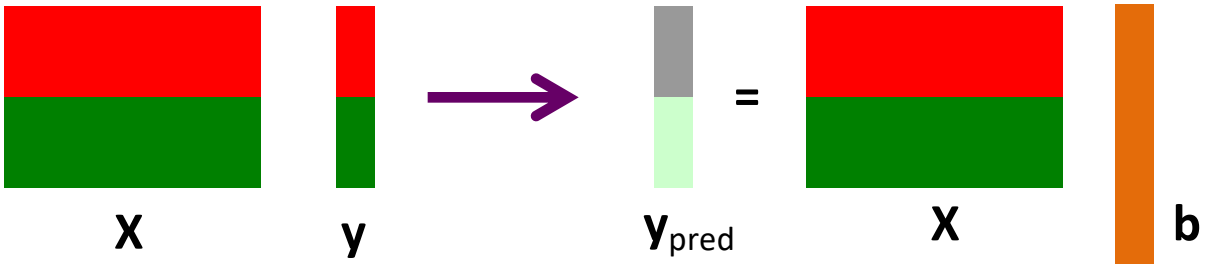


Partial Least Squares-Discriminant Analysis (PLS-DA)

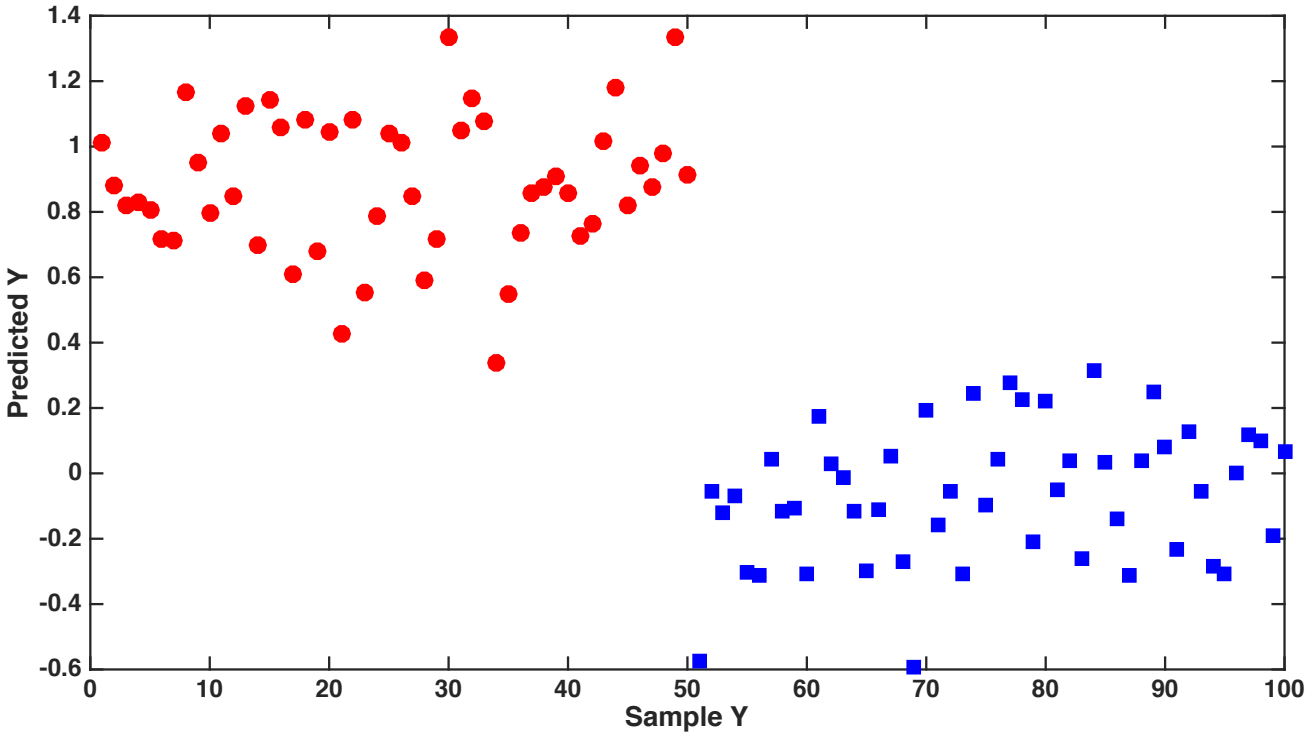
- Useful when the number of variables is higher than that of available samples and with correlated predictors
- Based on the PLS algorithm:
  - Classification problem should be re-formulated as regression
- Class in encoded in dummy Y vector



Partial Least Squares-Discriminant Analysis (PLS-DA) – 2

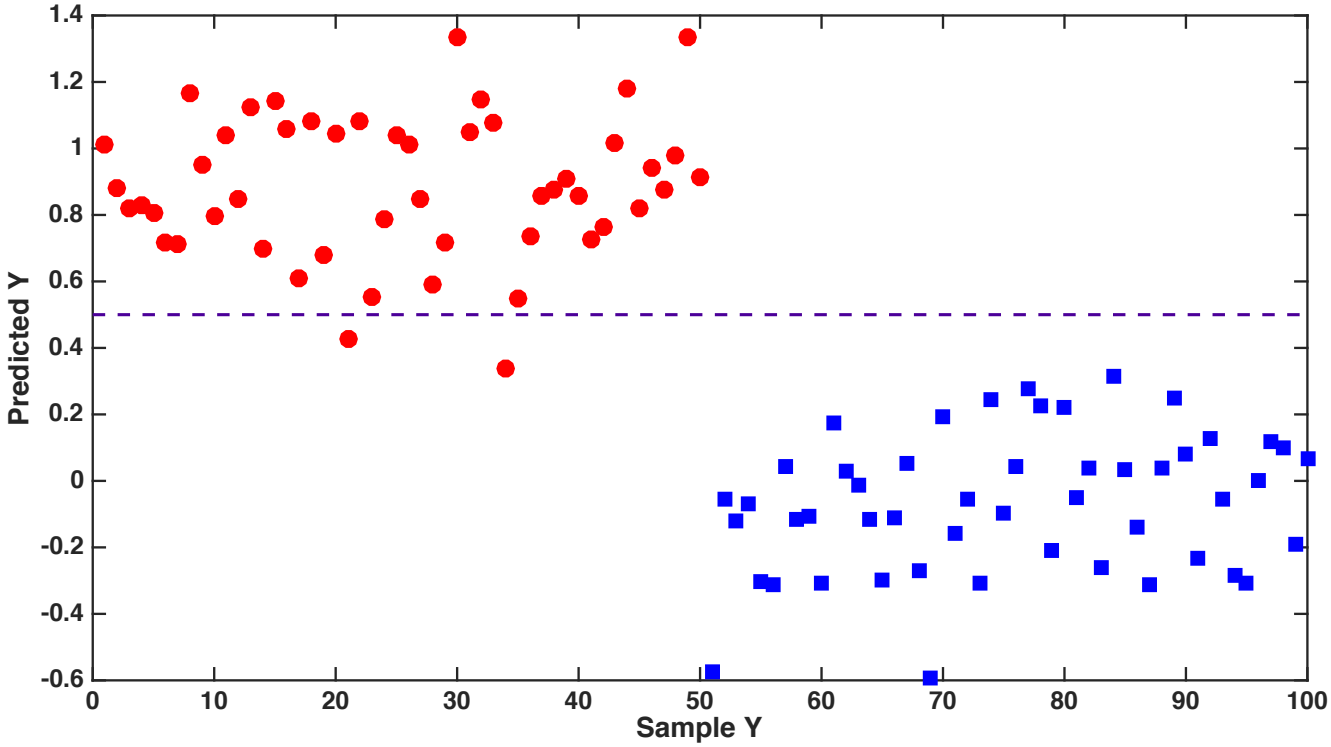


- While "true" Y values are binary-coded, predictions are real valued.



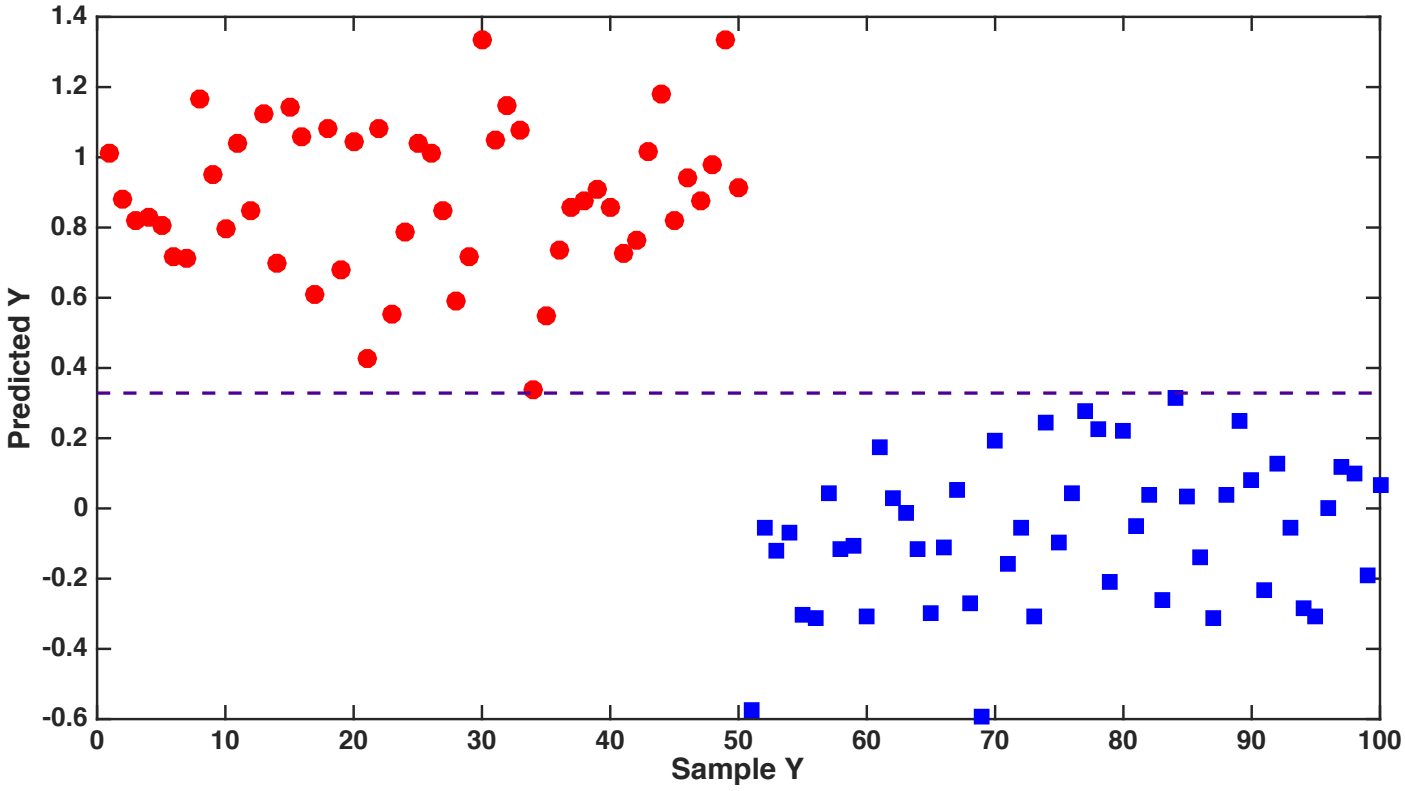
Partial Least Squares-Discriminant Analysis (PLS-DA) – 3

- Classification is accomplished by setting a proper threshold to the predicted Y values.
- The "natural" threshold is 0.5:
  - Y_{pred}>0.5 →class 1
  - Y_{pred}<0.5 →class 2



Partial Least Squares-Discriminant Analysis (PLS-DA) – 4

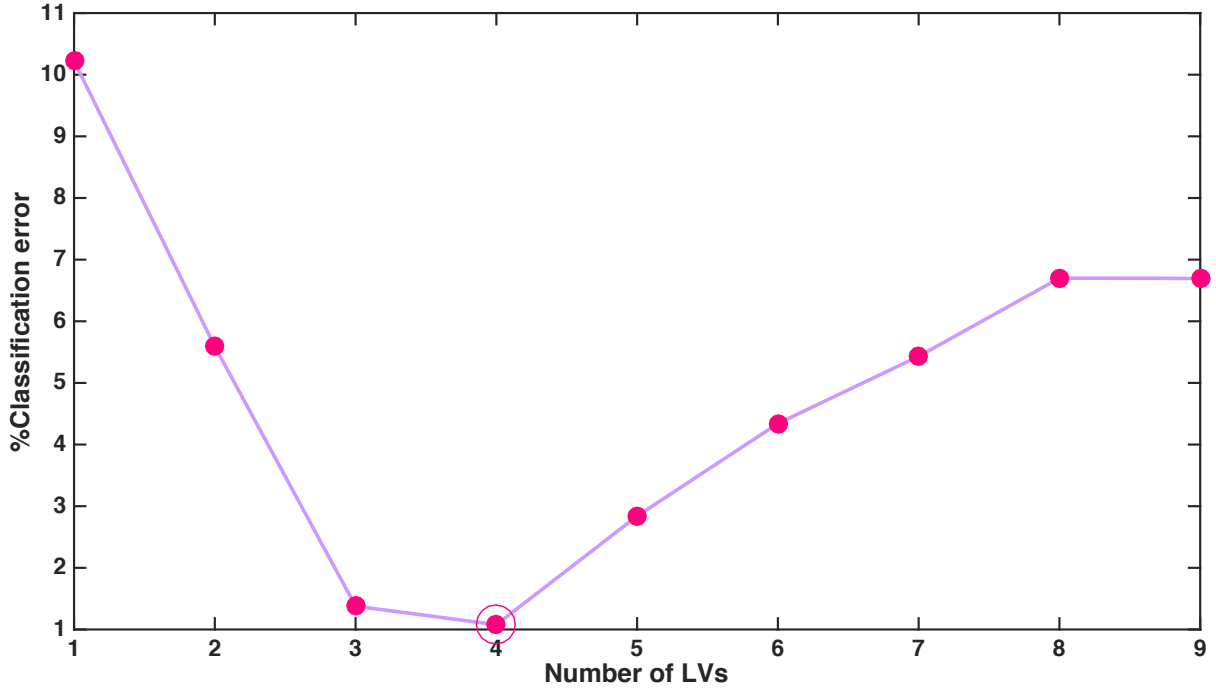
- Different methods have been proposed in the literature to find alternative "optimal" threshold values (see later).
- In the example, setting the value to 0.33 allows the correct classification of all samples:



Model complexity

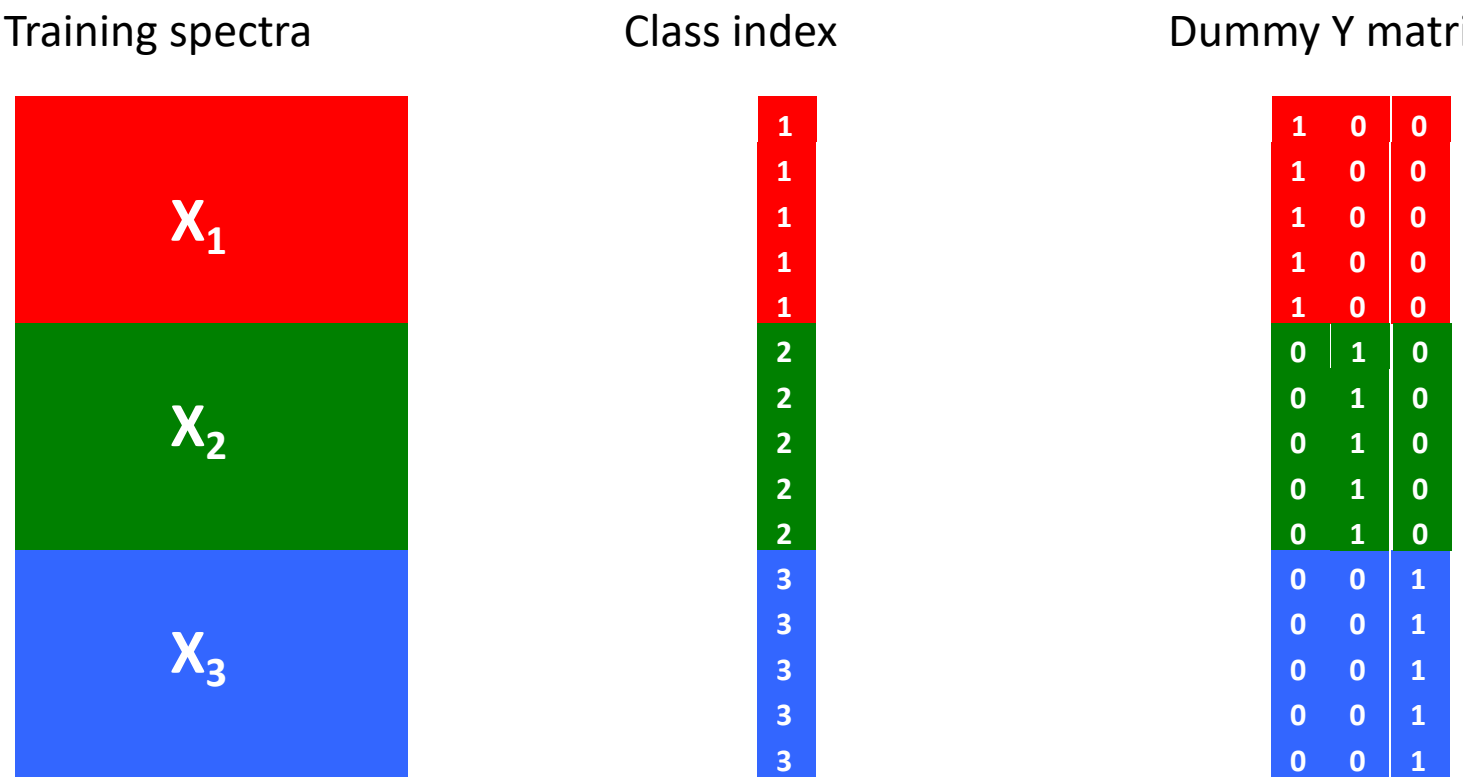
- PLS-DA is a bilinear model:
  - Optimal number of components (LVs) should be selected
- Error criterion in cross-validation is (usually) based on the number (percentage) of mis-classifications:

$$\%CE(tot) = 100 - \%NER(tot) = 100 \times \frac{\sum_{g=1}^{G} n_{g,misclassified}}{n_{tot}}$$



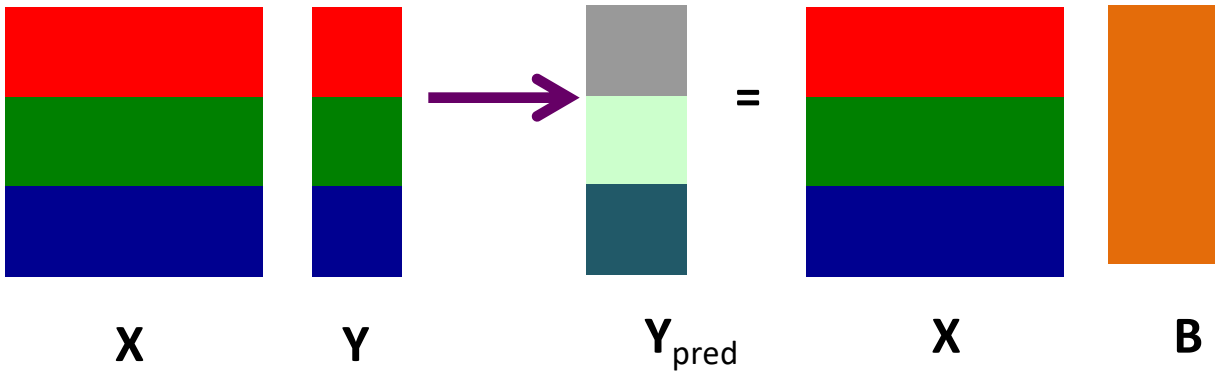
With more than two classes

- Instead of a binary vector, a dummy binary matrix is used to code for class belonging
- Y spans a G-1 dimensional space (G being the number of classes)



PLS-DA for more than two classes

- Model is built using PLS-2 algorithm
- A matrix of regression coefficient is obtained

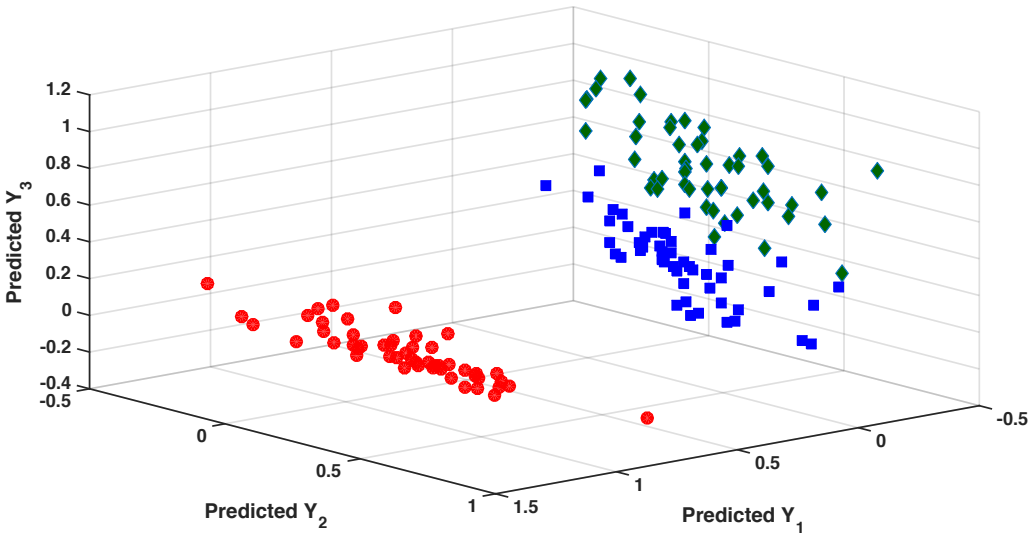


- For each sample, predicted Y is a row vector (real valued) having as many columns as the number of classes.
- Different options to achieve classification based on the values of predicted Y

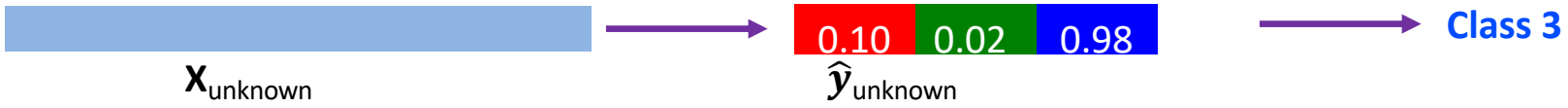
PLS-DA for more than two classes - 2

- Predicted y is real-valued:

"true" y			predicted y		
1	0	0	1.03	0.09	-0.10
1	0	0	0.68	0.21	0.08
1	0	0	0.99	-0.10	0.01
1	0	0	0.96	0.18	-0.14
1	0	0	0.79	0.02	0.25
0	1	0	0.14	0.94	0.07
0	1	0	-0.01	1.12	0.12
0	1	0	0.08	0.89	-0.02
0	1	0	0.33	0.45	0.25
0	1	0	0.15	0.72	0.06
0	0	1	0.13	-0.18	0.85
0	0	1	0.21	0.17	0.56
0	0	1	-0.09	0.32	0.69
0	0	1	0.12	0.06	1.01
0	0	1	0.02	-0.03	0.98



- In the simplest multiclass PLS-DA implementation, sample is assigned to the class corresponding to the highest value of predicted Y

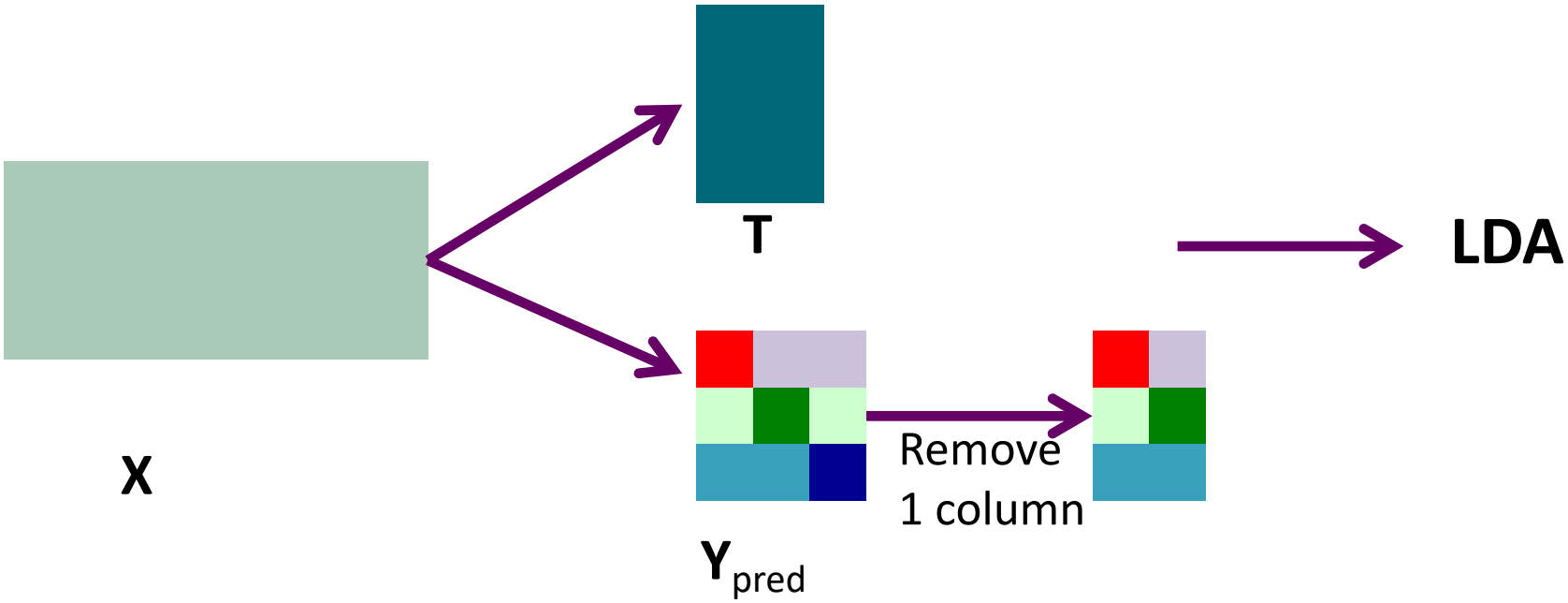


$\mathbf{x}_{unknown}$ $\hat{y}_{unknown}$

PLS-DA for more than two classes - 3

- Another alternative approach to perform classification based on discriminant PLS results is to apply LDA:
  - On the predicted **Y** (after removing one of the columns)
  - On the X scores **T**

$$\hat{Y} = TQ^T = XB$$



Digression: Validation

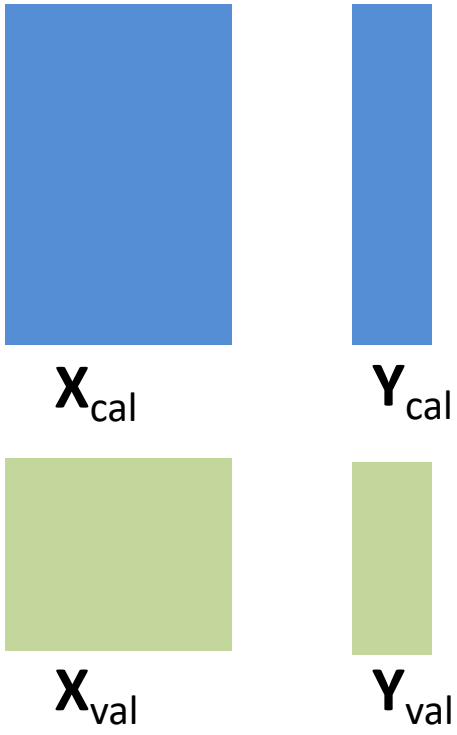
The concept of validation

- Verify if valid conclusions can be formulated from a model:
  - Able to generalize parsimoniously (with the smaller nr. of LV)
  - Able to predict accurately
- Define a proper diagnostics for characterizing the quality of the solution:
  - Calculation of some error criterion based on residuals
- Residuals can be used for:
  - Assessing which model to use;
  - Defining the model complexity in component-based methods;
  - Evaluating the predictive ability of a regression (or classification) model;
  - Checking whether overfitting is present (by comparing the results in validation and in fitting);
  - Residual analysis (model diagnostics).

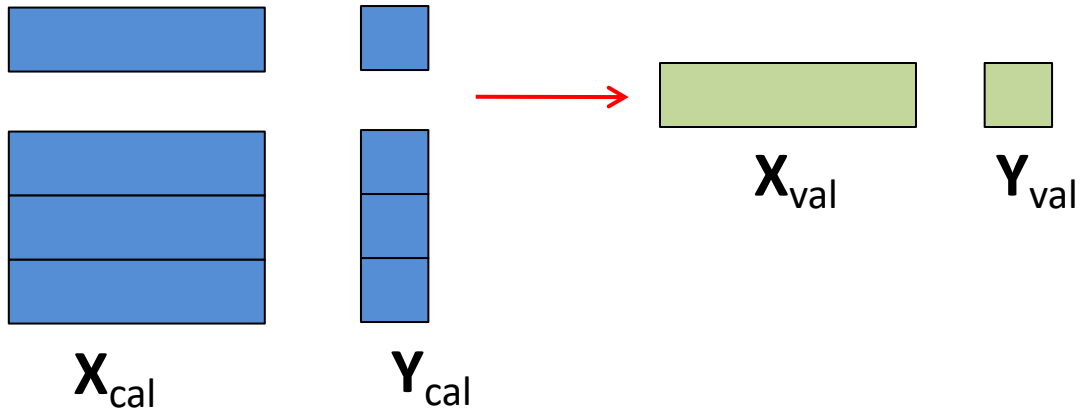
The need for “new” data

- The use of fitted residuals would lead to overoptimism:
  - Magnitude and structure not similar to the ones that would be obtained if the model were used on new data.

Test set validation

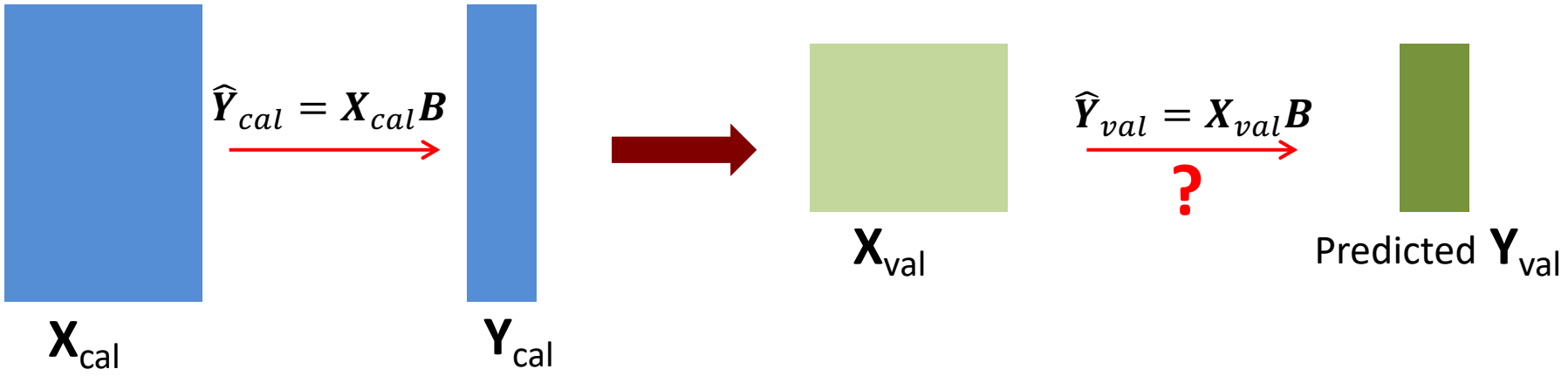


Cross-validation

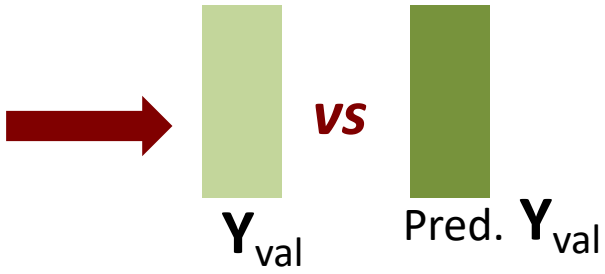


Test set validation

- Carried out by fitting the model to new data (test set):
  - Simulates the practical use of the model on future data.
  - Test set should be as independent as possible from the calibration set (collecting new samples and analysing them in different days…)
  - A representative portion of the total data set can be left aside as test set.



MODEL QUALITY

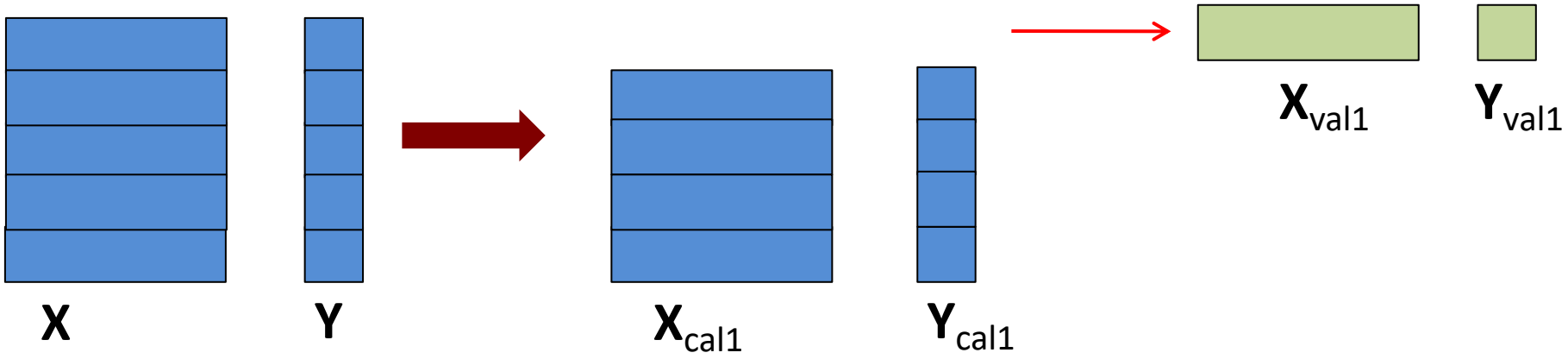


$$RMSEP = \sqrt{\frac{\sum_{i=1}^{N_{val}} \left(y_i^{val} - \hat{y}_i^{val}\right)^2}{N_{val}}}$$

Cross-validation

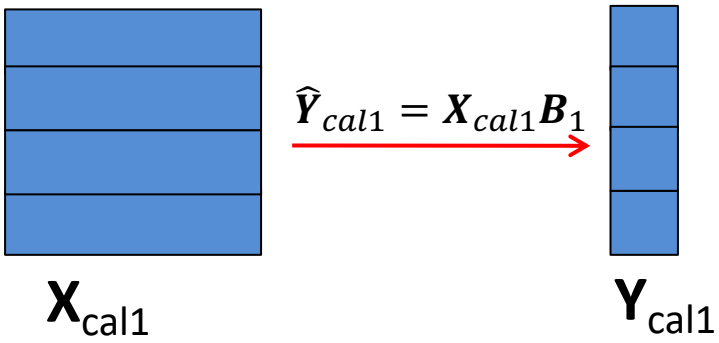
- Internal resampling method:
  - Simulates test set validation by repeating a data splitting procedure where different object are in turn placed in the validation set.
  - Particularly useful when a limited number of samples are available.
- Schematically, it consists of the following steps:

1. Leave out part of the data values

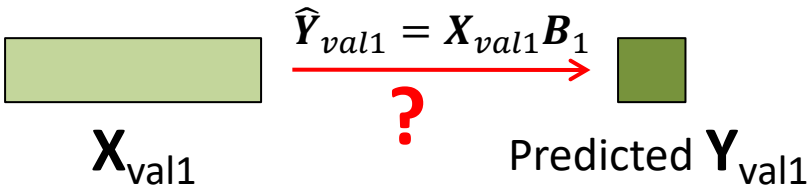


Cross-validation

2. Build the model without these data



3. Apply the model to the left out values and obtain predictions;

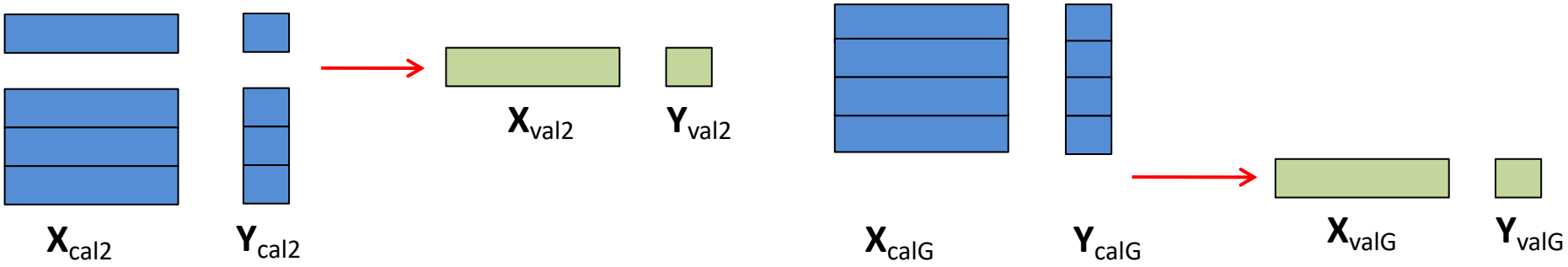


Cross-validation

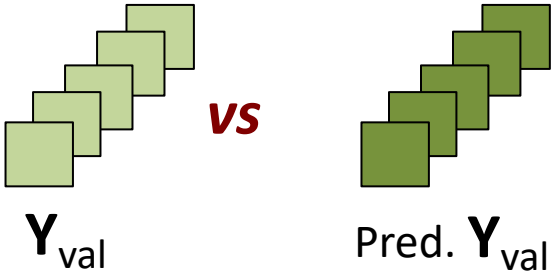
4. Calculate the corresponding residual error

$$PRESS_1 = \sum_{i=1}^{N_{val1}} \left(y_i^{val1} - \hat{y}_i^{val1}\right)^2$$

5. Repeat steps 1-4 until each data value has been left out once



6. Collect all the residuals into an overall error criterion



$$RMSECV = \sqrt{\frac{\sum_{j=1}^{G} PRESS_j}{N}} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_{-i})^2}{N}}$$

Cross-validation

- Number of objects is limited
- Understand the inherent structure of the system ↔ Estimating model complexity
- Objects in a data table can be stratified into groups based on background information:
  - Across instrumental replicates (repeatability)
  - Reproducibility (analyst, instrument, reagent...)
  - Sampling site and time
  - Across treatment/origin (year, raw material, batch…)