DISCRIMINATION AND CLASSIFICATION IN NIR SPECTROSCOPY

Federico Marini

[1]*Dept. Chemistry, University of Rome "La Sapienza", Rome, Italy*

SAPIENZA
UNIVERSITÀ DI ROMA

Hélio SPIR

Les 18èmes Rencontres HélioSPIR se dérouleront les 27 et 28 Novembre 2017

à Montpellier–Agropolis

sur le thème

Discrimination et classification par SPIR

Classification

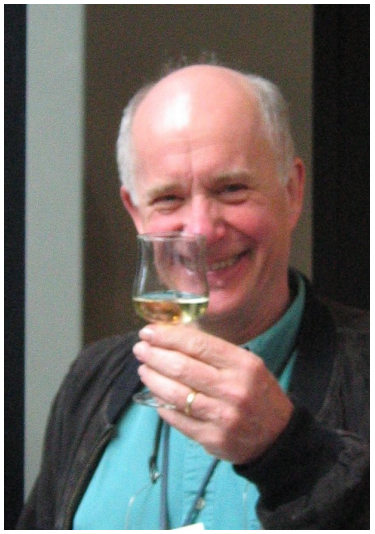
- “To find a criterion to assign an object (sample) to one category (class) based on a set of measurements performed on the object itself”
- Category or class is a (ideal) group of objects sharing similar characteristics
- In classification categories are defined a priori

Wine

Beer

Italian

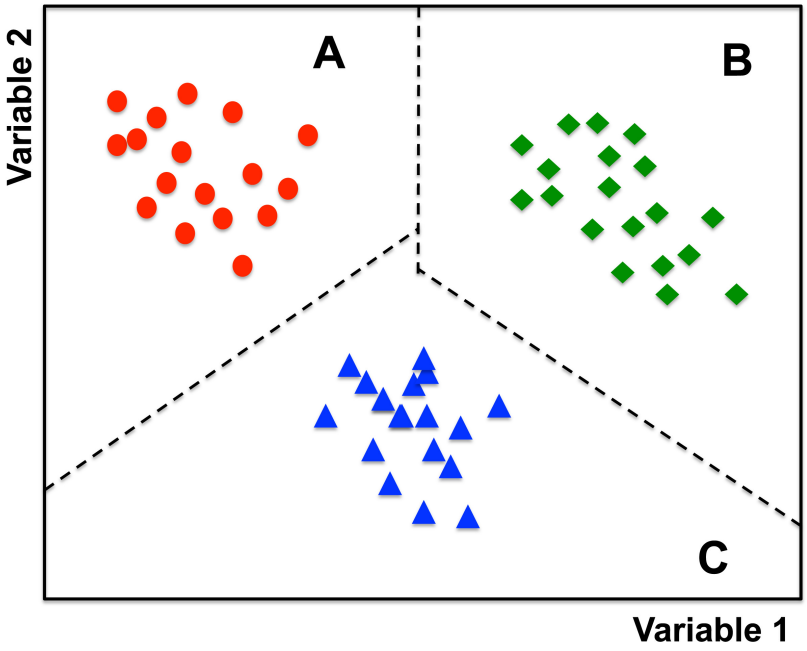
Scandinavian



Discrimination vs modeling

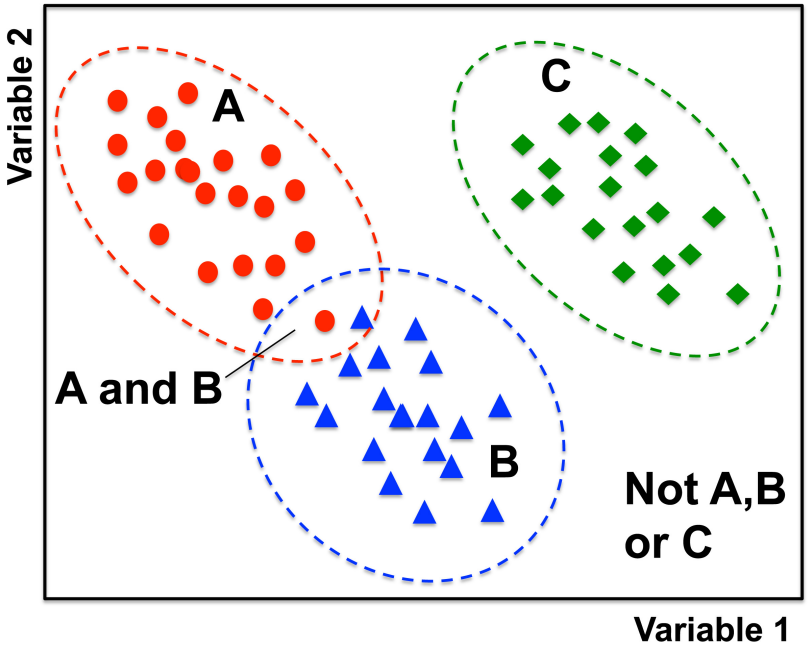
Discriminant

The corresponding outcome is always the classification to one of the G available categories.



Class modeling

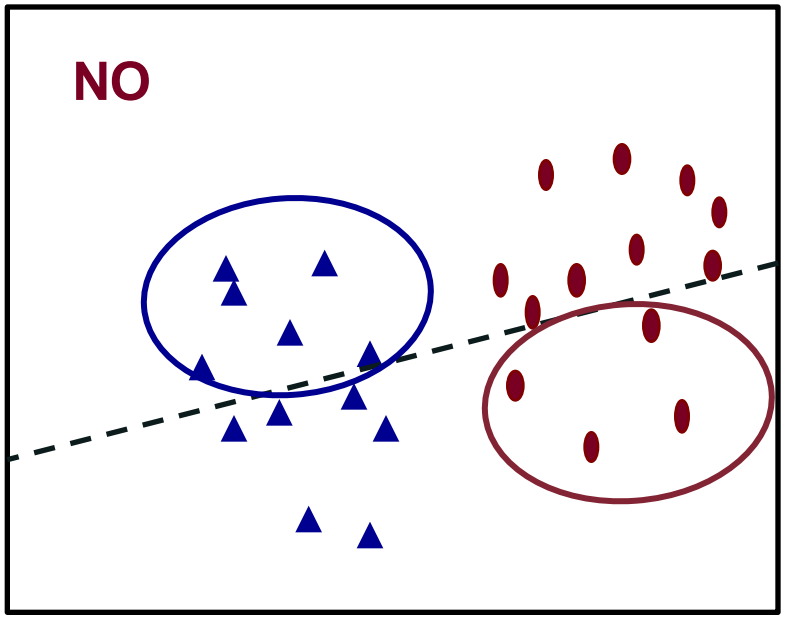
- Each category is modeled individually.
- A sample can be assigned to one class, to more than one class or to no class at all.

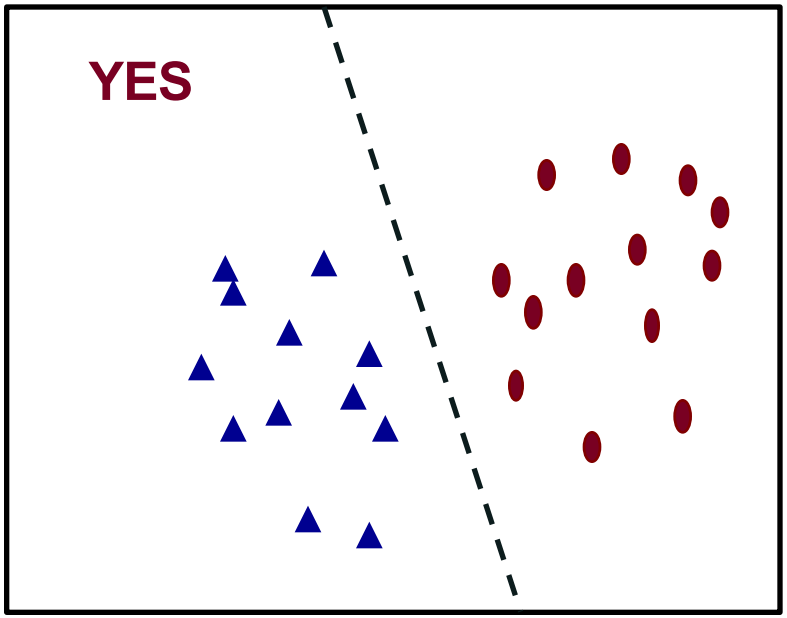


DISCRIMINANT METHODS

- Sample is ALWAYS assigned to one of the given classes.
- Class boundaries are built to minimize classification error:

$$E_T = \frac{\sum_{g=1}^{G} n_{g,wrong}}{N_T}$$





DISCRIMINANT METHODS

- The classification rule which minimizes E_T is the so-called Bayes' rule:

  "assign a sample to the category it has the highest probability of belonging to"

- This rule is satisfied also by non-probabilistic methods, even if not explicitly used to calculate the model parameters.
- Sometimes weighting can be introduced in the loss function to account for unequal costs of misclassification:

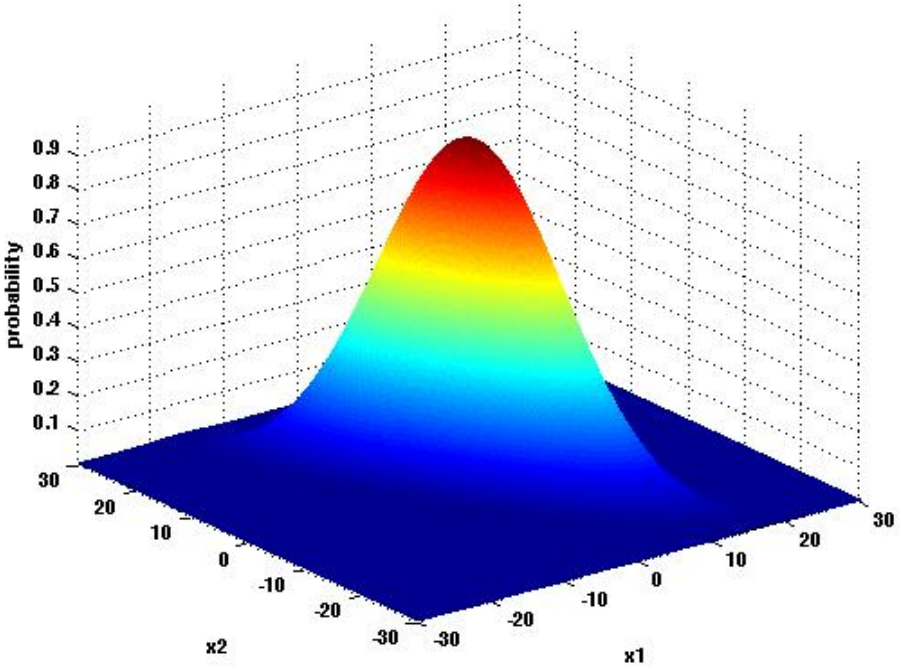
$$L_T = \sum_{g=1}^{G} c_g n_{g,wrong}$$

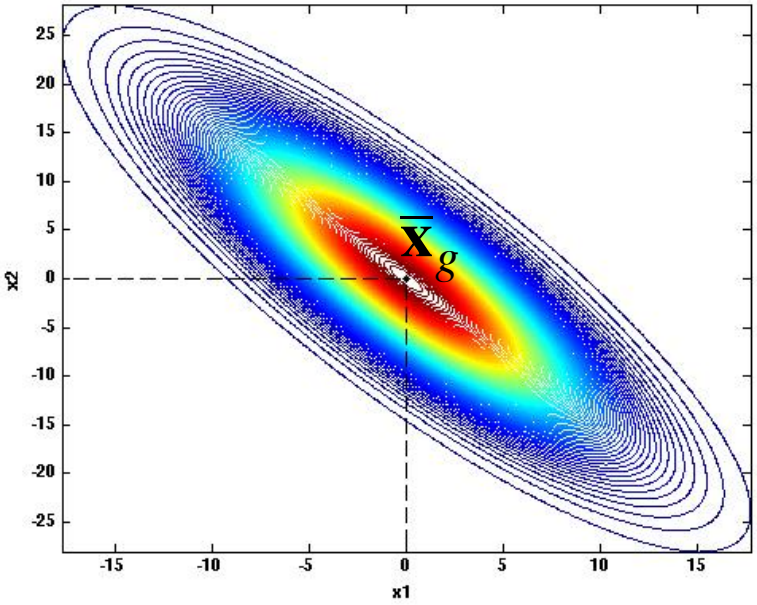
  – E.g., medical diagnosis.

Linear Discriminant Analysis (LDA)

PROBABILISTIC CLASSIFICATION: LDA

$$p(g|\mathbf{x}_i) \propto \frac{1}{(2\pi)^{n/2}|\mathbf{S}_g|} e^{-\frac{1}{2}(\mathbf{x}_i-\bar{\mathbf{x}}_g)^T \mathbf{S}_g^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}_g)}$$

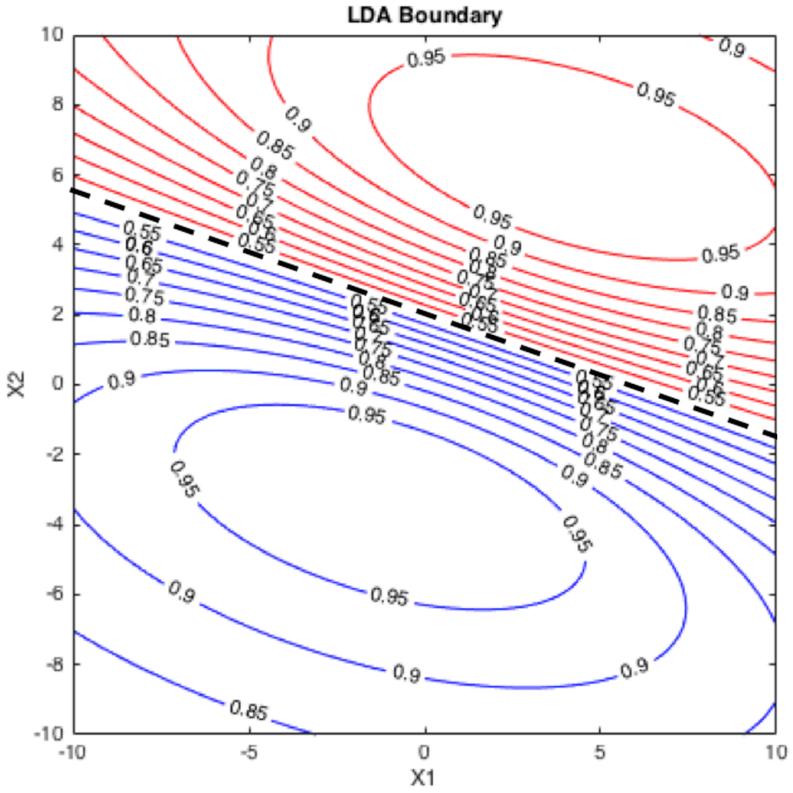




LDA

- Decision boundaries are defined by: $p(1|\mathbf{x}) = p((2|\mathbf{x})$
- Linear surfaces:

$$(c_1 - c_2) + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = 0$$



$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1}\mathbf{x} - w_o = 0$$

CANONICAL VARIATE(S)

$$\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)^T \mathbf{S}^{-1}\mathbf{x} - w_o = \mathbf{w}^T\mathbf{x} - w_o = 0$$

- The linear transformation identifies a direction along which separation among classes is maximum (Canonical Variate)

$$t_{CV} = \mathbf{w}^T\mathbf{x} \qquad \mathbf{w} = \left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)^T \mathbf{S}^{-1}$$

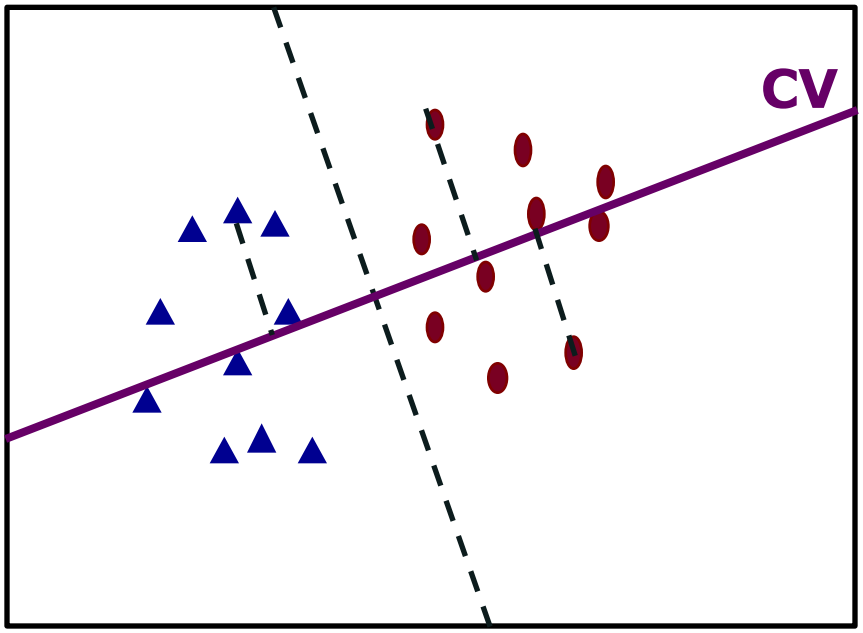
- Then classification rule becomes:
  - Assign the sample to class 1 if

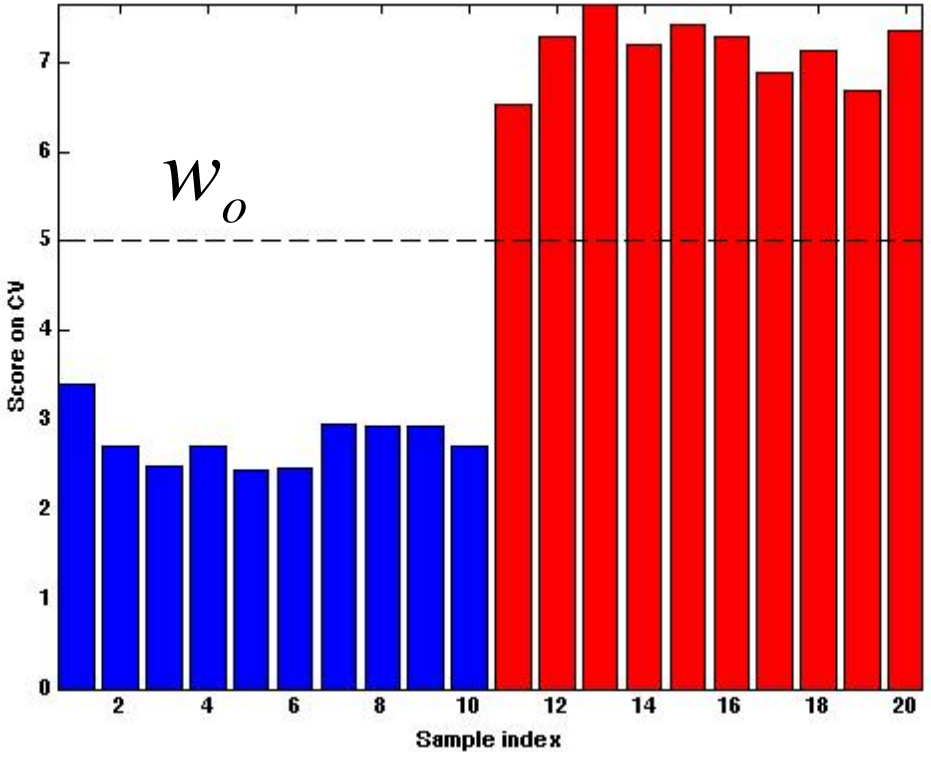
$$t_{CV} > w_0$$

  - Assign the sample to class 2 if:

$$t_{CV} < w_0$$

CANONICAL VARIATE(S)

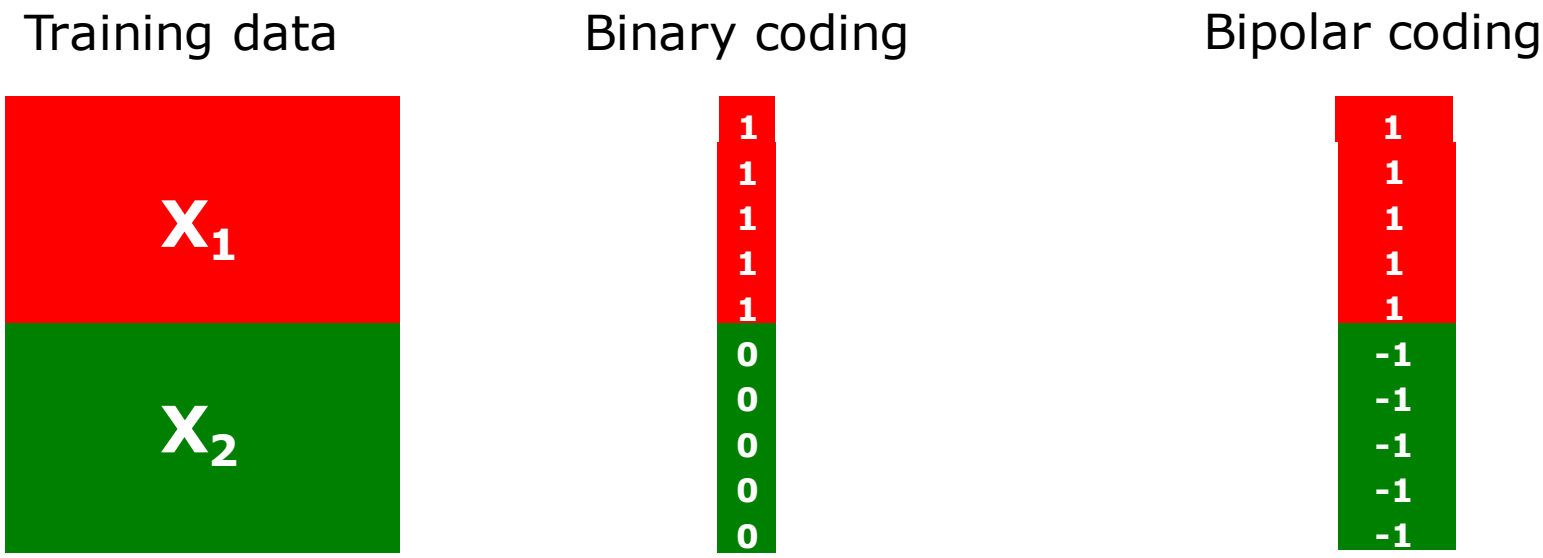




Partial Least Squares Discriminant Analysis (PLS-DA)

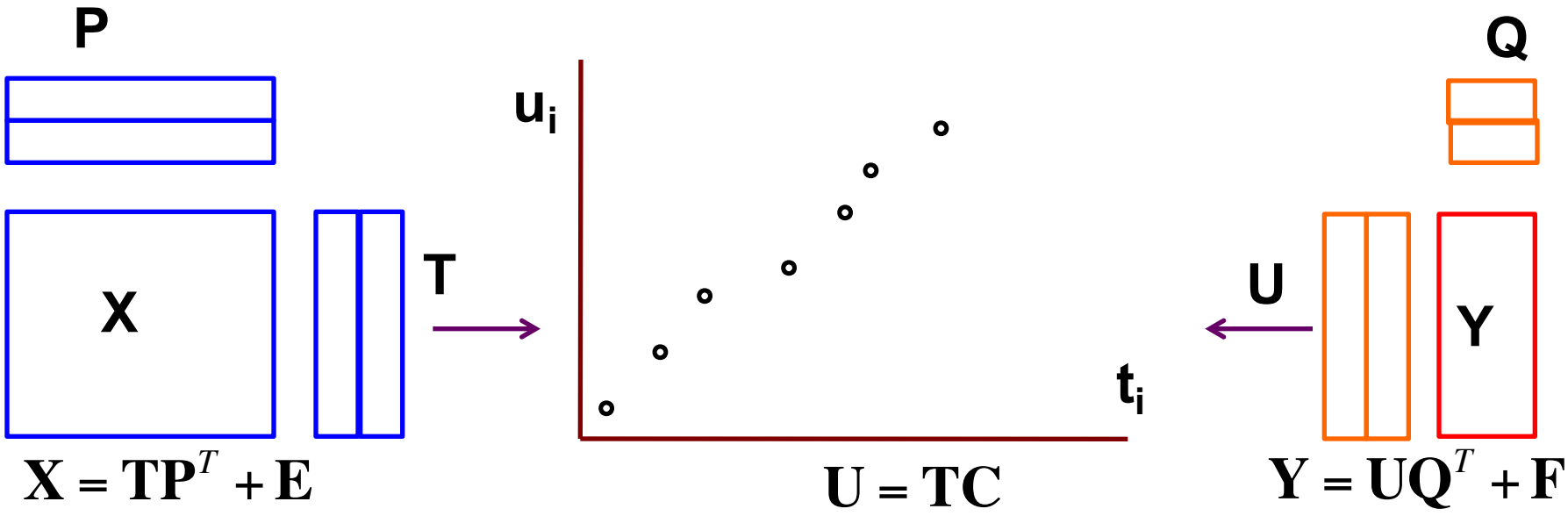
Partial Least Squares Discriminant Analysis (PLS-DA)

- 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



Discriminant-PLS (PLS-DA)

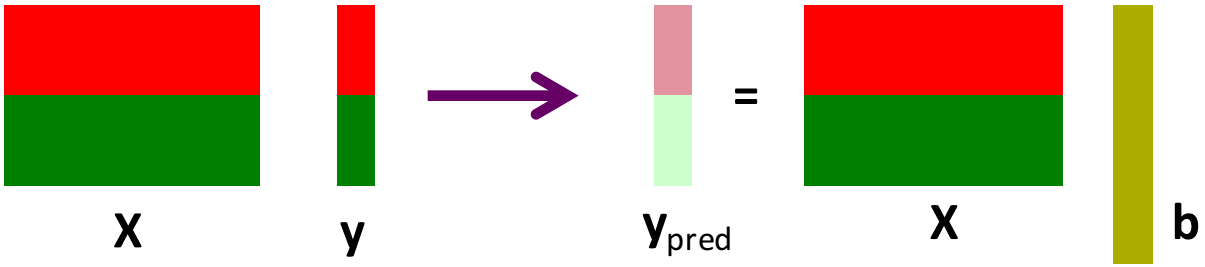
- As name suggests, regression is accomplished through PLS



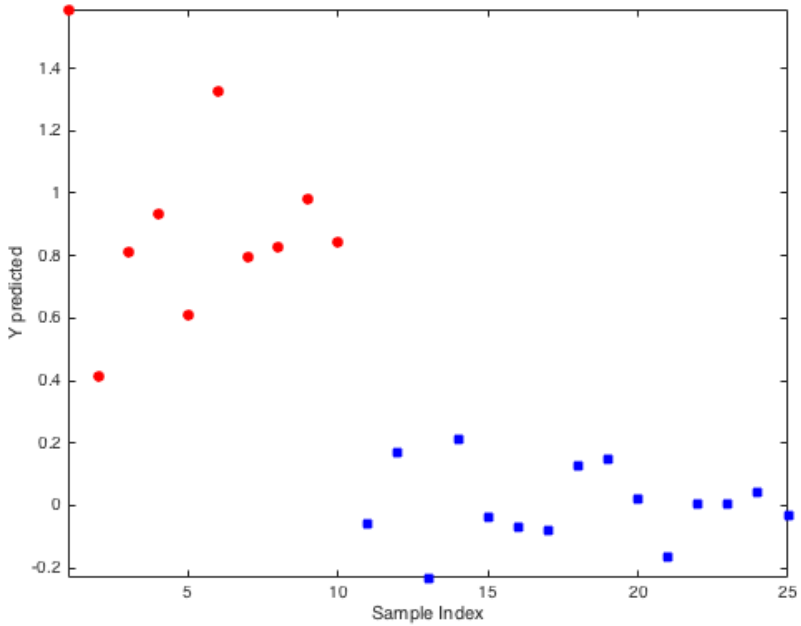
- Regression can be expressed in terms of original variables

$$Y = XB_{PLS} \quad \text{with:} \quad B_{PLS} = W\left(P^T W\right)^{-1} CQ^T$$

Partial Least Squares Discriminant Analysis (PLS-DA)

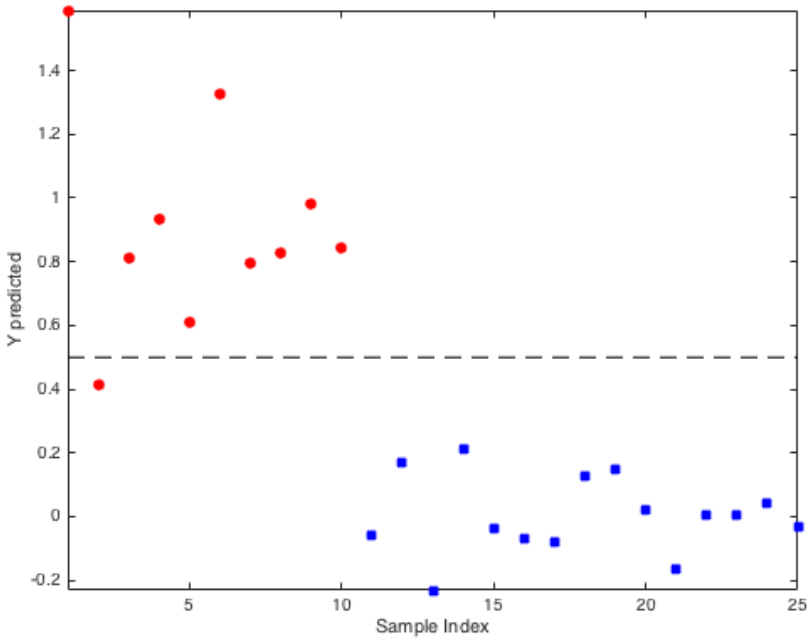


- While "true" Y values are binary-coded, predictions are real valued.



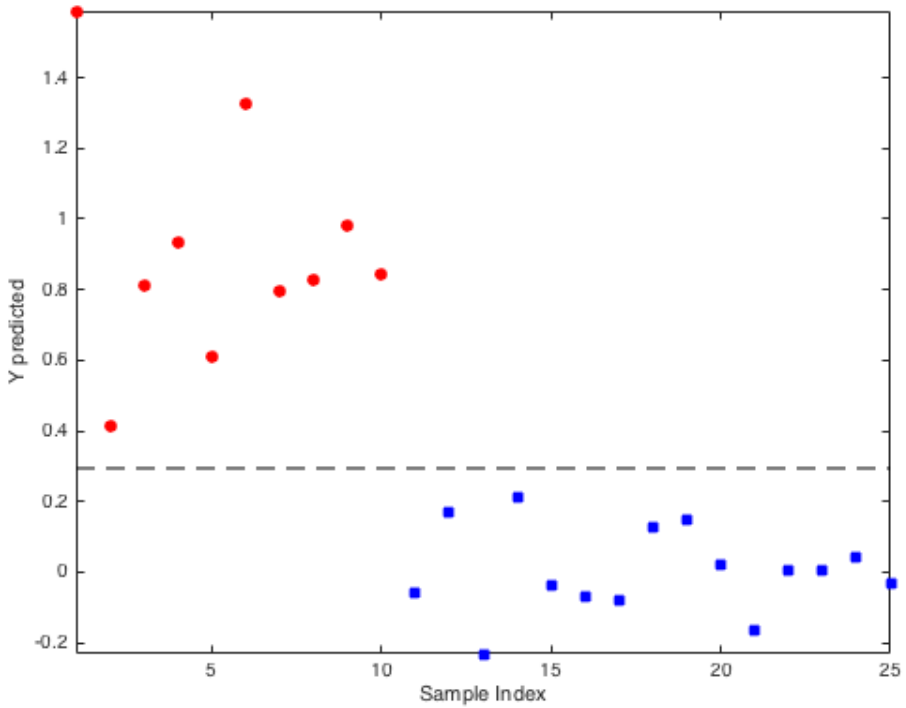
Partial Least Squares Discriminant Analysis (PLS-DA)

- 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



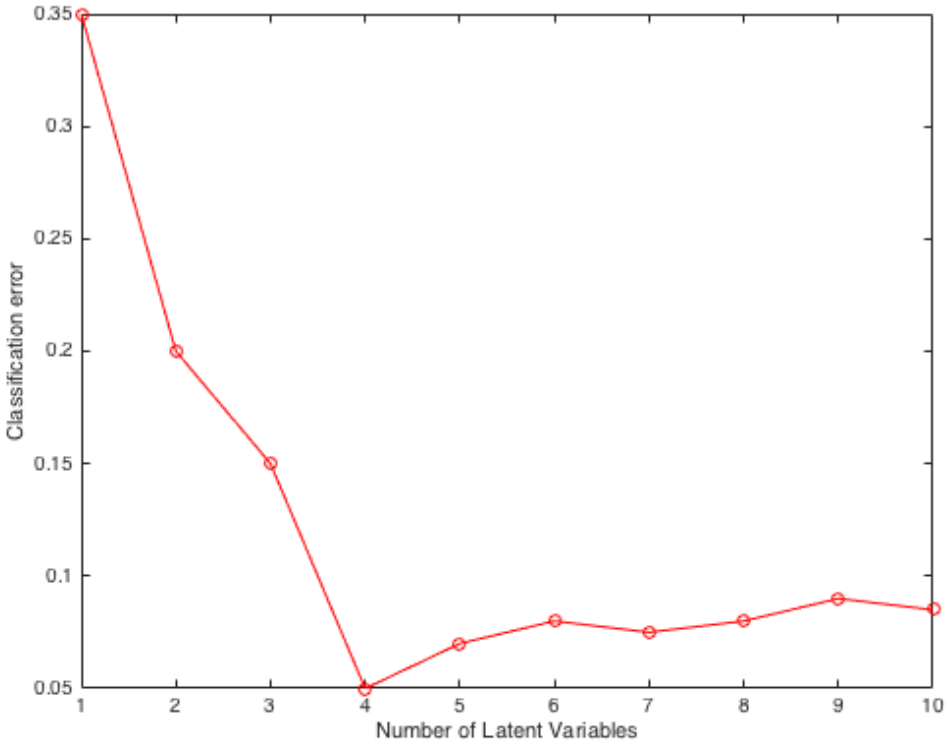
Partial Least Squares Discriminant Analysis (PLS-DA)

- Different methods have been proposed in the literature to find alternative “optimal” threshold values (see later).
- In the example, setting the value to 0.3 allows the correct classification of all samples:



Model complexity

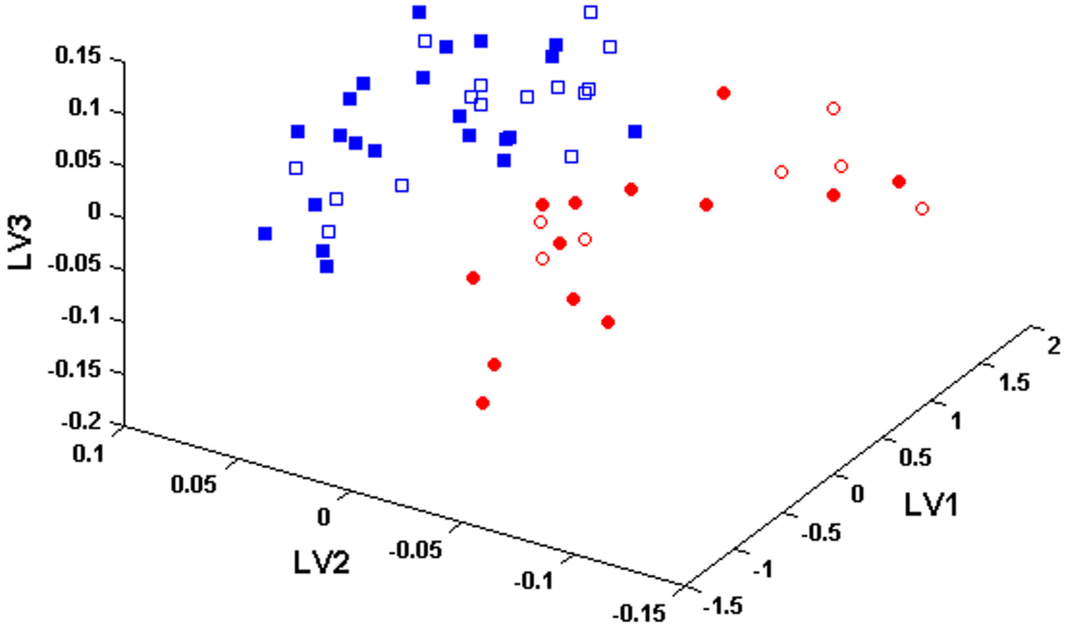
- PLS-DA is a bilinear model:
  - Optimal number of components (LVs) should be selected
- Error criterion in cross-validation is (usually) based on the number (percentage) of mis-classifications



$$Classification\ Error\ (CE) = \frac{n_{misclassified}}{n_{samples}}$$

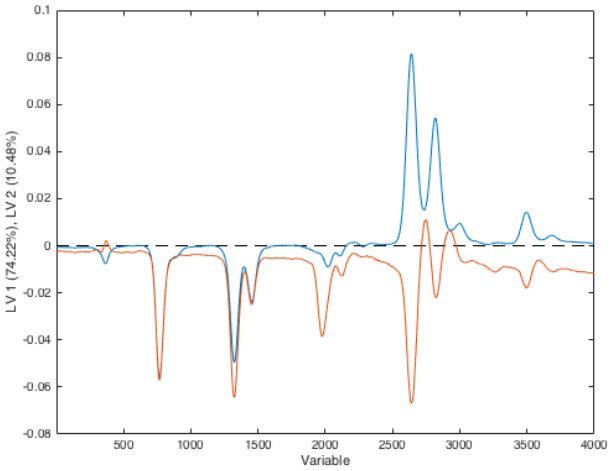
What we get (apart from class predictions)

- Parsimonious representation of the samples in the score space

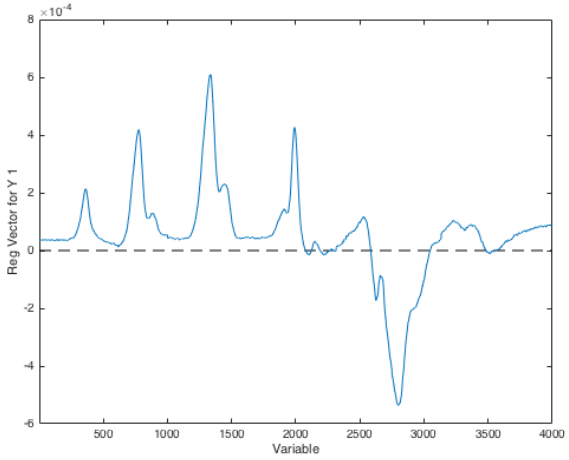


What we get (apart from class predictions)

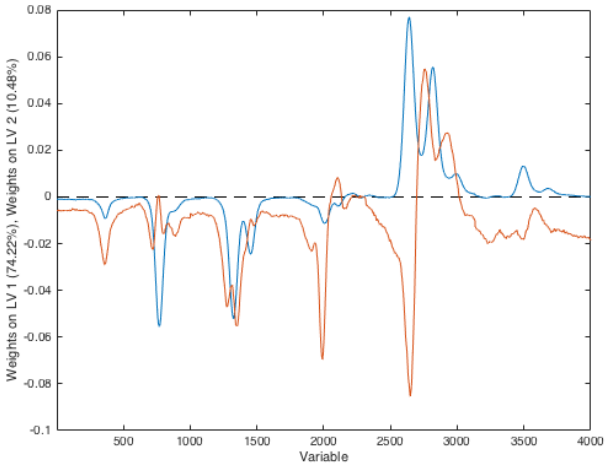
- Information about the contribution of variables:



X-Loadings **P**



Regression coefficients **b**

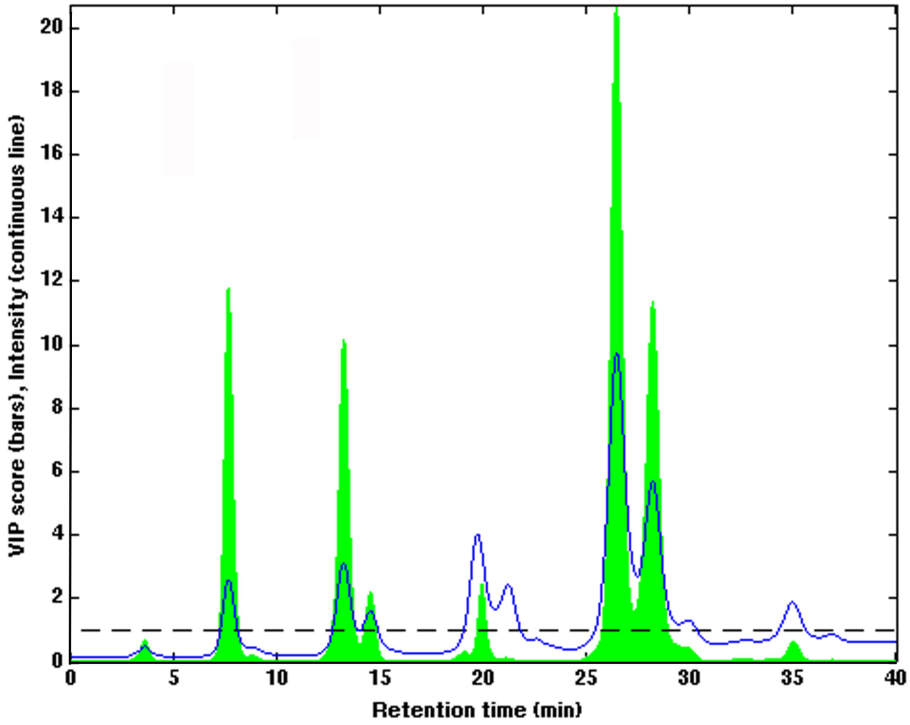


X-Weights **W**

Other tools for Interpretation

- Different tools for interpretation.
  - VIP (Variable importance in projection)

$$VIP_j = \sqrt{N_{\text{var}\,s} \frac{\sum_{k=1}^{F} \left(c_k^2 \mathbf{t}_k^T \mathbf{t} \right) \left(w_{jk} / \left\| \mathbf{w}_k \right\| \right)^2}{\sum_{k=1}^{F} \left(c_k^2 \mathbf{t}_k^T \mathbf{t} \right)}}$$

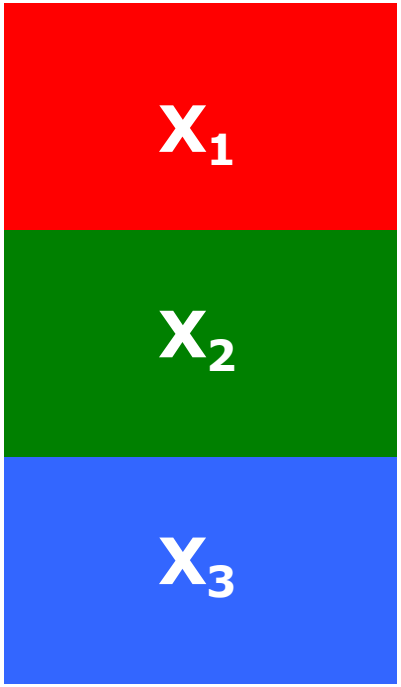


- But also:
  - Target Projection
  - Selectivity ratio
  - Stability (UVE-PLS)
  - O-PLS
  - …

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)

Training spectra



Class index

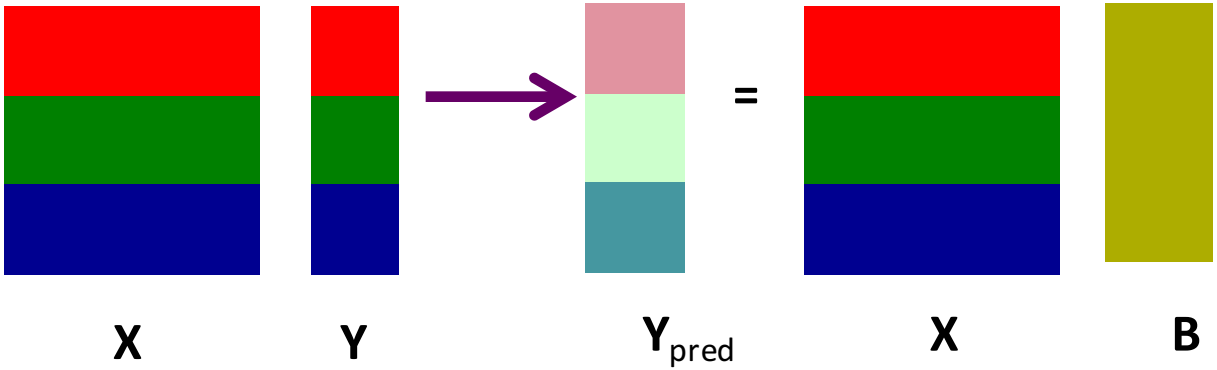
1
1
1
1
1
2
2
2
2
2
3
3
3
3
3

Dummy Y matrix

1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1

PLS-DA for more than two classes

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

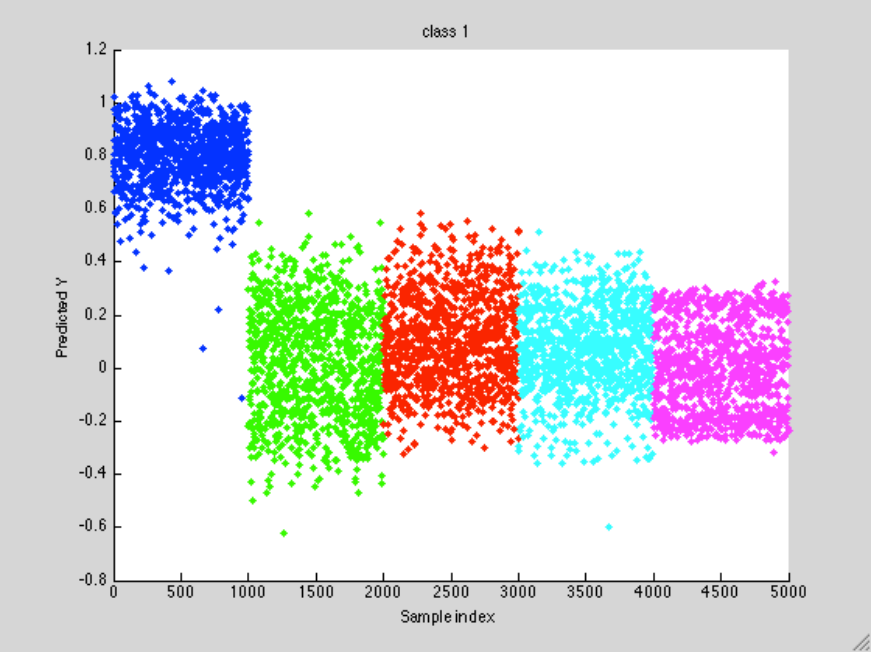


- For each sample, predicted Y is a row vector (real valued) having as many columns as the number of classes.
- Different options to achieve classification based on the values of predicted Y

PLS-DA for more than two classes

- Predicted y is real-valued:

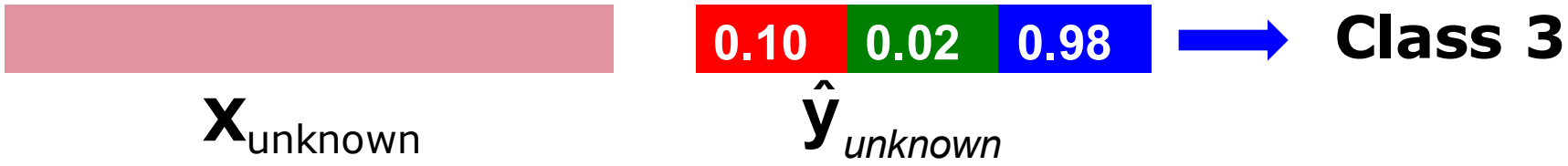
"true" y			predicted y		
1	0	0	1.03	0.09	-0.10
1	0	0	0.68	0.21	0.08
1	0	0	0.99	-0.10	0.01
1	0	0	0.96	0.18	-0.14
1	0	0	0.79	0.02	0.25
0	1	0	0.14	0.94	0.07
0	1	0	-0.01	1.12	0.12
0	1	0	0.08	0.89	-0.02
0	1	0	0.33	0.45	0.25
0	1	0	0.15	0.72	0.06
0	0	1	0.13	-0.18	0.85
0	0	1	0.21	0.17	0.56
0	0	1	-0.09	0.32	0.69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



- Sample is assigned to the class corresponding to the highest y component

More about PLS-DA

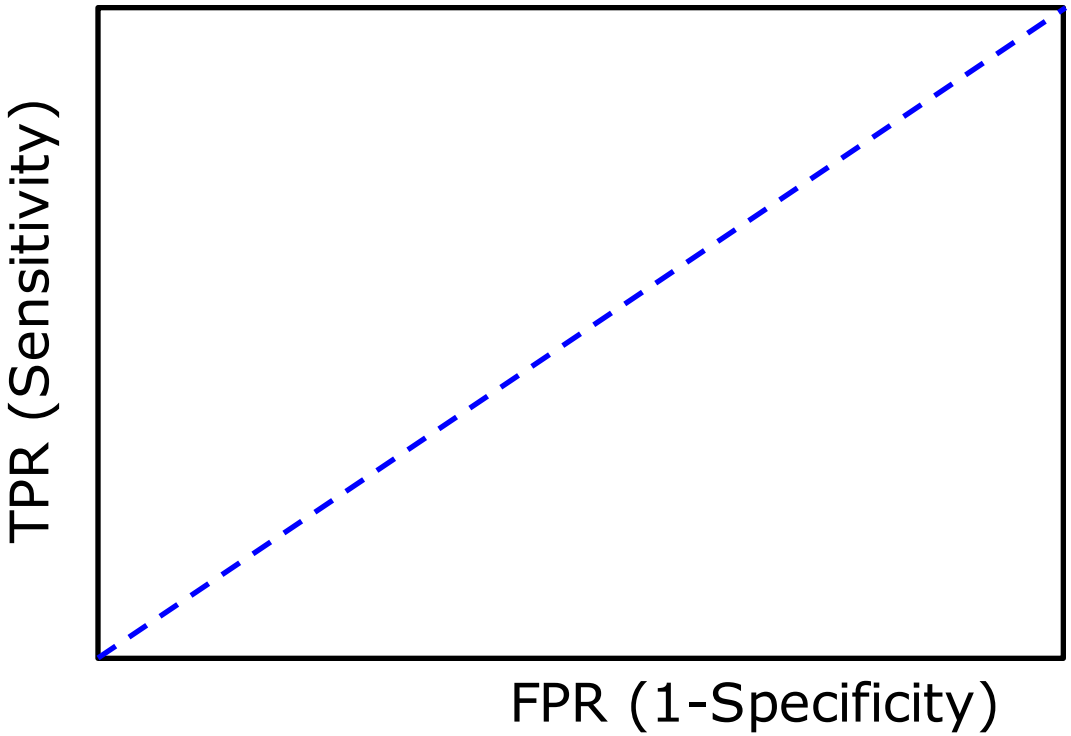
- Classification rule: "assign a sample to the class corresponding to the highest Y component"



- When there are only two classes, threshold is set at 0.5
- Alternative rules can be defined: "assign a sample to class 1 if y_1>0.68 or to class 2 if y_2>0.75"
- Category thresholds can be defined according to Bayes' rule or based on ROC curves.

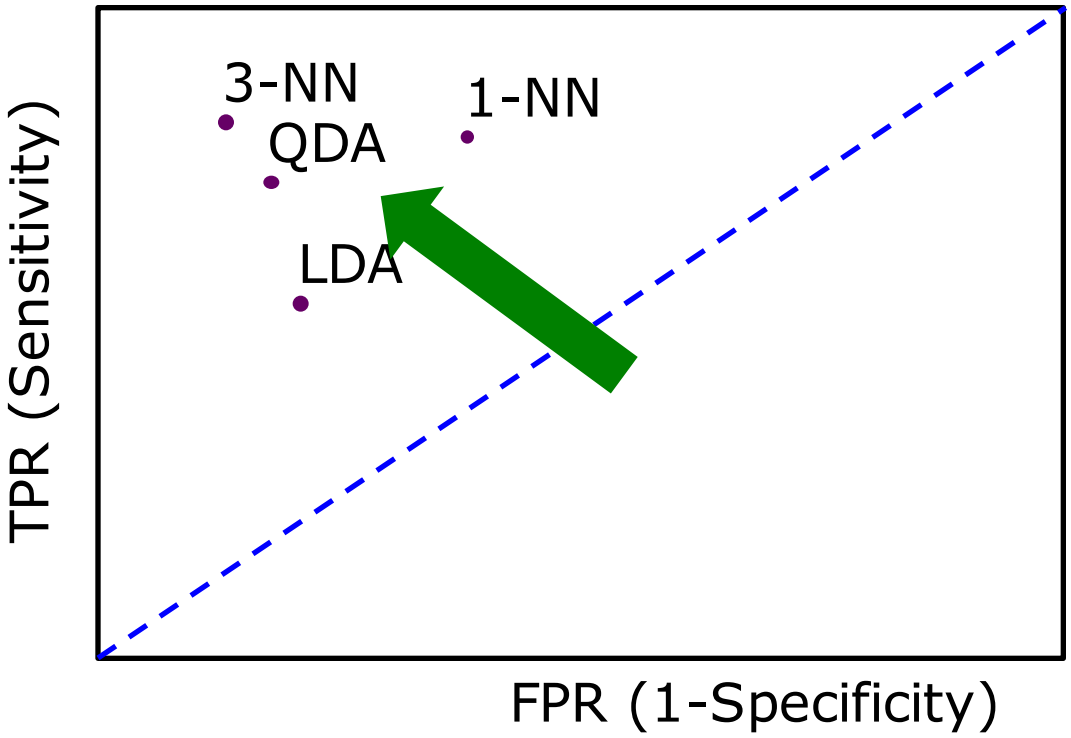
ROC

- Receiver Operating Characteristics (ROC) curves are a way of evaluating the performances of classifiers (two-class problems)
- It is a plot of TPR (true positive rate) vs FPR (false positive rate) or in other terms of sensitivity vs 1-specificity.



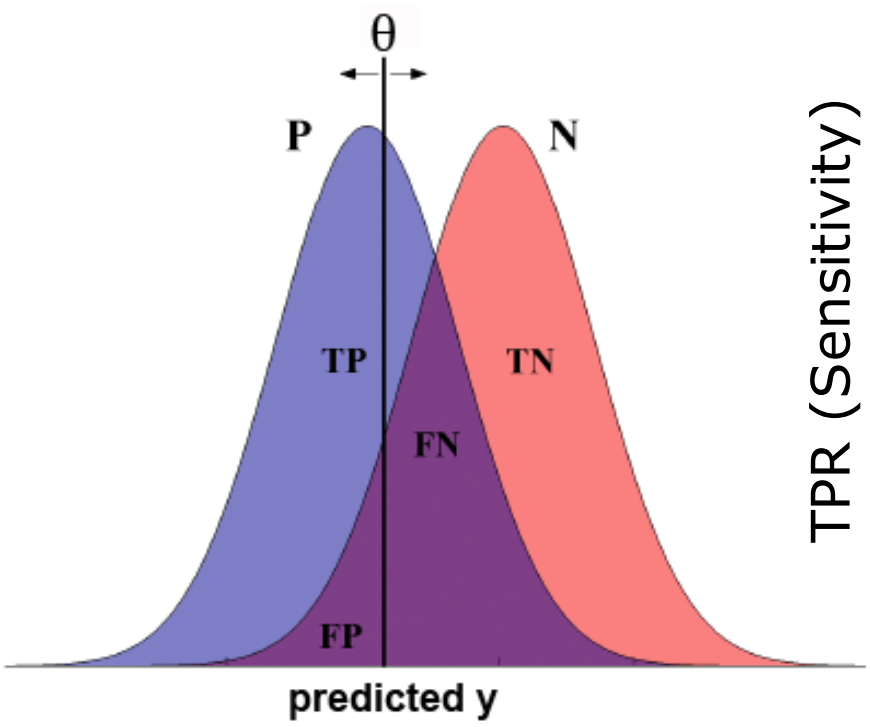
ROC - 2

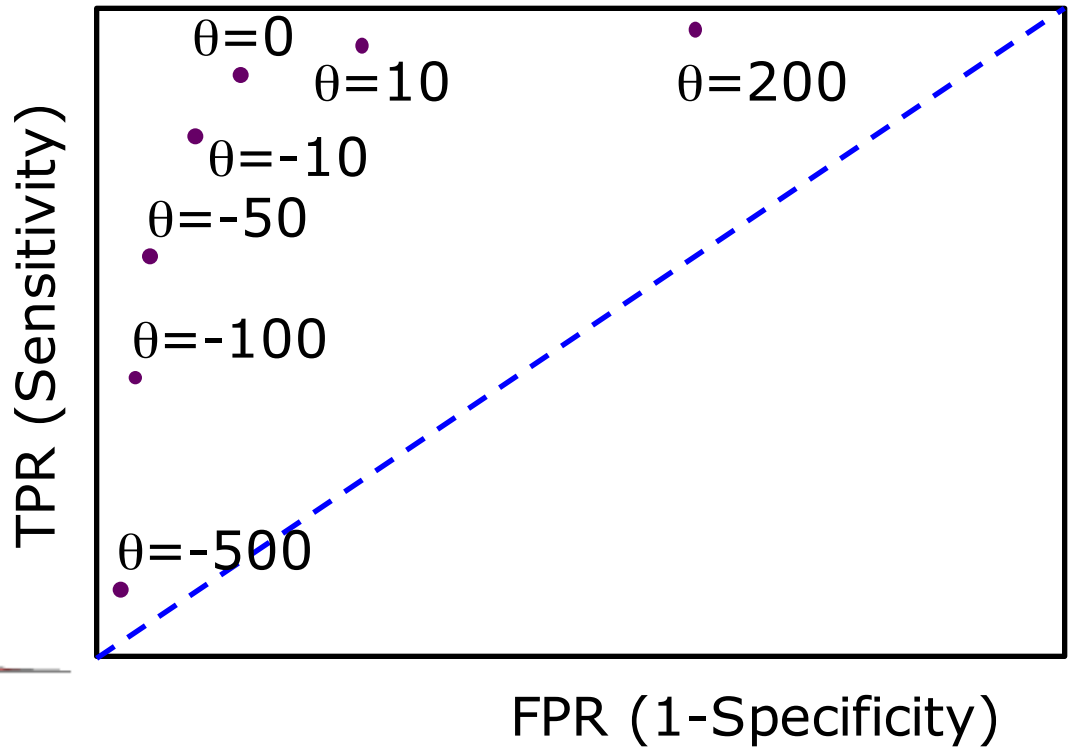
- Classifiers which don't depend on adjustable parameters are represented as points in the graph
- The classifier closest to the upper-most left corner of the graph has the best performances.



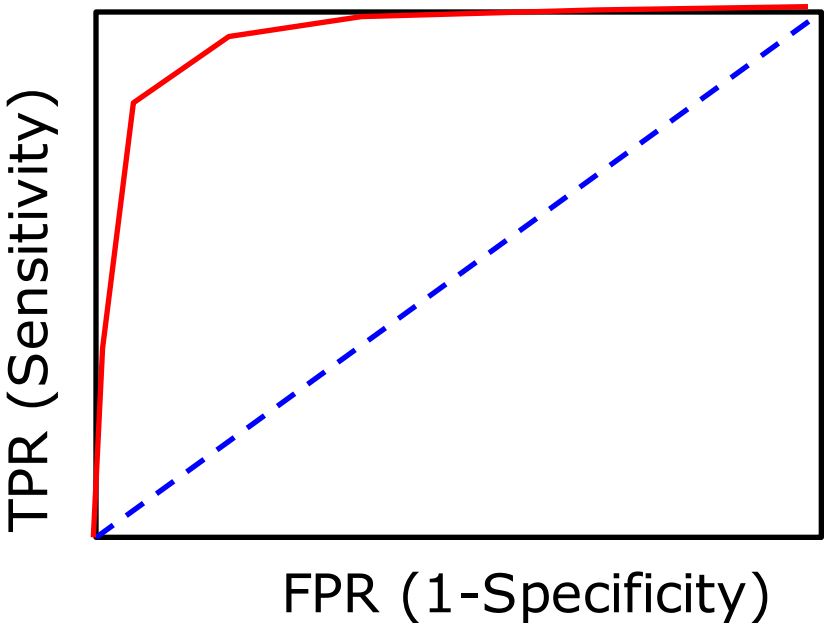
ROC - 3

- For models where a classification threshold can be defined, ROC graph corresponds to a curve
- Each point of the curve corresponds to the TPR and FPR values for a given value of the threshold.

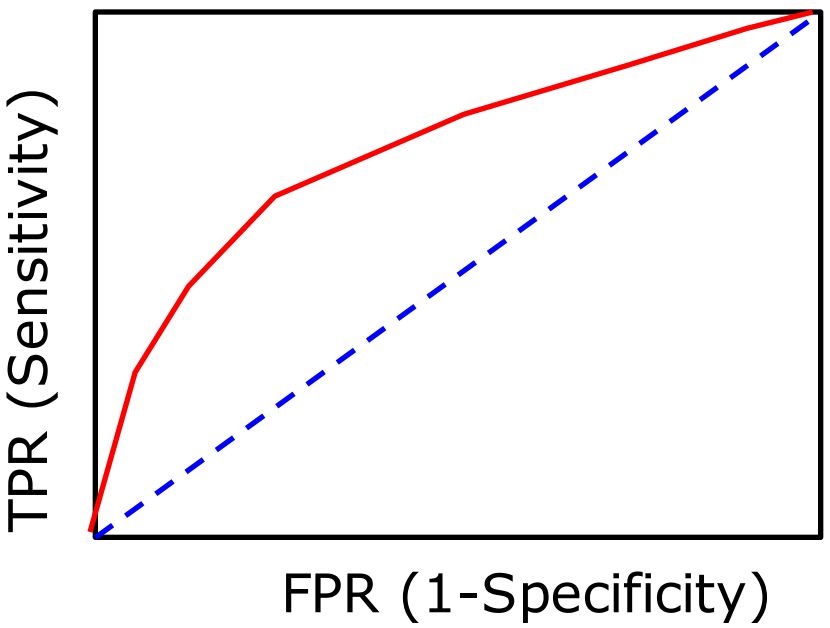




ROC - 4

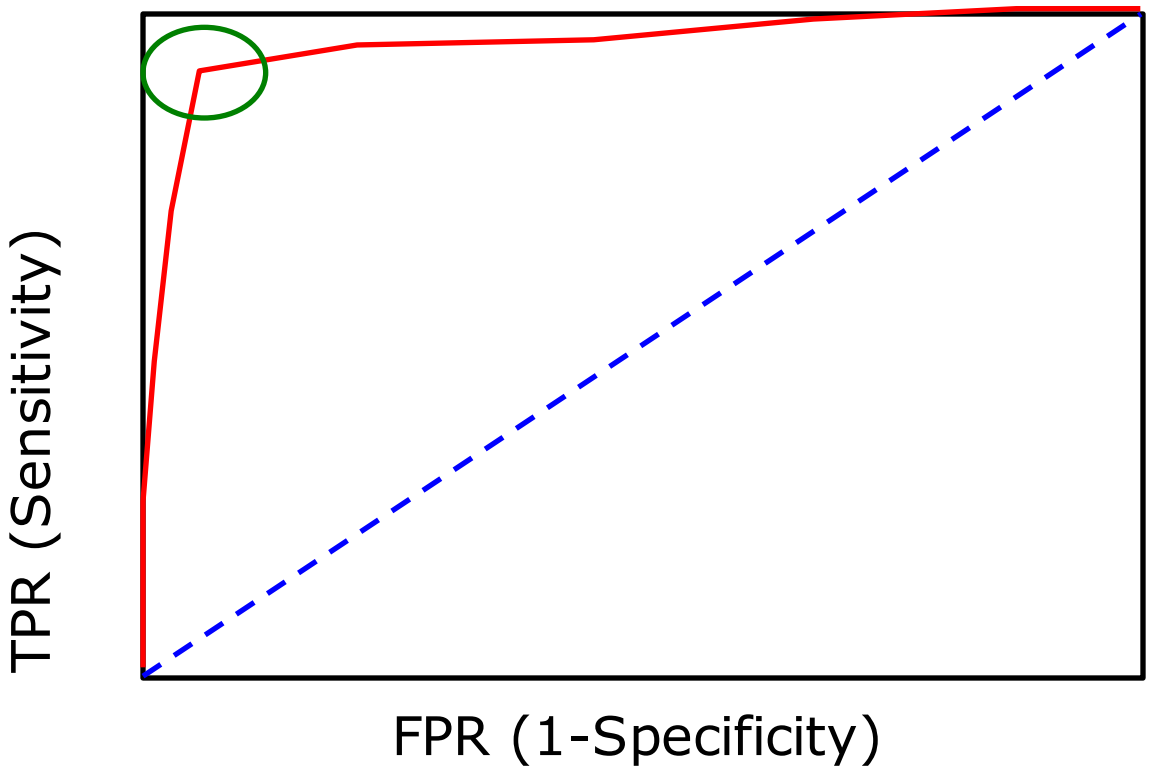


Good discrimination



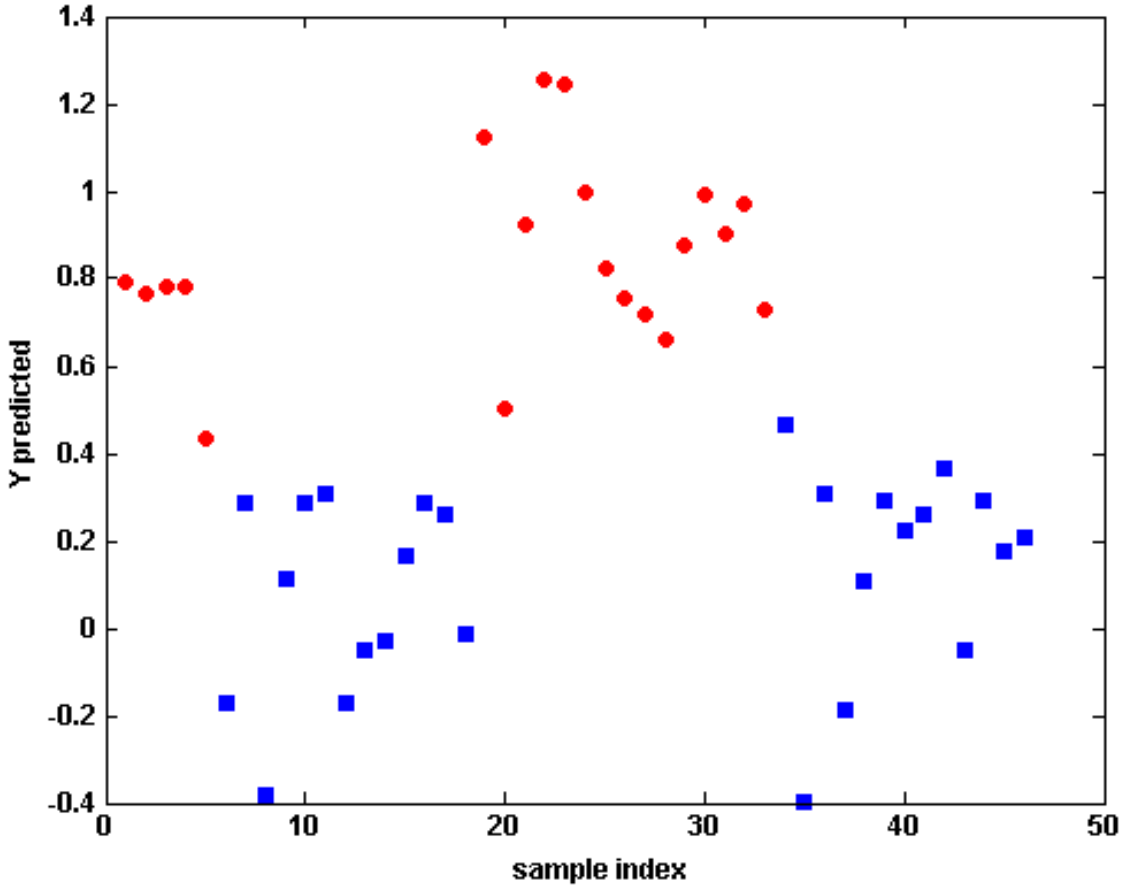
Medium/bad discrimination

ROC - 5



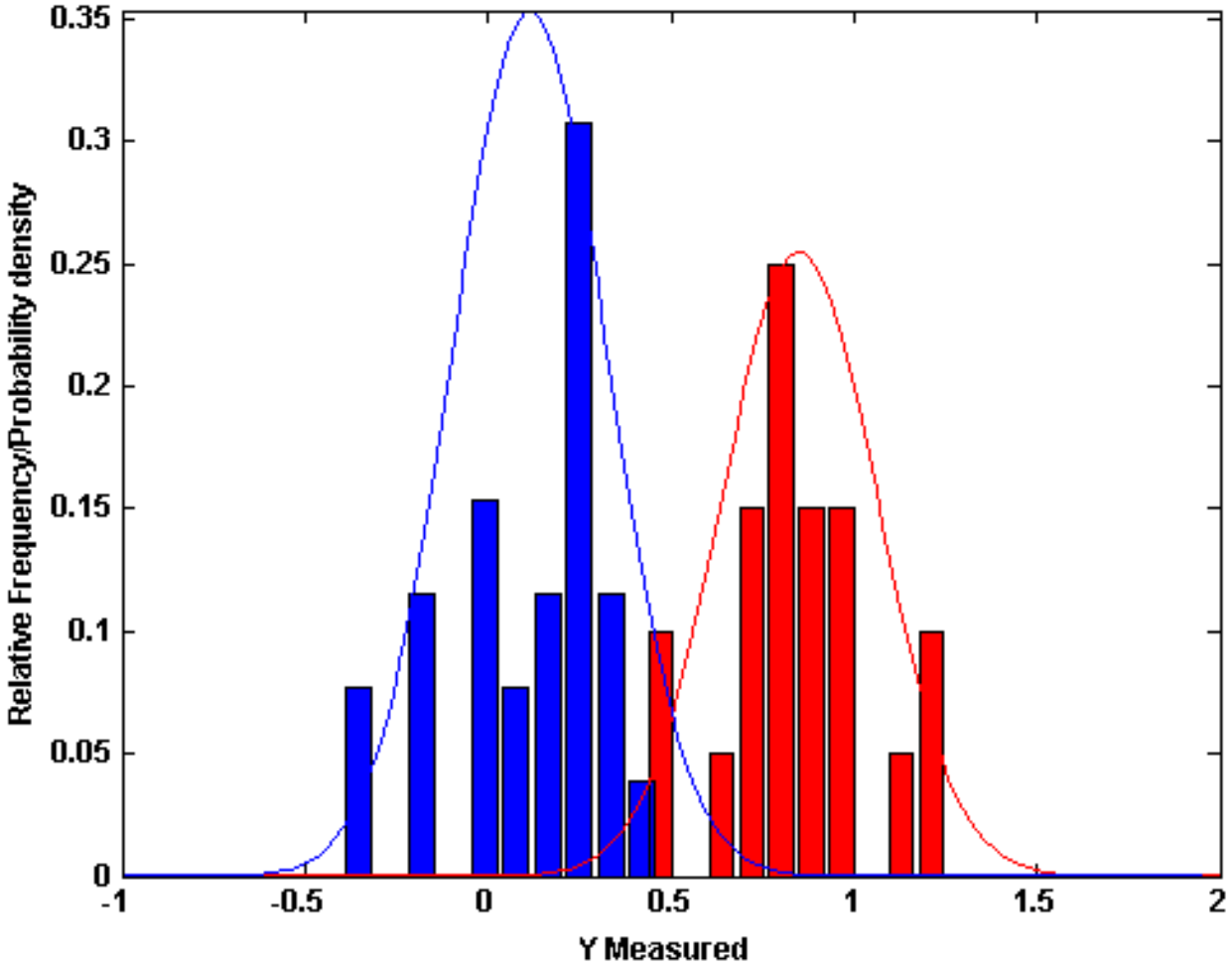
Threshold value is selected as the one corresponding to the point closest to the upper-left corner

Bayes' rule



Each column of the predicted Y is analyzed separately in terms of probability of belonging or not to that specific class.

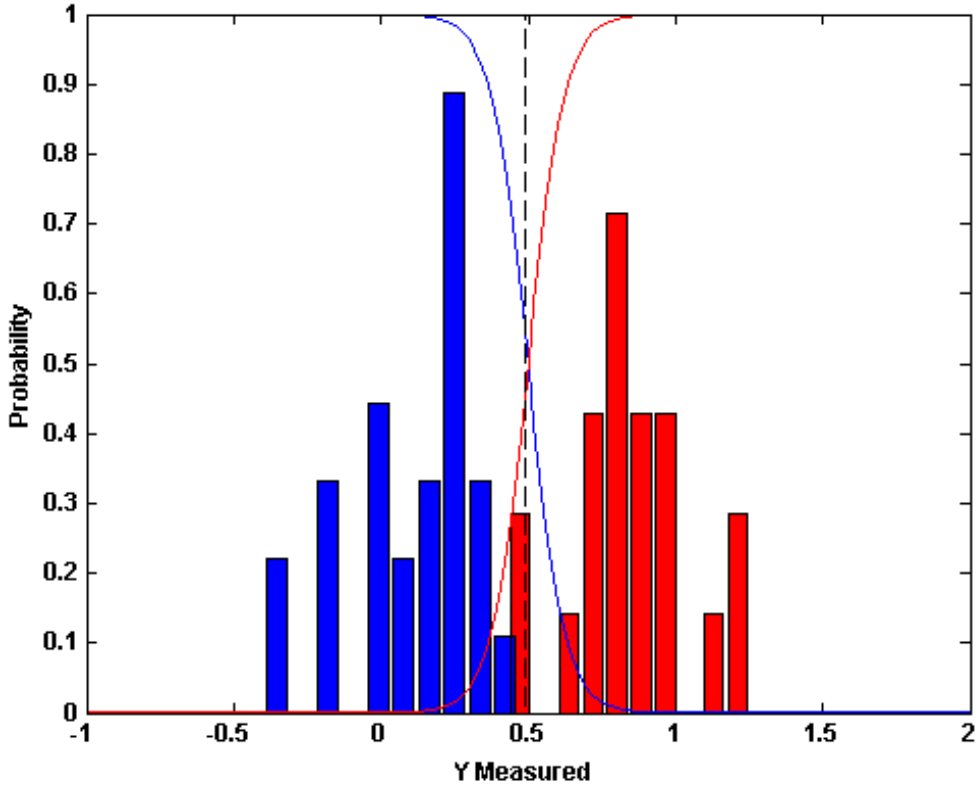
Bayes' rule



Two gaussian distributions (one for the samples belonging to the modeled class beloning and the other for the other samples) are fitted to the measured Y.

Bayes' rule

The probability of the samples belonging or not to the class are estimated bv Baves' rule:



$$p(C_i|y_i) = \frac{p(y_i|C_i)\pi(C_i)}{p(y_i|C_i)\pi(C_i) + p(y_{\neq i}|C_{\neq i})\pi(C_{\neq i})}$$

$$p(C_{\neq i}|y_i) = \frac{p(y_i|C_{\neq i})\pi(C_{\neq i})}{p(y_i|C_i)\pi(C_i) + p(y_i|C_{\neq i})\pi(C_{\neq i})}$$

The threshold is selected as that value y_i^* for which:

$$p(C_i|y_i^*) = p(C_{\neq i}|y_i^*)$$

A possible extension to the multi-class case (PLS toolbox)

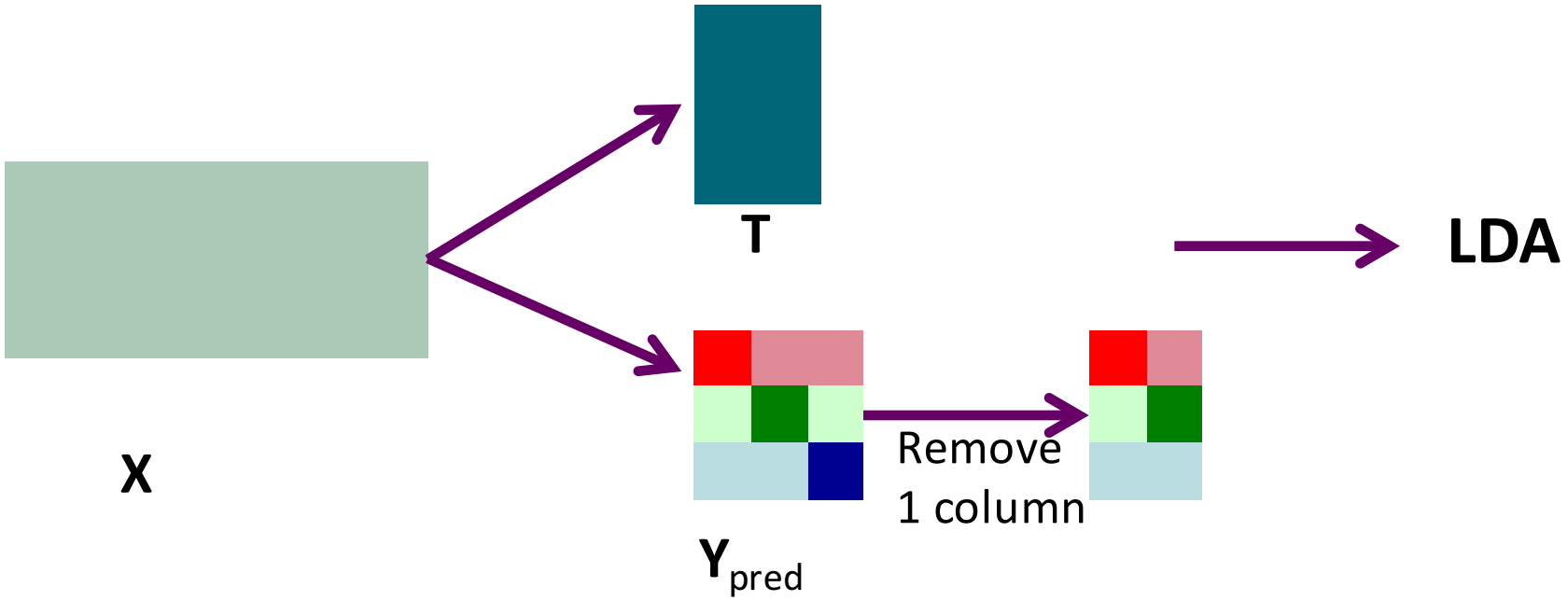
- **Strict**: a sample is assigned to a class if it has probability above threshold (>50%) for that class only.
- **Most probable**: a sample is assigned to the class it has the highest probability of belonging to (always provides an assignment).

P(class1)	P(class2)	P(class3)	Strict	Most probable
0.9503	0.1437	0.2356	class1	class1
0.3545	0.4550	0.1194	unassigned	class2
0.1102	0.6770	0.3200	class2	class2
0.3990	0.0377	0.3210	unassigned	class1
0.0056	0.0989	0.9812	class3	class3
0.1457	0.6769	0.7852	unassigned	class3

Another possible approach for more than 2 classes

- 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**

$$\hat{\mathbf{Y}} = \mathbf{T}\mathbf{Q}^T = \mathbf{X}\mathbf{B}$$



Contents lists available at SciVerse ScienceDirect

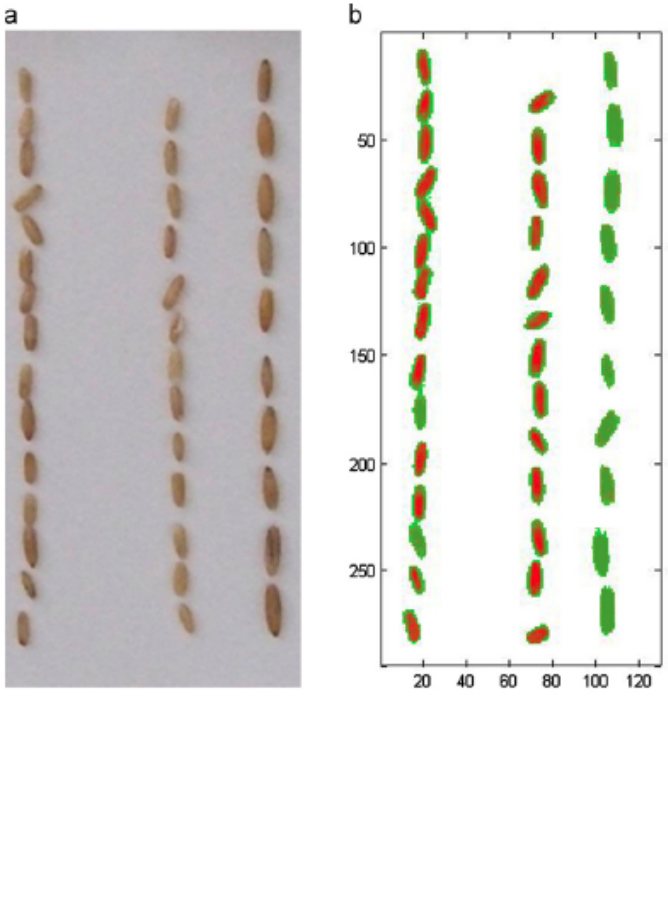
Talanta

journal homepage: www.elsevier.com/locate/talanta

Classification of oat and groat kernels using NIR hyperspectral imaging

Silvia Serranti [a,*], Daniela Cesare [a], Federico Marini [b], Giuseppe Bonifazi [a]

[a] Department of Chemical Engineering Materials & Environment Sapienza—Università di Roma Via Eudossiana 18, 00184 Rome, Italy
[b] Department of Chemistry Sapienza—Università di Roma P.le Aldo Moro 5, 00185 Rome, Italy



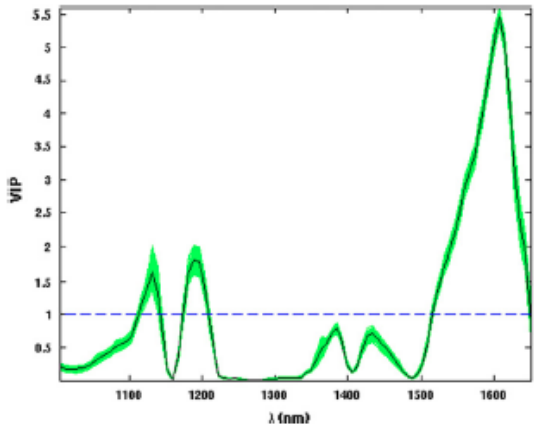
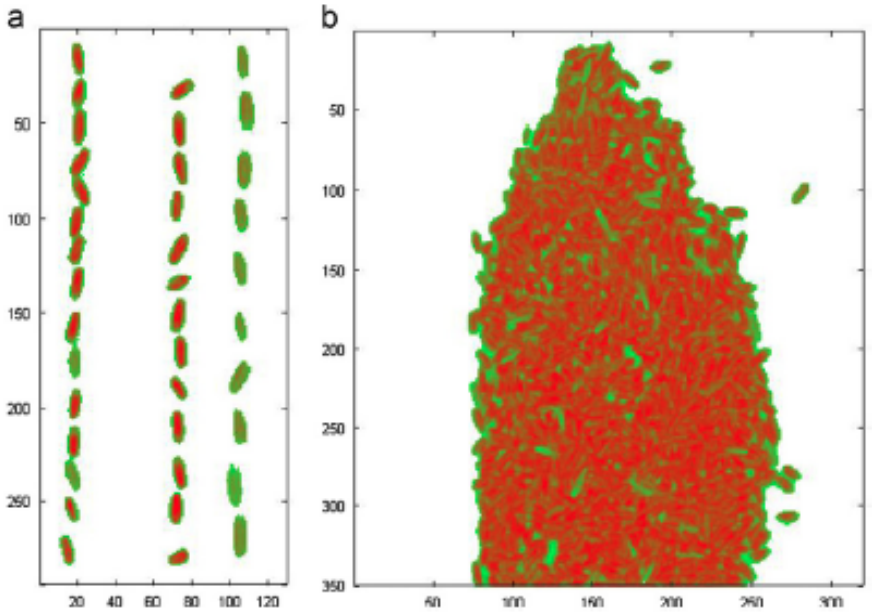


Fig. 7. PLS-DA modeling: VIP scores for the 93 spectral variables (black line) and their confidence intervals estimated by the described bootstrap procedure. The dashed horizontal line indicates the threshold value of 1.



SIMCA and class-modeling

SIMCA - 1

- Originally proposed by Wold in 1976
  - **SOFT**: No assumption of the distribution of variable is made (bilinear modeling)
  - **INDEPENDENT**: Each category is modeled independently
  - **MODELING of CLASS ANALOGIES**: Attention is focused on the similarity between object from the same class rather then on differentiating among classes.
- To build the individual category models, PCA is used.
  - The number of significant components A (defining the "inner space") can be different from class to class.
  - The remaining M-A components represent the residuals ("outer space")

SIMCA – original version

- A PCA model of A_C components is computed using the training samples from class C.

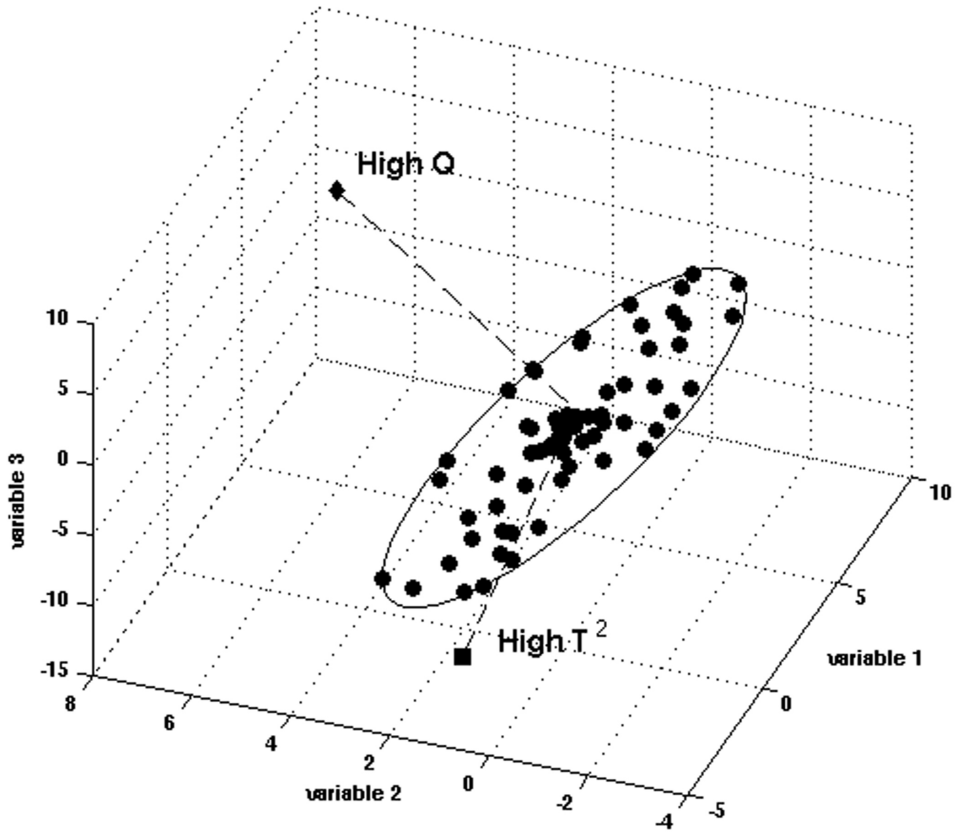
$$\underset{N\times A_C}{\mathbf{X}_C} = \underset{A_C\times M}{\mathbf{T}_C\mathbf{P}_C^T} + \mathbf{E}_C$$

- A residual standard deviation for the category is computed according to:

$$s_0 = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M} e_{ij,C}^2}{(M-A)(N-A-1)}}$$

- This rsd represent an indication of the typical deviation of samples belonging to the class to its category model.

SIMCA – MSPC version

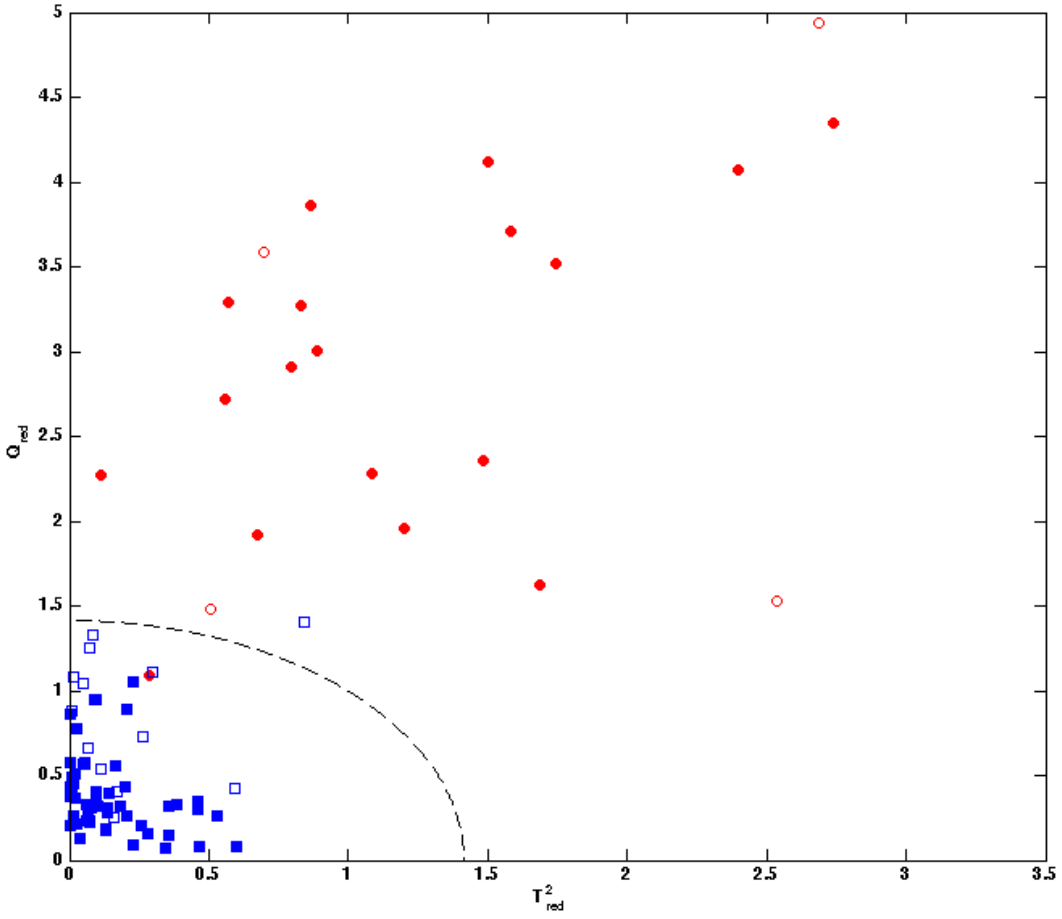


$$d_k^C = \sqrt{\left(T^2_{red,k}\right)^2_C + \left(Q_{red,k}\right)^2_C} = \sqrt{\left(\frac{T_k^2}{T^2_{0.95}}\right)^2_C + \left(\frac{Q_k}{Q_{0.95}}\right)^2_C}$$

SIMCA – MSPC version

- The corresponding criterion:

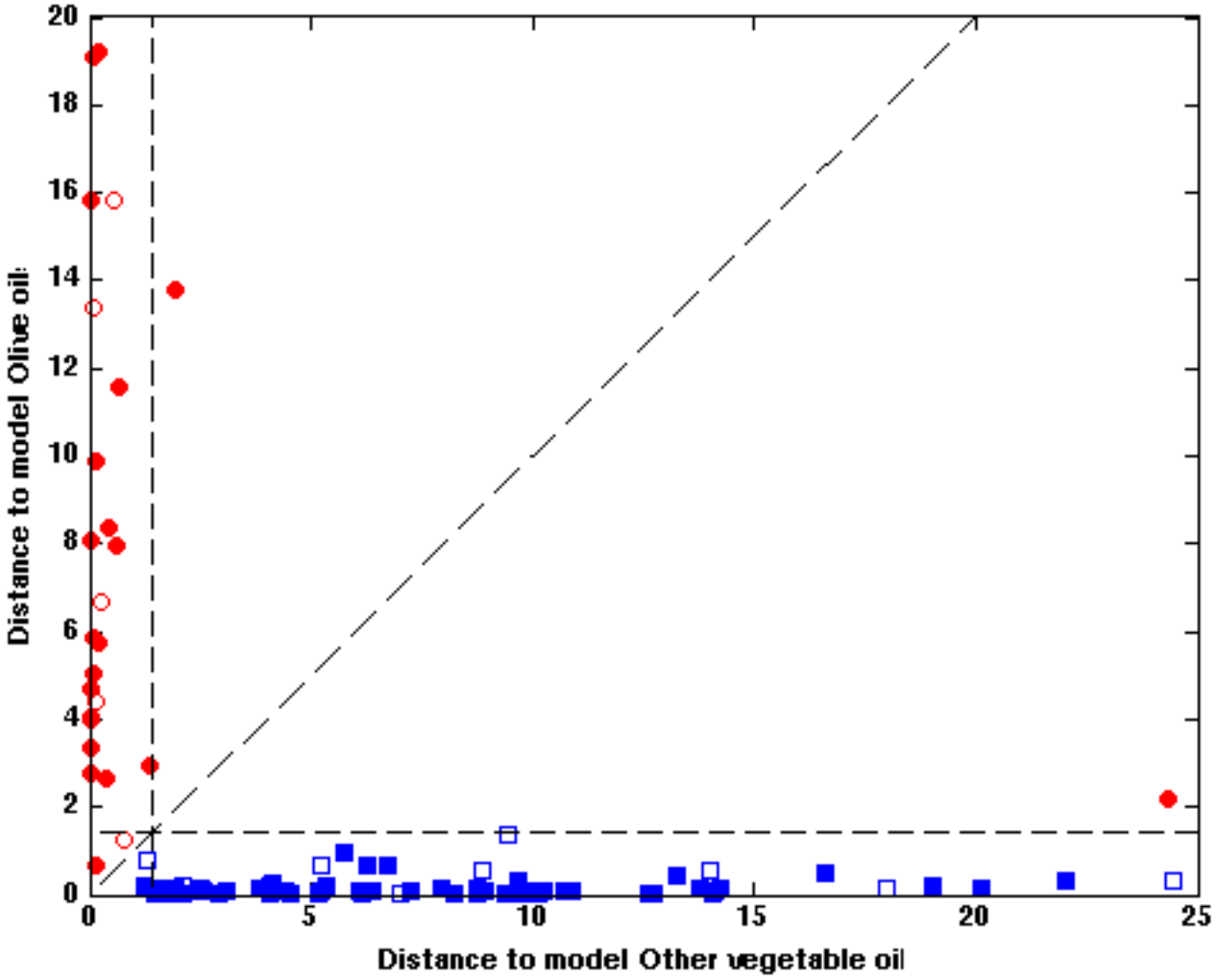
$$d_k^C < \sqrt{2}$$



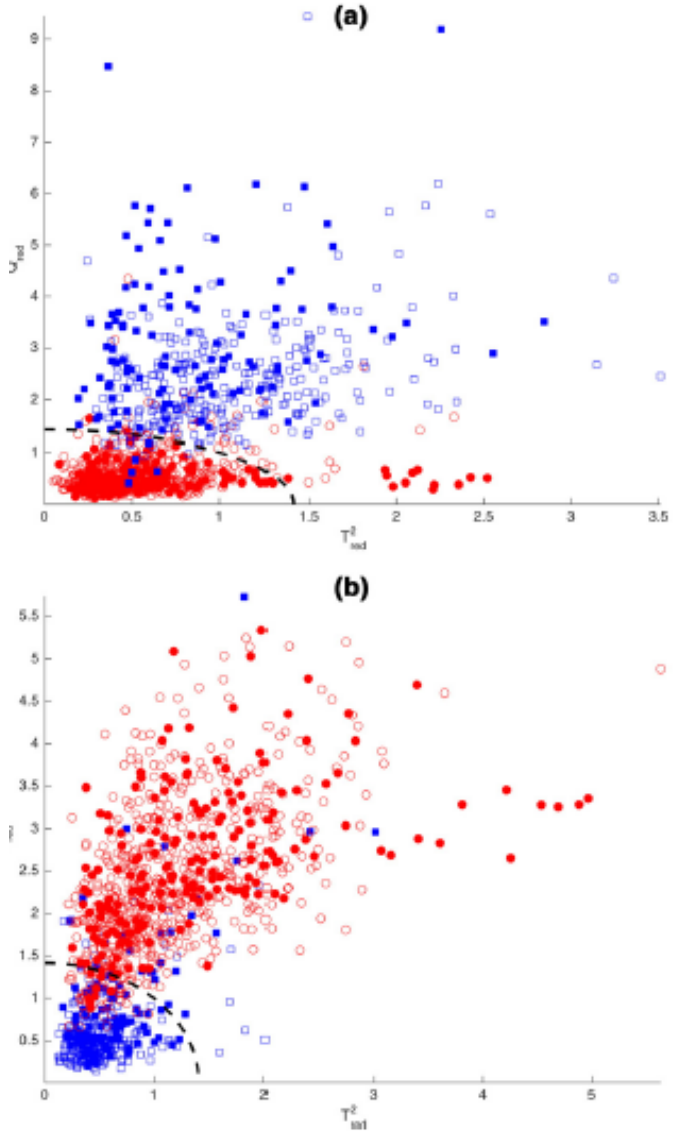
Characteristics

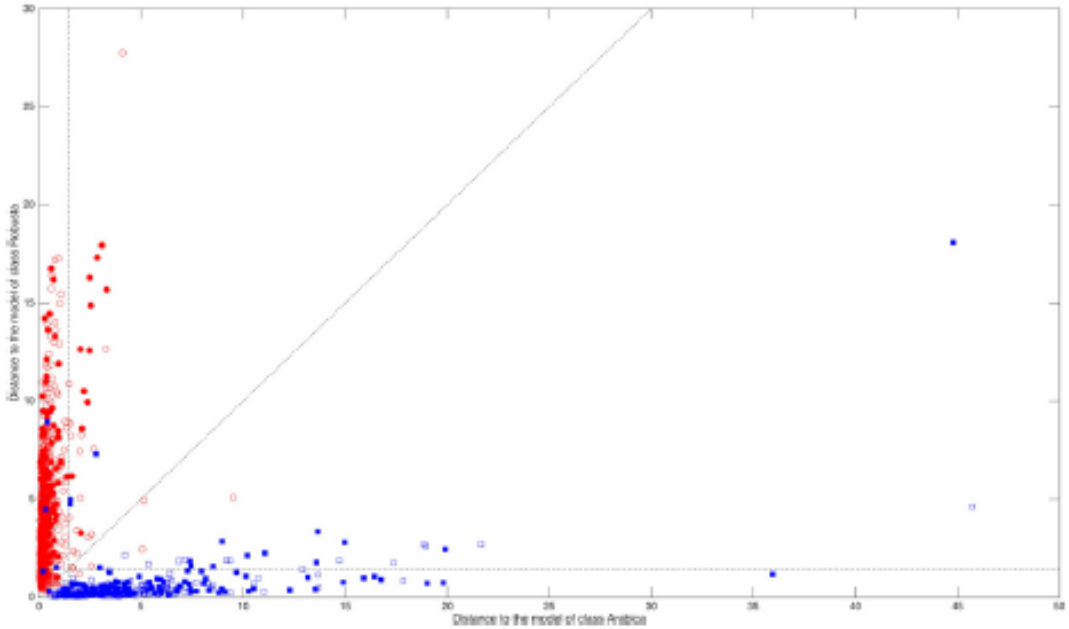
- Flexibility
- Additional information:
  - *sensitivity*: fraction of samples from category X accepted by the model of category X
  - *specificity*: fraction of samples from category Y (or Z, W….) refused by the model of category X
- No need to rebuild the existing models each time a new category is added.
- Can be easily transformed into a discriminant technique:
  - Sample is assigned to the class to the model of which it has the minimum distance.

Coomans plot



Discrimination of coffee varieties





Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

Characterization of the effects of different roasting conditions on coffee samples of different geographical origins by HPLC-DAD, NIR and chemometrics

Silvia De Luca, Martina De Filippis, Remo Bucci, Andrea D. Magrì, Antonio L. Magrì, Federico Marini *

Department of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro 5, I-00185 Rome, Italy

Validation


Ceci est un message
pour les voyageurs amateurs
de science-fiction
Obligatoire
la validation est
NE CÉDEZ PAS AU CÔTÉ OBSCUR !

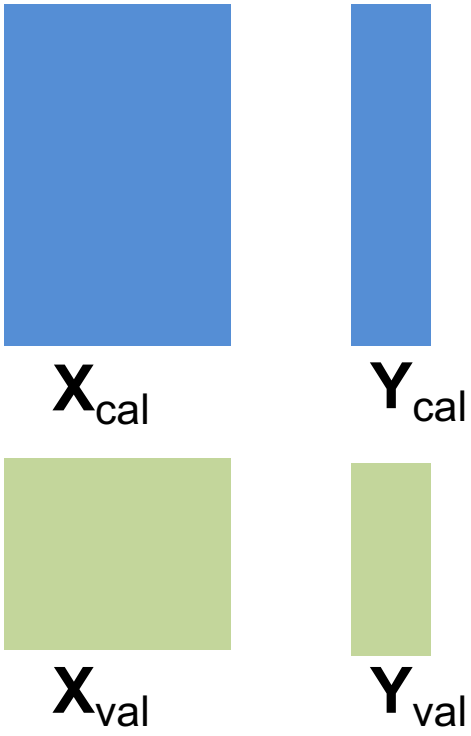
The concept of validation

- 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).

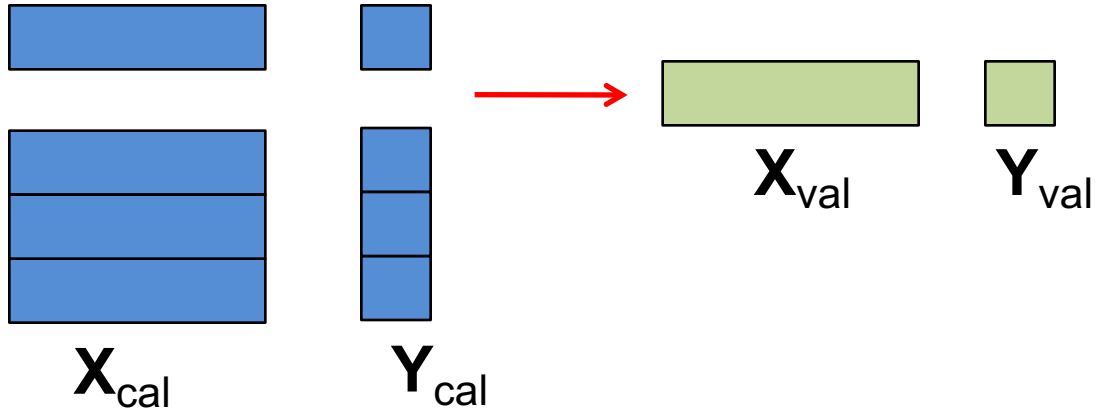
The need for “new” data

- 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

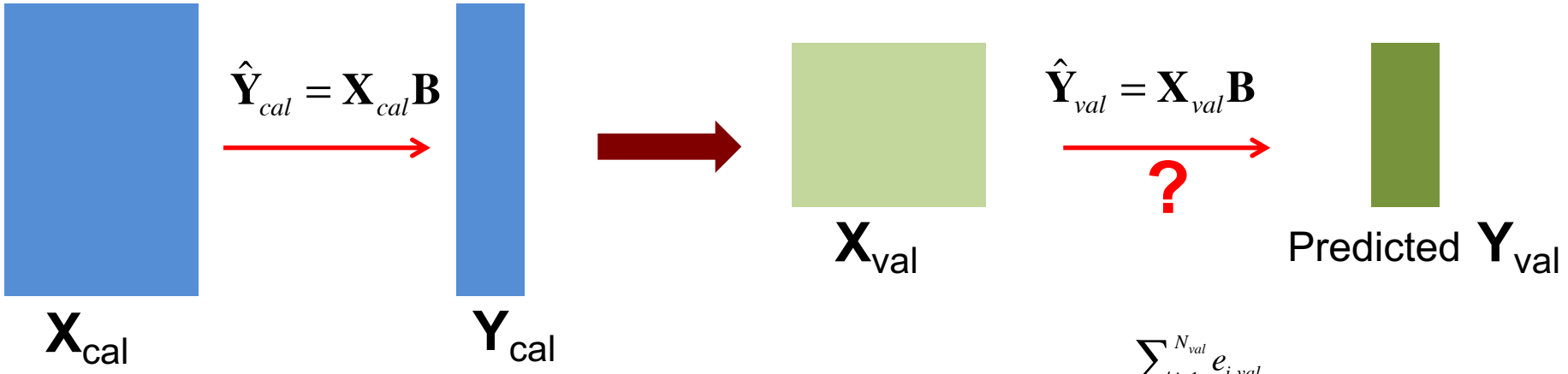


Cross-validation



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 → $\mathbf{Y}_{val}$ ***vs*** Pred. $\mathbf{Y}_{val}$

$$CE\% = 100 \times \frac{\sum_{i=1}^{N_{val}} e_{i,val}}{N_{val}}$$

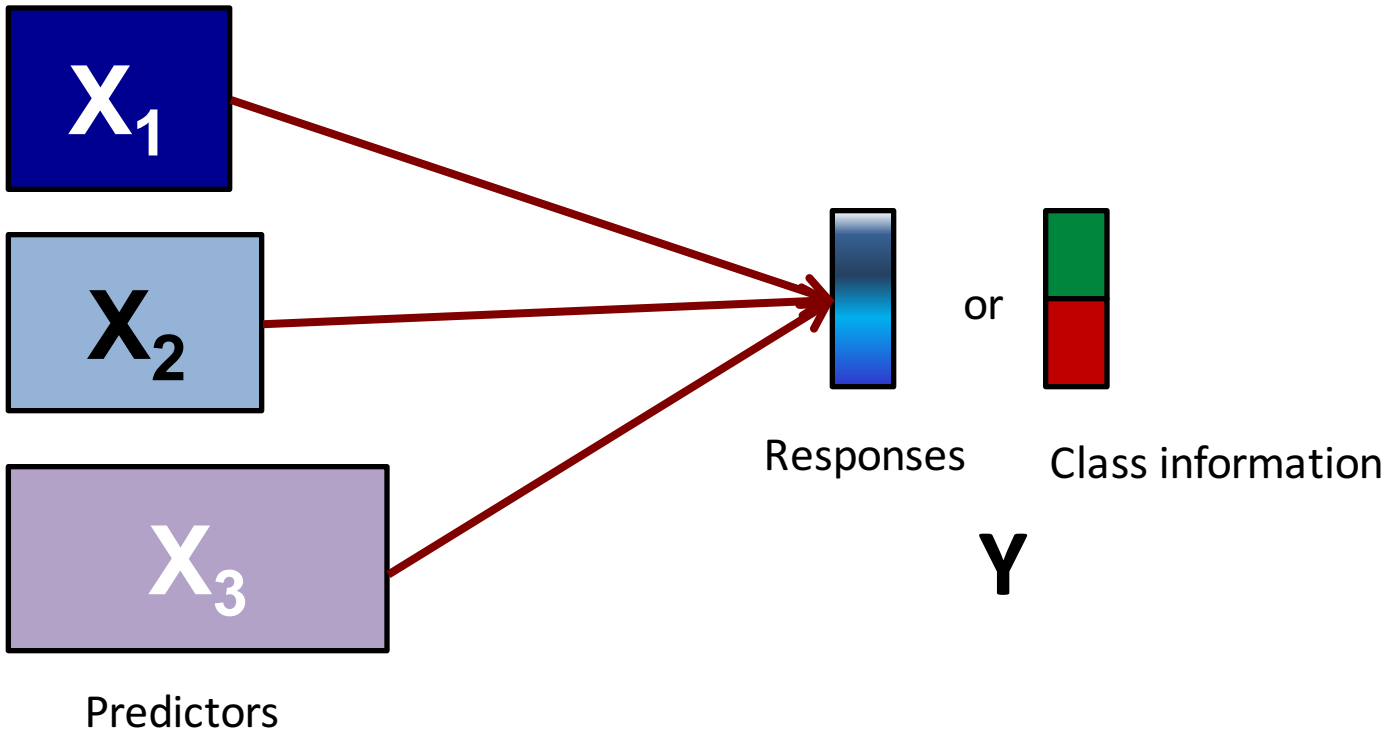
$$\text{where } e_{i,val} = \begin{cases} 1 \text{ if } PredClass(i) \neq TrueClass(i) \\ 0 \text{ if } PredClass(i) = TrueClass(i) \end{cases}$$

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…)

Integrating data from different blocks

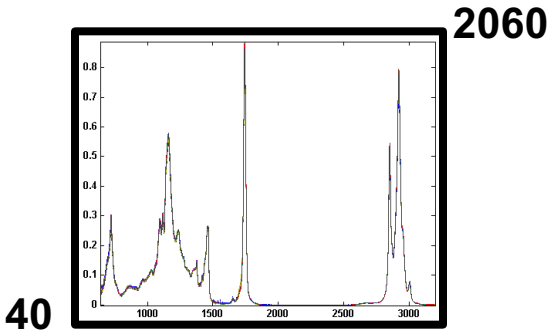
DATA FUSION STRATEGIES

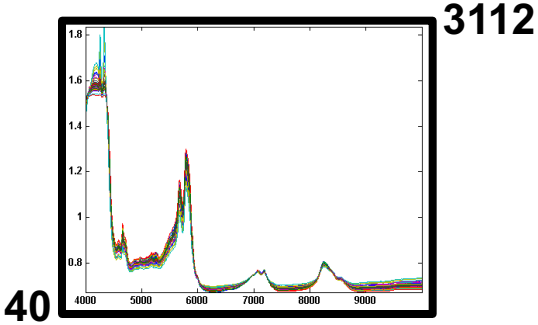


- LOW LEVEL → Data
- MID LEVEL → Features
- HIGH LEVEL → Decision rules

LOW LEVEL DATA FUSION

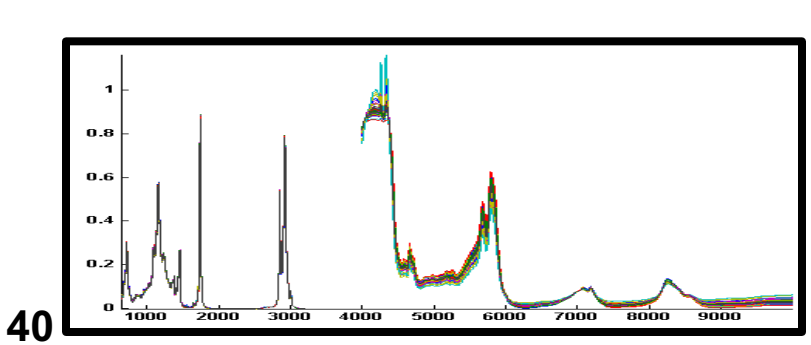
Data are concatenated and treated as they were a single fingerprint



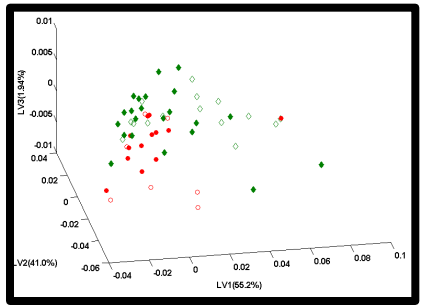


Preprocessing

Preprocessing



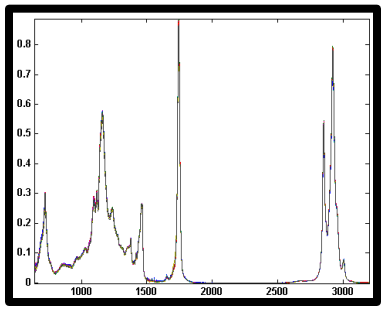
Preprocessing

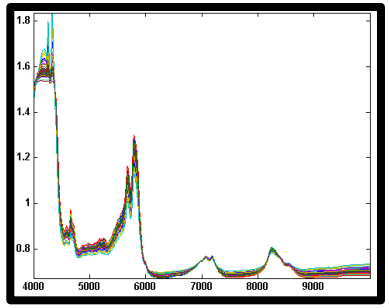


Modeling

MID LEVEL DATA FUSION

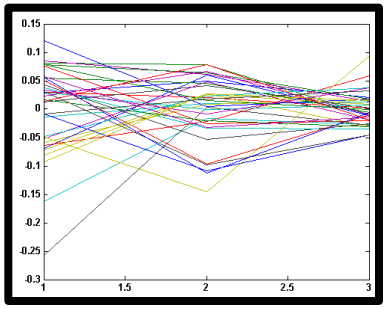
Features extracted from the data are concatenated

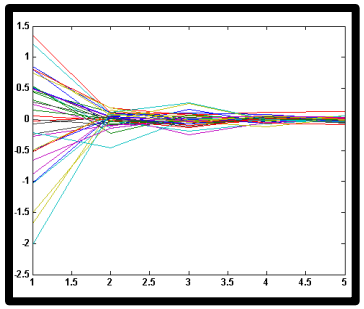




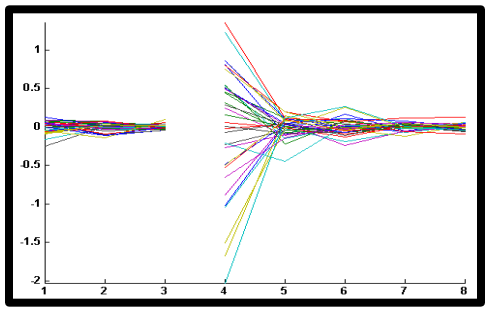
Preprocessing + Feature extraction

Preprocessing + Feature extraction



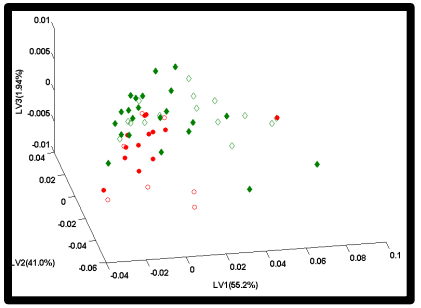


Concatenation



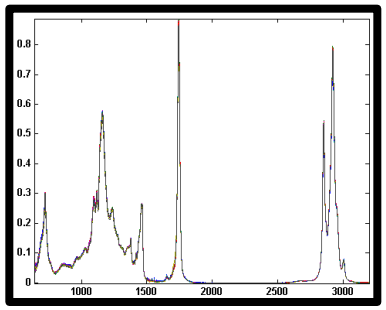
Preprocessing

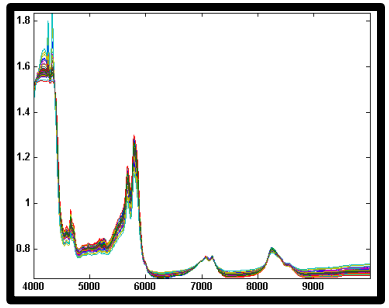
Modeling



HIGH LEVEL DATA FUSION

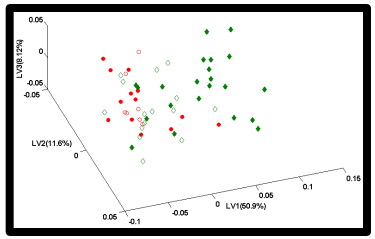
Fusion occurs at the decision level

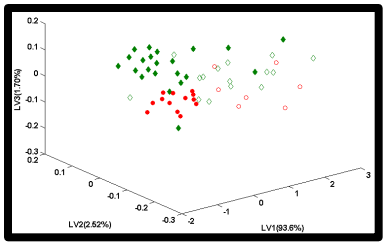




Preprocessing +
Feature extraction +
Modeling

Preprocessing +
Feature extraction +
Modeling





Decision 1 (e.g. Class A)

Decision 2 (e.g. Class B)

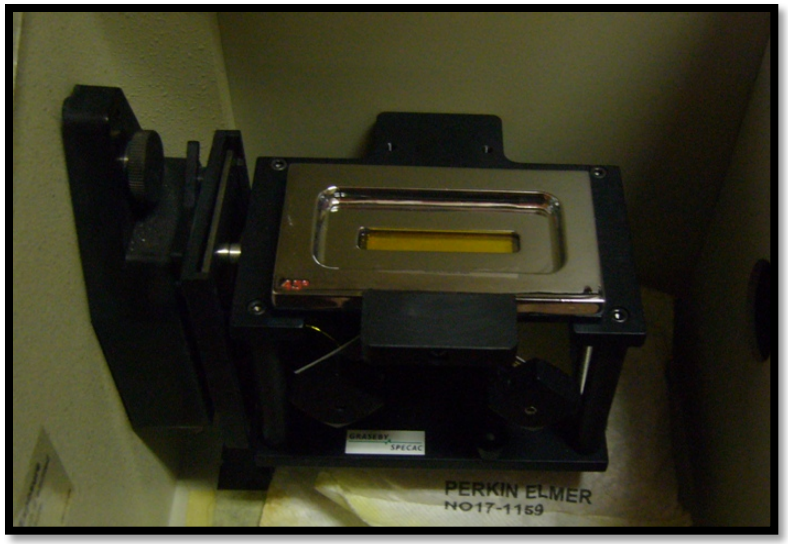
Majority vote
Bayes' theorem

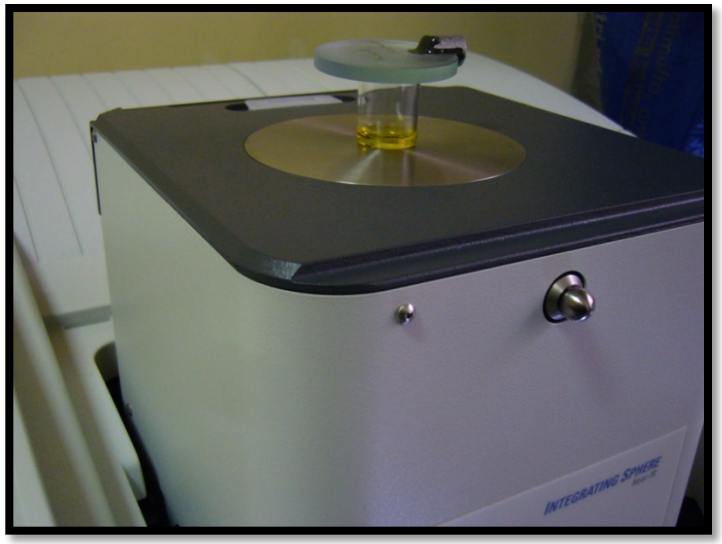
Final decision (e.g. Class A)

A first example: Authentication of extra virgin olive oils from PDO Sabina

OLIVE OIL DATA SET

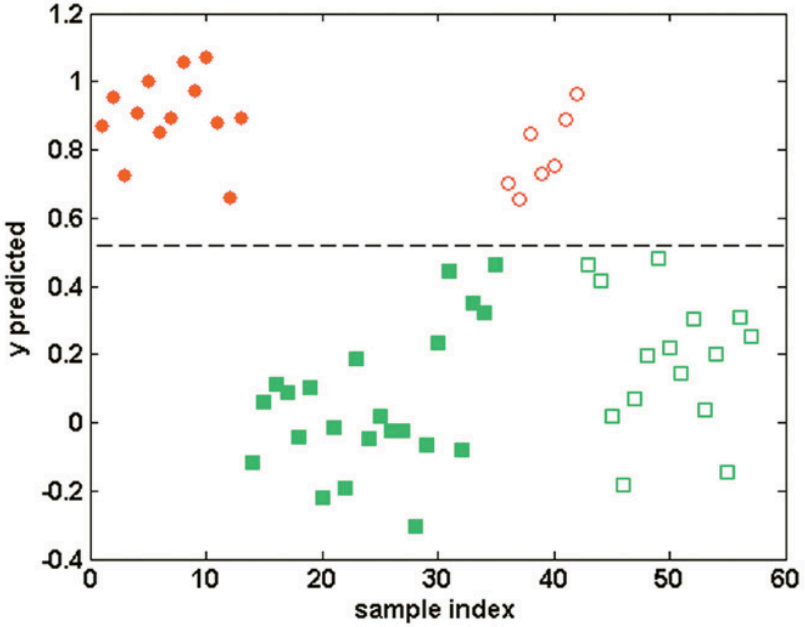
- Authentication of the origin of olive oil samples
- 57 extra virgin olive oil samples
  - 20 from Sabina, Lazio (13 harvested 2009, 7 harvested 2010)
  - 37 samples of different origin (22 from 2009, 15 from 2010
- MIR and NIR spectra recorded on each sample





DATA FUSION

- LOW LEVEL
  - Without block-scaling: Block with the highest variance (here MIR) governs the model
  - With block-scaling: Improved contribution of NIR but still poorer results than with NIR alone
- MID LEVEL (PLS-DA scores after autoscaling)



AUTHENTICATION OF BEER

Characterization of artesanal beer ‘‘Reale’’ and its authentication

BIRRA DEL BORGO

‘‘ **ReAle**’’ is an artesanal beer brewed by ‘‘*Birrificio del Borgo”*, an Italian microbrewery well recognized also abroad for its high quality products

Analytica Chimica Acta 820 (2014) 23–31

Contents lists available at ScienceDirect

Analytica Chimica Acta

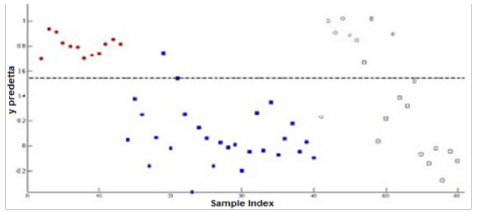
journal homepage: www.elsevier.com/locate/aca

Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication

Alessandra Biancolillo, Remo Bucci, Antonio L. Magrì, Andrea D. Magrì, Federico Marini*

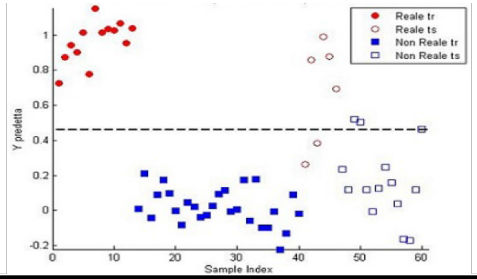
Department of Chemistry, University of Rome “La Sapienza”, Rome, Italy

Results of individual techniques



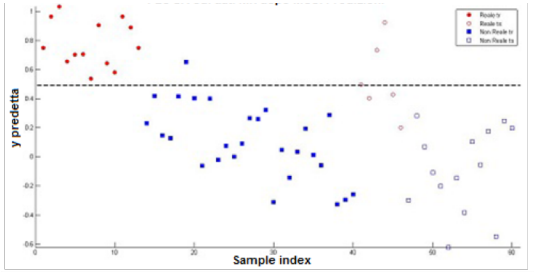
Predictions		
Pretreatment	%Correct Class. (Pred)	
	"Reale"	"Not Reale"
Deriv. I (+MC)	**83.3**	**71.4**

TG



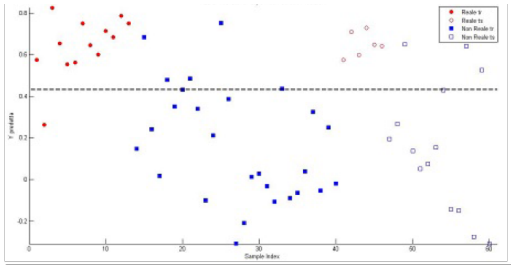
Predictions		
Pretreatment	%Correct Class. (Pred)	
	"Reale"	"Not Reale"
SNV+Detrending (+MC)	**66.7**	**78.6**

MIR



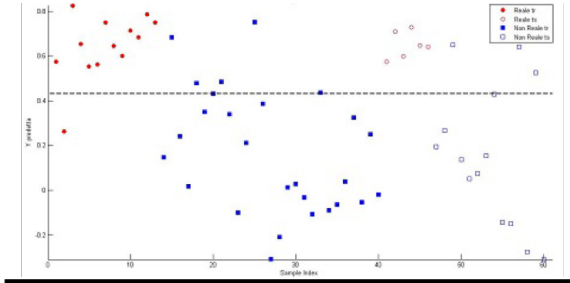
Predictions		
Pretreatment	%Correct Class. (Pred)	
	"Reale"	"Not Reale"
MSC (+MC)	**66.7**	**100.0**

NIR



Predictions		
Pretreatment	%Correct Class. (Pred)	
	"Reale"	"Not Reale"
Mean Centering	**82.3**	**77.8**

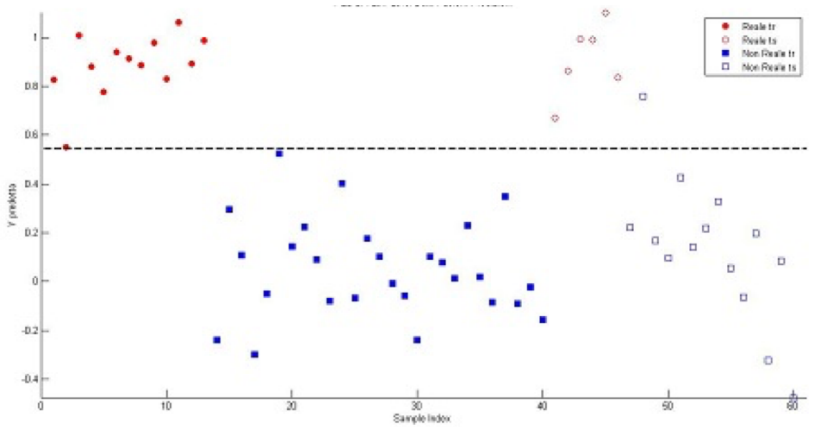
UV



Predictions		
Pretreatment	% Correct Class. (Pred)	
	"Reale"	"Not Reale"
Mean Centering	**100.0**	**85.7**

Vis

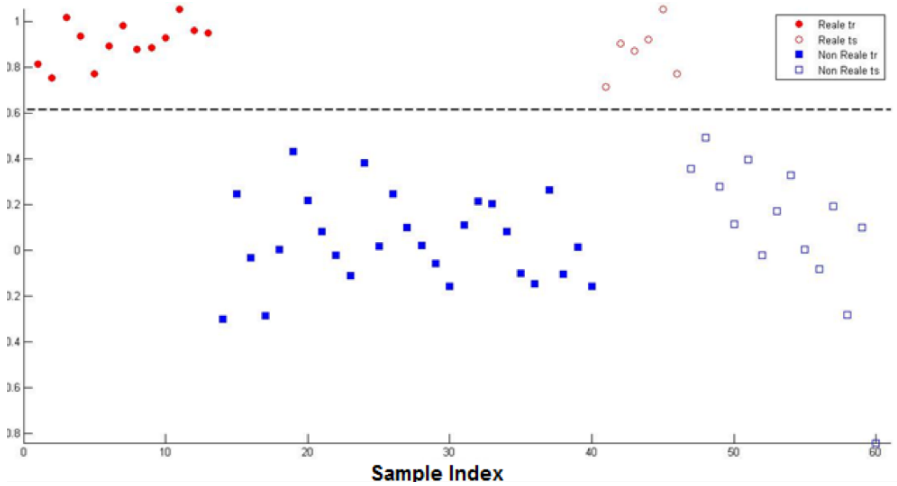
Data fusion



Predictions

Pretreatment	% Correct Class. (Pred)	
	"Reale"	"Not Reale"
Without block scaling	**100.0**	**92.3**
With block scaling	**100.0**	**78.6**

Low Level



Predictions

Pretreatment	% Correct Class. (Pred)	
	"Reale"	"Not Reale"
Mean Centering	**100.0**	**100.0**

Mid Level

Thank you for your attention

federico.marini@uniroma1.it