VALIDATION OF CHEMOMETRIC MODELS AND ASPECTS OF DATA FUSION

Federico Marini

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

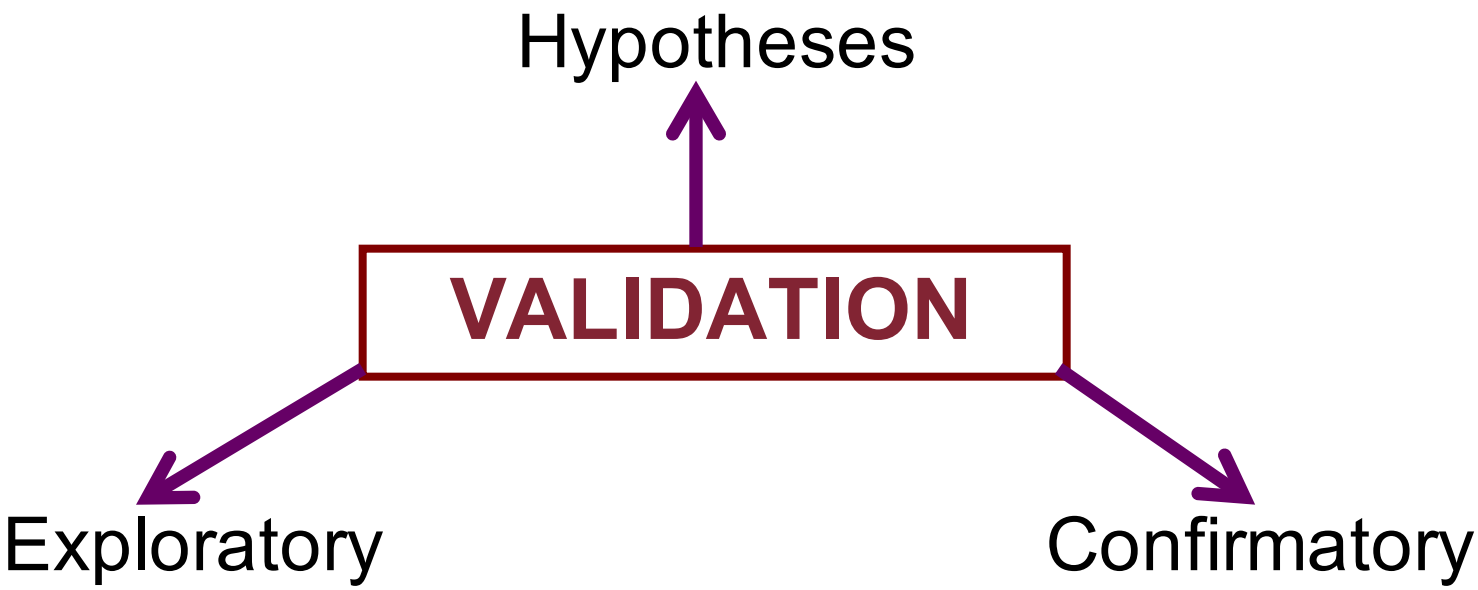
SAPIENZA
UNIVERSITÀ DI ROMA

Far too often, solutions obtained by multivariate procedures—including factor analysis, multidimensional scaling, and cluster analysis—are interpreted, and even published, without adequate evaluation of their reliability or validity. Particularly among inexperienced users, there is an uncritical and somewhat cavalier approach to determining what parts (or which version) of an analysis to accept. Clusters or dimensions are frequently taken to be "real" whenever an interpretation can be projected onto them by the imagination of the analyst. On the other hand, dimensions that don't fit preconceptions and are hard to interpret tend to be dismissed too easily. While some users may make a feeble attempt at justifying their choice of dimensionality by examining improvements in fit values, little effort is otherwise expended in determining whether clusters or dimensions are stable or reliable, whether the model is appropriate for the data, whether the algorithm achieved correct convergence, whether serious outliers are present in the data, and so forth.

"How can I know if it's 'real'?" A Catalog of Diagnostics for Use with Three-Mode Factor Analysis and Multidimensional Scaling

Richard A. Harshman

In H. G. Law, C. W. Snyder, Jr., J. Hattie, & R. P. McDonald (Eds.), Research methods for multimode data analysis (pp. 566-591). New York: Praeger.
Available from: http://psychology.uwo.ca/faculty/harshman/



BUT ALSO:

- was an appropriate model chosen?
- are outliers and/or highly influential points present?
- is the selected subspace stable?
- has the algorithm converged?

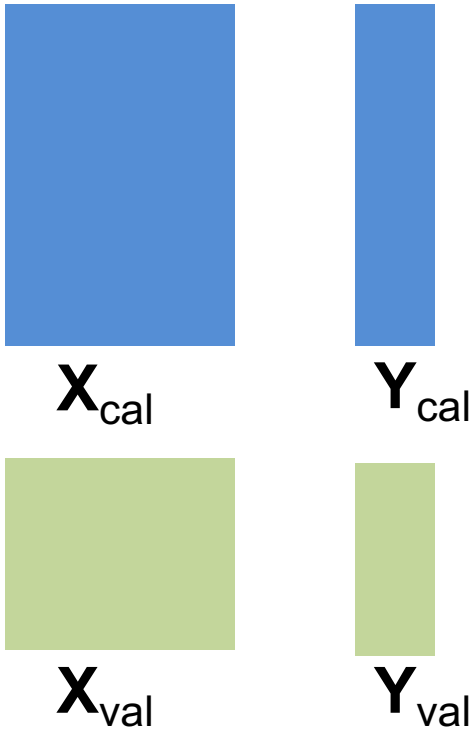
The concept of validation

- 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).

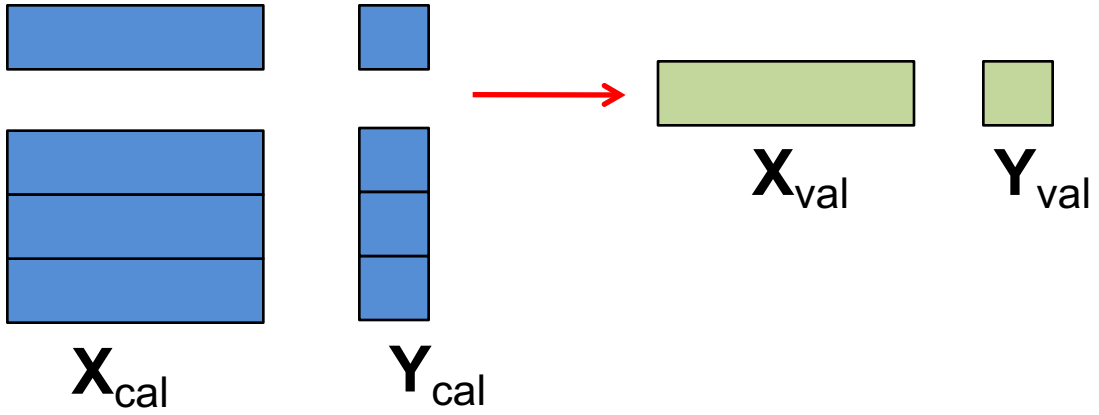
The need for “new” data

- 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

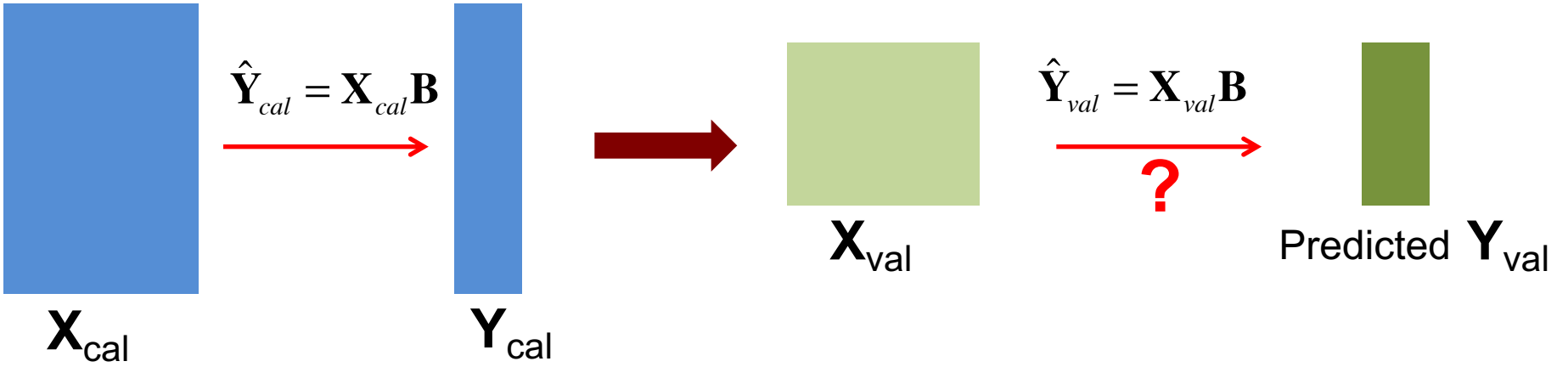


Cross-validation



Test set validation (regression)

- 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.

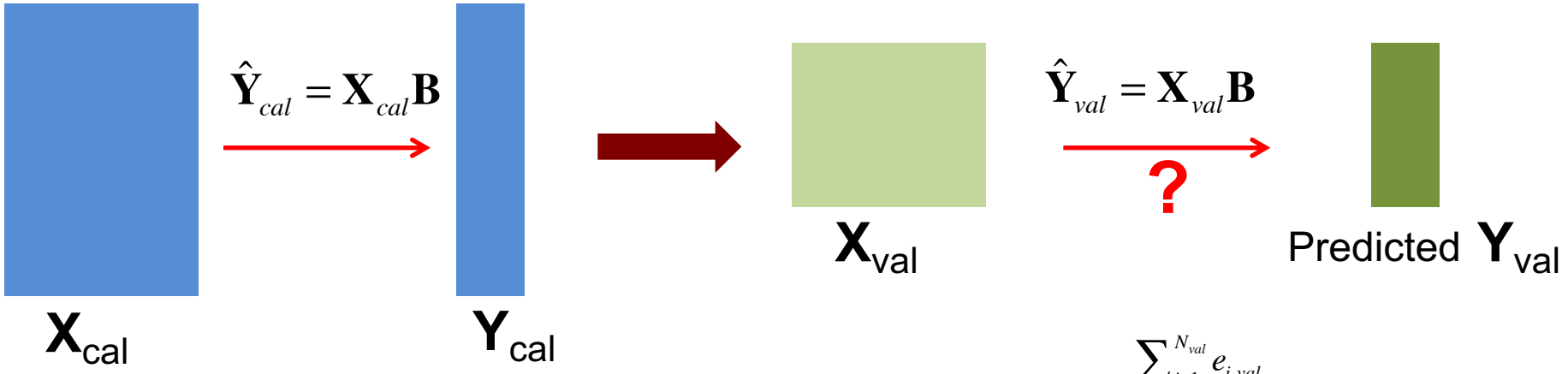


MODEL QUALITY → $\mathbf{Y}_{val}$ ***vs*** Pred. $\mathbf{Y}_{val}$

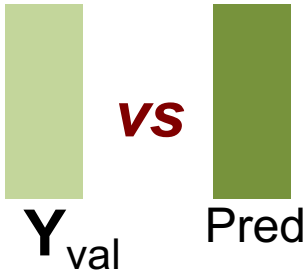
$$RMSEP = \sqrt{\frac{\sum_{i=1}^{N_{val}} (y_i^{val} - \hat{y}_i^{val})^2}{N_{val}}}$$

Test set validation (classification)

- 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



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

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

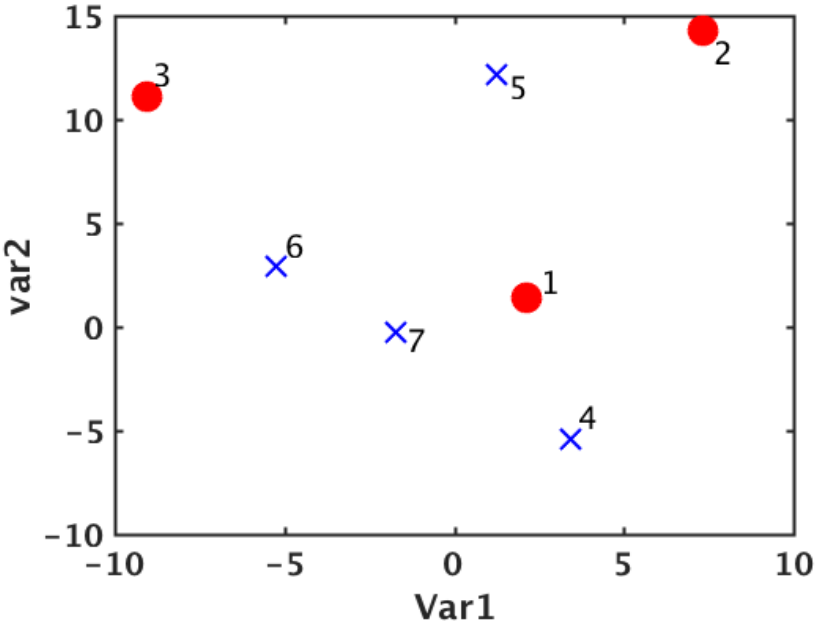
How to split the data?

- Intelligent choice of the samples to be put in each set → reliable considerations based on the obtained results.
- Different criteria have been proposed in the literature to operate an intelligent splitting
- They all share the same concept:
  - try to span the sample space as uniformly as possible.
- Just to cite a few:
  - Kennard-Stone
  - Duplex
  - D-optimal criterion
  - Kohonen-based

Kennard-Stone algorithm

R.W. Kennard, L.A. Stone, *Technometrics*, **11** (1969), 137-148

- The most diverse samples are placed in the training set
- All the remaining ones are left out as test set
- The "diversity" of a new samples from the ones already selected is defined by the ***maximin*** criterion:
  - The sample with the maximum value of the minimum distance to the ones already selected is added to the training set



Samples	1	2	3	Min distance
4	6.9	20.1	20.7	6.9
5	10.8	6.4	10.4	6.4
6	7.5	17.0	9.0	**7.5**
7	4.2	17.1	13.5	4.2

Sample 6 would be selected as the next one to be included

Duplex algorithm

R.D. Snee, *Technometrics*, **19** (1977), 415-428

- Kennard-Stone approach tries to concentrate as much of the data diversity in the training samples
- It can lead to overoptimistic results
- A modification of the algorithm aimed at maintaining a comparable diversity between the two sets was proposed by Kennard himself (even though it was left unpublished until it was discussed by Snee).

DUPLEX

D-optimal criterion

V.V. Fedorov, *Theory of optimal experiments*, Academic Press, NY, 1972

- Another possibility of uniformly sampling the sample space to build the training set is the use of optimal designs.
- Optimality is defined wrt some statistical criterion (usually related to minimizing the variance of the estimators).
- The definition of optimality requires a statistical model (e.g., multiple linear regression).
- Given the matrix of predictors **X**, the information matrix is defined as $\mathbf{X}^T\mathbf{X}$:
  - **A optimality**: mimimize $\text{tr}((\mathbf{X}^T\mathbf{X})^{-1})$
  - **D optimality**: maximize $\det(\mathbf{X}^T\mathbf{X})$
  - **E optimality**: maximize the miminum eigenvalue of $(\mathbf{X}^T\mathbf{X})$
  - **T optimality**: maximize $\text{tr}(\mathbf{X}^T\mathbf{X})$
- One could also focus on the variance of the predictions:
  - **G optimality:** minimize the maximum element of $\text{diag}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$, i.e. the max variance of the predicted values
  - **I optimality**: minimize the average prediction variance over the design space
  - **V optimality**: minimize the average prediction variance over a set of predefined points

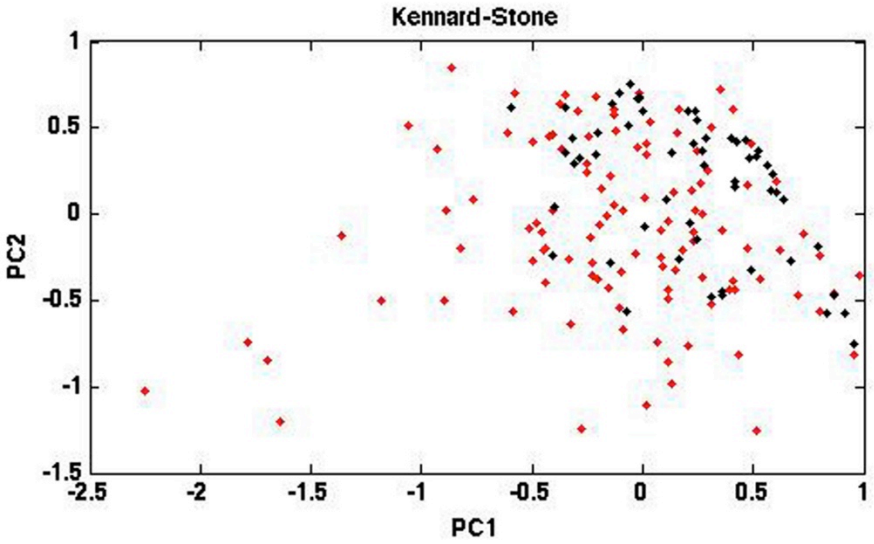
Using the D-optimal criterion for subset selection

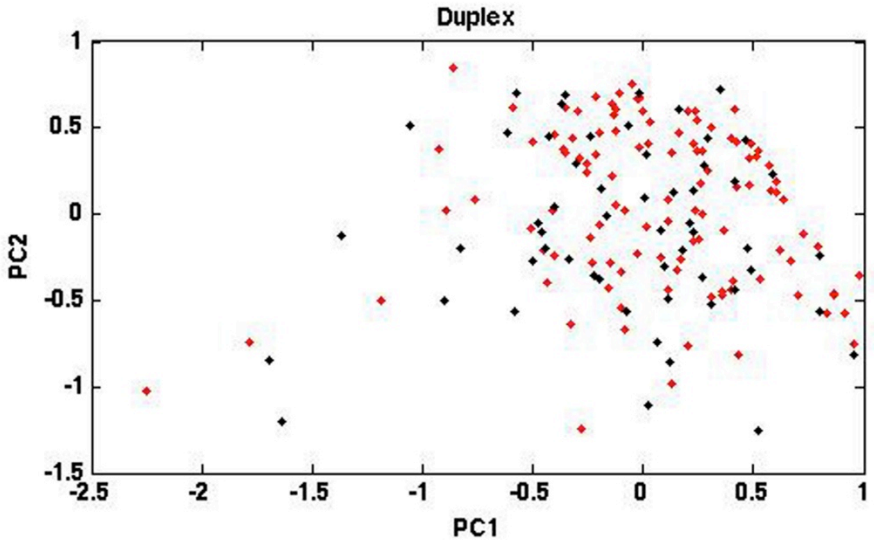
1. Generate a list of candidate points (in general, it is the whole data set)
2. Define the statistical model (usually a linear model without interactions):

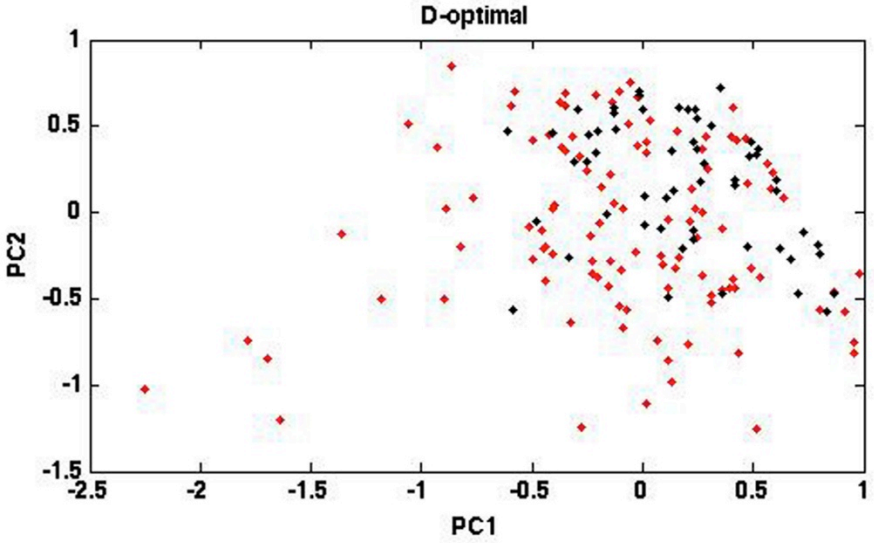
$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

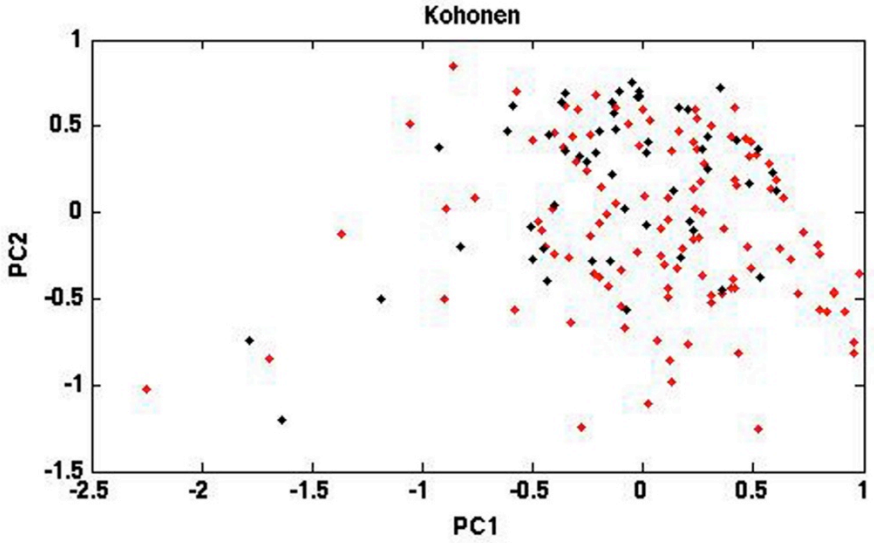
3. Select candidate subsets of N_{train} individuals and calculate their information matrix $\boldsymbol{X}_j^T \boldsymbol{X}_j$
4. Repeat the procedure until a subset is found which maximizes the determinant of $\boldsymbol{X}_k^T \boldsymbol{X}_k$
5. That subset will be the training set; all the other samples will be the test set.
6. If the matrix is ill-conditioned, calculate at most (n-1) PCs and build the information matrix using the scores $\boldsymbol{T}_j^T \boldsymbol{T}_j$

Data splitting

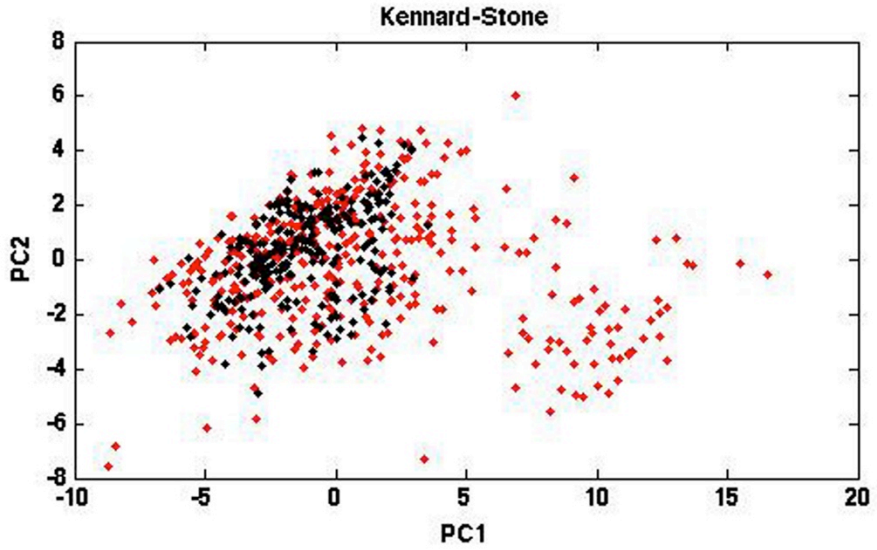


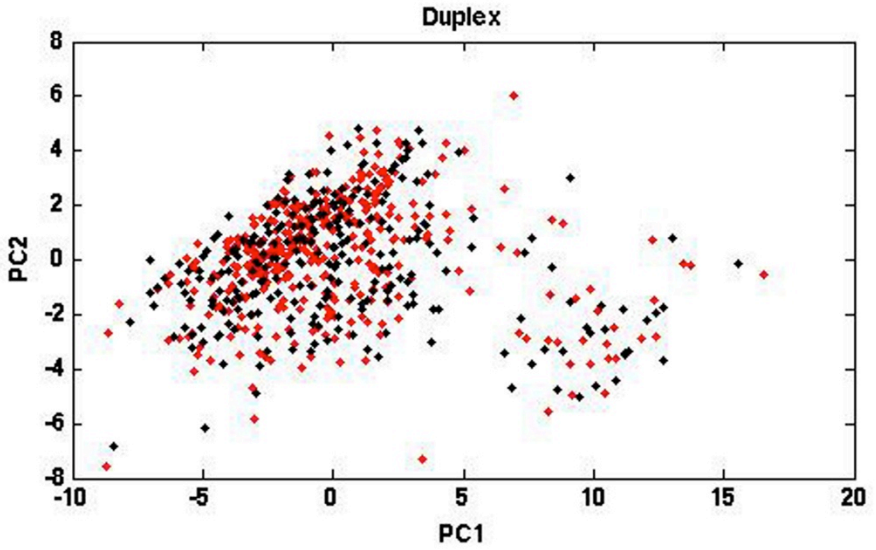


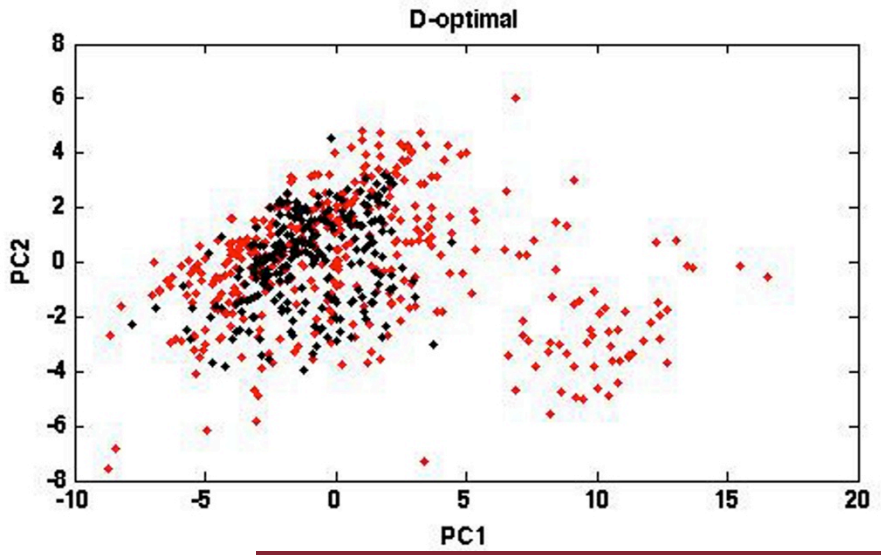


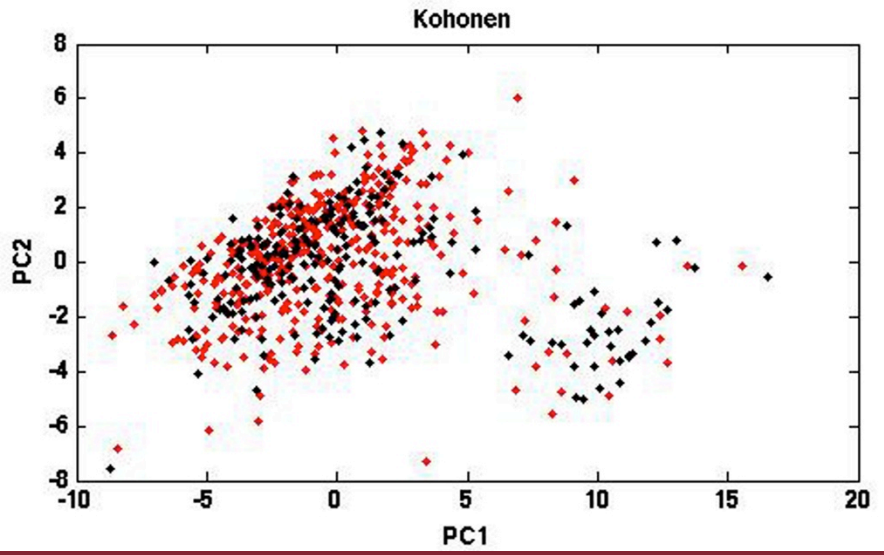


Data splitting - 2



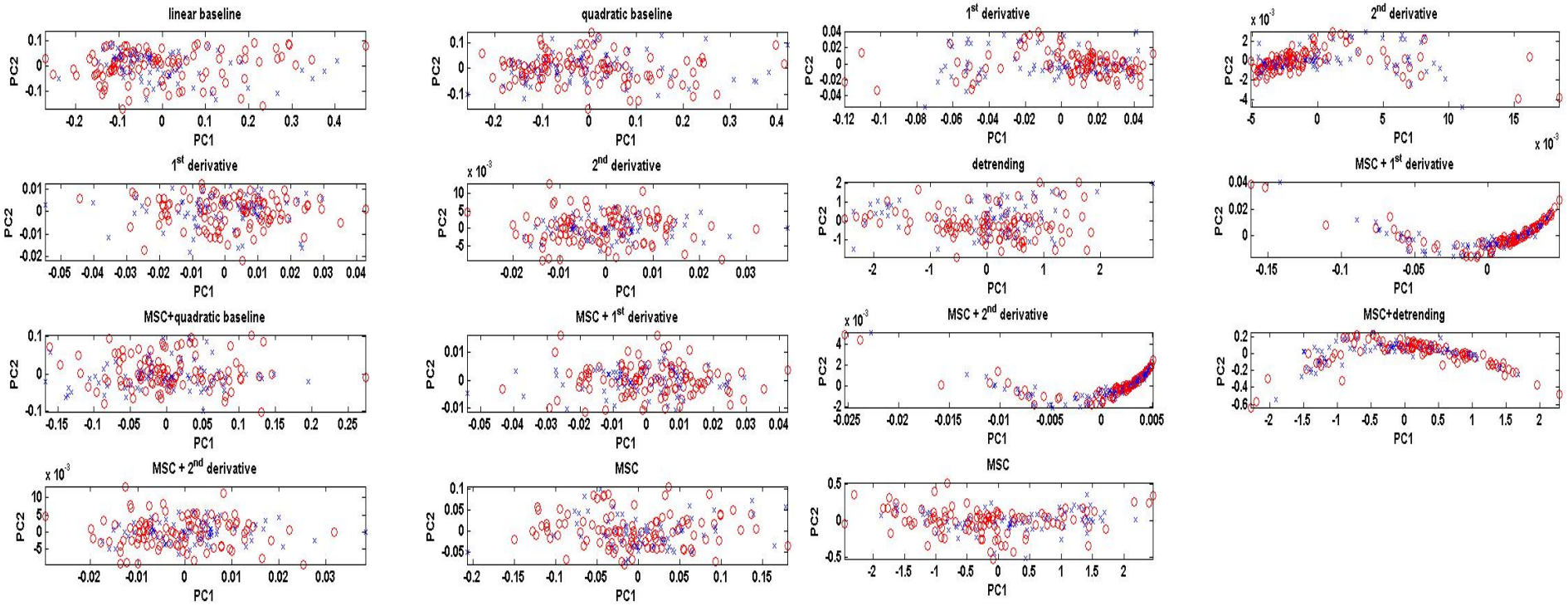






Training/test set selection

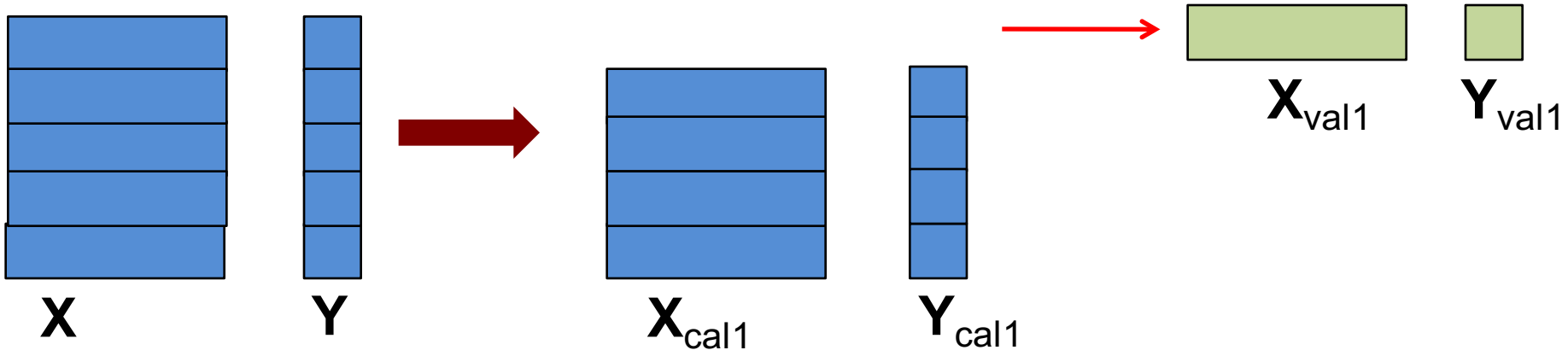
- Duplex algorithm repeated class-wise on each pretreatment separately (Split ratio: 2/1)
- Data selected more than 10 times (out of 15) in test set



Cross-validation

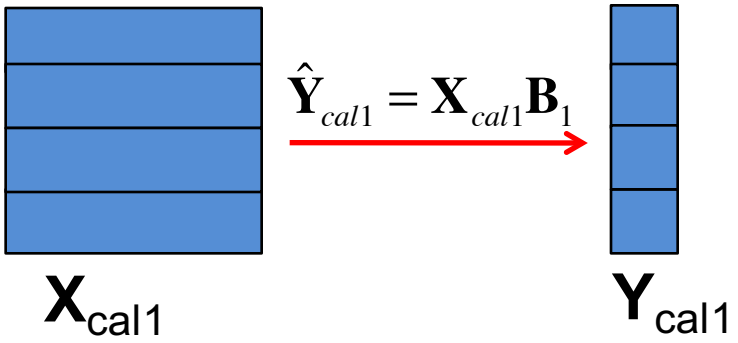
- Internal resampling method:
  - Simulates test set validation by repeating a data splitting procedure where different object are in turn placed in the validation set.
  - Particularly useful when a limited number of samples are available.
- Schematically, it consists of the following steps:

1. Leave out part of the data values

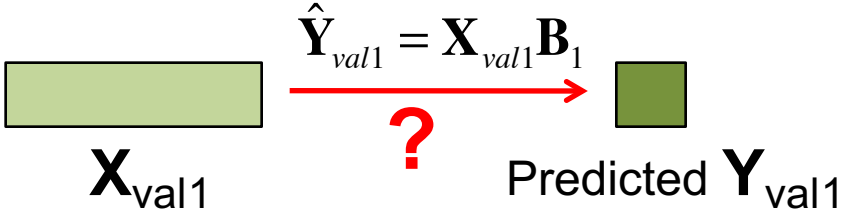


Cross-validation

2. Build the model without these data

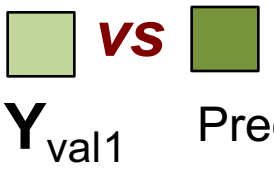


3. Apply the model to the left out values and obtain predictions;



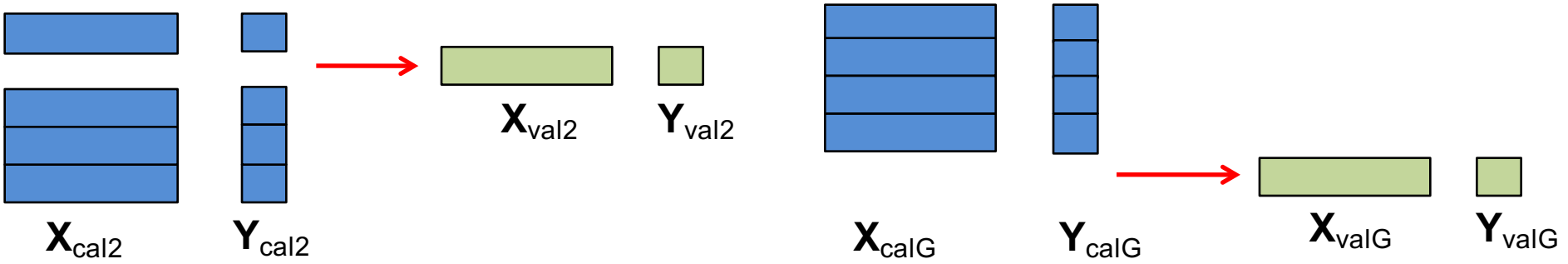
Cross-validation

4. Calculate the corresponding residual error

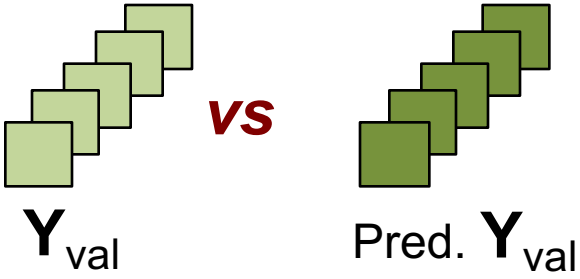


$$PRESS_1 = \sum_{i=1}^{N_{val1}} (y_i^{val1} - \hat{y}_i^{val1})^2$$

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



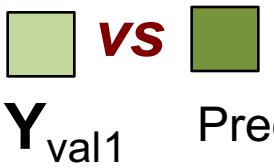
6. Collect all the residuals into an overall error criterion



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

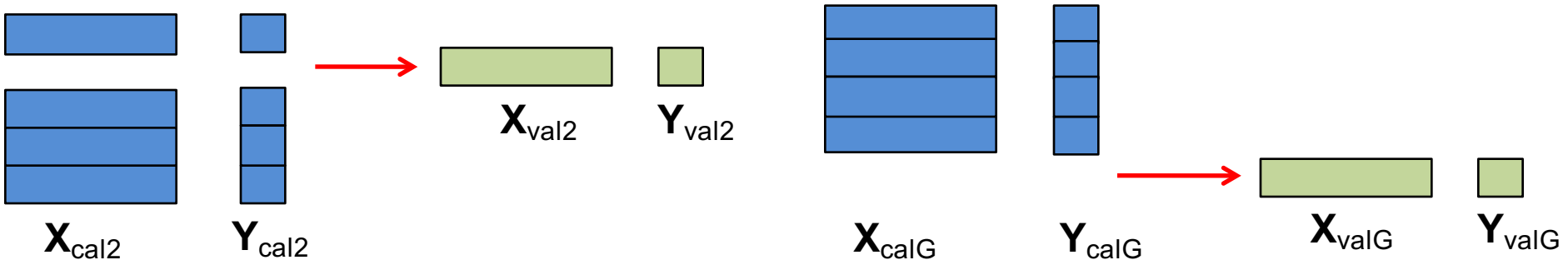
Cross-validation (classification)

4. Calculate the corresponding residual error

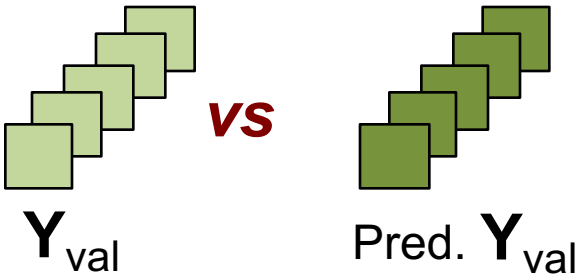


$$CE_1 = \sum_{i=1}^{N_{val1}} e_i^{val1}$$

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



6. Collect all the residuals into an overall error criterion



$$CE_{CV}\% = 100 \times \frac{\sum_{j=1}^{G} e_j}{N}$$

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…)

Cross-validation

Validation scheme	No. of objects	No. of factors	RMSEC	RMSECV	RMSEP
A: Random calibration and test	210/122	7	0.35	0.37	0.38
B: Keeping replicates out	332	8	0.35	0.37	-
C: Keeping sample out	166	8	0.35	0.44	-
D: Model based on 9 cultivars; test set 3 cultivars	118/47	7	0.39	0.44	0.58
E1: Model validated randomly year 2006-2007; test 2008	113/53	11	0.83	1.11	4.49
E2: Model validated across year 2006-2007; test 2008	113/53	2	1.44	2.09	1.38

Take home message ☺

Aspects of Data fusion

TOWARDS THE USE OF MULTIPLE BLOCKS

- Food quality control is a complex problem often requiring the interplay of more analytical platforms.
- Benefit of the specific advantages and characteristics of the different techniques ➔ more reliable and stable model

DATA FUSION

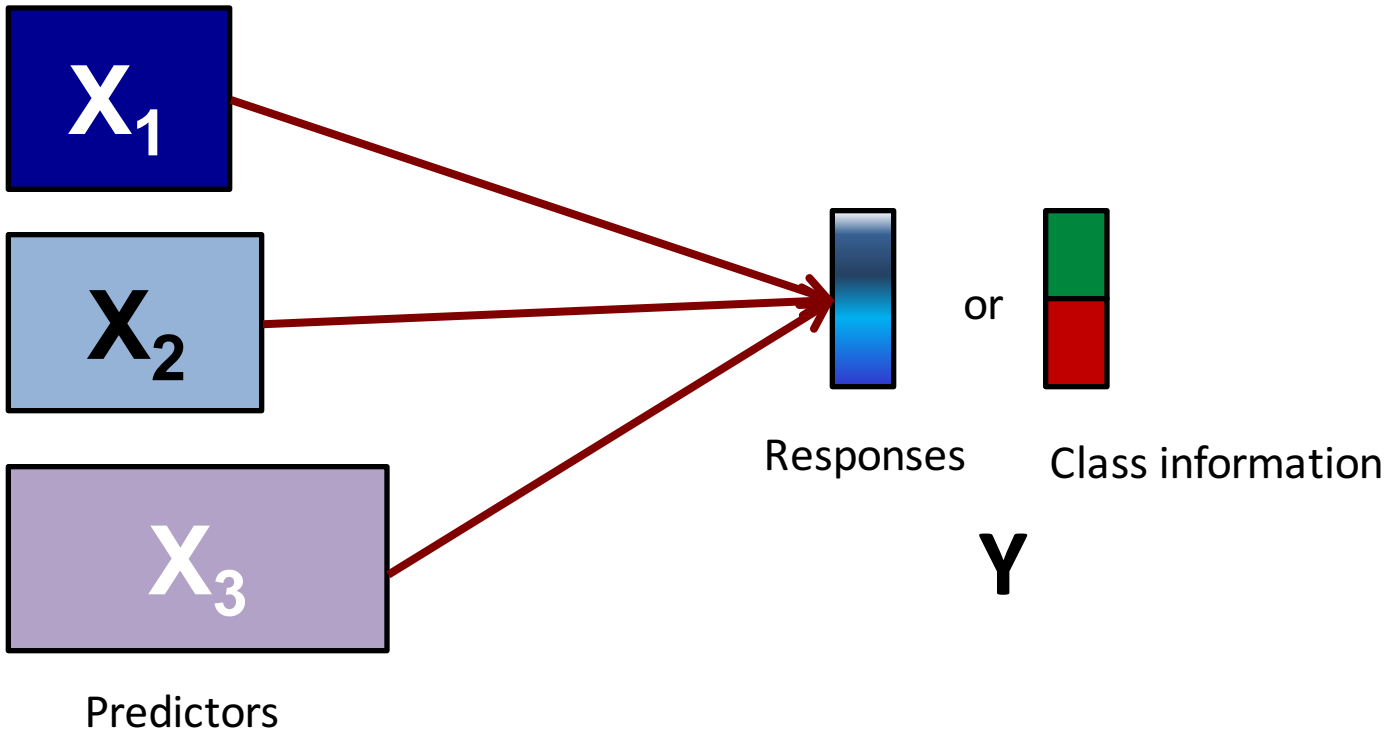
How to combine the different information coming from the various analytical platforms?

Multi-block data and models

- Blocking can occur naturally within the data:
  - Signals collected using different techniques
  - Directionality induced by the problem (dependent vs independent variables)
  - Sample groupings (categories)
- Ignoring the block structure may blur the final results

- Multiblock models:
  - Keep the natural ordering of the data
  - Explain relation between blocks
  - Describe variation within blocks
  - Assess block contribution to the overall variability

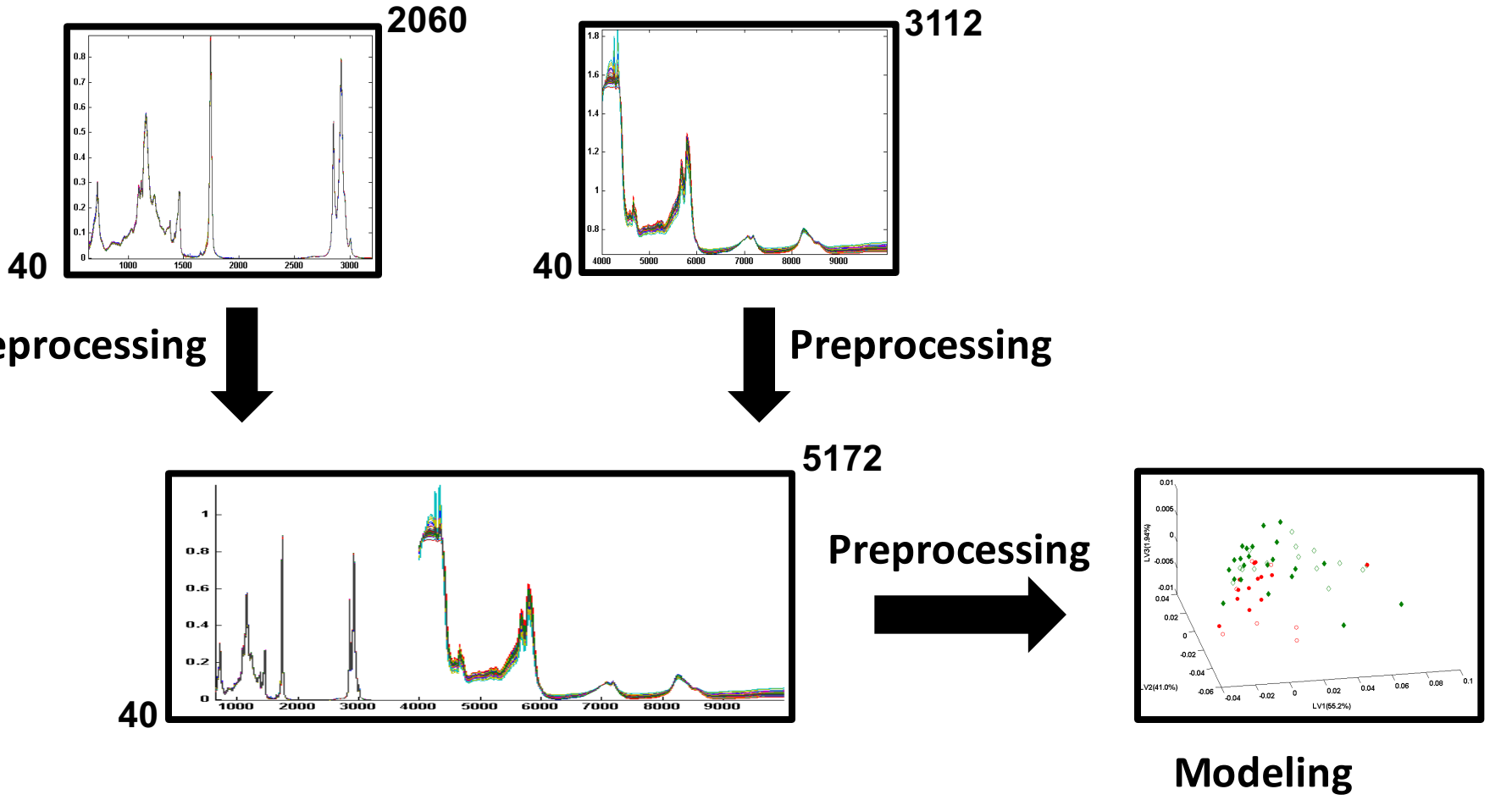
DATA FUSION STRATEGIES



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

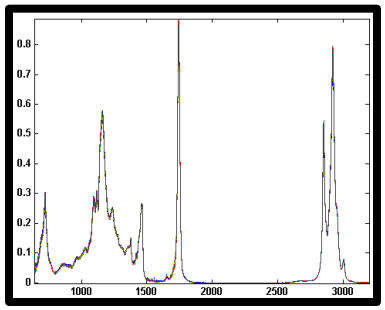
LOW LEVEL DATA FUSION

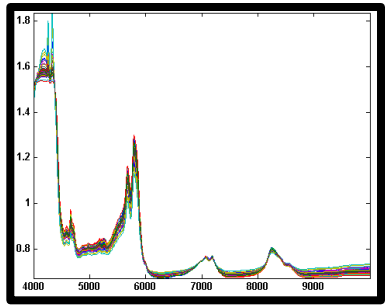
Data are concatenated and treated as they were a single fingerprint



MID LEVEL DATA FUSION

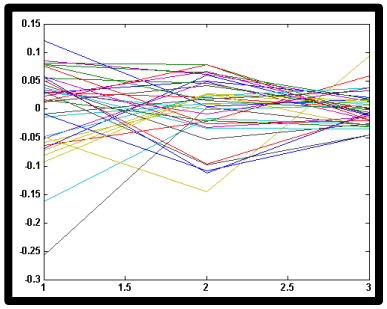
Features extracted from the data are concatenated

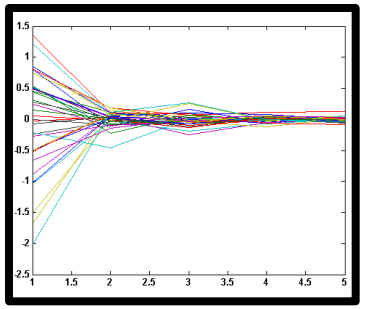




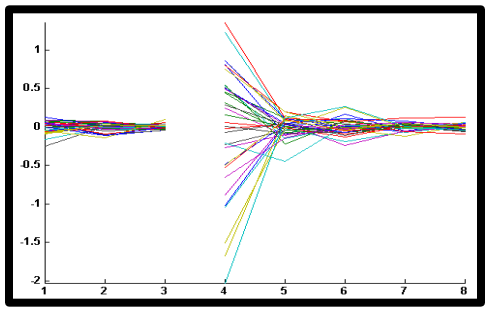
Preprocessing + Feature extraction

Preprocessing + Feature extraction



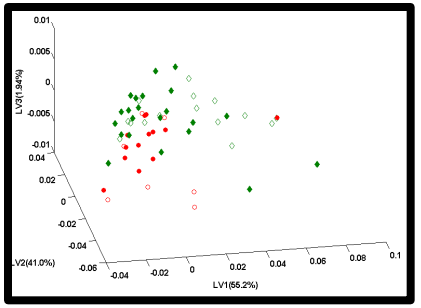


Concatenation



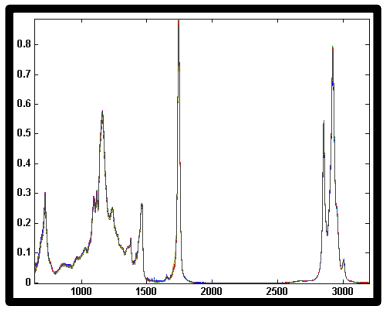
Preprocessing

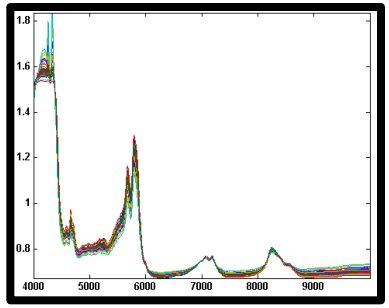
Modeling



HIGH LEVEL DATA FUSION

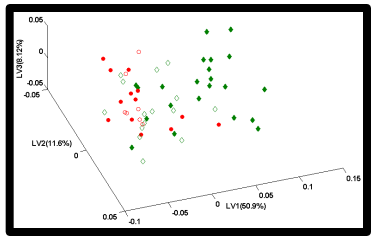
Fusion occurs at the decision level

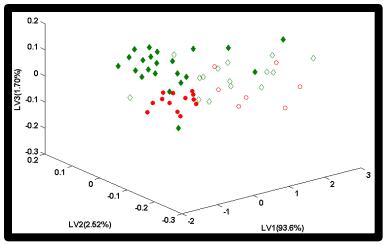




Preprocessing +
Feature extraction +
Modeling

Preprocessing +
Feature extraction +
Modeling





Decision 1 (e.g. Class A)

Decision 2 (e.g. Class B)

Majority vote
Bayes' theorem

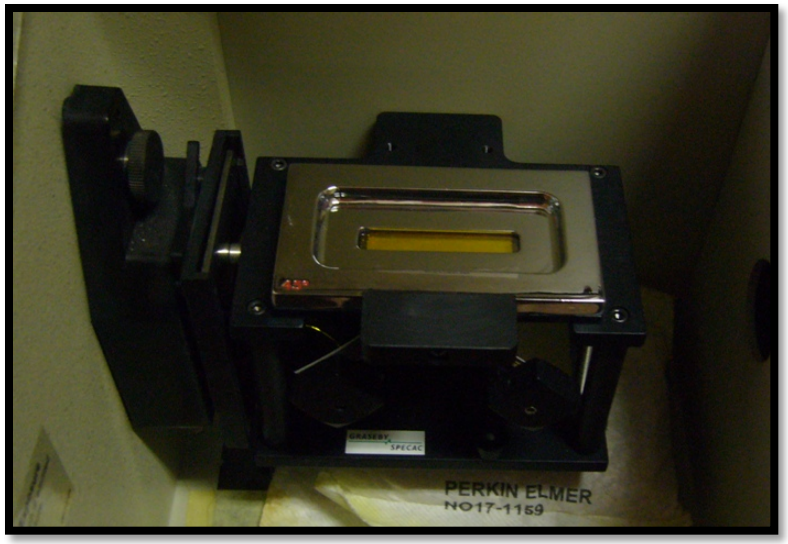
Final decision (e.g. Class A)

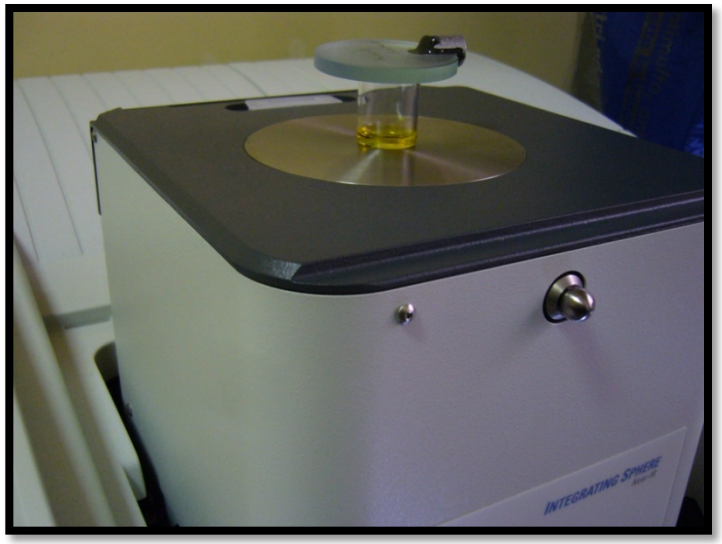
EXAMPLES:

I. Traceability of PDO oils with infrared spectroscopic techniques

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

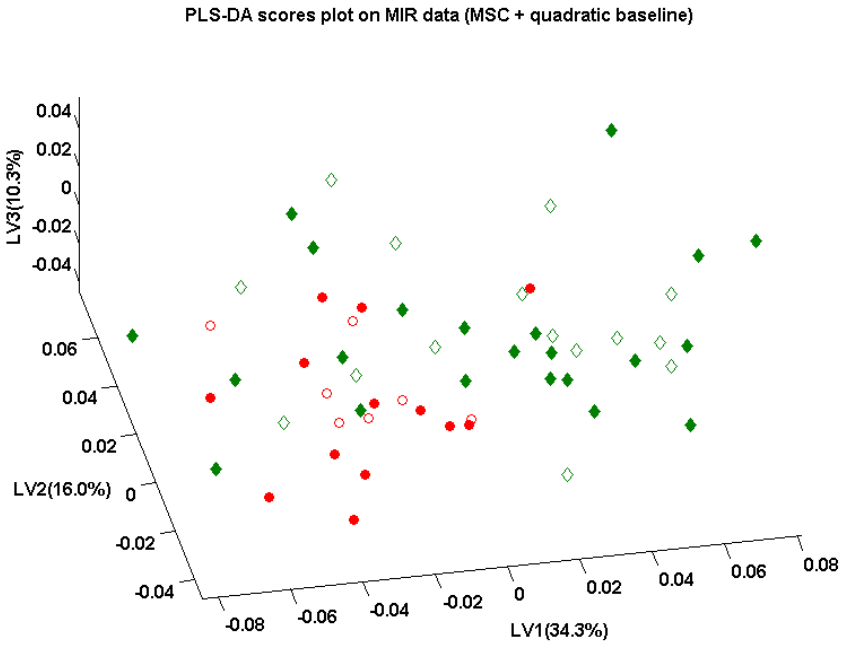


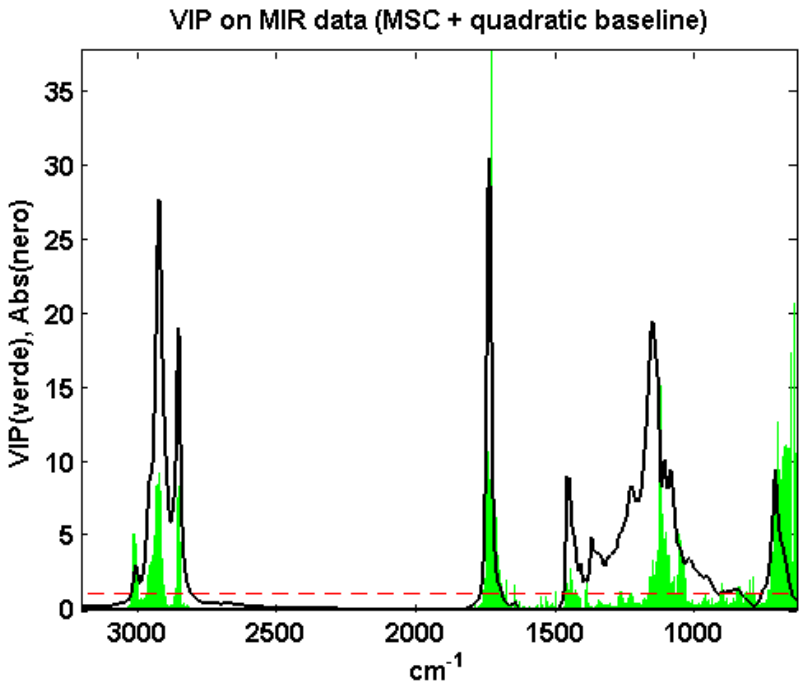


PLS-DA on MIR data

Pretreatment	LV	% Correct Classification Calibration		% Correct Classification Cross-validation	
		Sabina	Other origins	Sabina	Other origins
Linear baseline	6	100.0	100.0	92.3	86.4
Quadratic baseline	6	100.0	100.0	92.3	86.4
1st derivative (SG)	7	100.0	100.0	84.6	86.4
2nd derivative (SG)	3	84.6	86.4	84.6	72.7
MSC	3	100.0	95.5	84.6	95.5
MSC + quadratic baseline	4	100.0	95.5	92.3	95.5
MSC + 1st derivative	6	100.0	100.0	84.6	86.4
MSC + 2nd derivative	3	84.6	86.4	84.6	68.2

- Best results with MSC + quadratic bl.
- %cc on test set: 85.7% (sabina); 86.7% (other origins)

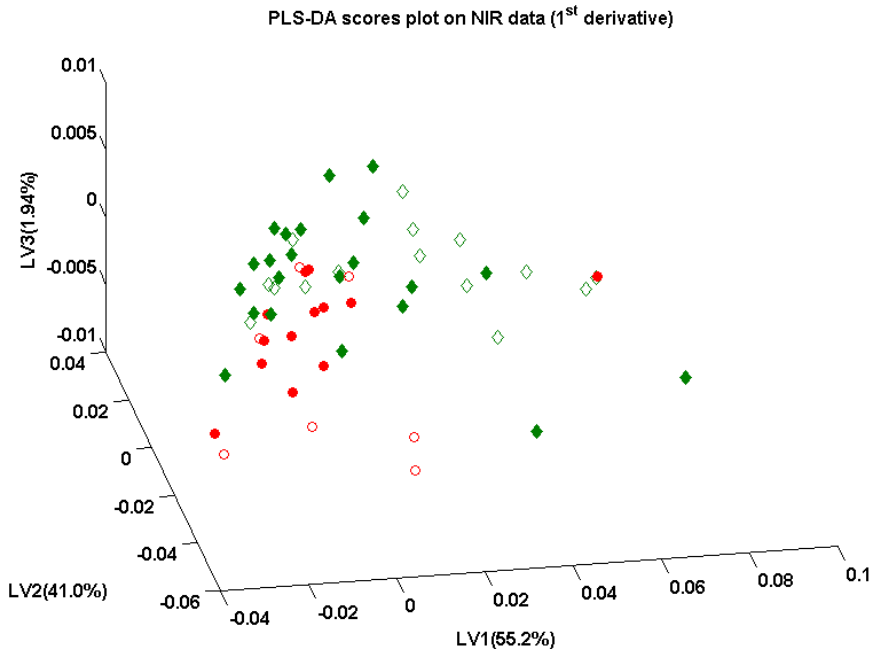


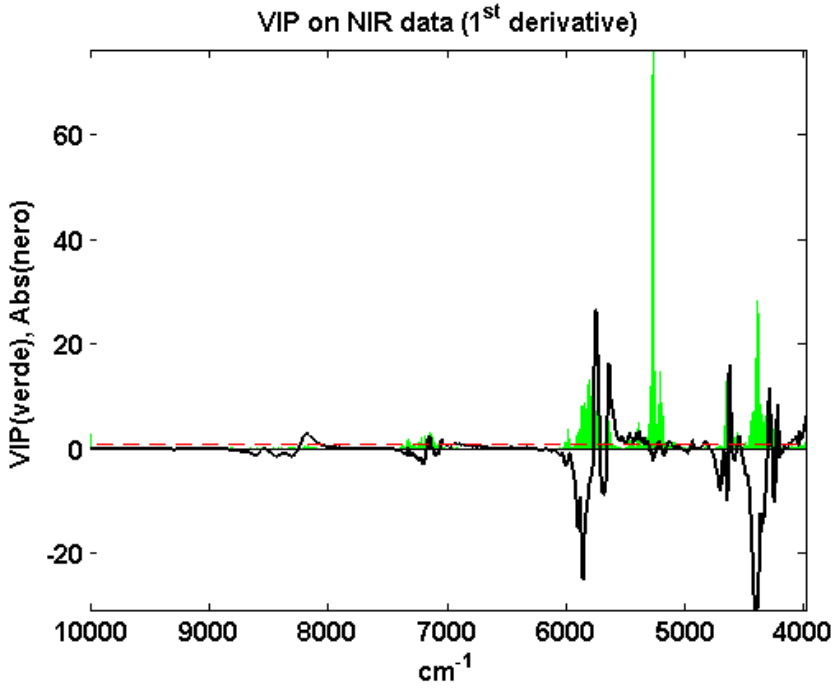


PLS-DA on NIR data

Pretreatment	LV	% Correct Classification Calibration		% Correct Classification Cross-validation	
		Sabina	Other origins	Sabina	Other origins
MSC	3	100.0	95.5	100.0	95.5
Detrending	4	100.0	95.5	100.0	95.5
1st derivative (SG)	5	100.0	95.5	100.0	95.5
2nd derivative (SG)	3	92.3	81.8	76.9	86.4
MSC + detrending	4	100.0	95.5	100.0	95.5
MSC + 1st derivative	4	92.3	95.5	92.3	90.9
MSC + 2nd derivative	4	84.6	90.9	84.6	86.4

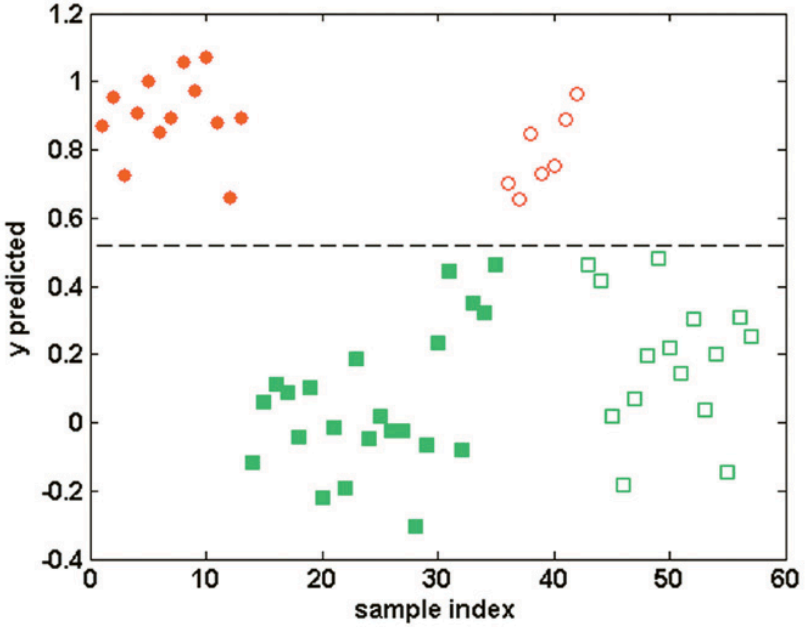
- Best results in CV with 4 pretreatments.
- %cc on test set (d1): 100% (sabina); 100% (other origins)
- %cc on test set (other 3): 100% (sabina); 93.3% (other origins)





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)



EXAMPLES:

II. Traceability of PDO oils with HPLC-DAD of polyphenols

Contents lists available at ScienceDirect

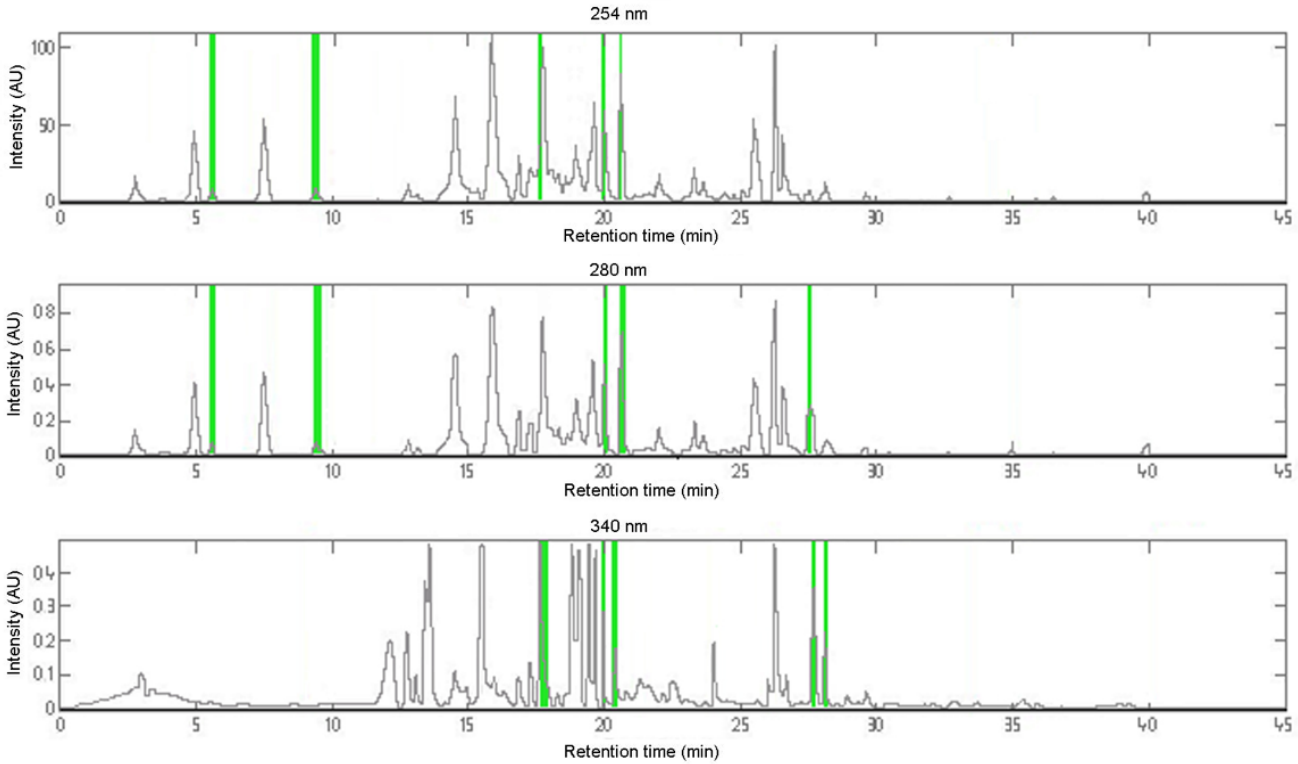
Chemometrics and Intelligent Laboratory Systems

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

Geographical traceability of extra virgin olive oils from Sabina PDO by chromatographic fingerprinting of the phenolic fraction coupled to chemometrics

Riccardo Nescatelli, Rossana Claudia Bonanni, Remo Bucci, Antonio L. Magrì, Andrea D. Magrì, Federico Marini *

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



Predictions

Wavelength	LVs	%Correct class.		%Correct class.		%Correct class.	
		Calibration		CV		Validation	
		Sabina	Others	Sabina	Others	Sabina	Others
254nm+280nm	4	92.3%	91.9%	91.2%	88.0%	85.7%	80.0%
254nm+340nm	1	92.3%	86.5%	88.8%	85.4%	85.7%	85.0%
280nm+340nm	3	100%	91.9%	91.2%	91.4%	85.7%	90.0%
254nm+280nm+340nm	2	100%	97.3%	87.7%	85.0%	85.7%	85.0%

Interpretation

Retention time	Compound	Ion mode	m/z	Fragments (Rel. abundance)	Identification[c]
5.4	vanillic acid	negative	167.1	108.0(100);151.8(10)	t_R & standard
9.1	p-coumaric acid	negative	163.1	119.1(100);167.1(27);91.1(13)	t_R & standard
18.2	luteolin	positive	287.2	287.2(100);153.2(77);135.2(24)	t_R & standard
19.9	pinoresinol	positive	359.1	359.1(100);327.1(10)	t_R & standard
21.0	acetoxypinoresinol	positive	417.4	417.4(100);358.4(10)	Literature
26.8	apigenin	negative	269.0	117.0(100);107.0(17);151.0(12)	t_R & standard
27.9	methoxyluteolin	negative	299.4	299.4(100);199.4(25);191.4(20)	Literature

EXAMPLES:

III. Traceability of an Italian craft beer Reale (from Birra del Borgo)

AUTHENTICATION OF BEER

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

BIRRA DEL BORGO

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

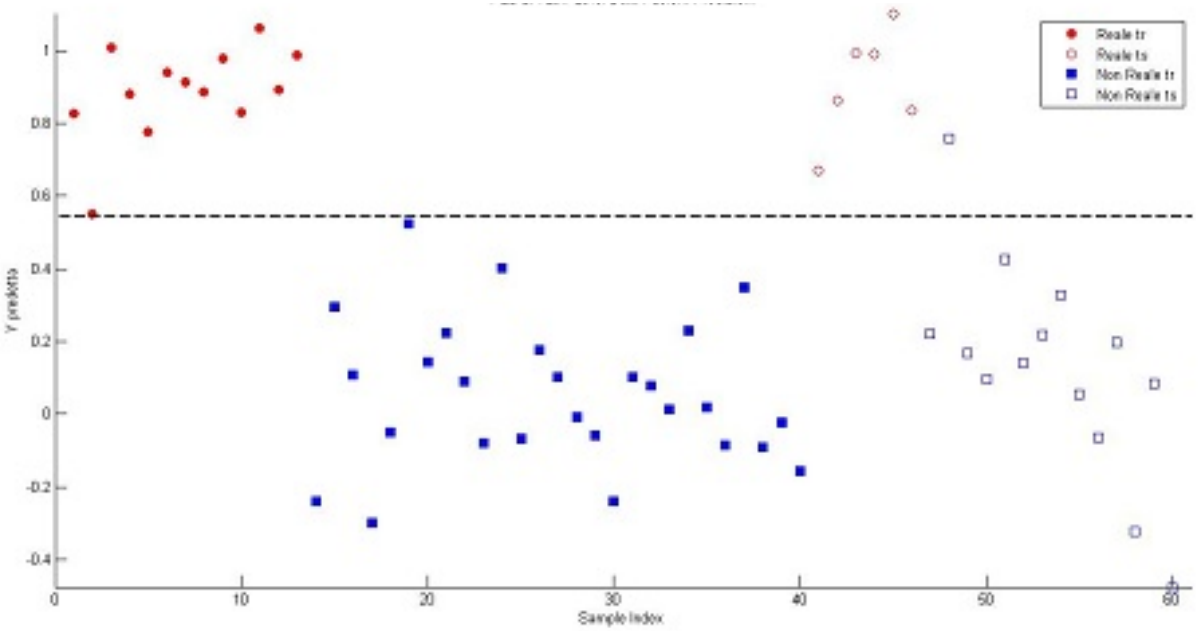
SAMPLES and TECHNIQUES

- A total of 60 samples were analyzed:
  - 19 Reale
  - 12 beers from Birra del Borgo
  - 29 beers from other breweries
- Samples were split into training and test sets using duplex algorithm:
  - 40 training (13 Reale and 27 not Reale)
  - 20 test (6 Reale and 14 not Reale)

- The following fingerprints were recorded:
  - TG
  - UV
  - Vis
  - NIR
  - MIR

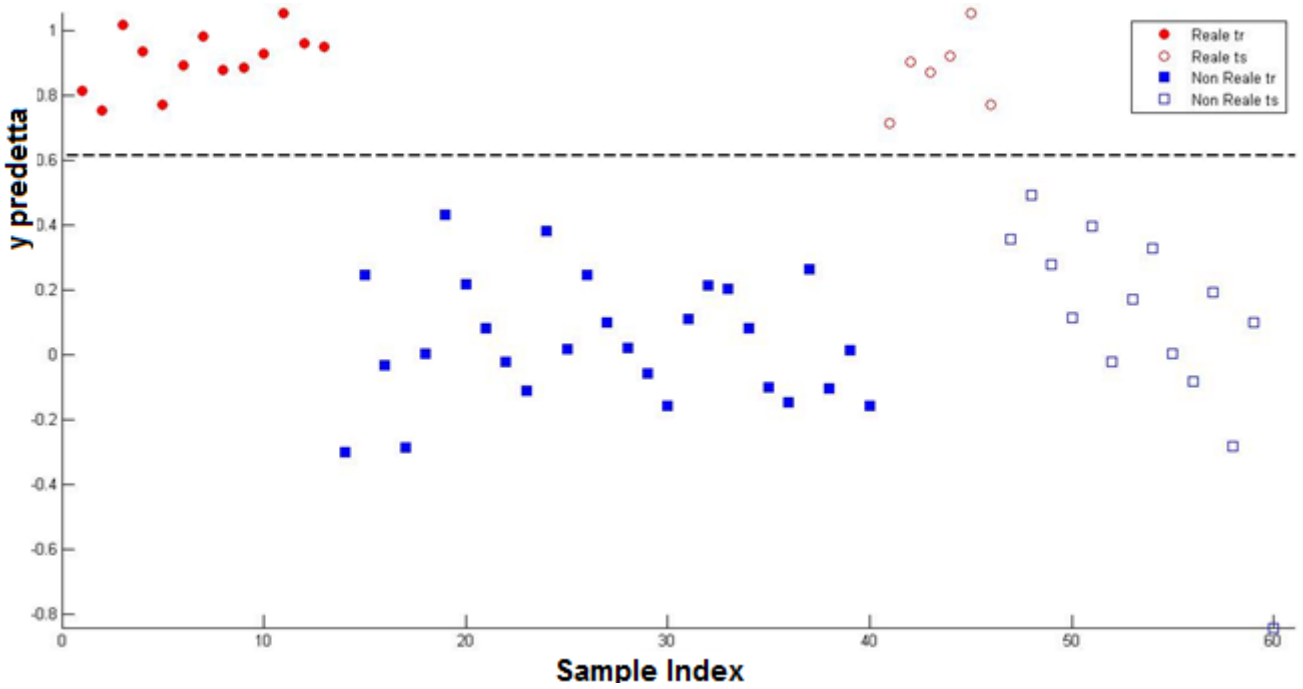
Low Level – Results

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

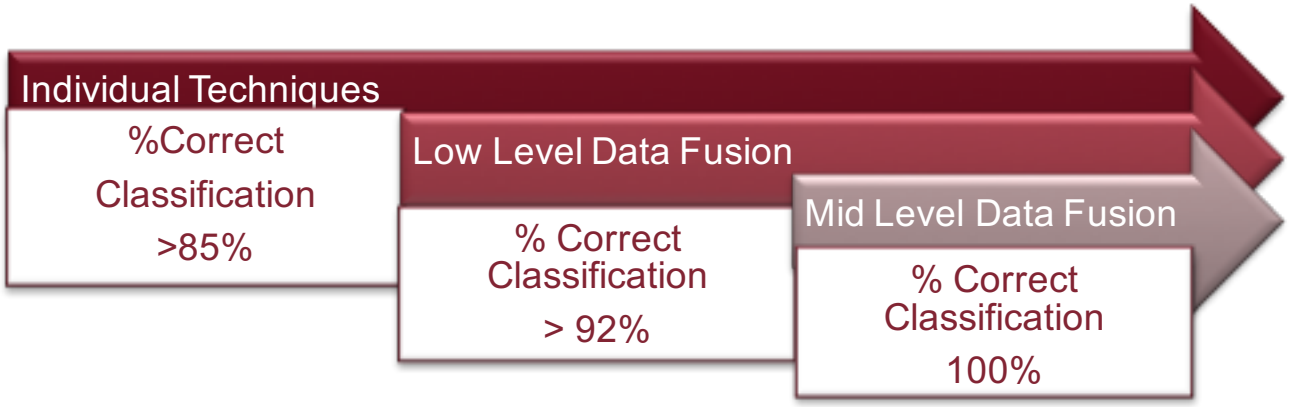


Mid Level – Results

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



Conclusions



Data fusion helps!

Thank you for your attention

federico.marini@uniroma1.it