Extension of SIMCA classification method to higher order array: N-SIMCA

Marina Cocchi*, C. Durante, A. Marchetti, R. Bro[1], N. Gallagher[2]

[1]KVL, Frederiksberg, DK [2]Eigenvector Inc., Manson WA, USA

UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

*Department of Chemistry
University of Modena and Reggio Emilia
Via Campi 183, Modena I-41100
MO_RE Chemometrics Research Group
cocchi.marina@unimore.it

Outline

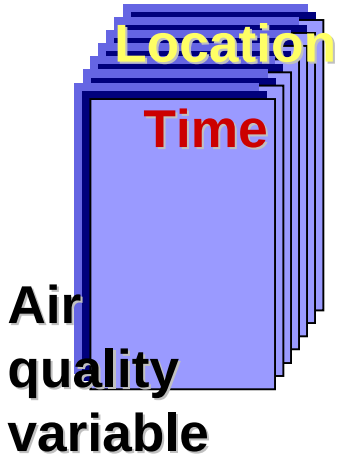
- context
- overview of 2-way SIMCA
- SIMCA extension to 3-way
- Case studies
- future perspective

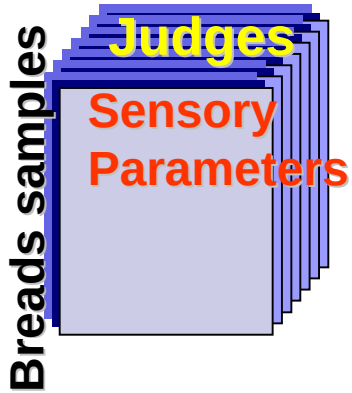
Why multi-way SIMCA

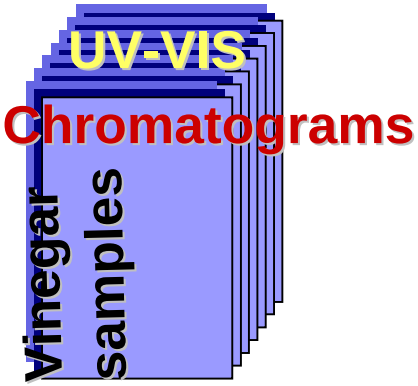
- Data characterized by more that two sources of variability occur in many different research fields

- in both explorative analysis and calibration/regression context the use of multi way data analysis tools on multi way data lead to highly improved models/results

- a true multi way classification tool is still missing







Outline

- context
- **overview of 2-way SIMCA**
- SIMCA extension to 3-way
- Case studies
- future perspective

1th Class

2th Class

3rd Class

original SIMCA

distance from a class, i.e q, of a test object, i.e. p :

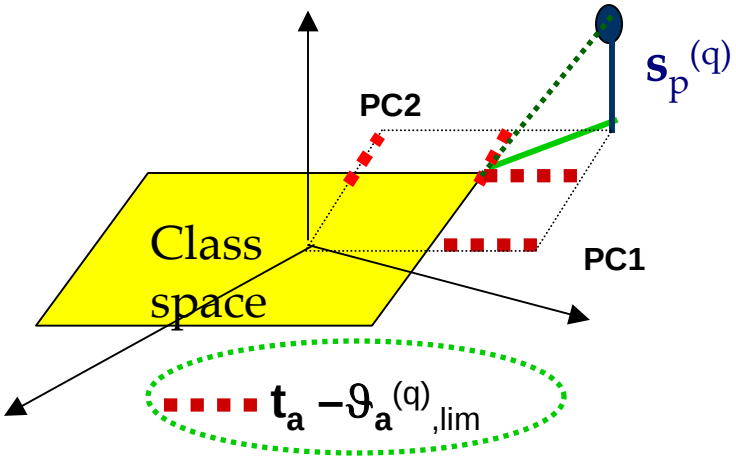
$$\mathbf{d}_p^{(q)} = \mathrm{sqrt}[\mathbf{s}_p^{(q)2} + \Sigma_a \Phi_a^2 (\mathbf{t_a} - \vartheta_a^{(q)}{}_{,lim})^2]$$

$$\mathbf{s}_p^{(q)} = \mathrm{sqrt}(\Sigma_k e_{pk}^2/(M\text{-}A))$$

total RSD of a class, i.e. q:

$$\mathbf{s}_0^{(q)} = \mathrm{sqrt}(\Sigma_{ik} e_{ik}^2/(N\text{-}A\text{-}1)(M\text{-}A))$$

i=1:M M=number of variables; k= 1:N objects belonging to class *q;*



Classification rule

$$\mathbf{F} = \mathbf{d}_p^{(q)2} / \mathbf{s}_0^{(q)2} < \mathbf{F}_{crit}\,(M\text{-}A),(N\text{-}A\text{-}1)(M\text{-}A)$$

If true for both *q* and *r* Unique assignment only if

$$\mathbf{F} = \mathbf{d}_p^{(q)2} / \mathbf{d}_p^{(r)2} > \mathbf{F}_{crit}\,(M\text{-}A_r),(M\text{-}A_q)$$

S.Wold, *Pattern Recognition,* 1976 (8), 127; S. Wold, M. Sjostrom, in ACS Symposium Series, 1977 (52), 243-281

1th Class

2th Class

3rd Class

- **a somehow alternative approach of** original SIMCA
  (coming from MSPC realm)

Distance from a class model:

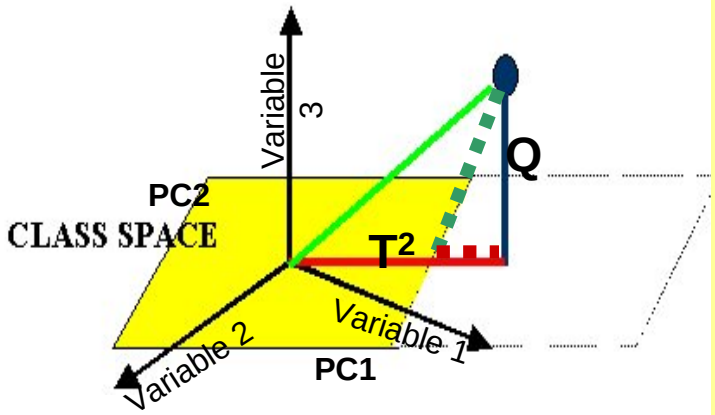
Q is the sum of squared residuals: $\Sigma_k e_{pk}^2$ (same role as $\mathbf{s}_p^{(q)2}$)

Q_{lim} is calculated by assuming a χ^2 distribution (approximation of Jackson and Mudholkar)

Distance in scores space (inside class model) **:**

use Hotelling's T-square ($\mathbf{T^2}$)

calculates $\mathbf{T^2_{lim}}$ as $[A*(N-1)/N*(N-A)]*F_{crit}\ A, (N-A)$



Classification rule

Assign an object to a class if its reduced combined distance satisfies :

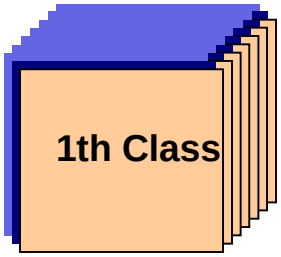
PLS-Toolbox SIMCA

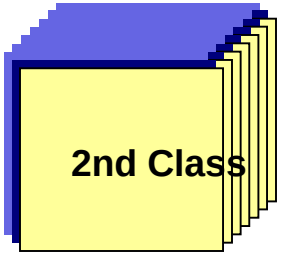
$$\sqrt{\left(Q/Q_{\lim ite}\right)^2 + \left(T^2/T^2_{\lim}\right)^2} < \sqrt{2}$$

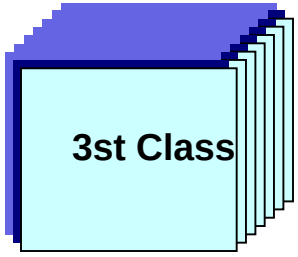
Outline

- context
- overview of 2-way SIMCA
- **SIMCA extension to 3-way**
- Case studies
- future perspective

- **Extension to 3-way**







PARAFAC

$$\hat{X} = A(C \odot B)' + E$$

TUCKER3

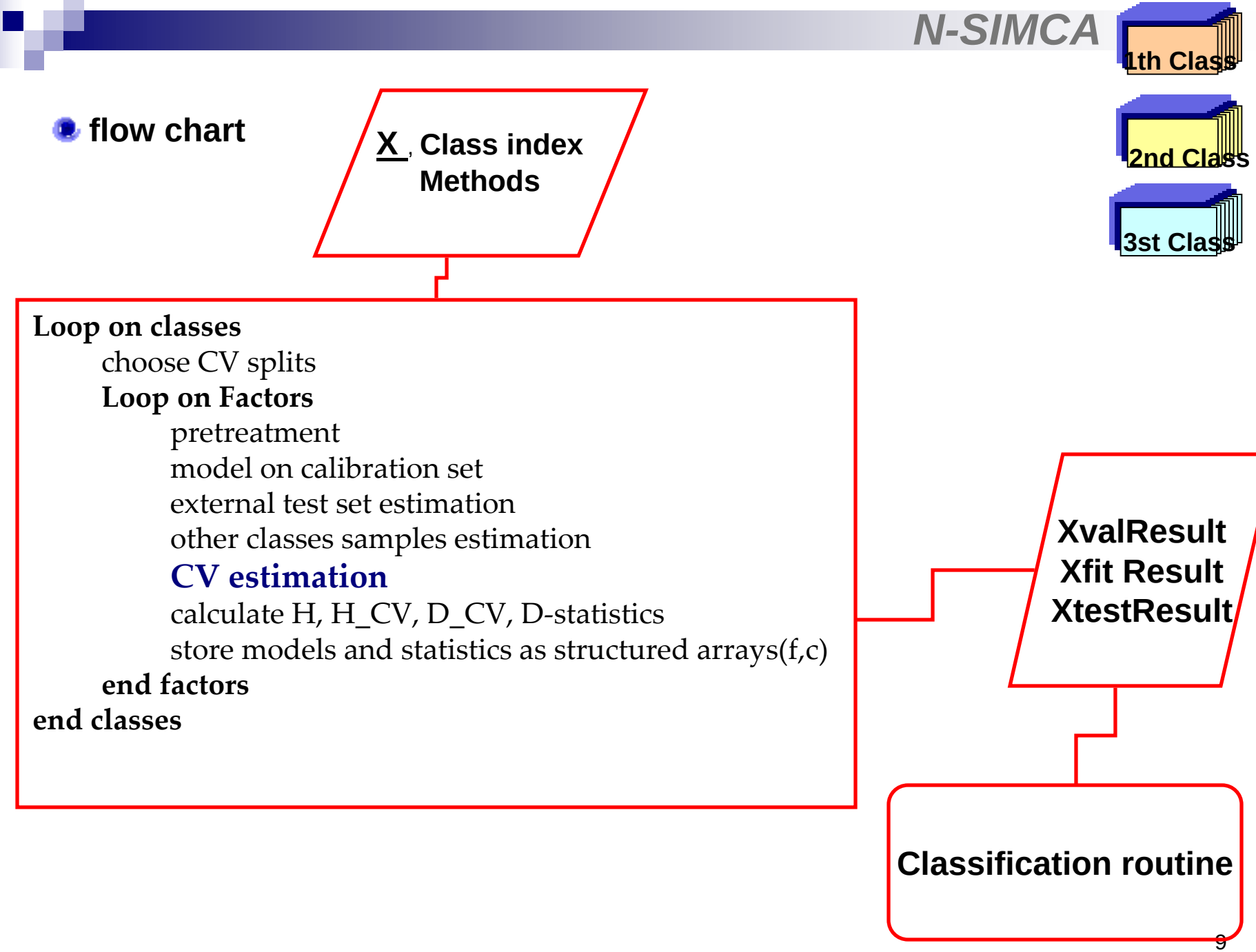
$$\hat{X} = AG_X(C \otimes B)' + E$$

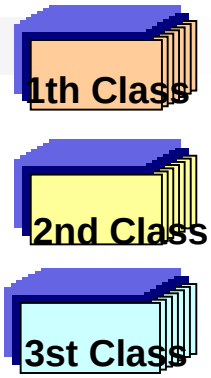
X_1 Model

X_2 Model

X_3 Model

- flow chart


1th Class
2nd Class
3st Class
X , Class index
Methods
Loop on classes
choose CV splits
Loop on Factors
pretreatment
model on calibration set
external test set estimation
other classes samples estimation
CV estimation
calculate H, H_CV, D_CV, D-statistics
store models and statistics as structured arrays(f,c)
end factors
end classes
XvalResult
Xfit Result
XtestResult
Classification routine



classification rules

At this stage we adopt the “alternative” SIMCA approach and try different rules:

Distance from Model:

Q-value

$$q_i = \sum_{jk=1}^{JK} (e_{ijk})^2 \qquad \textit{fit}$$

$$q_{new} = \sum_{jk=1}^{JK} (e^*_{ijk})^2 \qquad \textit{CV, test}$$

I: number of objects of the calibration set

J,K: number of mode 2 and 3 variables

e_{ijk} : Residual matrix’s elements

e^*_{ijk}: residual for a left-out object or a predicted object

scores Distance (inside model space):

H-value (Leverage)

$\mathbf{H}_{fit} = \text{diag} [\mathbf{A}_{fit}(\mathbf{A}_{fit}\text{'}\mathbf{A}_{fit})^{-1}\mathbf{A}_{fit}\text{'}]$

$\mathbf{H}_{CV} = \text{diag} [\mathbf{A}_{CV}(\mathbf{A}_{fit}\text{'}\mathbf{A}_{fit})^{-1}\mathbf{A}_{CV}\text{'}]$

$\mathbf{H}_{new} = \text{diag} [\mathbf{A}_{new}(\mathbf{A}_{fit}\text{'}\mathbf{A}_{fit})^{-1}\mathbf{A}_{new}\text{'}]$

D-statistic

$\mathbf{D}_{fit} = \text{diag} [\mathbf{A}^T_{fit}\,\mathbf{S}_{fit}^{-1}\,\mathbf{A}_{fit}]$

$\mathbf{D}_{CV} = \text{diag} [\mathbf{A}^T_{CV}\,\mathbf{S}_{fit}^{-1}\,\mathbf{A}_{CV}]$

$\mathbf{D}_{new} = \text{diag} [\mathbf{A}^T_{new}\,\mathbf{S}_{fit}^{-1}\,\mathbf{A}_{new}]$

S: mode 1 scores Covariance matrix;

A: mode 1 Score matrix;

O-limit :

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

where $h_0 = 1 - ((2\theta_1\theta_3)/(3\theta_2^2))$, $\theta_1 = tr(\mathbf{V})$, $\theta_2 = tr(\mathbf{V}^2)$ and $\theta_3 = tr(\mathbf{V}^3)$, $\mathbf{V}$ is the covariance matrix of $\mathbf{E}$, and z_α is the standardized normal variable with $(1-\alpha)$ confidence limit, having the same sign as h_0.

$\mathbf{Q_{lim\_fit}}$: **E** residual matrix

$\mathbf{Q_{lim\_CV}}$: **E** cross-validated residuals

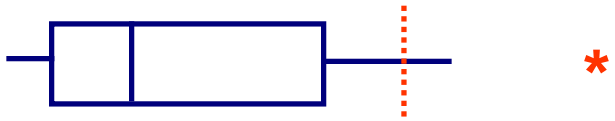
D-limit[1] :

$\mathbf{D_{lim\_fit}}$:

$$\sim \frac{R(I^2-1)}{I(I-R)} F(R, I-R, \alpha)$$

R = 1st mode factors, I = samples

$\mathbf{D_{lim\_CV}}$ → **95** percentile of D_{CV}



Classification rules

$$\sqrt{\left(\frac{Q_X}{Q_{X\_\lim}}\right)^2 + \left(\frac{H_X \text{ or } D_X}{H_{X\_\lim} \text{ or } D_{X\_\lim}}\right)^2} \langle = \sqrt{2}$$

X = fit or CV

H-limit :

$\mathbf{H_{lim\_fit}}$ **= 1**

$\mathbf{H_{lim\_CV}}$ **= 95** percentile of $\mathbf{H_{cv}}$

[1]*Generalized contribution plots in multivariate statistical process monitoring"* J.A. Westerhuis, S.P. Gurden and A.K. Smilde, *Chemolab. 51 (2000) 95-114*.

Outline

- context
- overview of 2-way SIMCA
- SIMCA extension to 3-way
- **Case studies**
- future perspective

Classification of dry-cured PARMA ham of different ageing through Fluorescence analysis* and Multi-way SIMCA methods

* data from J.M. Moller et al. *J. Agr. Food. Chem.* **2003** (51), 1224

Data Sets

Jens K.S. Møller et al. evaluated surface autofluorescence spectroscopy in order to measure age-related quality index of Parma ham during processing.

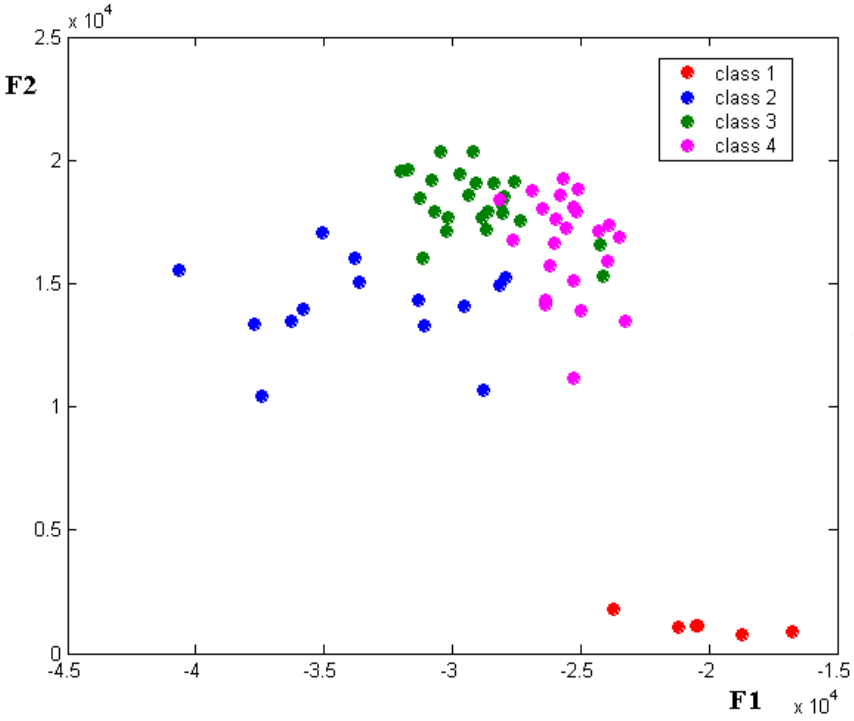
- **67 Parma ham samples:**

According to literature, the data array was reduced:

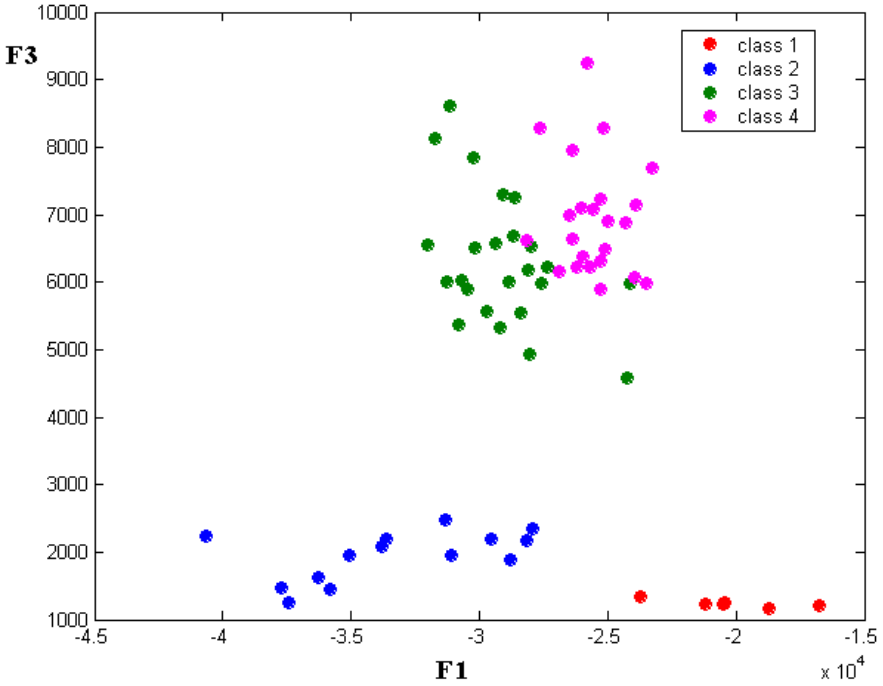
67×13×11 (samples × emission × excitation).

Samples category	Characteristics	N° sample	Train/test
Raw meat	fresh meat, just prepared 0 months	6	4 / 2
Salted	3 months ageing	14	9 / 5
matured	11-12 months ageing	24	17 / 7
Aged	15-18 months ageing	23	16 / 7

exploratory analysis:



3 Factors PARAFAC model

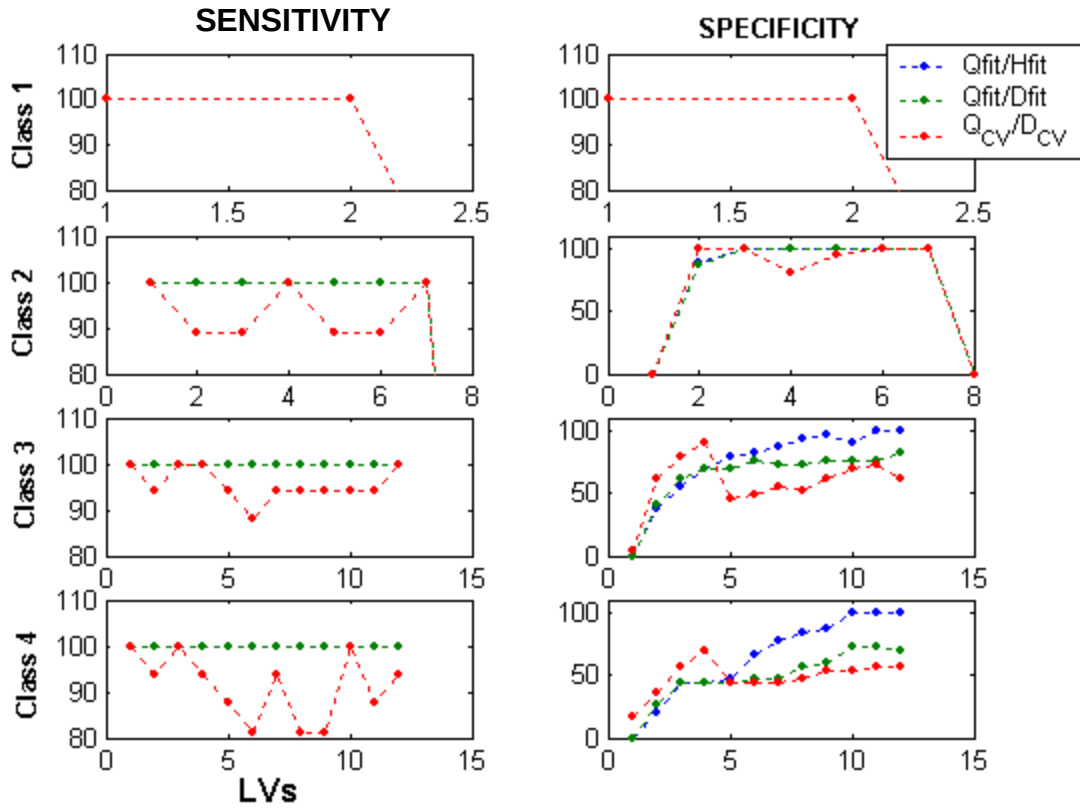


raw meat (red) is well distinguishable

The last two classes overlaps

SIMCA 3-way analysis:

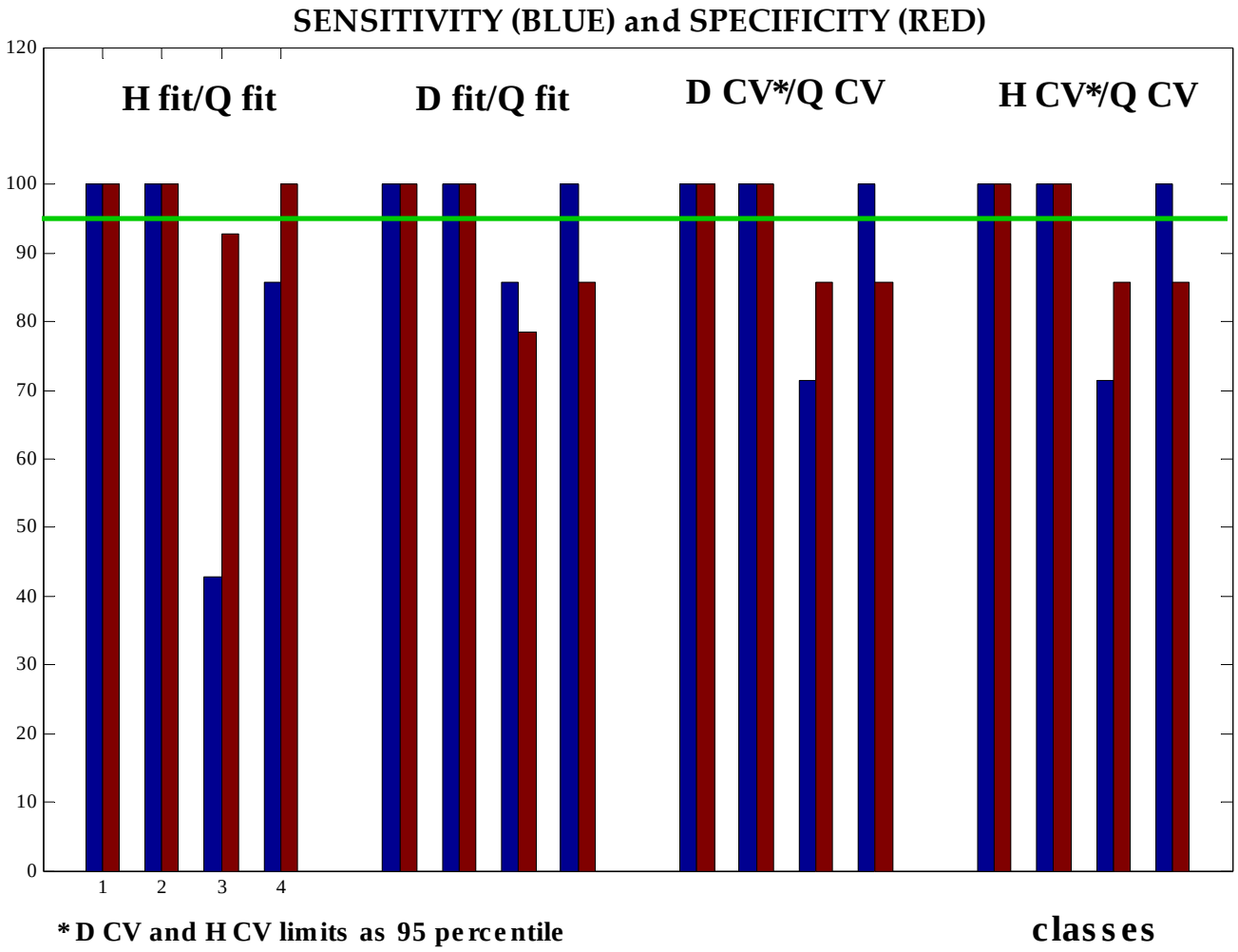
- **in general CV criteria lead to more parsimonious models in the case of overlapping classes**



LV's corresponding to Best compromise of Sensitivity and Specificity

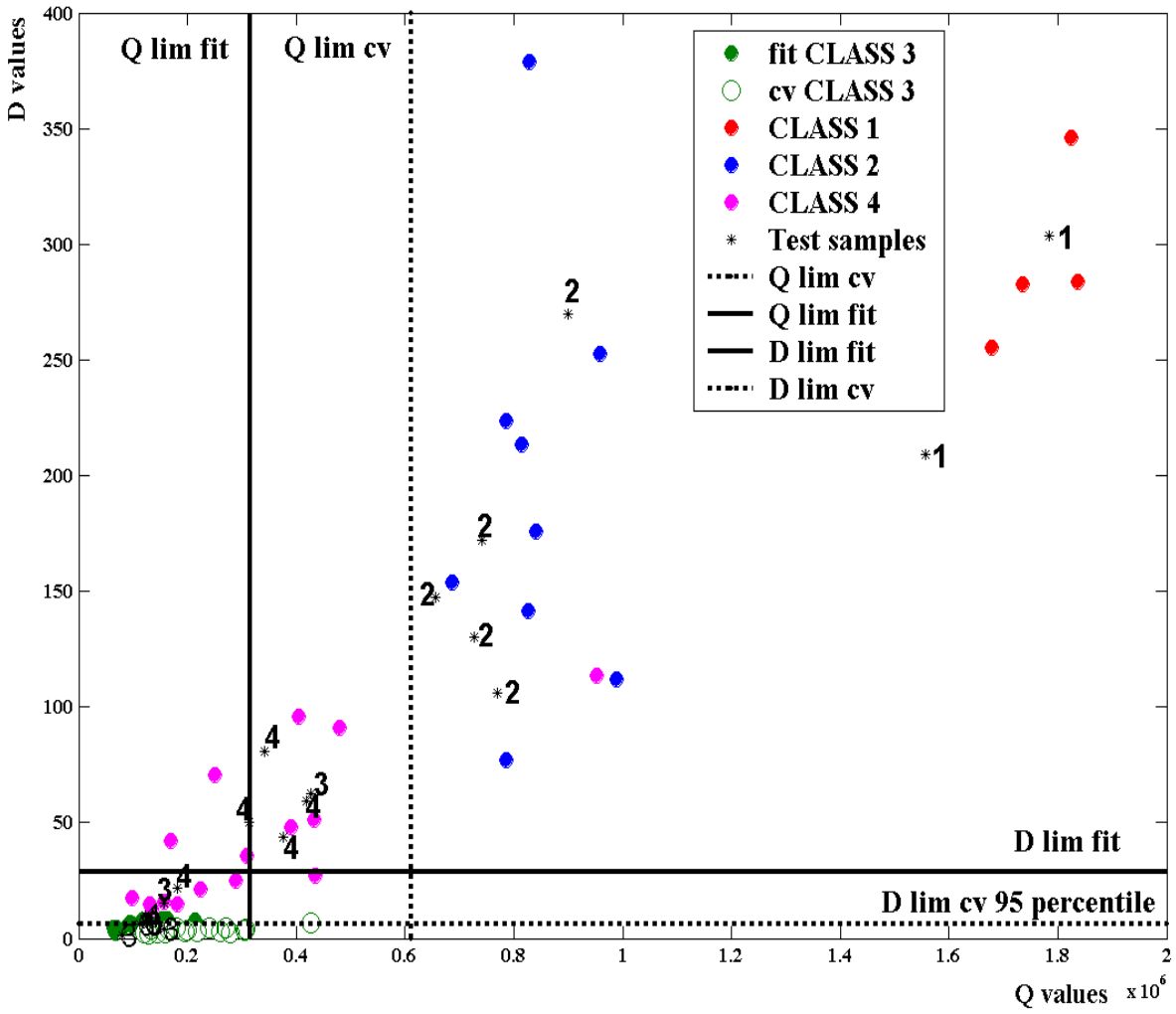
	hTfit/Qfit	Dfit/Qfit	Dcv/Qcv	hTcv/Qcv
class 1	1	1	1	1
class 2	3	3	2	2
class 3	9	6	4	4
class 4	10	10	4	4

Classification of test set



Model of 3rd class

- Q_lim_{CV} seems too wide
- D_lim_{fit} results too high
- empirical D_lim_{CV} seems ok



summary of SIMCA 3D results:

	H_fit/Q_fit	D_fit/Q_fit	D_CV/Q_CV	H_CV/Q_CV
SENS	100 (100)	100 (100)	100 (100)	100 (100)
	100 (100)	100 (100)	89 (100)	89 (100)
	100 (43)	100 (86)	100 (71)	100 (71)
	100 (86)	100 (100)	94 (100)	94 (100)
SPEC	100 (100)	100 (100)	100 (100)	100 (100)
	100 (100)	100 (100)	100 (100)	100 (100)
	97 (93)	76 (79)	90 (86)	90 (86)
	100 (100)	73 (86)	70 (86)	70 (86)

(test set)

- **overall, classification estimated by using empirical 95 percentile limit seems to have best performance**
- **H_fit/Q_fit gives the best results as far as specificity is concerned**
- **higher sensitivities for both the 3rd and 4th classes are obtained in CV**

Classification and characterization of LIGURIAN Extra-Virgin Olive Oil (EVOO) through HS-SPME/GC analysis and Multi-way SIMCA methods

Original, not yet published data

- Evaluation of a new multi-way method (N-SIMCA) for classification purposes

- Classification and differentiation of the LIGURIAN products, more important for economic reasons, from the other Italian and Foreign olive oils.

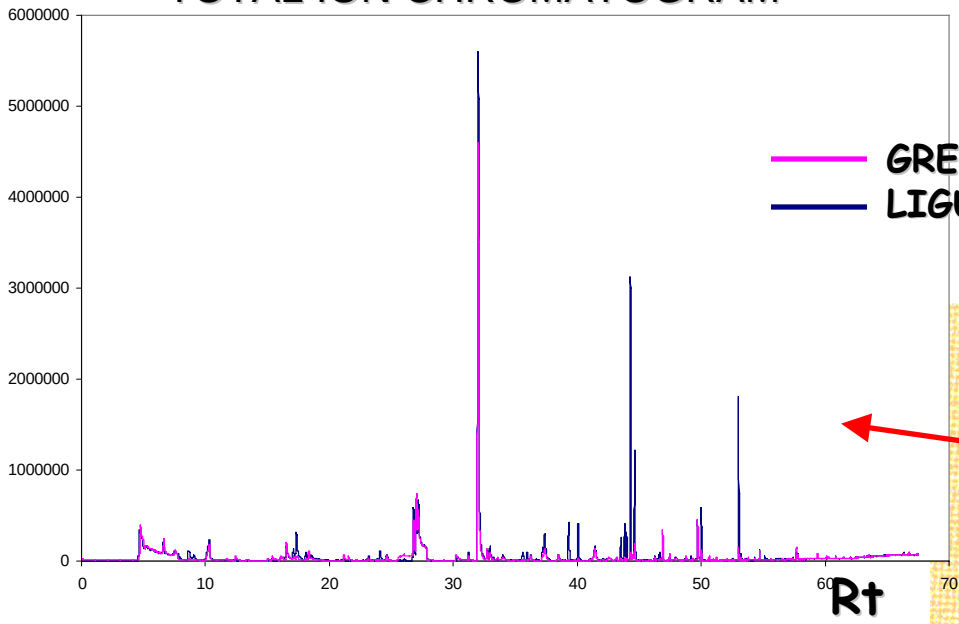
NB. WORK IN PROGRESS

Extra-virgin Olive Oils samples* :

ORIGINE	Characteristics	N° sample
LIGURIA	All samples belong to the same PDO but comes from two different areas both in the Liguria region	24
APULIA	Mixture of different variety.	25
FOREIGN (Greece, Spain, Tunisia)	Different zones	31

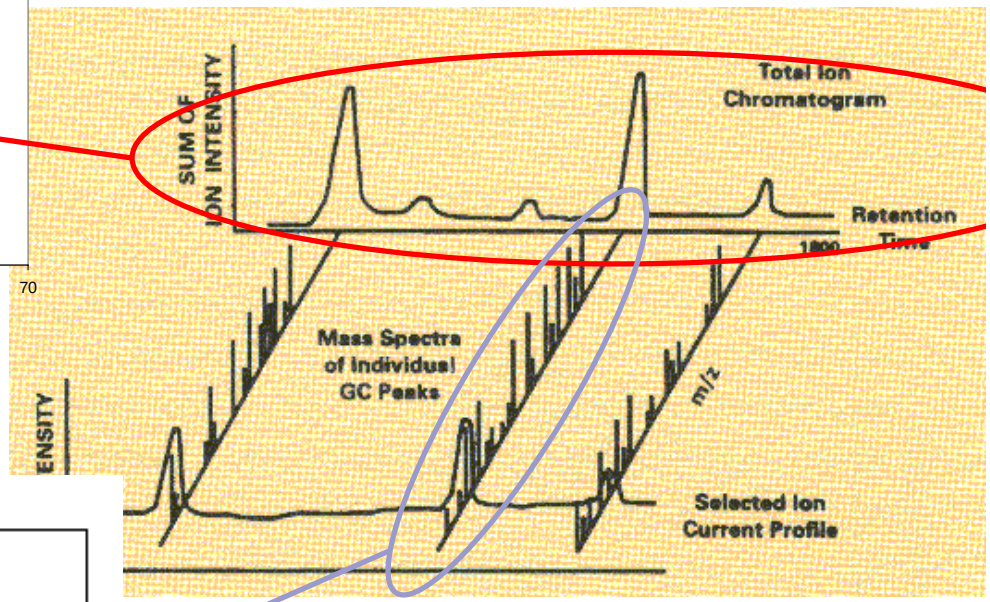
* Samples collected in collaboration with Food Chemistry group University of Genova

TOTAL ION CHROMATOGRAM

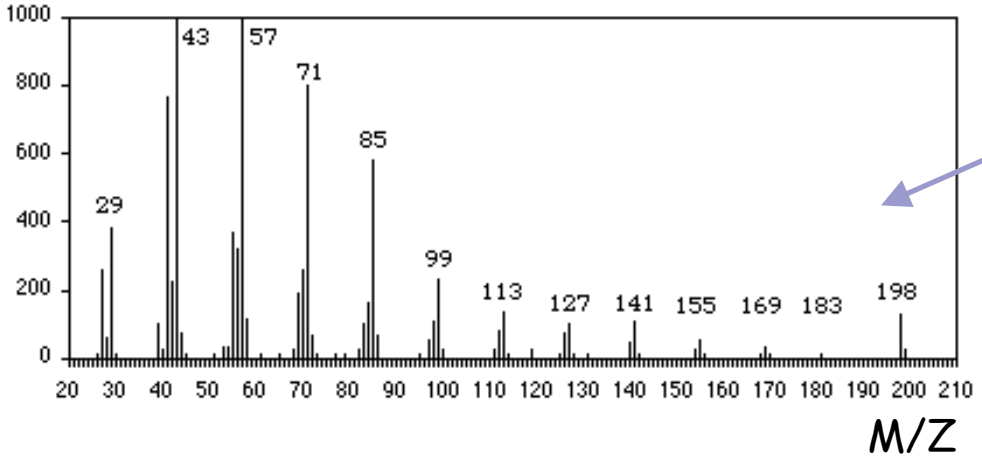


HS-SPME/GC-MS Signals

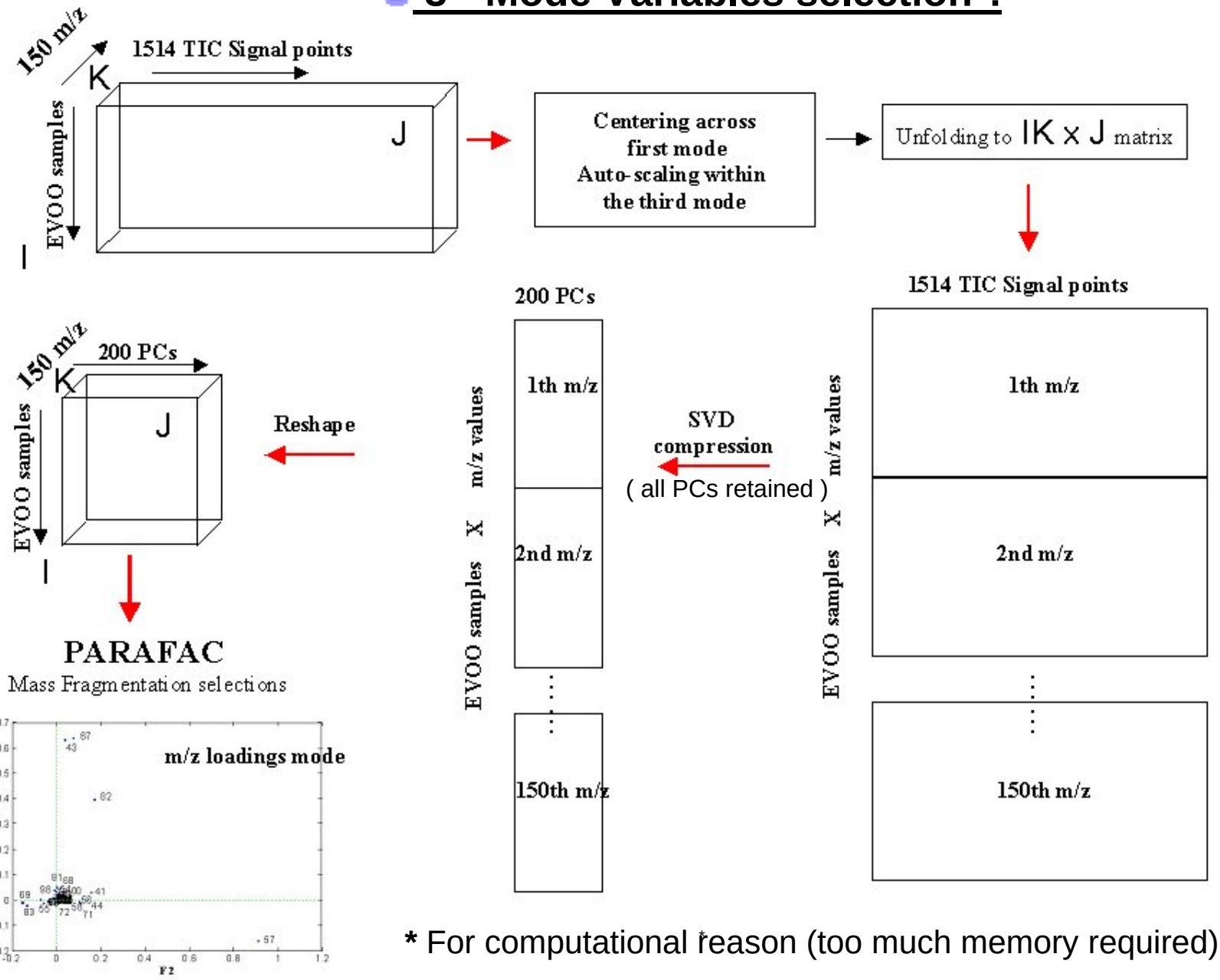
Fiber: DVB-CARBOXEN-PDMS



MASS SPECTRUM

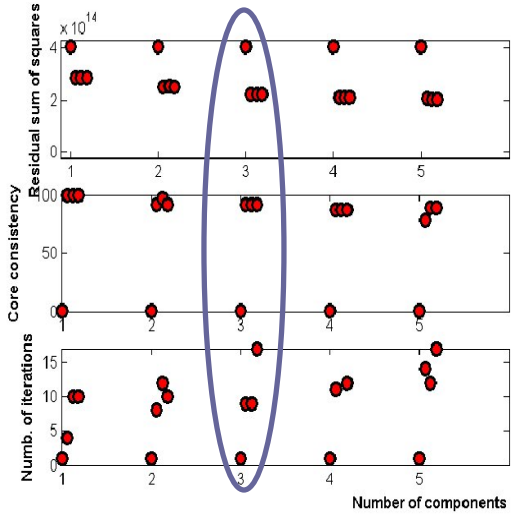


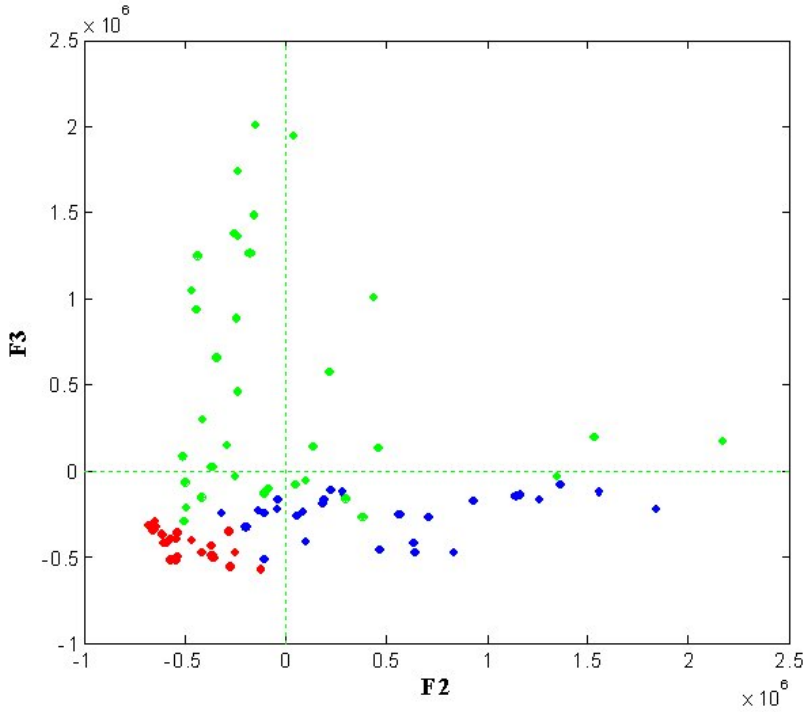
3rd Mode Variables selection*:

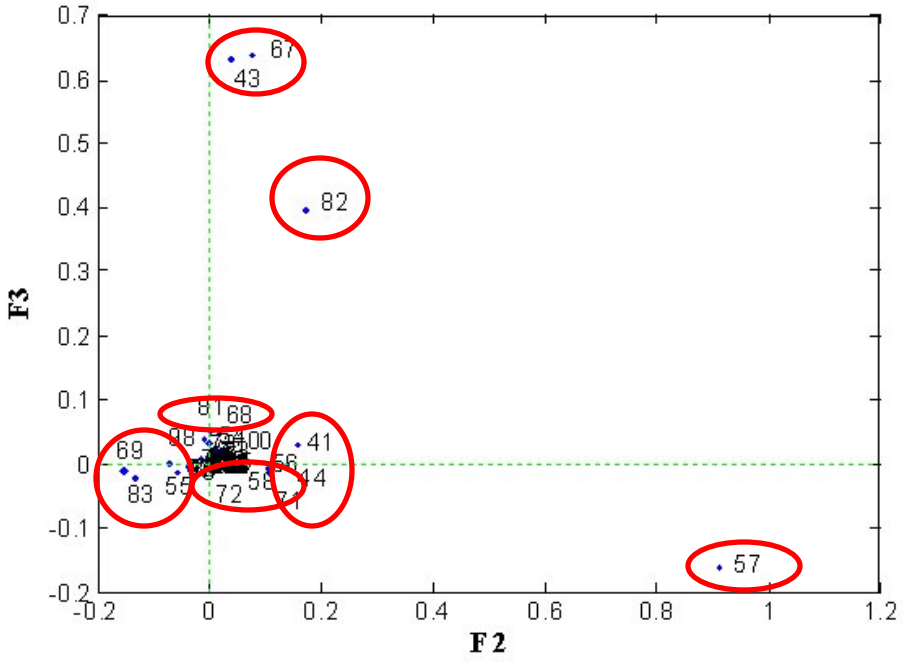


* For computational reason (too much memory required)

Variables selection:





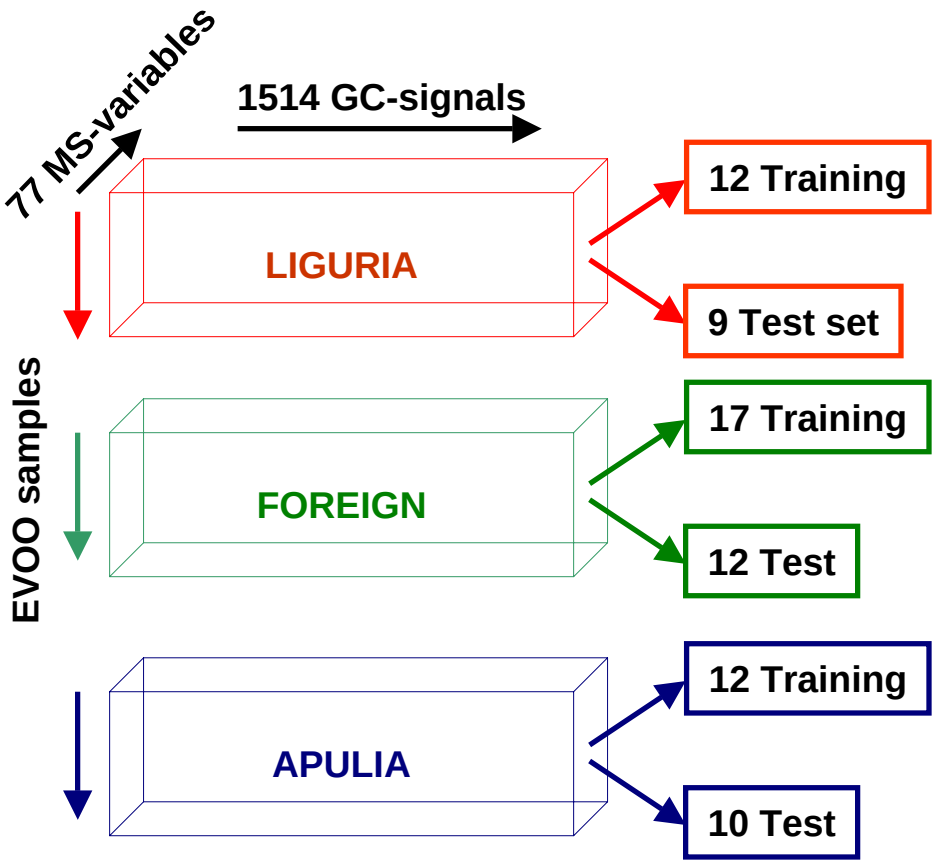


Ligurian olive oils class

Puglia olive oils class

Foreign olive oils class

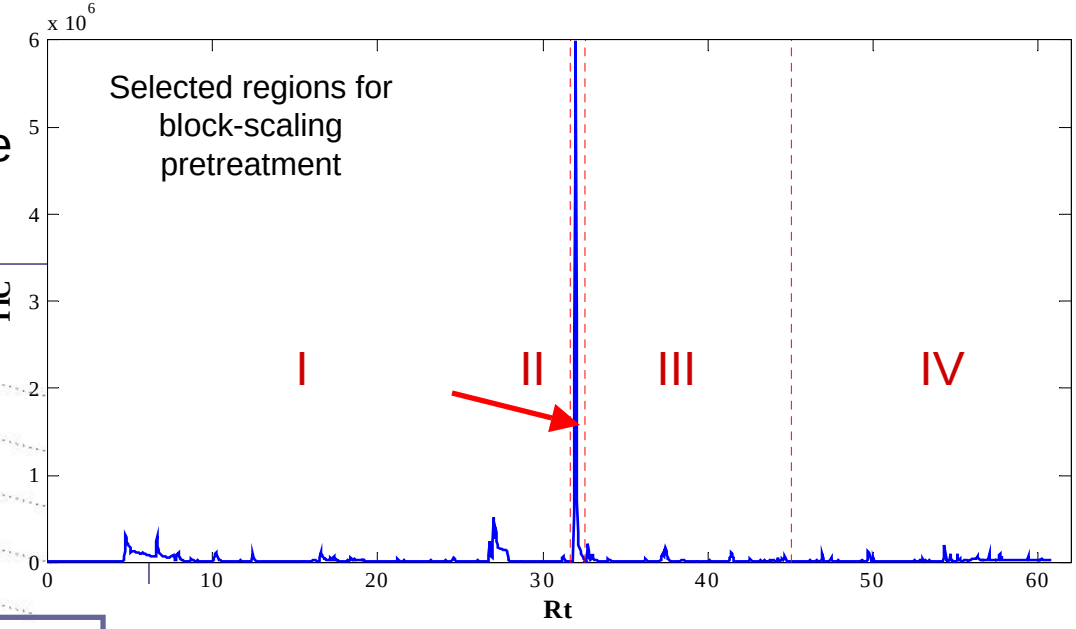
77 Selected variables

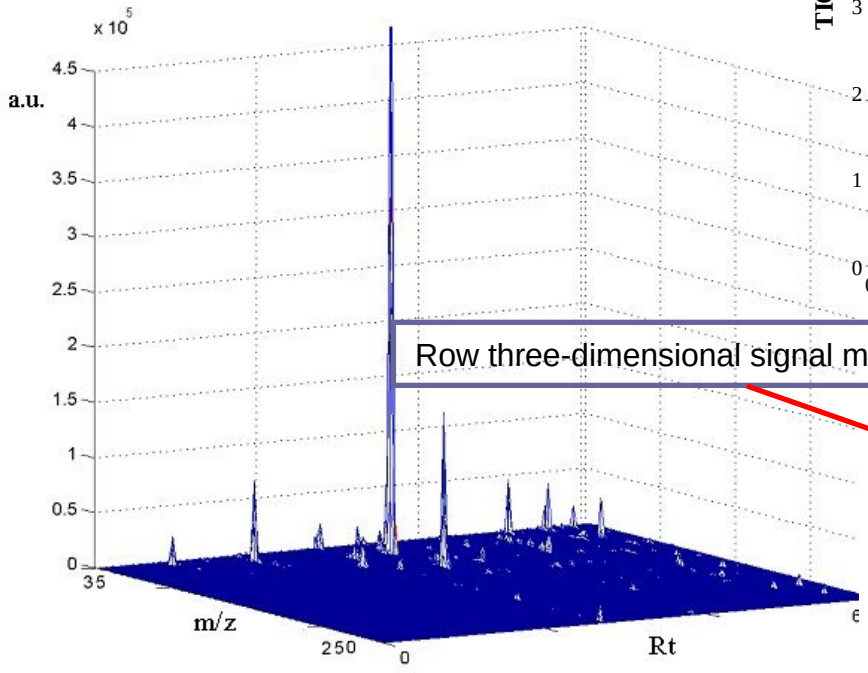


- exploratory single class data analysis (TUCKER3) for outliers detection
- random split in training and test sets for each class

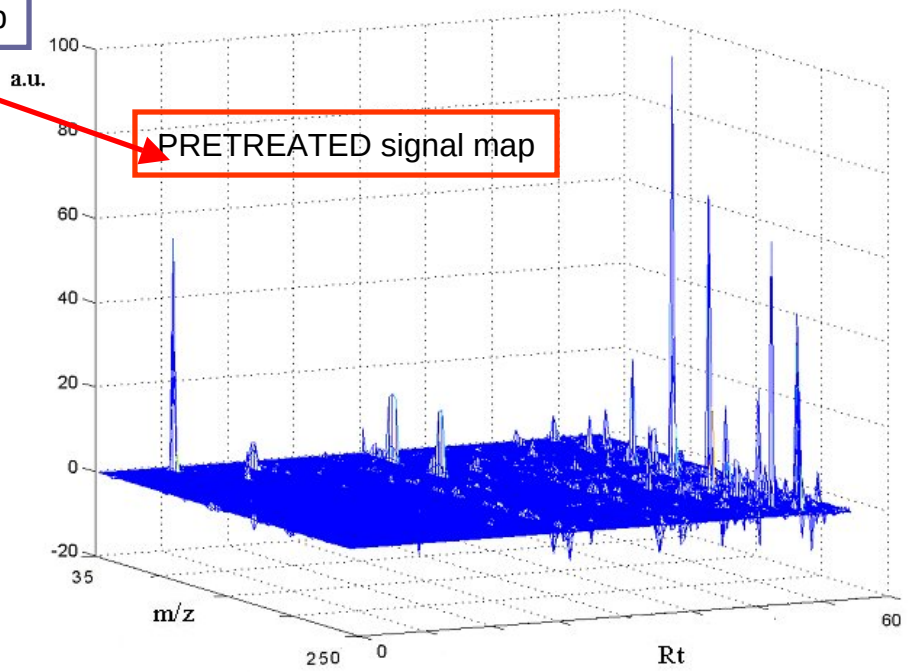
- **PRETREATMENT :**
  - Autoscaling within the third mode
  - Centering across the first mode
  - Block-scaling within the second mode

- Block-scaling within the second mode



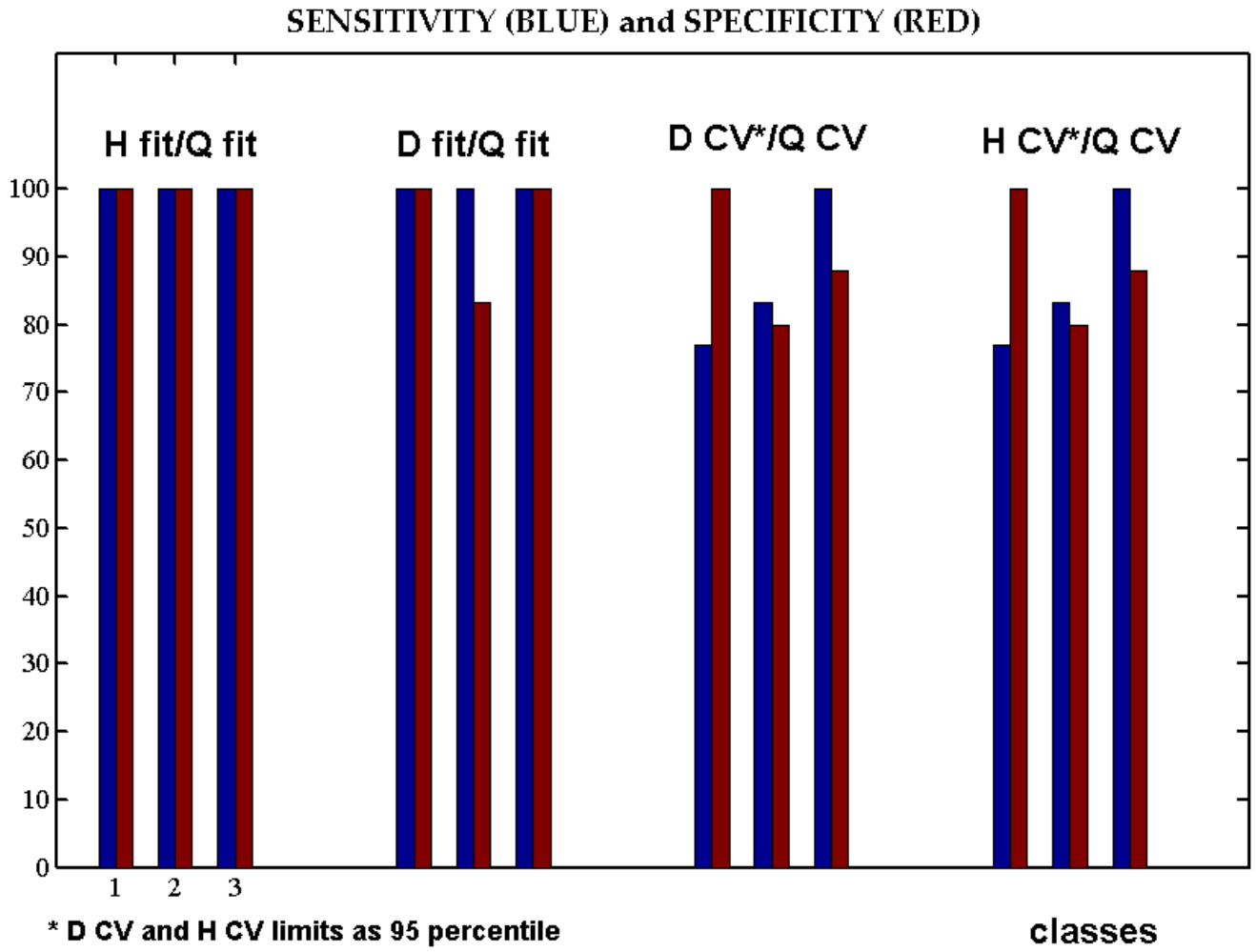


Row three-dimensional signal map

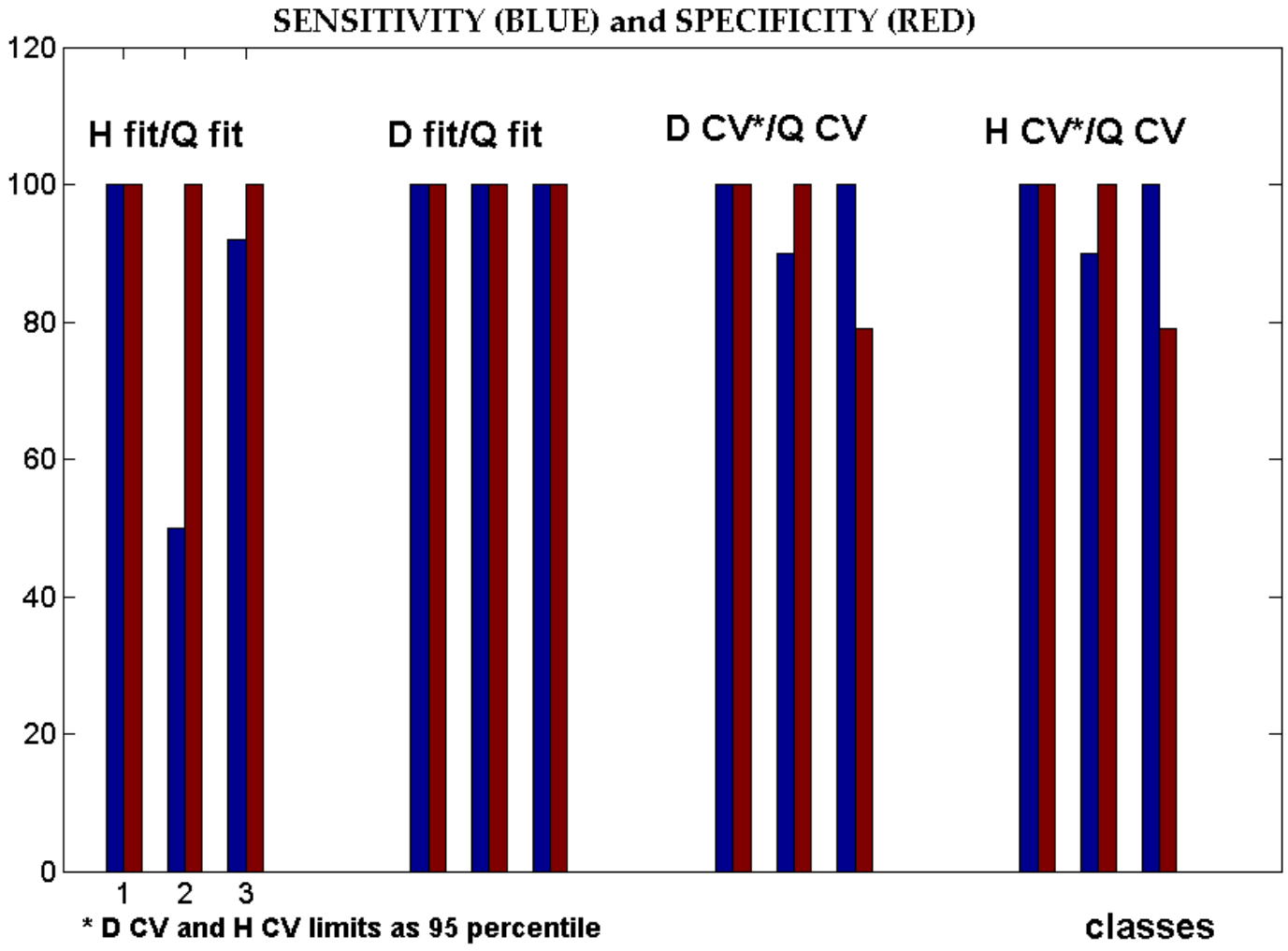


PRETREATED signal map

SIMCA 3D analysis: calibration set

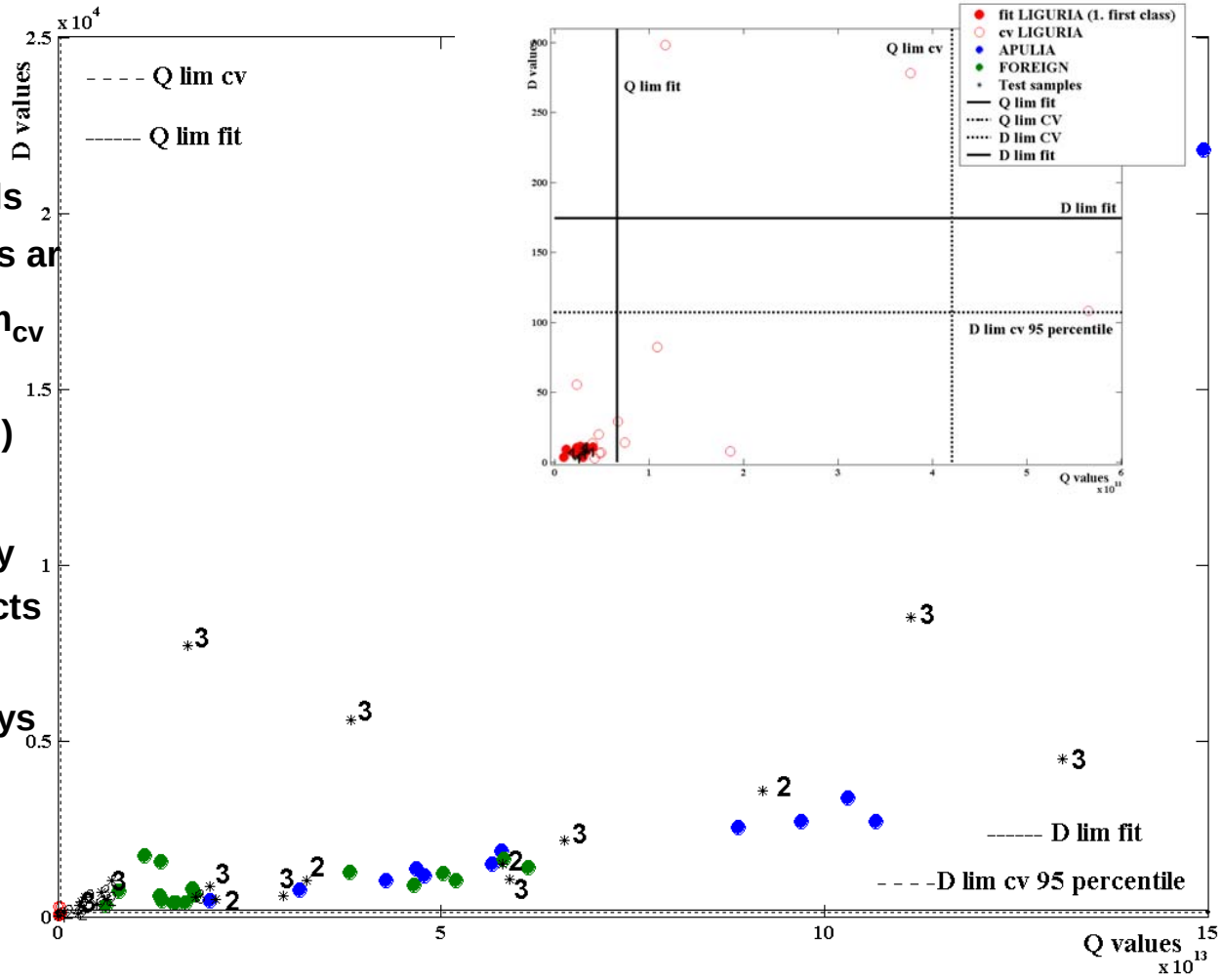


SIMCA 3D analysis: test set



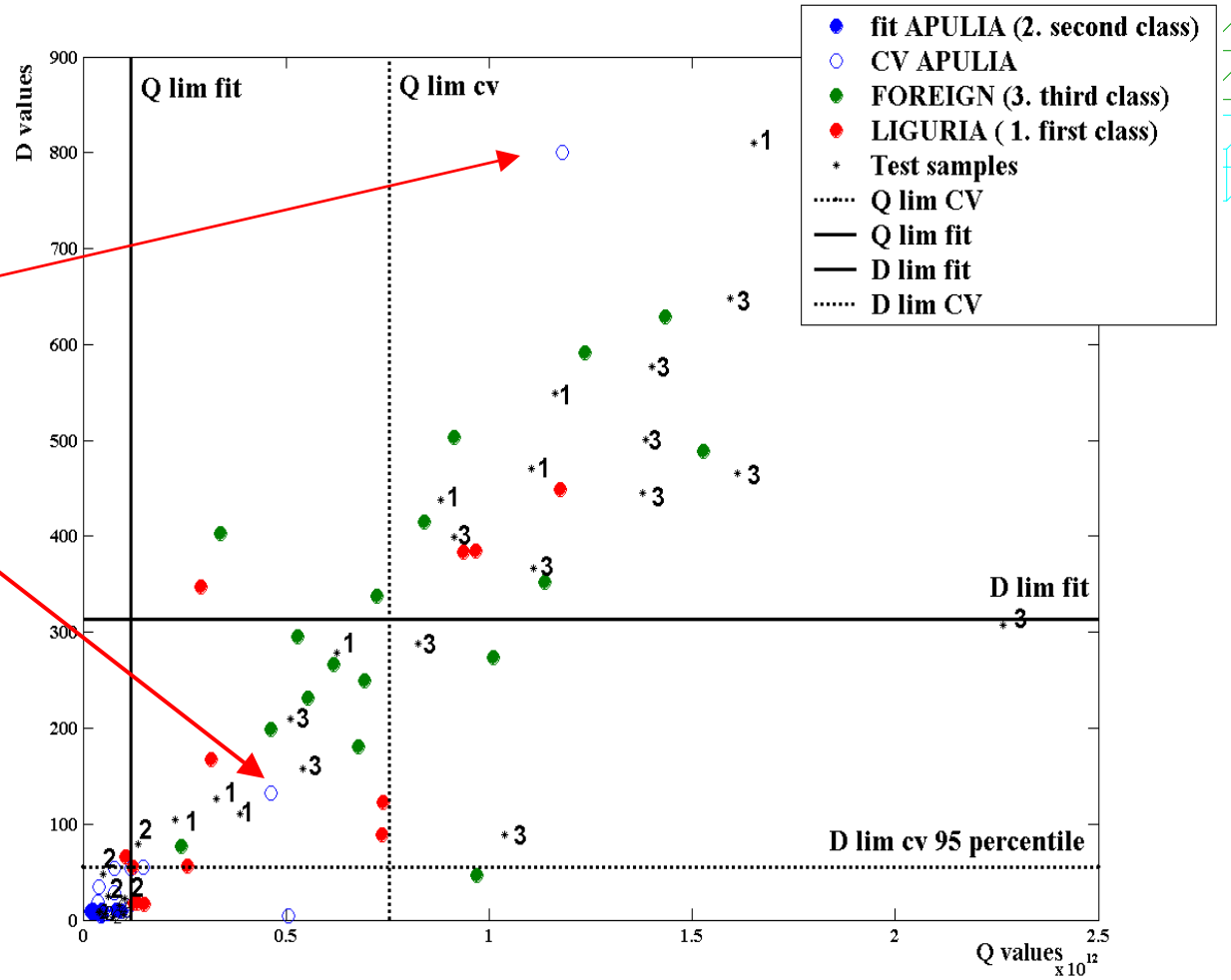
Model of Liguria class

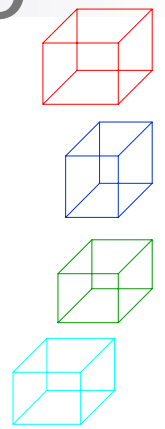
- CV predicted residuals and CV estimated scores ar far from fitted, thus $Qlim_{cv}$ is wide (instability of the model)
- $Dlim_{CV}$ is calculated by excluding outlying objects
- test objects are always well predicted



Model of Apulia class

- Q_lim_{CV} too wide
- D_lim_{fit} results too high
- empirical D_lim_{CV} seems ok





summary:

SENS	H_fit/Q_fit	D_fit/Q_fit	D_CV/Q_CV	H_CV/Q_CV
liguria	100 (100)	100 (100)	77 (100)	77 (100)
apulia	100 (50)	100 (100)	83 (90)	83 (90)
foreign	100 (92)	100 (100)	100 (100)	100 (100)
SPEC				
liguria	100 (100)	100 (100)	100 (100)	100 (100)
apulia	100 (100)	83 (100)	80 (100)	80 (100)
foreign	100 (100)	100 (100)	88 (79)	88 (79)

- Optimal results for LIGURIA model considering both CV and fit confidence limits for training and test set, lower sensitivity in CV due to extreme objects

- Good results for the other classes considering fit confidence limits

- refinement of the models is needed after checking the peculiar samples

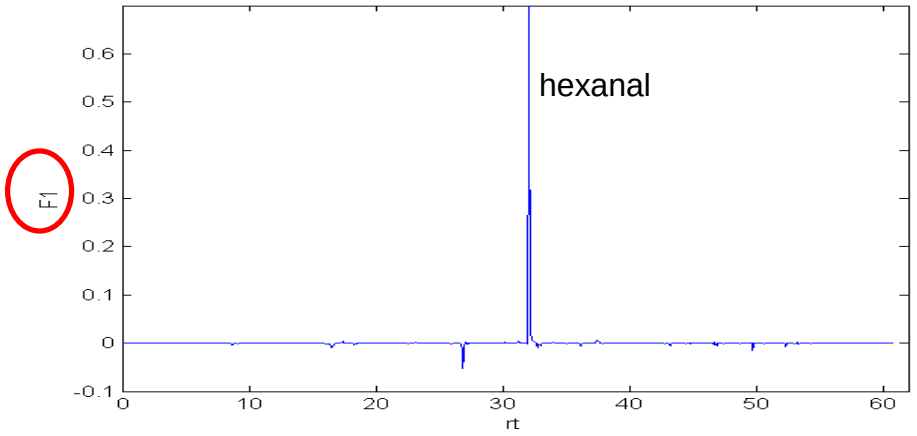
Second mode,
i.e GC-signals mode

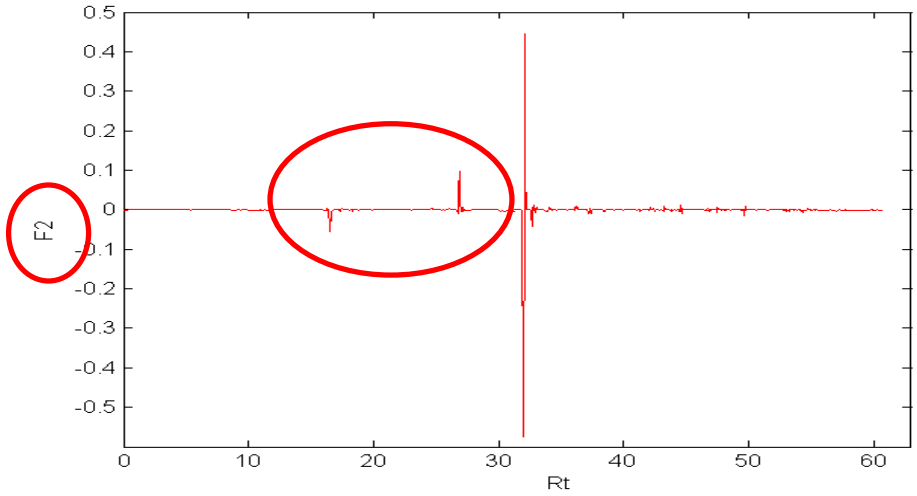
Largest square elements of Core array:

(1 1 1)

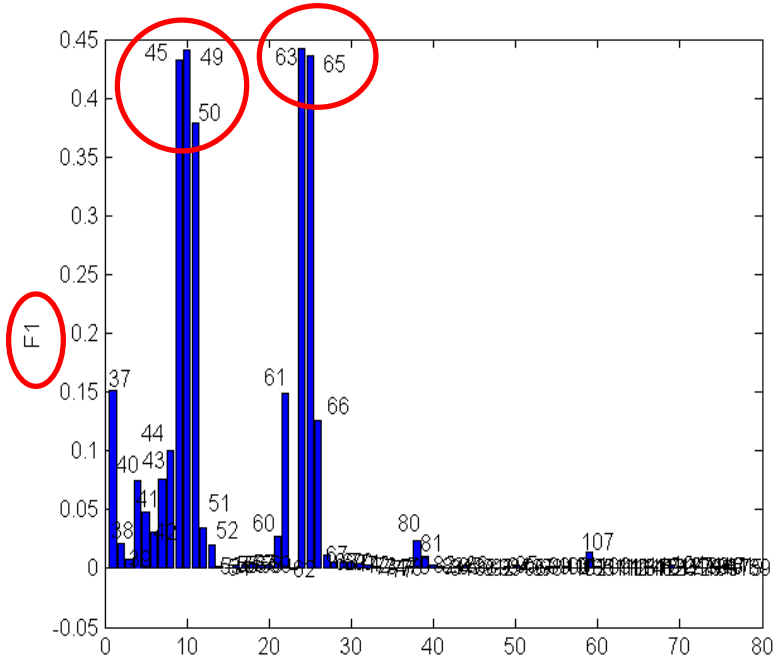
(2 2 1)

(2 1 1)





Third mode,
i.e MS-variables mode



Conclusion:

- Efficacy of the application of 3-way SIMCA models for recognition of Ligurian Extra-Virgin Olive Oils samples and discrimination of Aged Parma Ham samples

- The classification rule based on combintation of Q and D- statistics by using the empirical D_{CV_lim} seems promising

- if Specifity is the main issue H_fit seems a good solution

Future Perspectives:

- test on more data sets and applications needed
- code refinement and availability
- implement a parameter analogous to the variable discriminant power, for model interpretation
- use only a reference dstribution for combined Q and D

THANKS FOR YOUR ATTENTION