

A classification tool for N-way arrays based on SIMCA methodology

M. Cocchi¹, C. Durante¹ and R. Bro²¹University of Modena and Reggio Emilia, Chemistry
Department, Modena IT

marina.cocchi@unimore.it

²University of Copenhagen, Faculty of Life Sciences, Dept. of Food Science, Frederiksberg C
DK

Outline

- ▶ context
- ▶ overview of 2-way SIMCA
- ▶ SIMCA extension to 3-way
- ▶ Simulated data / Case studies
- ▶ future perspective

▶ Data characterized by more than 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. Up to my knowledge:

✓ unfolding +kohonen maps+CART

[D. Ballabio, V. Consonni, R. Todeschini, *Anal. Chim. Acta*, 605 (2007) 134-146.]

✓ PARAFAC scores (whole model)

+ Fisher's LDA [Guimet F., J. Ferré, R. Boqué, *Chemom. Intell. Lab. Syst.*, 81 (2006) 94-106]

or + SIMCA [G.J. Hall, J.E. Kenny, *Anal. Chim. Acta*, 581 (2007) 118-124.]

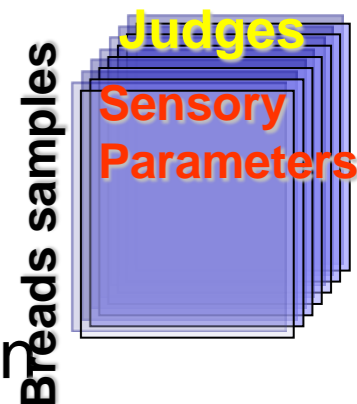
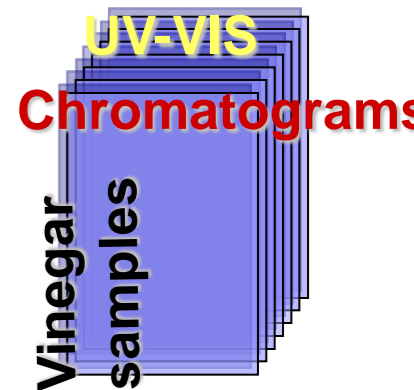
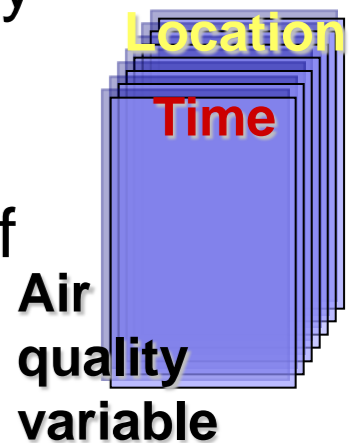
✓ discriminant multi-way partial least squares regression

[Guimet F., J. Ferré, R. Boqué. *Anal. Chim. Acta*, 544 (2005) 143-152.]

[Evrim Acar Ataman, Doctoral Thesis 2008, Rensselaer Polytechnic Institute Troy, New York]

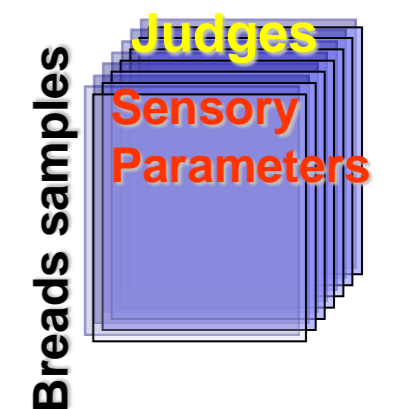
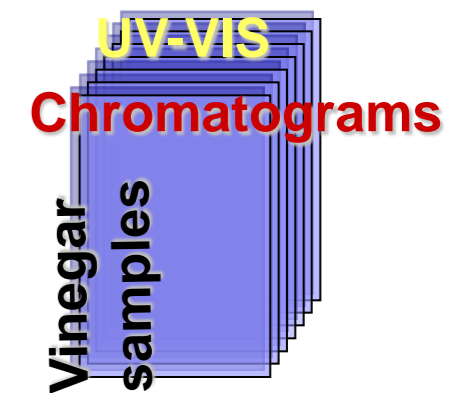
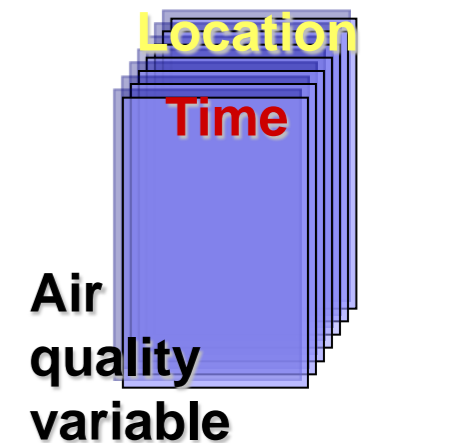
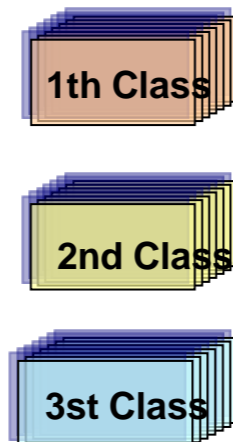
✓ first attempt to use TUCKER3 in conjunction with SIMCA classification

[G. R. Flåten, B. Grung, O. M. Kvalheim. *Journal of Chemometrics*, 18 (2004) 173 – 182.]



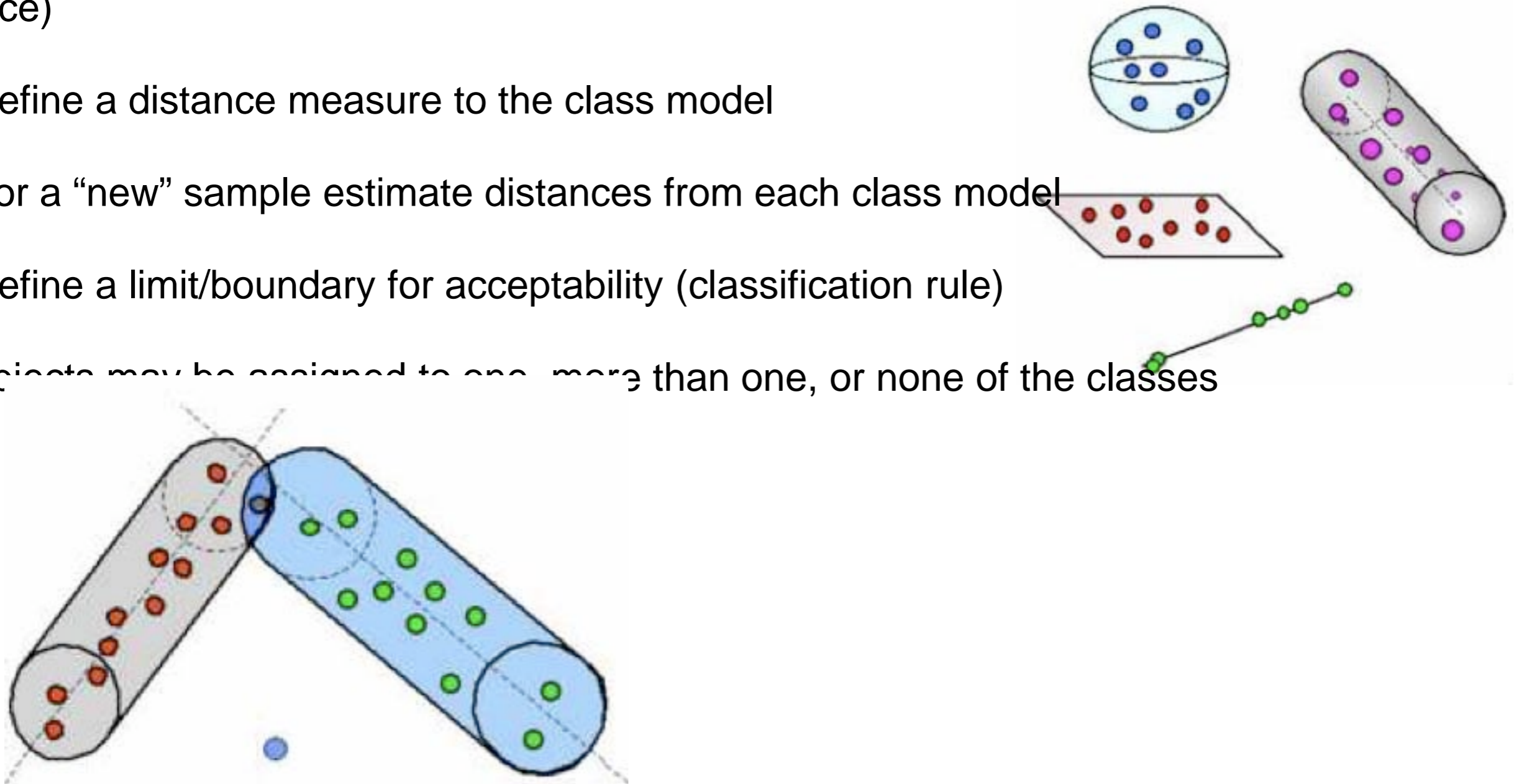
- ▶ a true multi way classification tool is still missing
- ▶ to retain 3D information
- ▶ to deal with one class modeling issue

disjoint category models



Overview of 2-way SIMCA

1. Build a distinct PCA model for each class (separate centering/scaling & dimensionality choice)
2. Define a distance measure to the class model
3. For a “new” sample estimate distances from each class model
4. Define a limit/boundary for acceptability (classification rule)
5. objects may be assigned to one, more than one, or none of the classes



Overview of 2-way SIMCA

SIMCA original

S.Wold, *Pattern Recognition*, 1976, 127;
S. Wold, M. Sjostrom, *ACS Symposium Series*, 1977, 243-281.

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

$$\mathbf{d}_p^{(q)} = \sqrt{\mathbf{s}_p^{(q)2} + \sum_a \sqrt{\lambda_a}^2 (\mathbf{t}_a - \nabla_{\mathbf{a}}^{(q)},_{lim})^2}$$

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

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

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

Text

i=1:M M=number of variables;

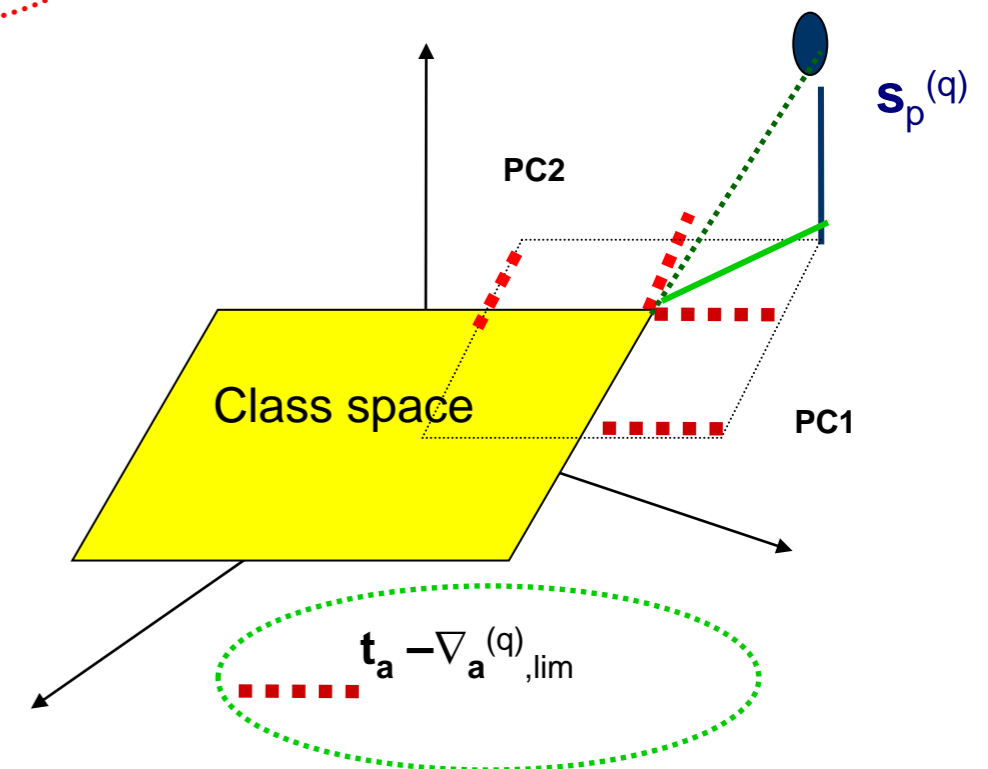
k= 1:N objects belonging to class q;

Classification rule

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

If true for both q and r Unique assignment only if

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



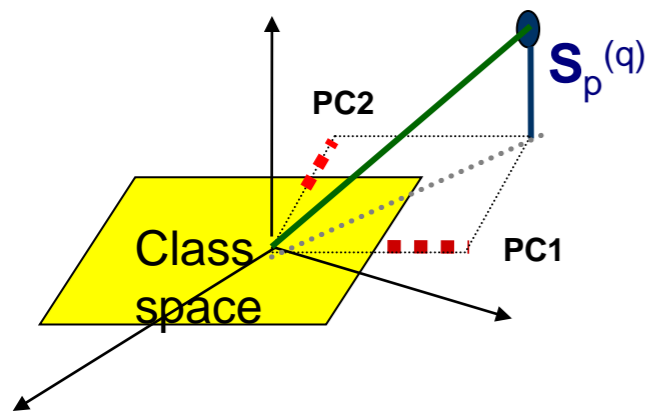
Overview of 2-way SIMCA

SIMCA evolution

@ evolution of original SIMCA

Distances estimation:

– use Mahalanobis distance



– use cross-validated residuals in Orthogonal Distance [1, 2]

– use only Orthogonal distance

– use only Leverage values [3]

Degrees of freedom – correction procedures:

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

› use $\frac{1}{2}[(N-A-1)(M-A)]$ in F_{crit}

› correct $F * N / (N-A-1)$

› use $F_{\text{crit}} (a-A)(N-A-1)/N, (N-A-1)(a-A)$

› correct $F * (N-1) / (N-A-1)$

› correct $F * N / (N-A-1)$ and use $F_{\text{crit}} 1, (N-A-1)$

See references 15-18 in [1]

1. R.D. Maesschalck, A. Candolfi, D.L. Massart, S. Heuerding, Chemom. Intell. Lab. Syst. 47 (1999) 65-77
2. G.R. Flaten, B. Grung, O.M. Kvalheim, Chemom. Intell. Lab. Syst. 72 (2004) 101
3. Todeschini et al., Chemom. Intell. Lab. Syst. In press

@ **robust SIMCA** [1,2]

- use robust PCA for class models
- use a weighted sum of “reduced” orthogonal (OD, i.e. $\mathbf{s}_p^{(q)2}$ in previous slide) and Mahalanobis (MD, in scores space) distances
- thus classification rule will be to assign a new sample to the class for which R is minimal:

$$R = \textcircled{c} (OD/OD_{lim}) + (1-\textcircled{c}) (MD/MD_{lim}) \quad \textcircled{c} \text{ in the range } 0-1$$

where MD_{lim} is estimated by using as reference the chi-squared distribution with *dof* equal to the number of retained components (A) and OD_{lim} is estimated by assuming that a scaled version of chi-squared distribution is appropriate (Wilson-Hilferty approximation)

1. K.V. Branden and M. Hubert, Chemolab 79 (2005) pp.10-21
2. M. Daszykowski, K. Kaczmarek, I. Stanimirova, Y. Vander Heyden, B. Walczak Chemom. Intell. Lab. Syst. (2006)

Overview of 2-way SIMCA

SIMCA PLS-toolbox

• a somehow alternative approach (coming from MSPC realm)

Distance from a class model:

Q is the sum of squared residuals: $\sum_k e_{pk}^2$ (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 (**T²**)

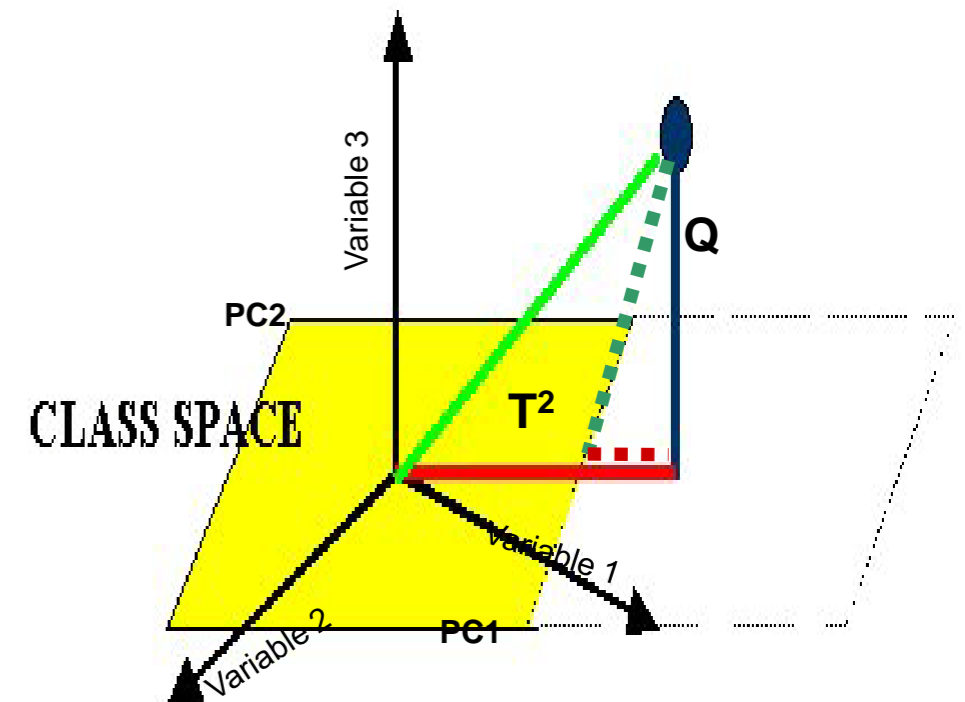
calculates **T²_{lim}** as $[A*(N-1)/N*(N-A)]*F_{crit}(A, (N-A))$

Classification rule

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

PLS-Toolbox SIMCA

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



Overview of 2-way SIMCA

SIMCA comparison

original SIMCA:

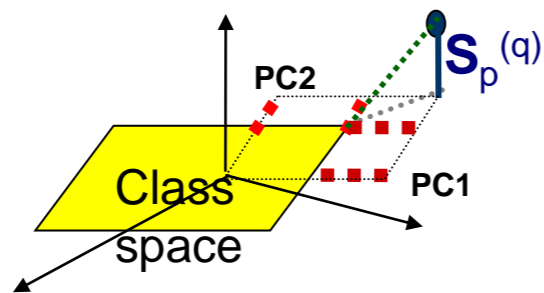
➤ score and orthogonal distances are combined as a “rsd” measure

➤ square distance/rsd of a “new” sample is compared to square rsd of class by F-test

➤ Score distance is computed from the “box” boundaries (but can be taken from average **problematic**)

➤ degree of freedom for F-test

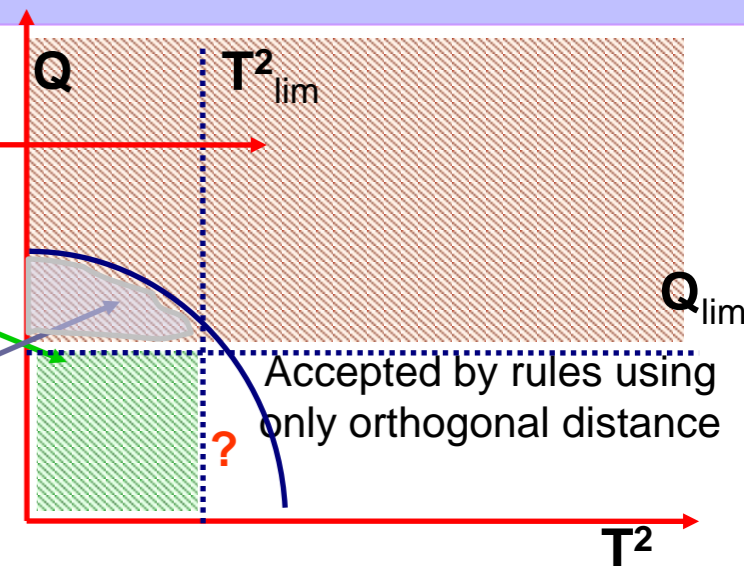
➤ scaling factor needed in order to combine score distance with orthogonal one



Rejected by both approaches

Accepted by both approaches

Rejected by original SIMCA
Accepted by this rule



alternative SIMCA:

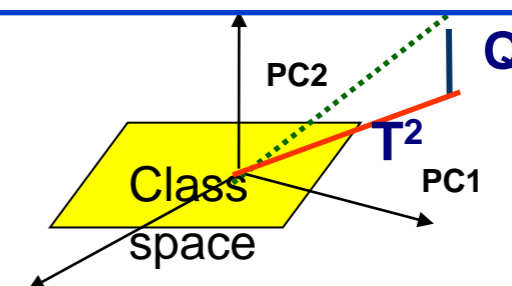
➤ reduced score distance and orthogonal distance are combined, thus being directly comparable

➤ Score distance is computed from the average **problematic**

➤ two different reference distribution are assumed

➤ degree of freedom (?)

➤ giving the same weight to both type of distances

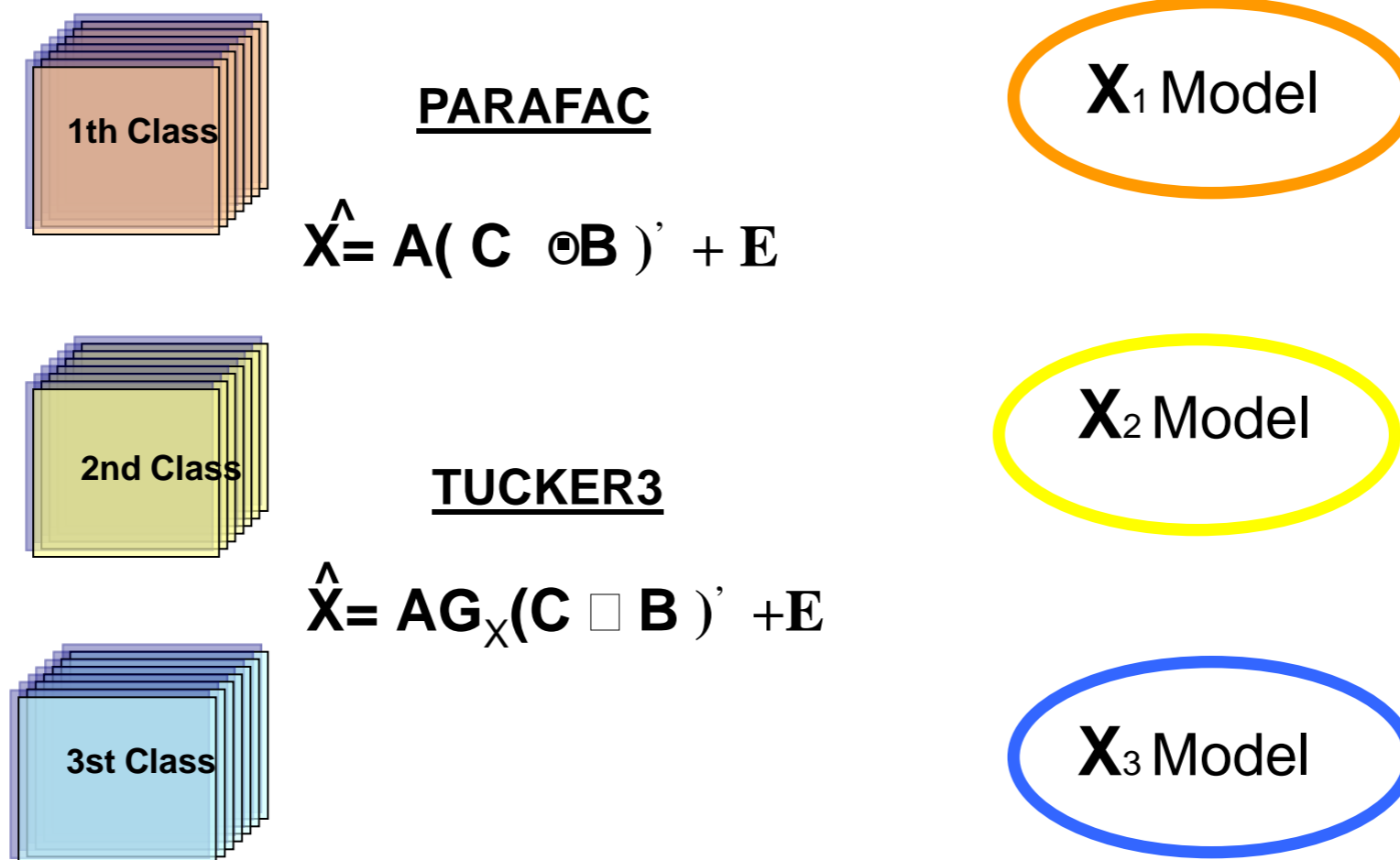


Evolution

- using T^2 avoids to define “box” boundaries
- using only Orthogonal distances does not take into account deviations inside class space
- using MD may avoid scaling factor

SIMCA extension to N-way

- ▶ a disjoint N- decomposition model for each class (both PARAFAC & Tucker3 implemented)
- ▶ separate multiway centering/scaling
- ▶ dimensionality (distinct for each class): *CV (ncrossdecomp)* or according to best SENS/SPEC



SIMCA extension to N-way

DIFFERENT CLASSIFICATION RULES implemented at this stage to evaluate

performances

\underline{X} (I x J x K) Factors: L, M, N (in case of PARAFAC L=M=N)

▶ SIMCA original

Orthogonal distance (defined as in Flaten et al.)

$$\text{RSD}_{\text{new}} = \sqrt{\mathbf{e}^T \mathbf{e} / (J-M) \cdot (K-N)}$$

$$\text{RSD}_{\text{ref}} = \sqrt{\mathbf{e}_r^T \mathbf{e}_r \cdot (I-1) / ((I-L-1) \cdot (J-M) \cdot (K-N))}$$

or dof = n. of data entry - n. free parameters

(SIMCA original fp)

Scores distance (defined as in original Wold et al. 1978) using T (Mode 1 loadings) and from boundary

Classification rule (defined as in original Wold et al. 1978)

$$D^2_{\text{new}} = \text{RSD}^2_{\text{new}} + \sum \alpha \Phi \alpha^2 (T_{\alpha} - \vartheta_{\alpha}(q), \text{lim})^2 \quad D^2_{\text{new crit}} = \text{RSD}^2_{\text{ref}} \cdot F_{\text{crit}} \quad F_{0.95 (J-M)(K-N), (I-L-1)(J-M)(K-N)}$$

N), (I-L-1)(J-M)(K-N)

▶ SIMCA original CV

same as above but with residual (\mathbf{e}) and scores (T) estimated in CV (one slab out in Mode 1)

SIMCA extension to N-way

DIFFERENT CLASSIFICATION RULE implemented

▶ SIMCA alternative

Orthogonal distance: Q

Scores distance:

Leverage:

$$H_{fit} = \text{diag} [A_{fit}(A_{fit}^T A_{fit})^{-1} A_{fit}^T]$$

$$H_{cv} = \text{diag} [A_{cv}(A_{fit}^T A_{fit})^{-1} A_{cv}^T]$$

$$H_{new} = \text{diag} [A_{new}(A_{fit}^T A_{fit})^{-1} A_{new}^T]$$

S: mode 1 scores Covariance matrix;

A: mode 1 Score matrix;

D-statistic

$$D_{fit} = \text{diag}[A_{fit}(S_{it})^{-1} A_{fit}^T]$$

$$D_{cv} = \text{diag}[A_{cv}(S_{it})^{-1} A_{cv}^T]$$

$$D_{new} = \text{diag}[A_{new}(S_{it})^{-1} A_{new}^T]$$

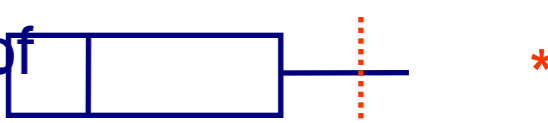
Classification rule (reduced distance as in PLS toolbox)

▶ Q, H, D limits:

$Q_{lim\ fit}$ χ^2 distribution JM approximation
 $H_{lim\ fit}$ $\sqrt{\frac{h_{ig} \cdot n_g}{n_g - 1} \cdot \frac{\beta(V/2, (n_g - V - 1)/2)}{\beta(V/2, (n_g - V - 1)/2)}}$ $n_g = n^{\circ}obj$ $V = n^{\circ}var$
 M. Forina *Chemometrics and Intelligent Laboratory Systems* 96 (2009) 239-245

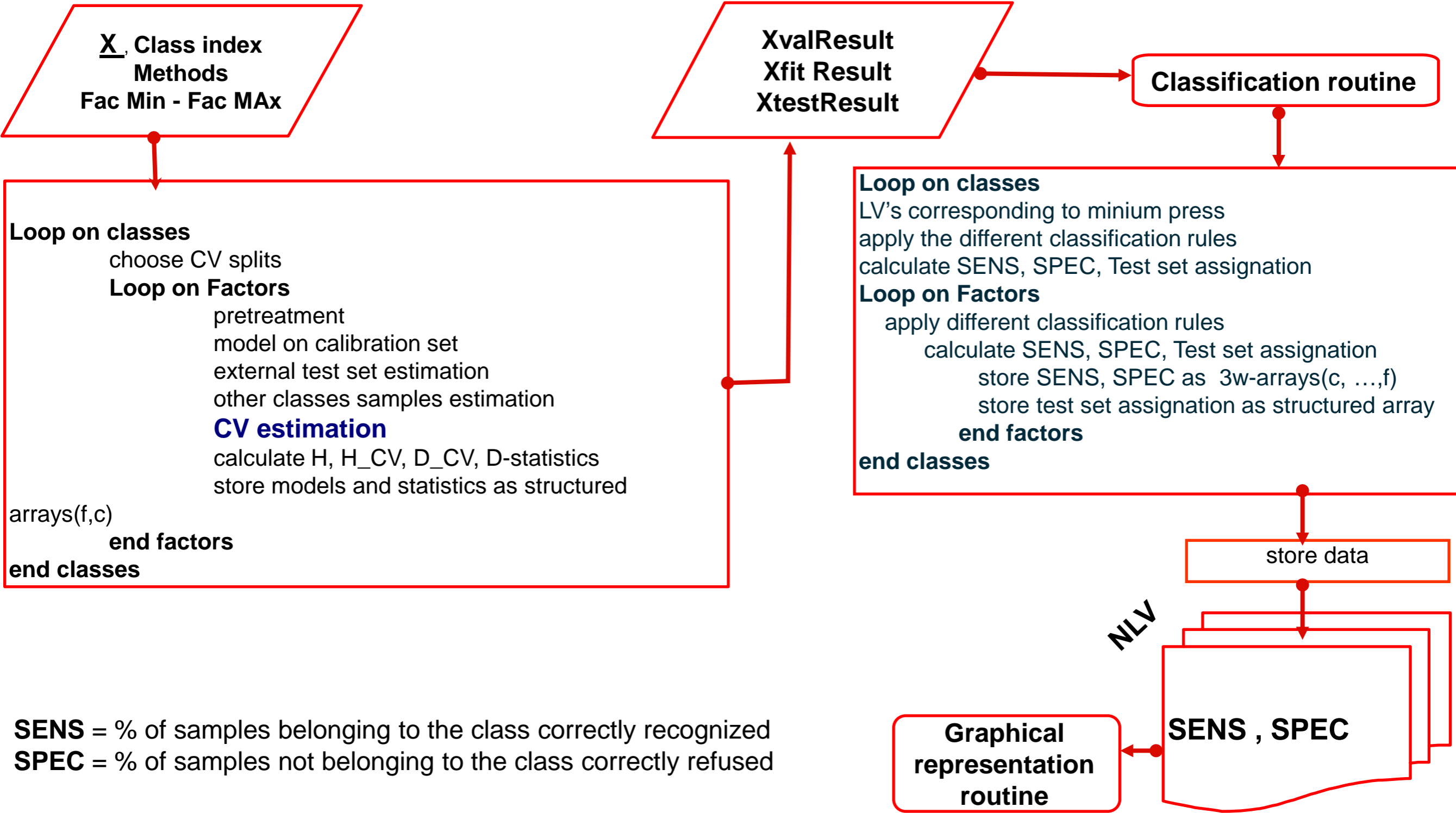
$D_{lim\ fit} \sim \frac{R(I^2 - 1)}{I(I - R)} F(R, I - R, \alpha)$
 J.A. Westerhuis, S.P. Gurden and A.K. Smilde, *Chemolab.* 51 (2000) 95-114.
 $R = 1^{st}$ mode factors, $I =$ samples

$H_{lim\ fit\ AP}$
 $Q_{lim\ fit\ AP}$
 also $H_{lim\ CV\ AP}$
 $Q_{lim\ CV\ AP}$
 $N_h \frac{h}{h_0} \sim \chi^2(N_h)$ $\frac{1}{N_h} [\chi^{-2}(N_h, 0.75) - \chi^{-2}(N_h, 0.25)] = \frac{1}{h_0} IQR(h_1, \dots, h_l)$
 Alexey L. Pomerantsev *J. Chemometrics* (2008)

H CV 95%: the 95 percentile of H_{cv}
 Q CV 95%: the 95 percentile of Q_{cv}


SIMCA extension to N-way

algorithm flow chart



SENS = % of samples belonging to the class correctly recognized
SPEC = % of samples not belonging to the class correctly refused

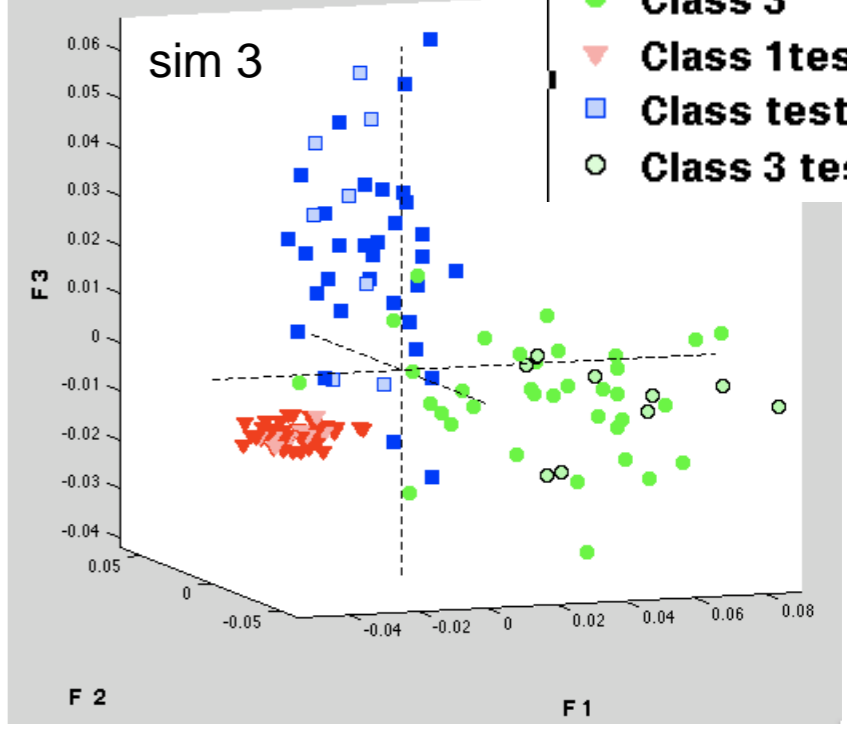
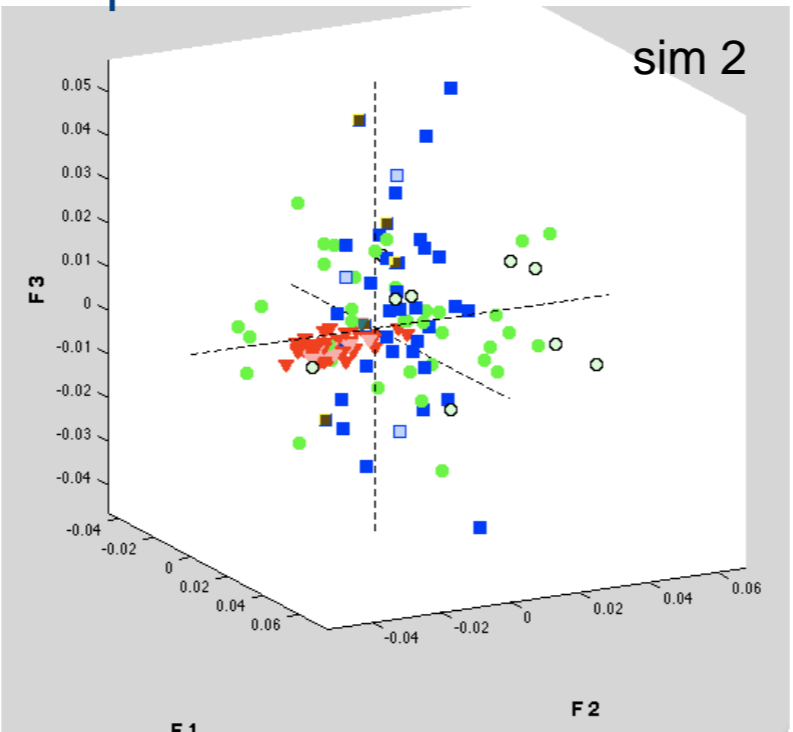
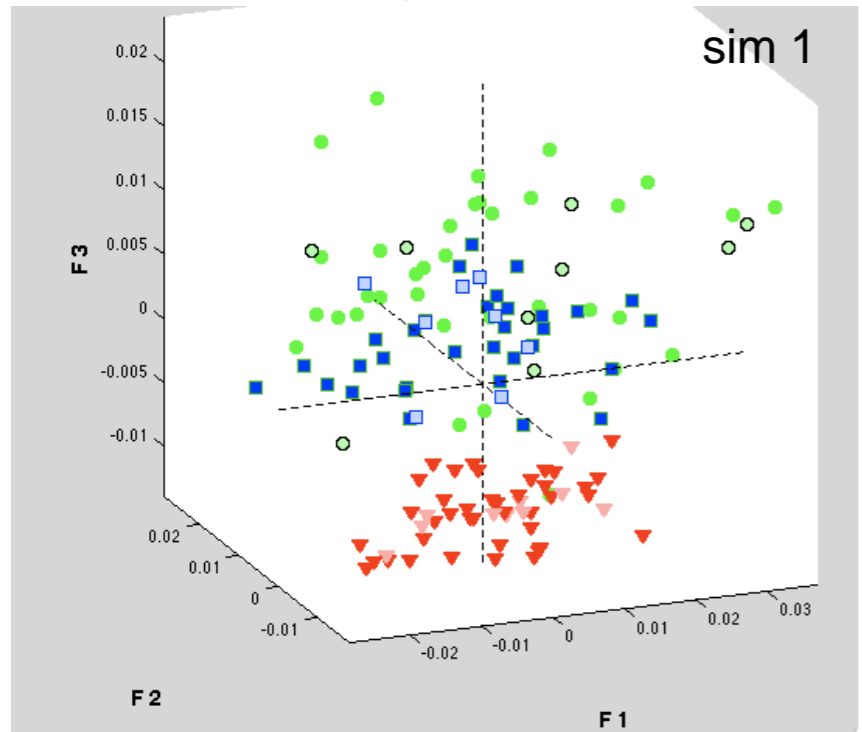
Simulated data

3 (+ 2) Data sets n x 40 x 40

- ▶ 3 classes
- ▶ generated by a pseudo 4 Factors PARAFAC models varying means and std of each class
- ▶ two dat sets have the same means and std but lower number of objects per class
- ▶ for sim1-sim3 each class random split

	class1	class2	class3
sim 1, sim 2, sim3	40	32	36
sim 1 small1	6	7	8
sim 1 small2	12	12	15
test same for all sets	10	8	9

- ▼ Class 1
- Class 2
- Class 3
- ▼ Class 1 test
- Class test2
- Class 3 test

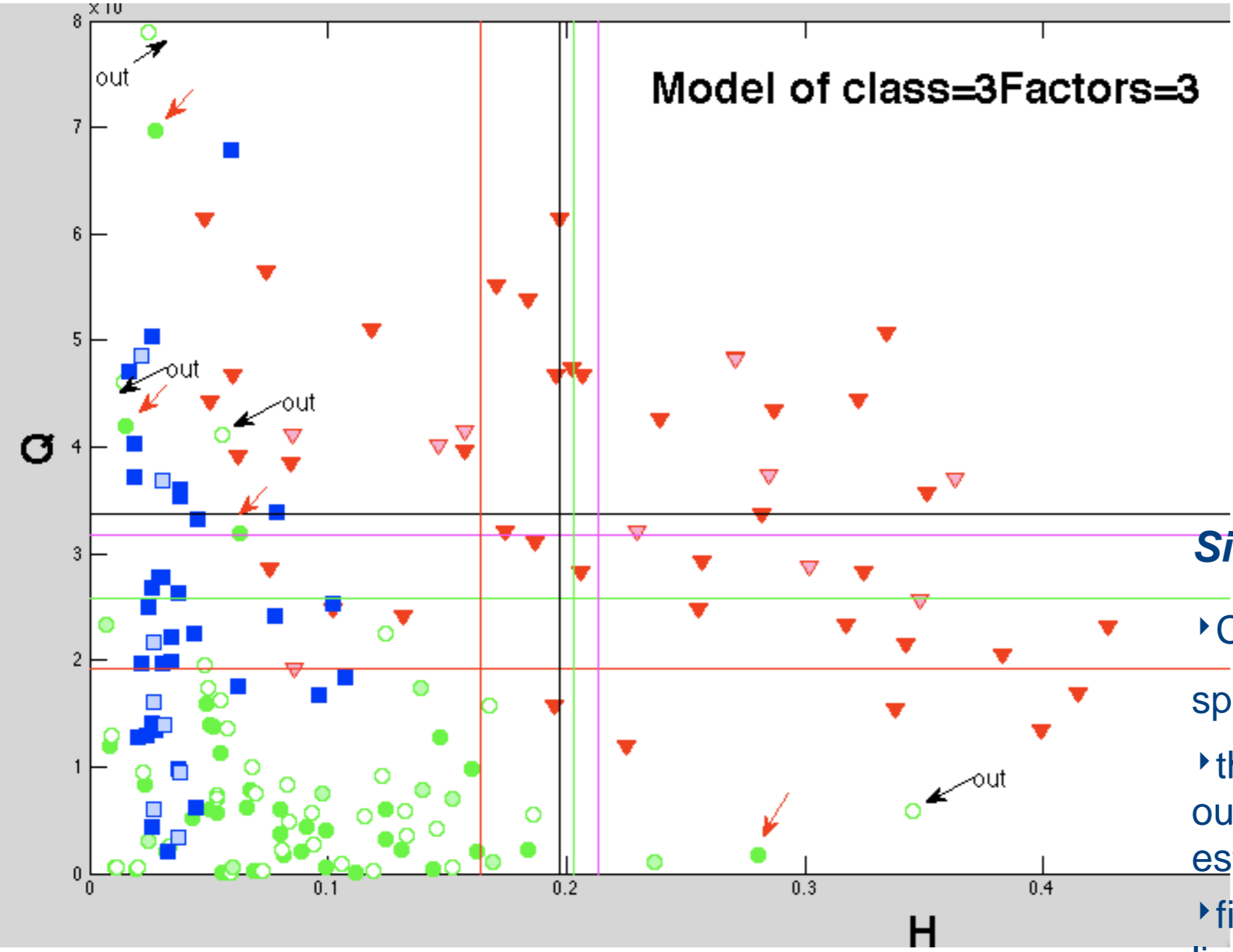
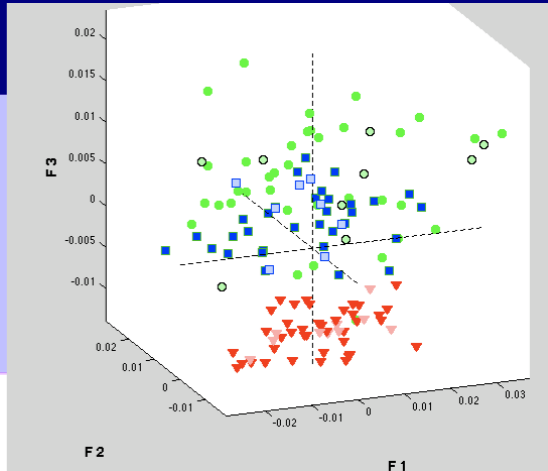


Simulated data

SENS	SIM 1						SIM 2						SIM 3					
	class1		class2		class3		class1		class2		class3		class1		class2		class3	
	train	test	train	test	train	test	train	test	train	test	train	test	train	test	train	test	train	test
H/Q Fit	100	100	94	100	94	100	100	100	91	75	97	44	98	100	97	100	97	100
H/Q FitAP	100	100	97	100	94	100	100	100	94	75	94	33	100	100	97	100	97	100
H/Q CV 95%	95	100	97	100	89	89	95	100	97	100	97	100	98	100	94	100	97	100
H/Q CV AP	100	100	97	100	94	100	100	100	100	100	97	100	100	100	97	100	97	100
D/Q Fit	100	100	100	100	94	89	100	100	94	75	97	56	98	100	100	100	100	100
D/Q CV	100	100	97	100	97	100	100	100	100	100	100	100	98	100	100	100	97	100
D/Q CV 95 %	95	100	97	100	89	89	95	100	97	100	95	100	98	100	88	100	94	100
SIMCA orig	100	100	100	100	100	89	100	100	100	100	100	67	100	100	100	100	100	78
SIMCA orig CV	100	100	100	100	100	89	100	100	100	100	100	78	100	100	100	100	100	78
SPEC	non SPEC						non SPEC											
H/Q Fit	100	100	100	100	43	44	99	100	76	68	26	17	100	100	100	100	79	83
H/Q FitAP	100	100	99	100	56	61	99	94	79	68	31	22	100	100	100	100	82	83
H/Q CV 95%	100	100	100	100	72	61	99	100	64	53	6	11	100	100	100	100	89	83
H/Q CV AP	100	100	99	100	55	61	99	94	51	42	6	11	100	100	100	100	79	83
D/Q Fit	100	100	99	100	35	28	99	94	67	63	26	10	100	100	100	100	65	72
D/Q CV	100	100	99	100	21	17	99	94	39	37	3	0	100	100	99	100	63	72
D/Q CV 95 %	100	100	100	100	72	61	99	100	64	53	6	11	100	100	100	100	89	83
SIMCA orig	98	100	100	100	26	22	96	88	21	26	3	6	100	100	97	100	70	83
SIMCA orig CV	98	100	100	100	26	22	93	88	21	26	1	0	100	100	97	100	70	83
SIMCA orig fp	98	100	100	100	26	33	96	88	21	26	3	6	100	100	97	100	70	83

ometrics, Mode

Sim 1 PARAFAC [3 4 4]

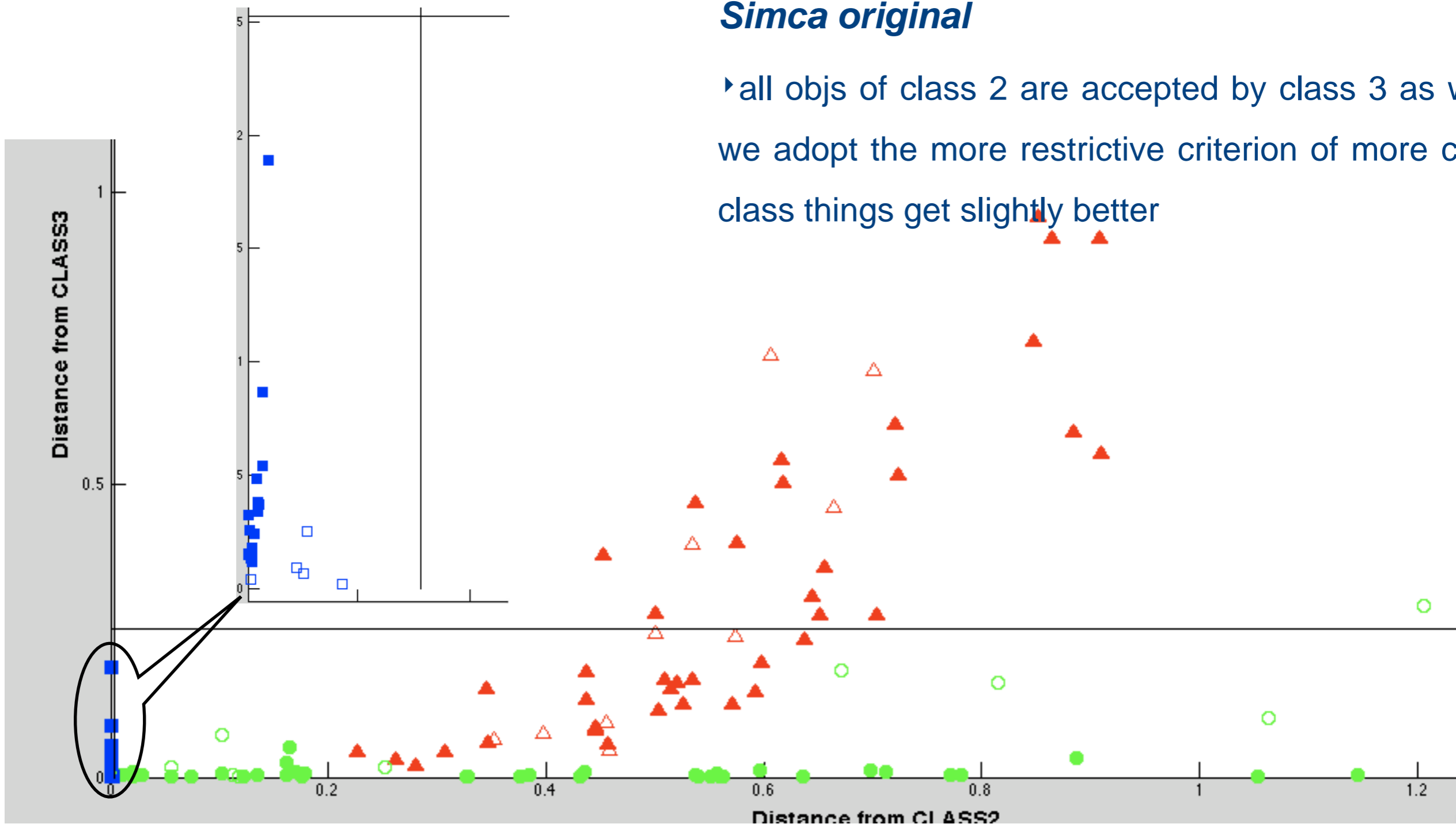


- train3
- CV
- test3
- ▼ train1
- ▼ test1
- train2
- test2
- QI-fit
- QI-fitAP
- QI-CYper
- QI-CYAP
- HI-fitF
- HI-fitAP
- HI-CYper
- HI-CYAP

Sim1 class 3: 36 x 40 x 40

- ▶ CV 95% limits are more specific
- ▶ the arrow out indicates box-plot outliers omitted from CV 95% limit estimation
- ▶ fitAP limits seems better than fit limits

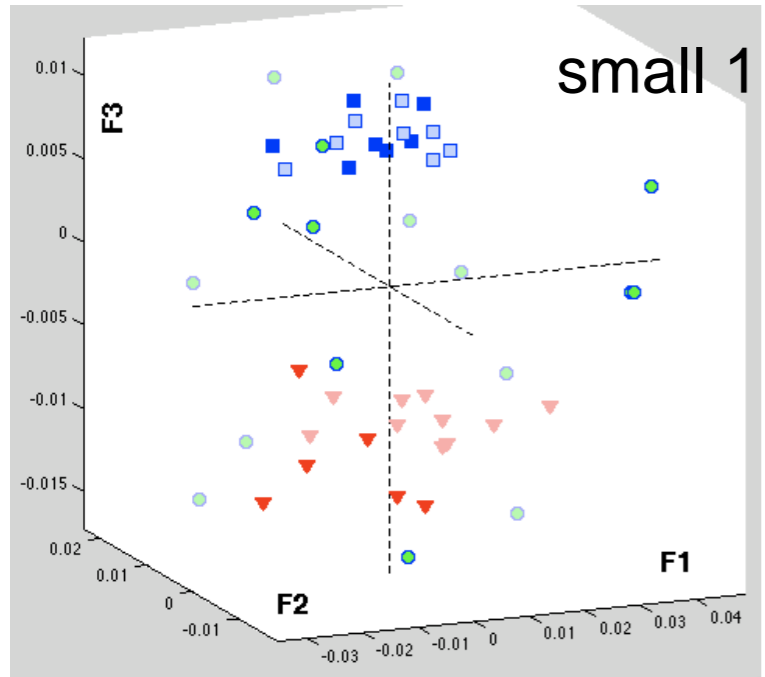
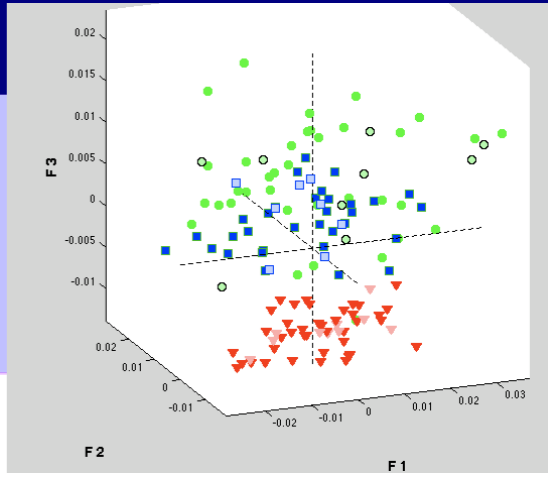
▶ Anyhow overlap 2-3 is in the data



Simca original

▶ all objs of class 2 are accepted by class 3 as well. If we adopt the more restrictive criterion of more closest class things get slightly better

Sim 1 small

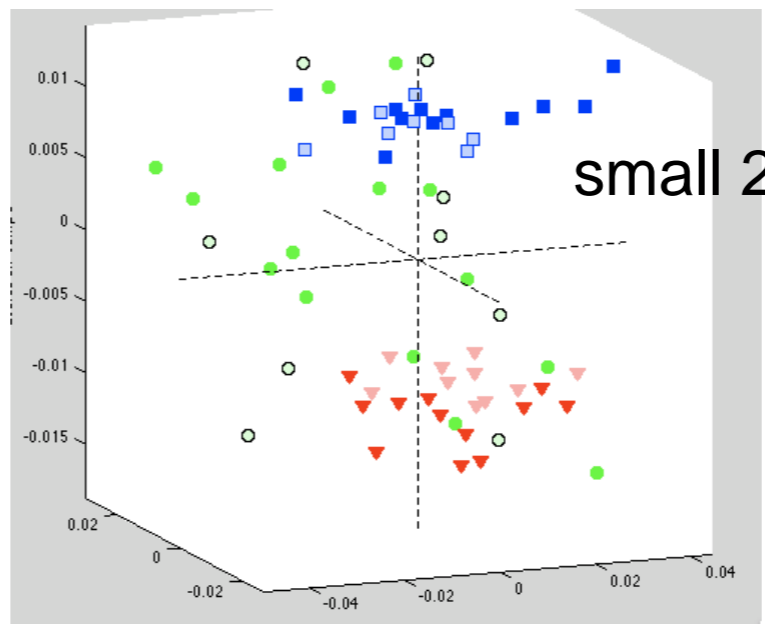


SENSITIVITY		H/Q Fit					
		class1		class2		class3	
		train	test	train	test	train	test
[2 3 4]	SIM 1_small	100	20	100	25	100	32
[2 3 3]	SIM 1_small2	100	80	100	62	93	33
		H/Q CV 95%					
SIM 1_small		100	50	100	38	100	33
SIM 1_small2		100	100	100	88	87	33
		SIMCA orig CV					
SIM 1_small		100	80	100	75	67	100
SIM 1_small2		100	100	100	100	100	100

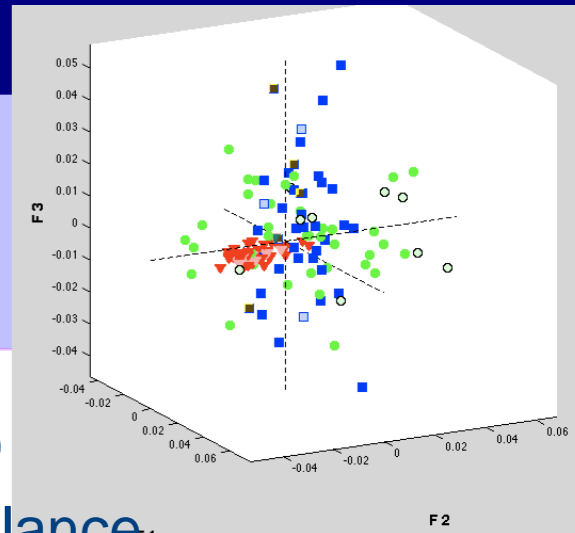
small 1 & small 2

- ▶ similar performance of different limits
- ▶ undersampling - overfitting
- ▶ SIMCA higher sensitivity for class 3 small 2 but low specificity (H/Q CV 95% opposite)

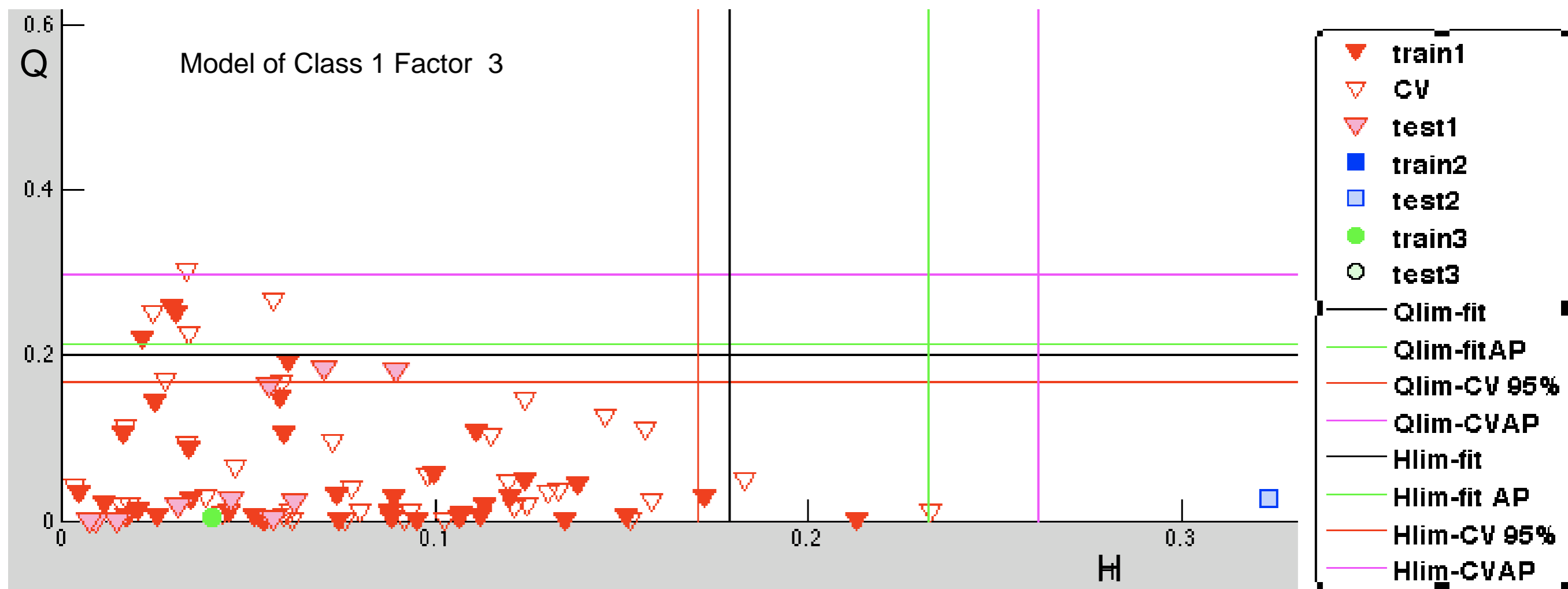
SPECIFICITY		H/Q Fit					
		class1		class2		class3	
		train	test	train	test	train	test
[2 3 4]	SIM 1_small	100	100	100	100	100	100
[2 3 3]	SIM 1_small2	100	100	100	100	96	89
		H/Q CV 95%					
SIM 1_small		100	100	100	100	92	100
SIM 1_small2		100	100	100	100	96	89
		SIMCA orig CV					
SIM 1_small		100	100	100	100	23	50
SIM 1_small2		100	94	92	100	33	33



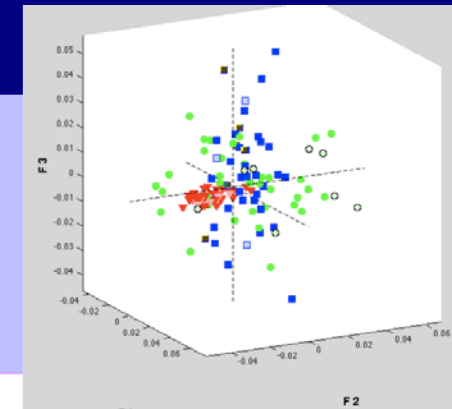
Sim 2 PARAFAC [3 4 4]



- ▶ class 1 can be modeled because of small rsd (even if lies among the other two)
- ▶ CV 95% limits seem too tight (however only 2 train sample are rejected, i.e. balance Q/H)
- ▶ CV AP seems best in this case

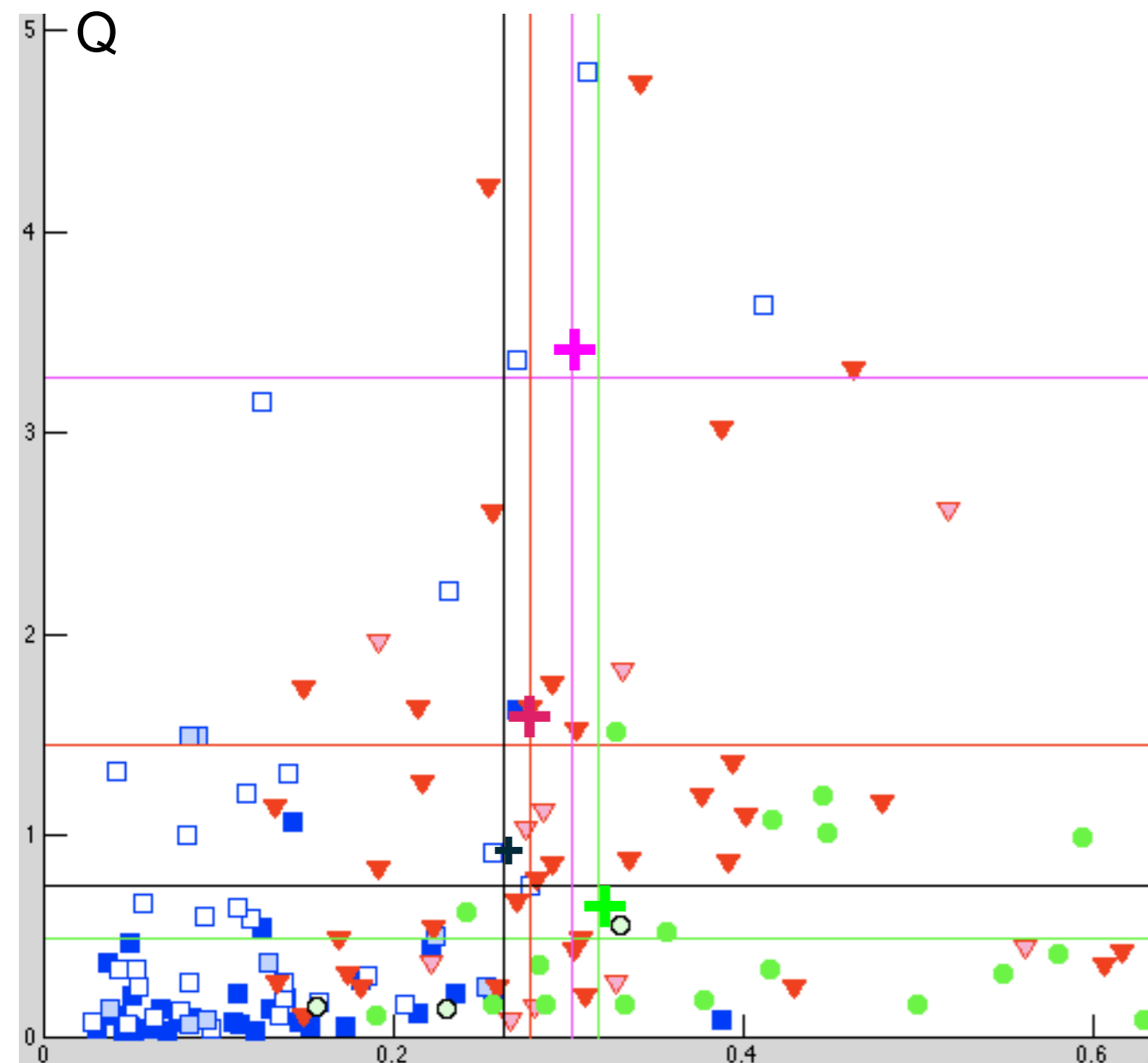


Sim 2 PARAFAC [3 4 4]

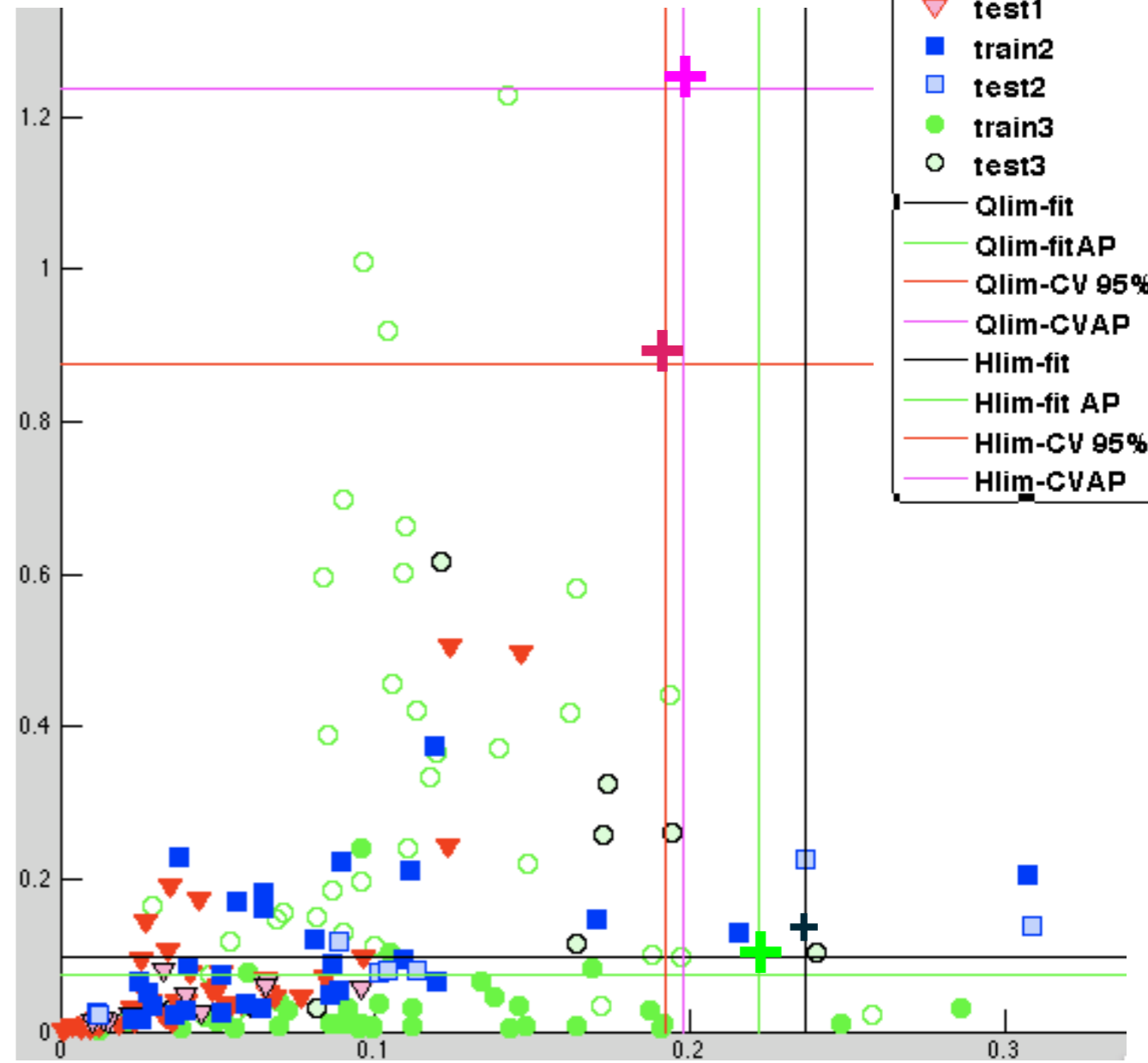


▶ in this case with higher class rsd and strong overlap fit limits seems better, difference is in Qlim not much in Hlim

Model of Class 2 Factors 4

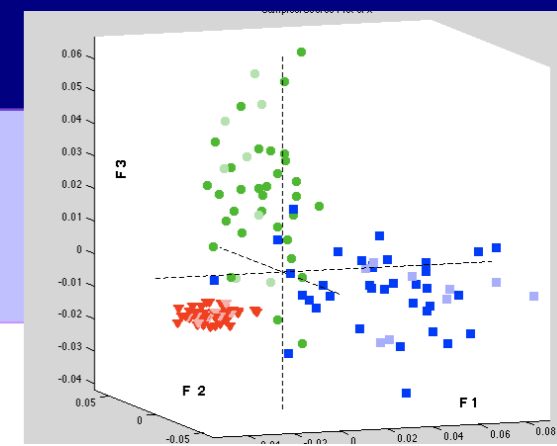


Model of Class 3 Factors 4

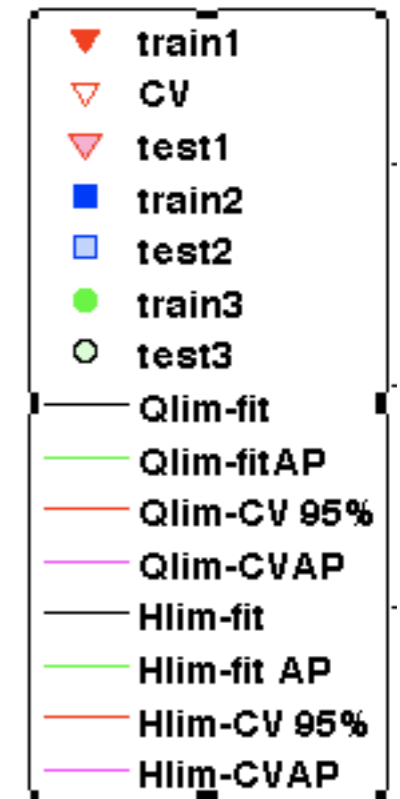
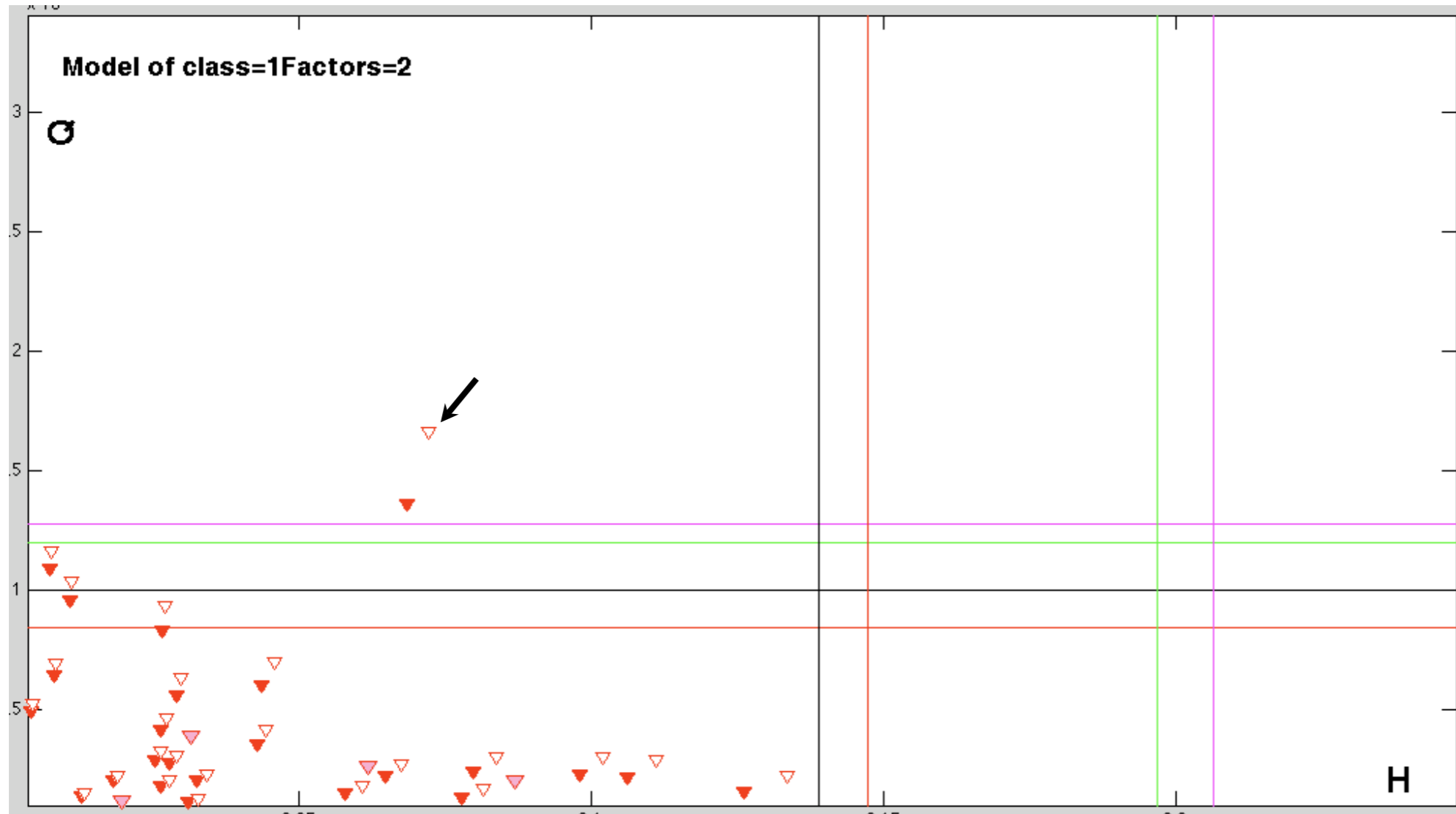


Sim 3

PARAFAC [2 3 2]

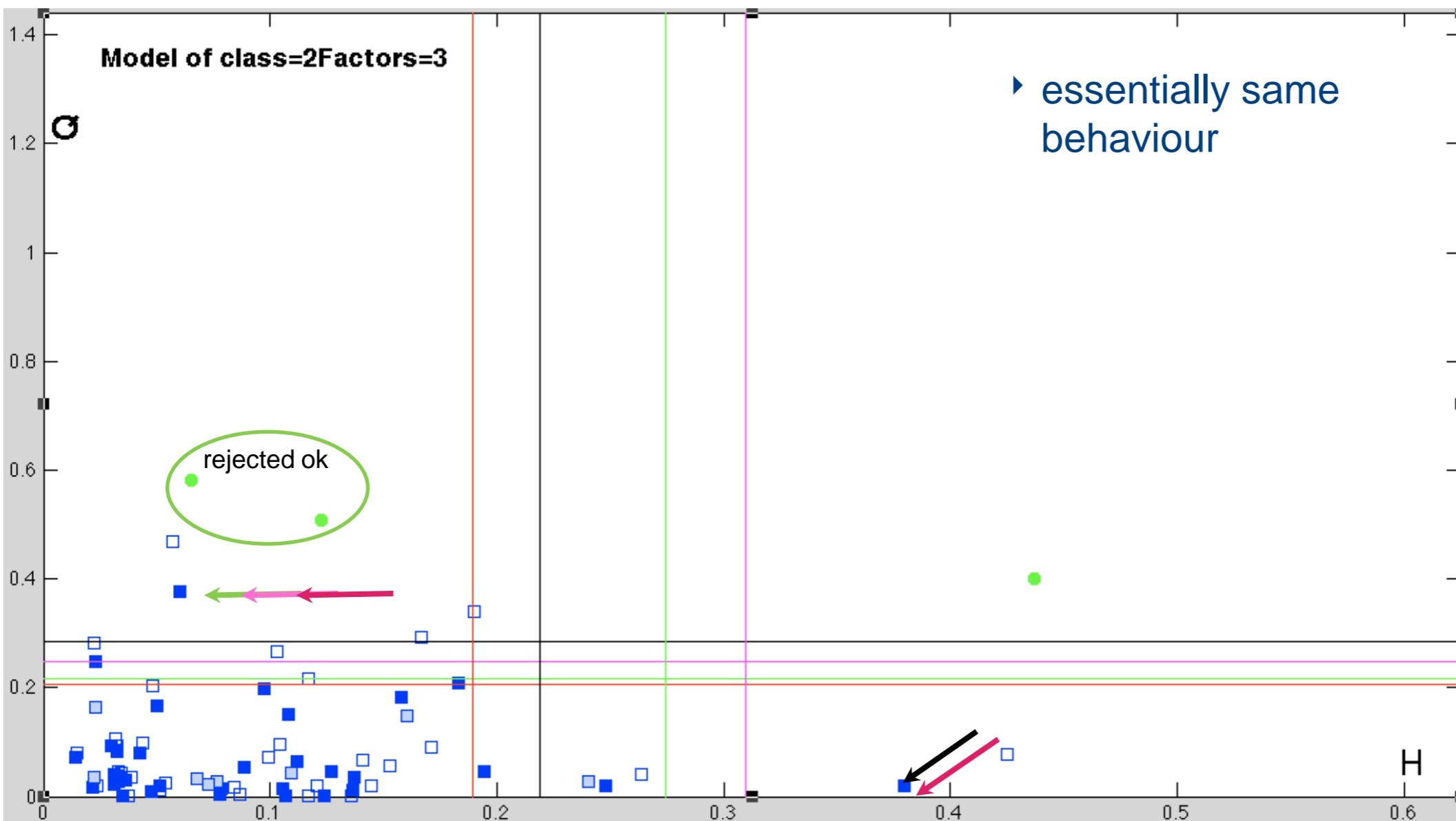
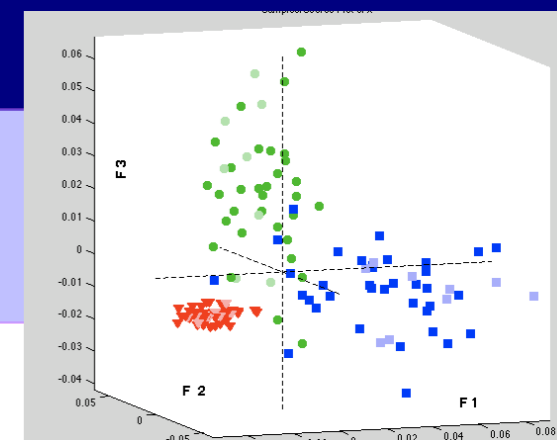


▶ it is ok only one sample rejected by CV 95%



Sim 3

PARAFAC [2 3 2]

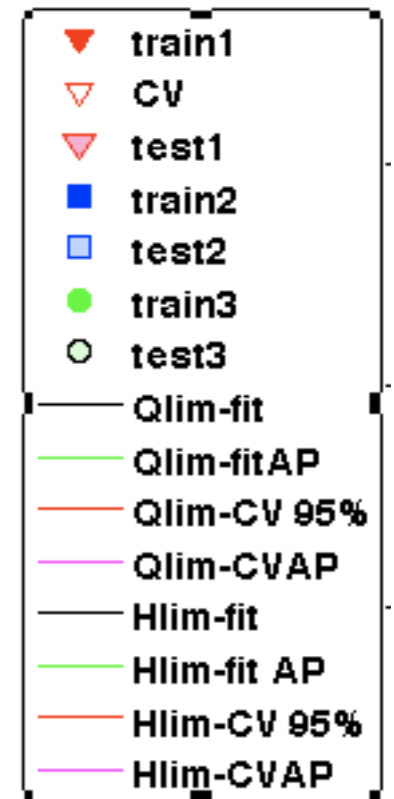
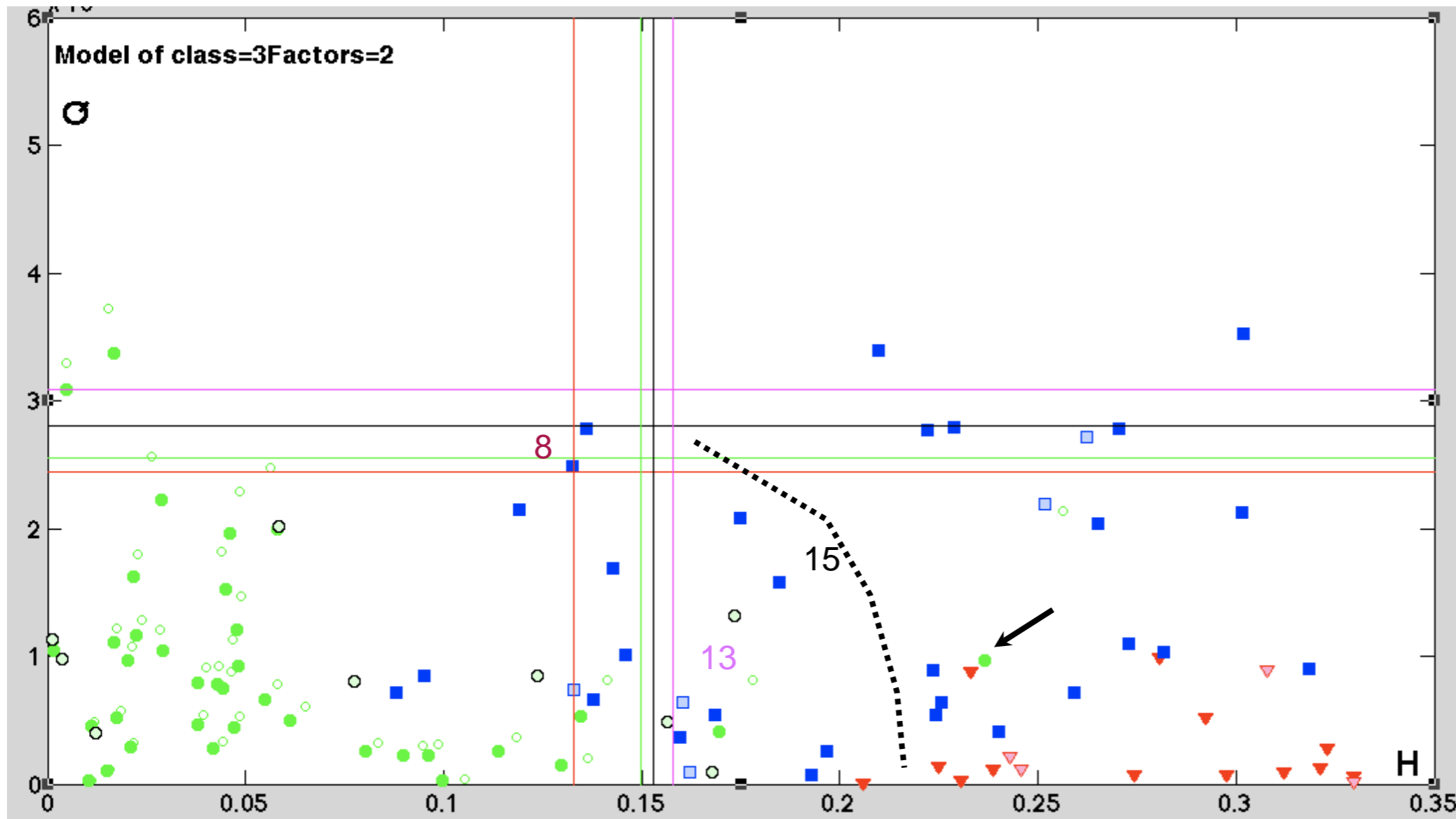
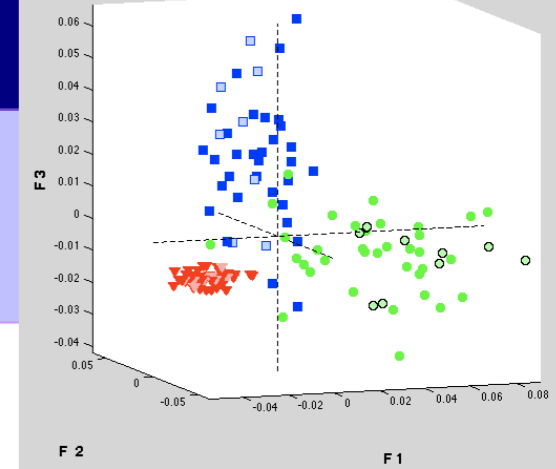


essentially same behaviour

- train1
- CV
- test1
- train2
- test2
- train3
- test3
- Qlim-fit
- Qlim-fit AP
- Qlim-CV 95%
- Qlim-CVAP
- Hlim-fit
- Hlim-fit AP
- Hlim-CV 95%
- Hlim-CVAP

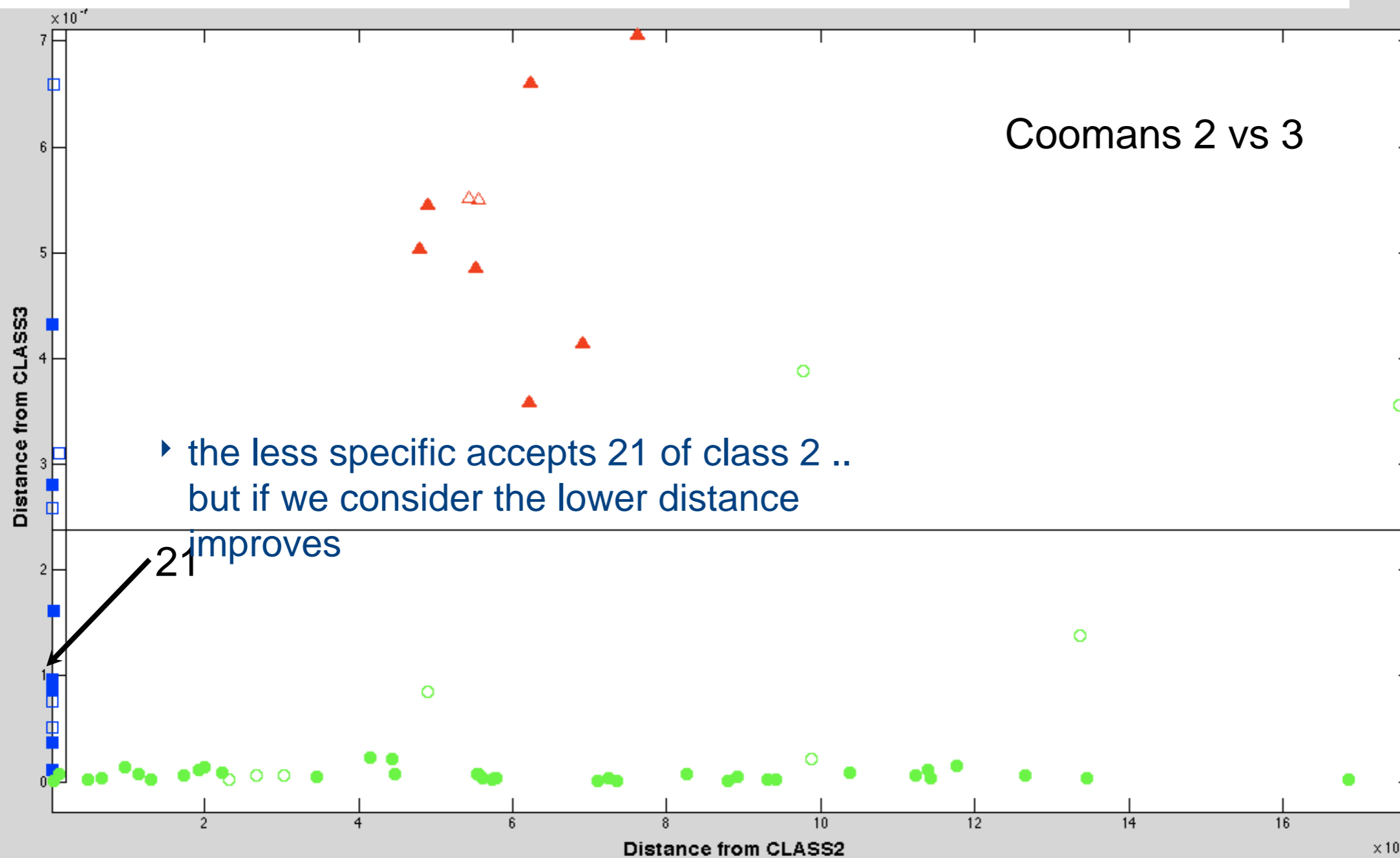
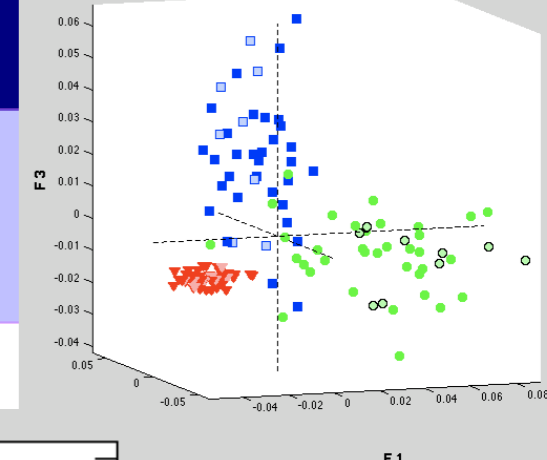
Sim 3

PARAFAC [2 3 2]



▶ class 3 is not specific CV 95% & fit AP accept 8 of tr class 2 samples, CV AP 13, fit 15

Sim 3 PARAFAC [2 3 2]



- ▶ not bigger difference among criteria; more in Q than H
- ▶ D in these sets was behaving as H
- ▶ SIMCA original more sensible less specific
- ▶ CV limits ok for not too high class spread
- ▶ fit limits better for heterogeneous sample
- ▶ there may be overfitting without validation independently of the used criteria

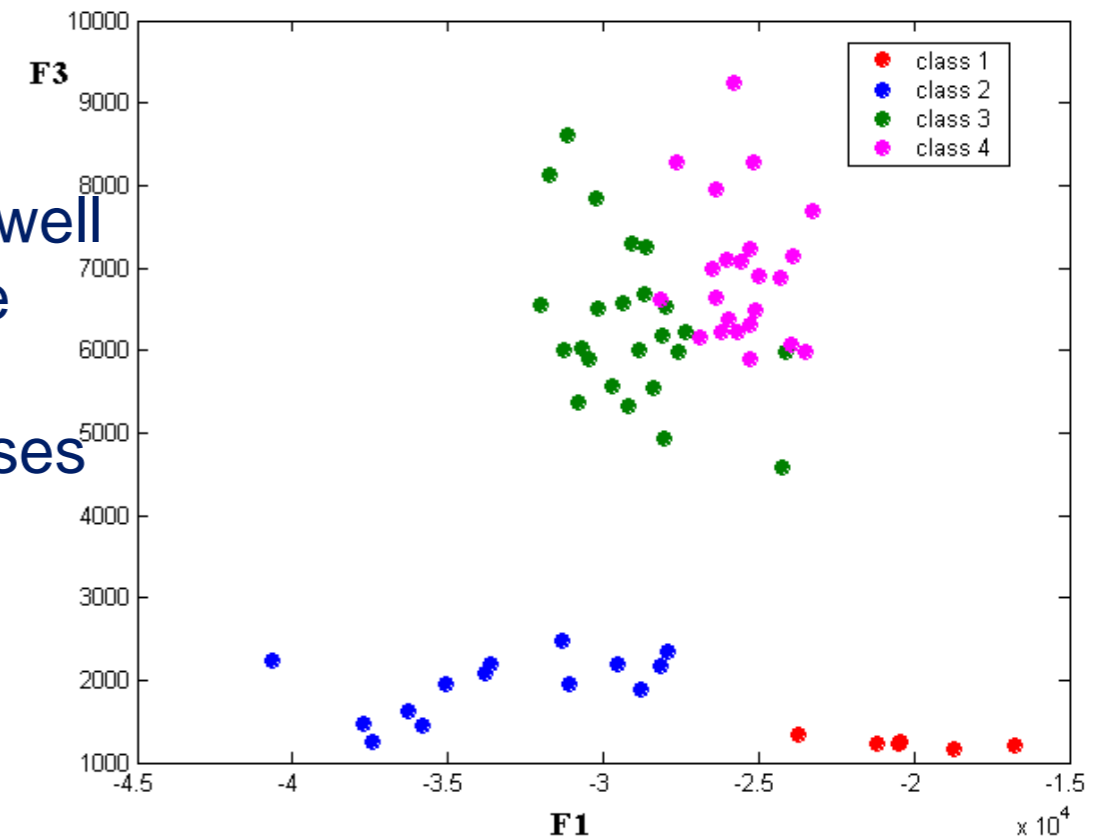
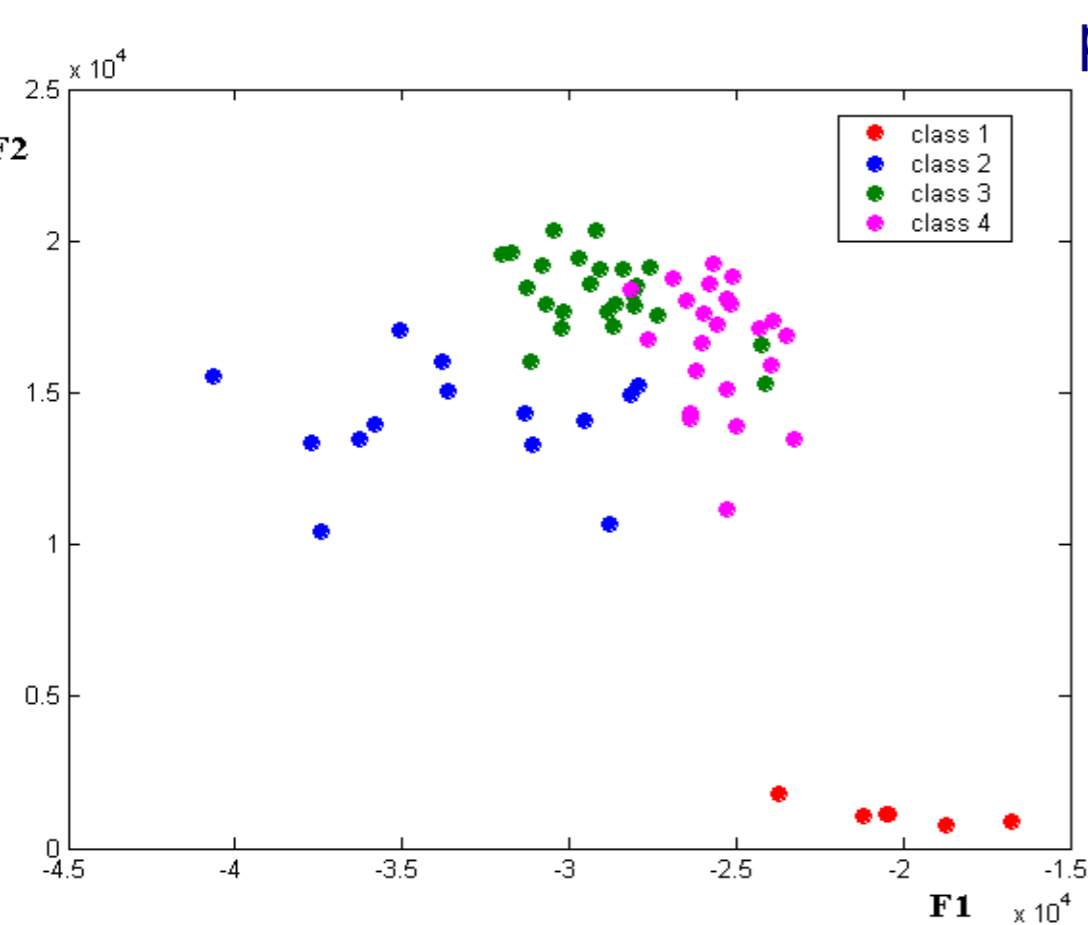
Parma Ham Data set

Jens K.S. Møller et al. *J. Agr. Food. Chem.* **2003** (51), 1224 evaluated surface autofluorescence spectroscopy in order to measure age-related quality index of Parma ham during processing.

Data Array consists of: 67x13x11 (samples x emission x 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

analysis

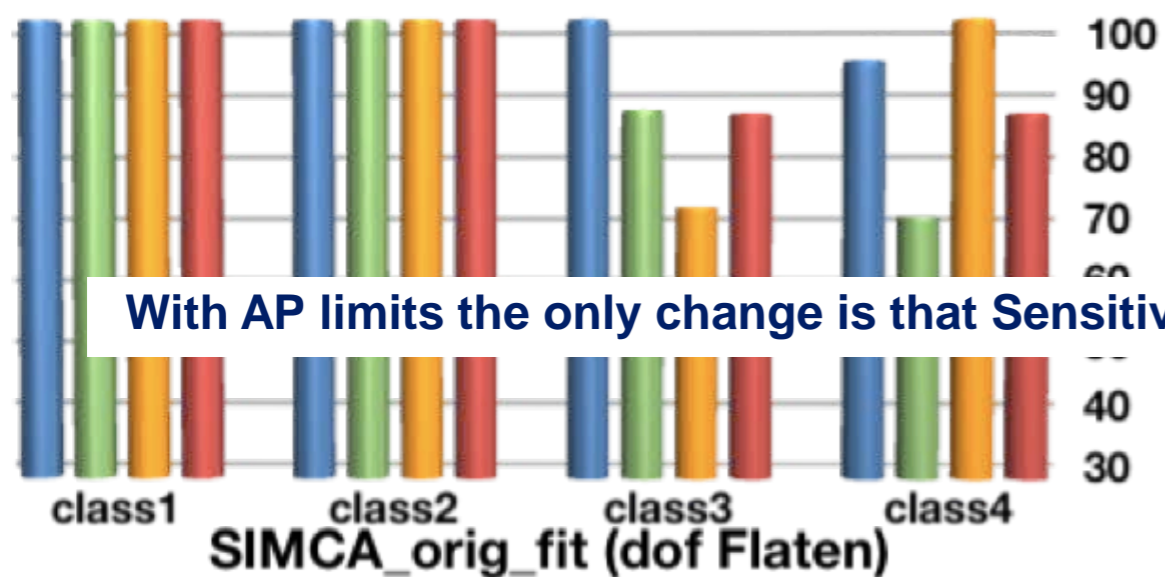


To choose the number of factors to explore in NSIMCA pftest was applied to each separate class: **Fac max [1 4 4 5]**

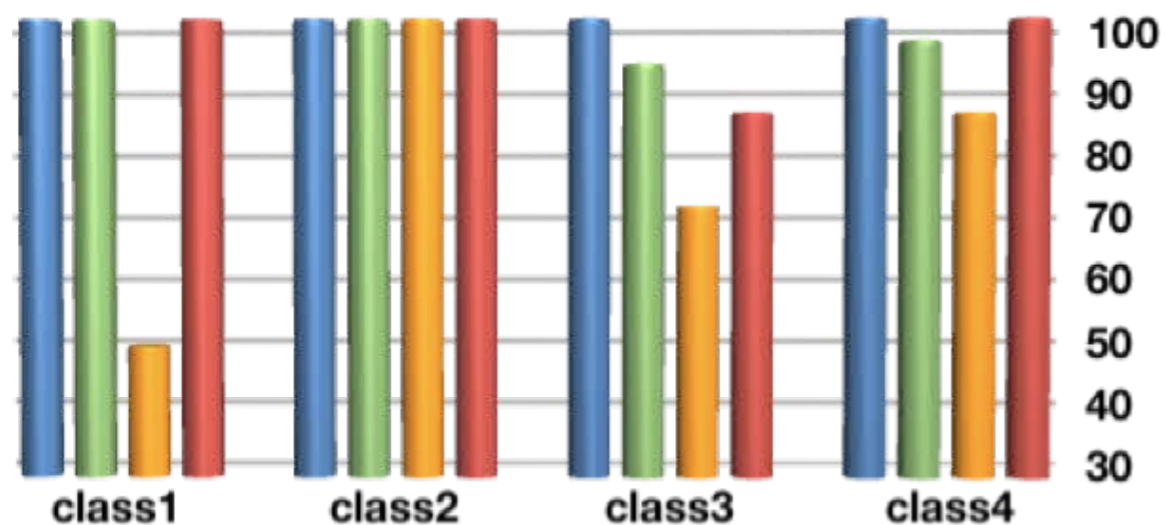
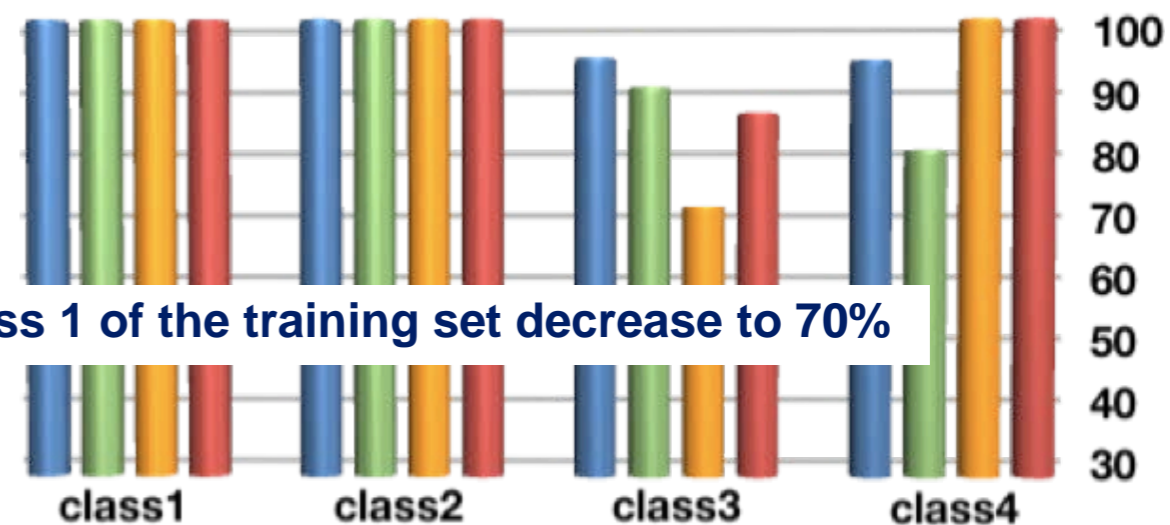
Parma Ham

Hlim_fit, Qlim_fit

■ SENS ■ SPEC ■ SENS TEST ■ SPEC TEST

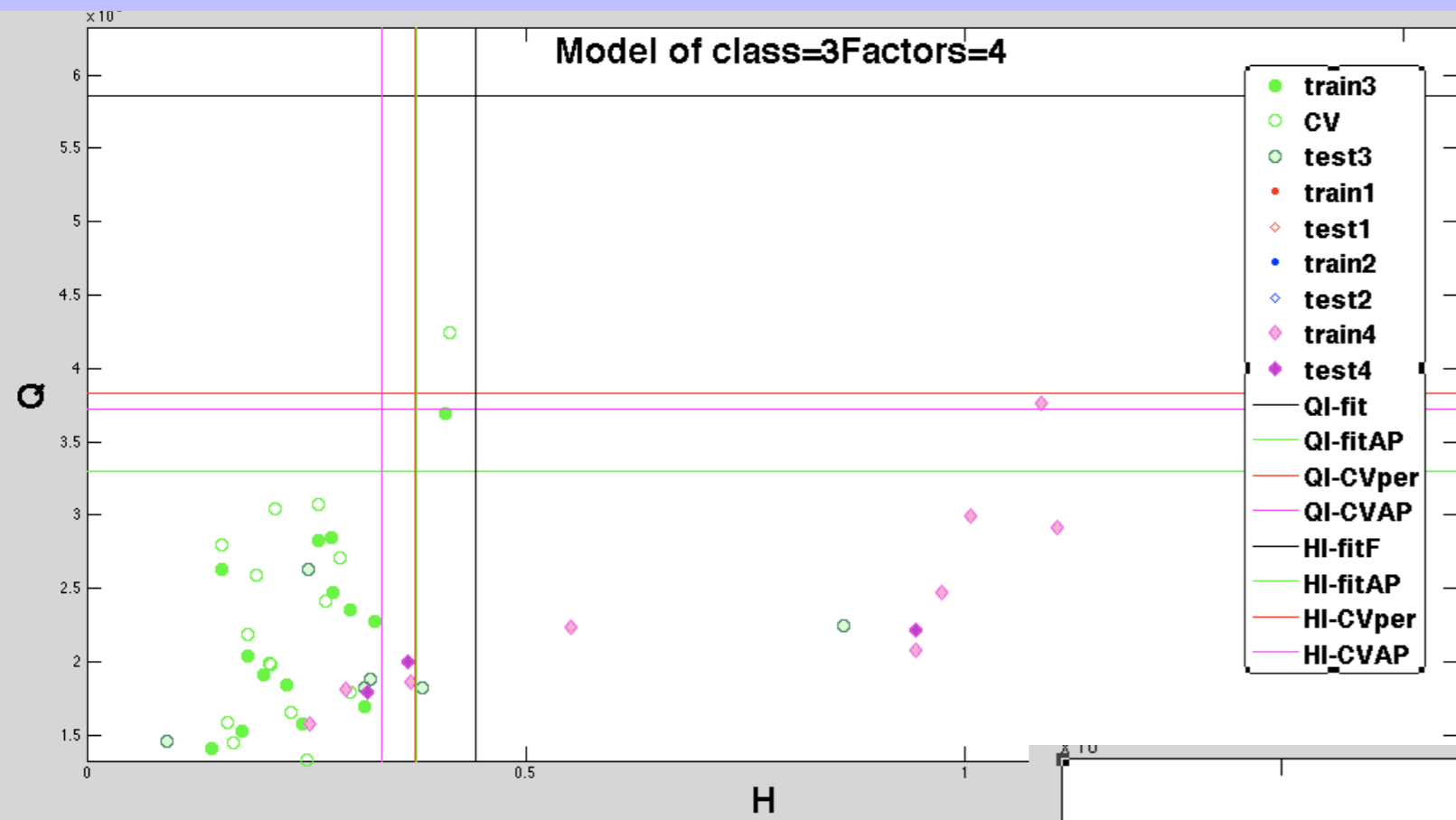


Hlim_95%, Qlim_95%

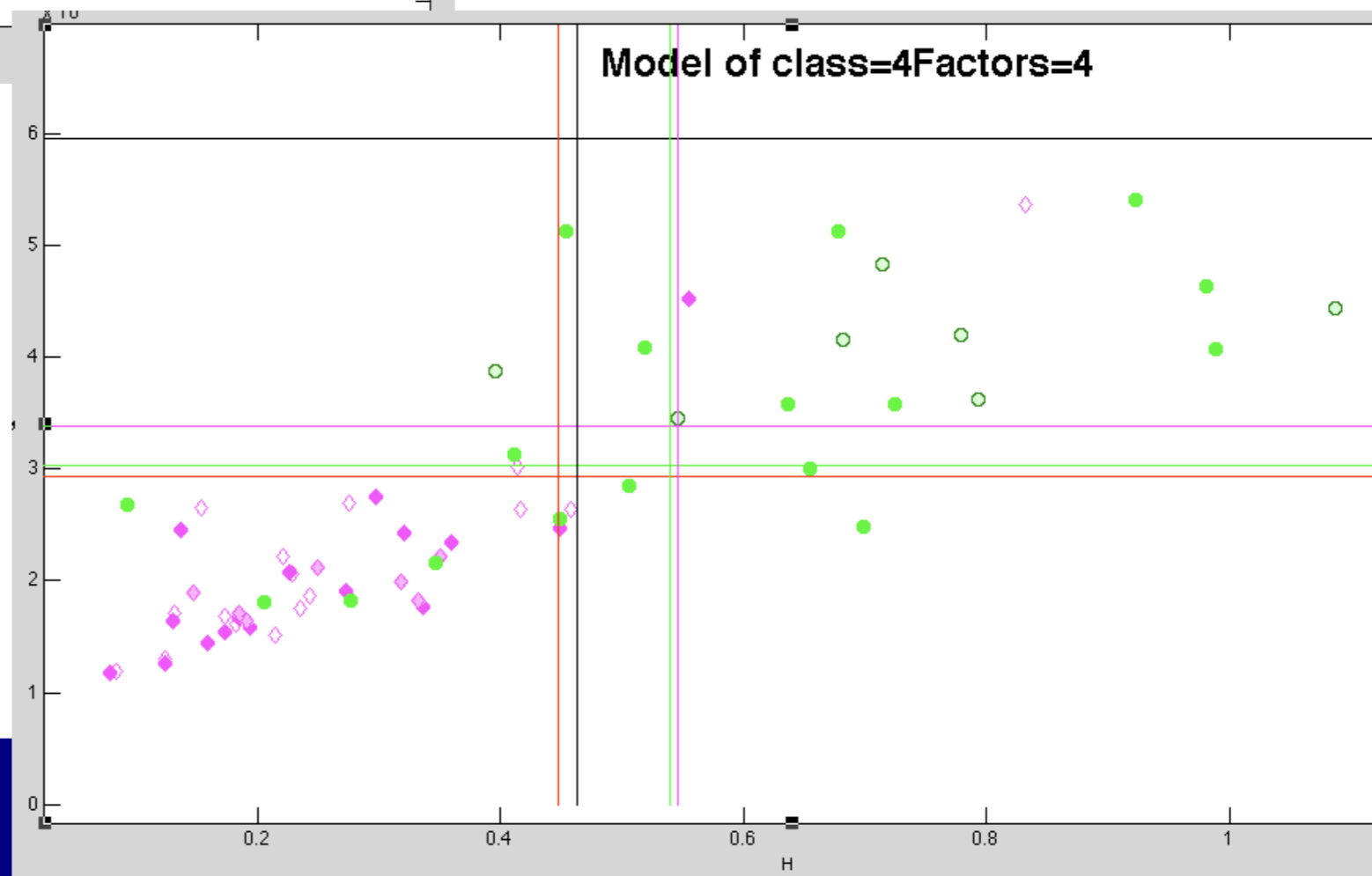


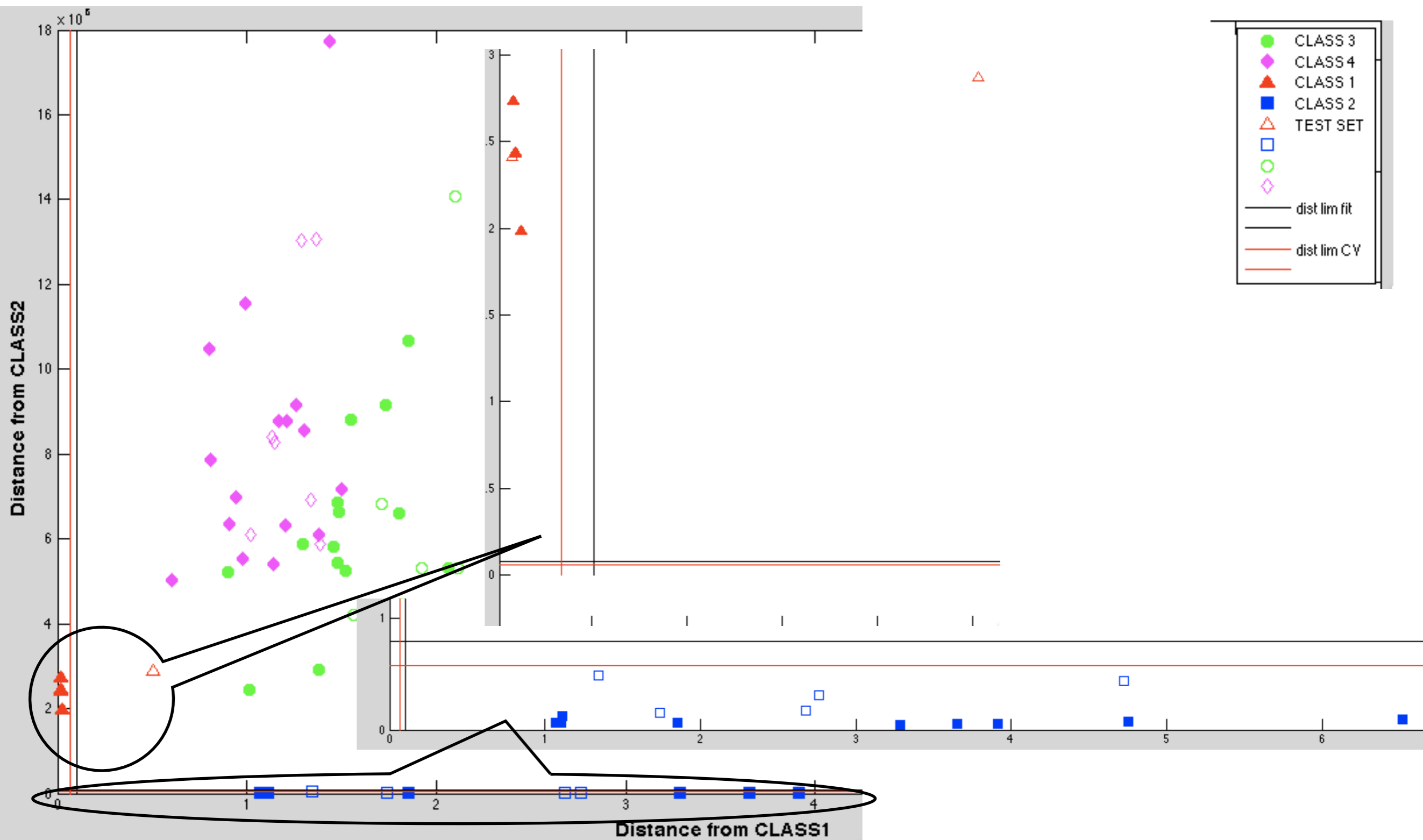
With SIMCA orig CV and with dof calculated as: data entry - free parameters, almost same the results; only the Sensitivity for class 3 of the test set decrease to 57%

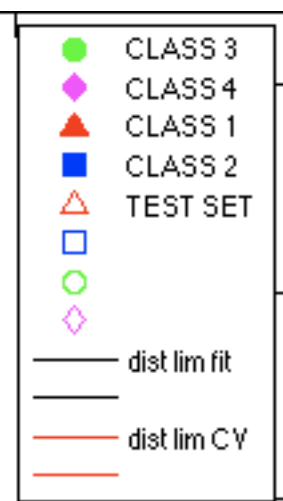
Parma Ham



only few mistakes from overlapping classes







Extra Vergin Oliv Oil (EVOO) Data set

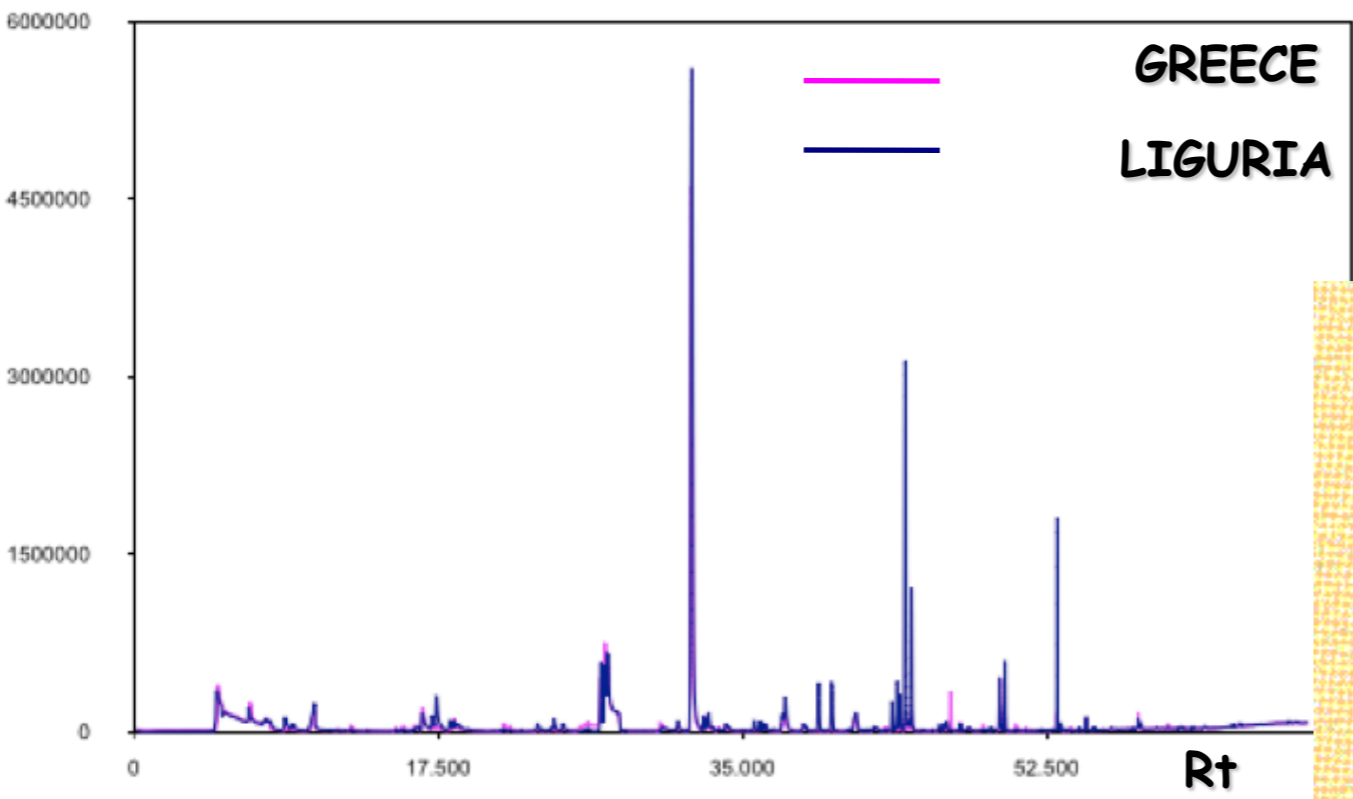
Original, not yet published data

- ▶ Recognition of the LIGURIAN products (PDO denomination) high added value from other Italian and Foreign olive oils (typical Class Modeling Issue)

Data Array consists of: $n \times 1514 \times 150$ (samples \times GC \times MS)

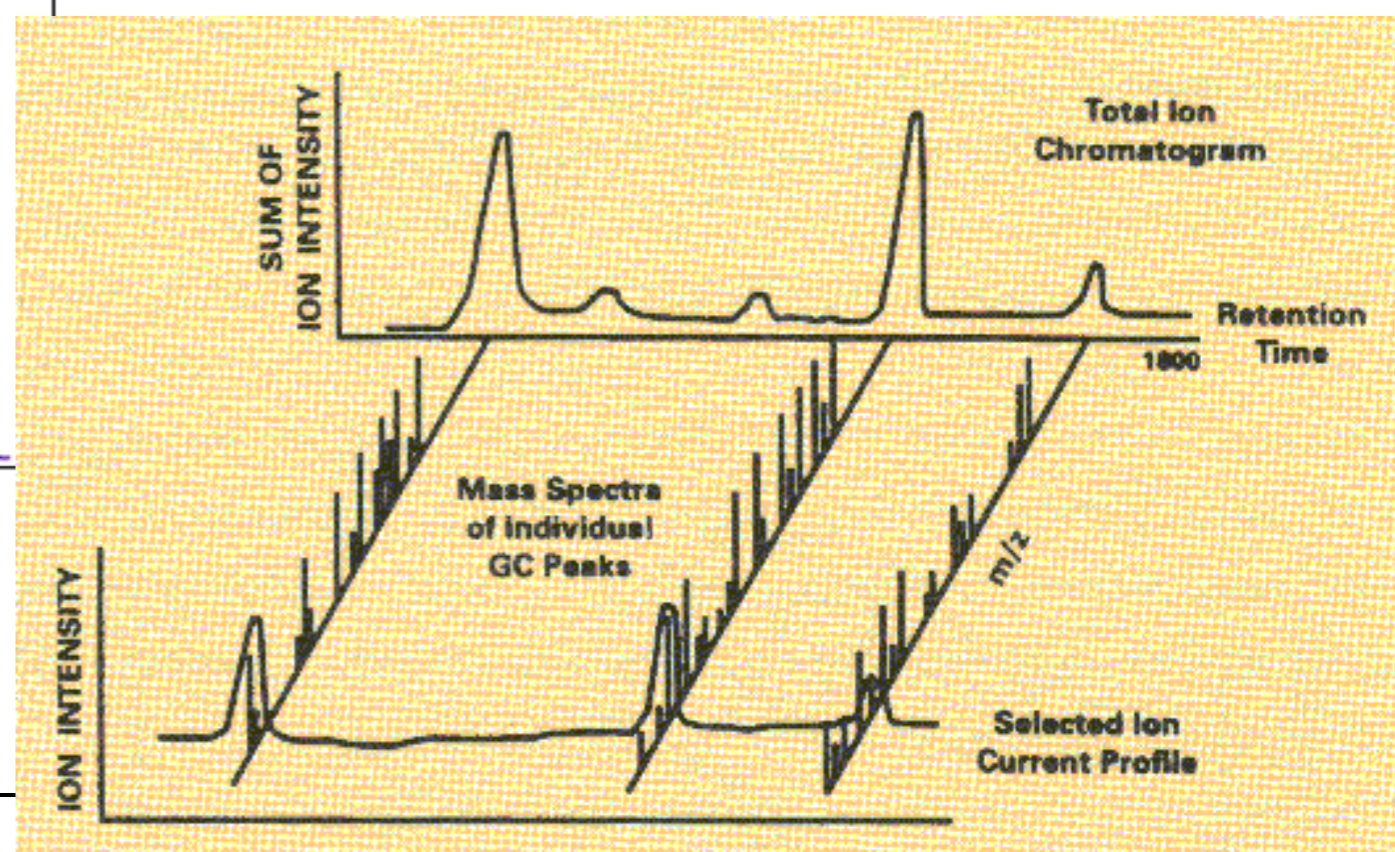
ORIGIN	Characteristics	N° sample
LIGURIA	All samples belong to the same PDO but comes from two different areas both in the Liguria region	24
APULIA	Mixture of different variety.	25
FOREIGN (Greece, Spain, Tunisia)	Different zones	31

TOTAL ION CHROMATOGRAM

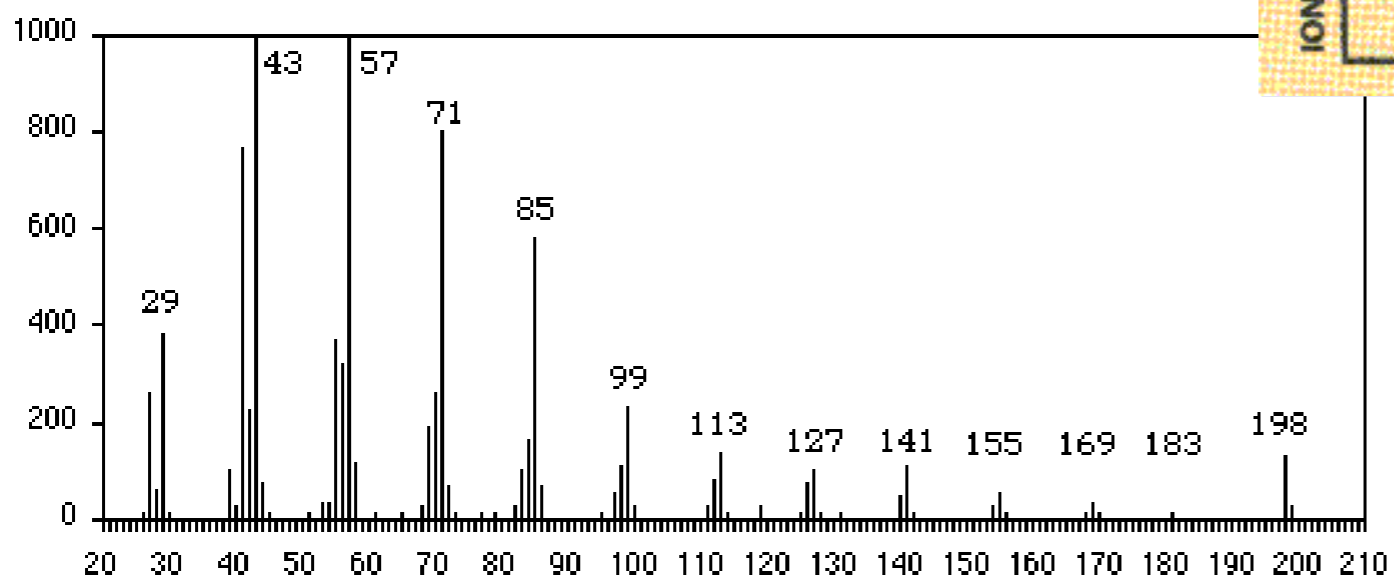


HS-SPME/GC-MS Signals

Fiber: DVB-CARBOXEN-PDMS



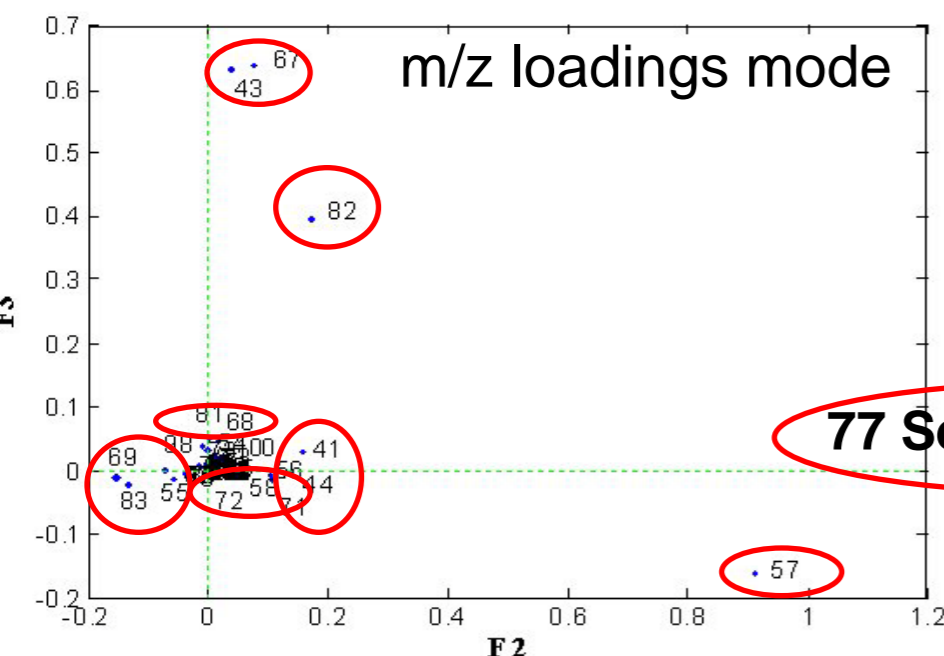
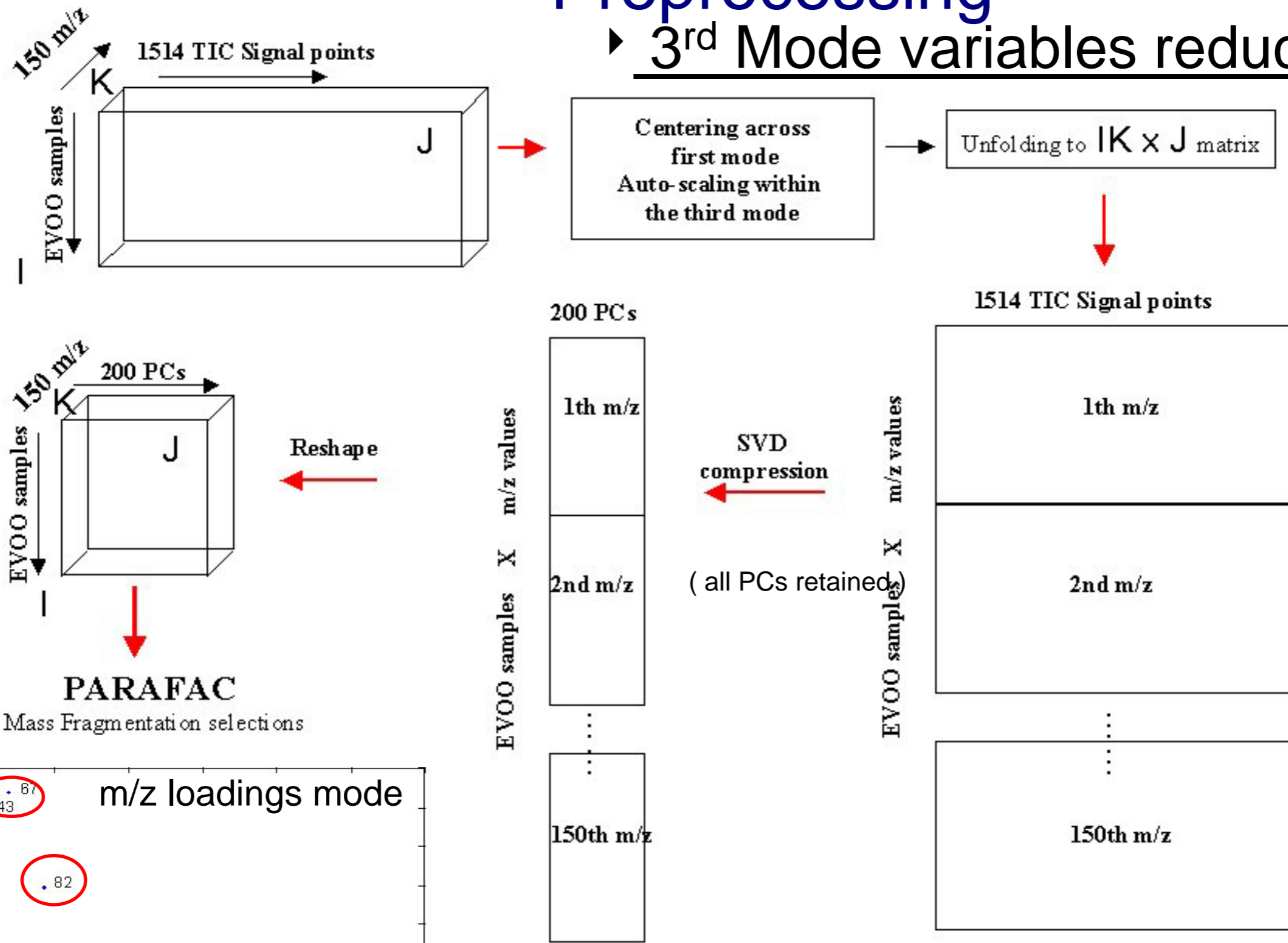
MASS SPECTRUM



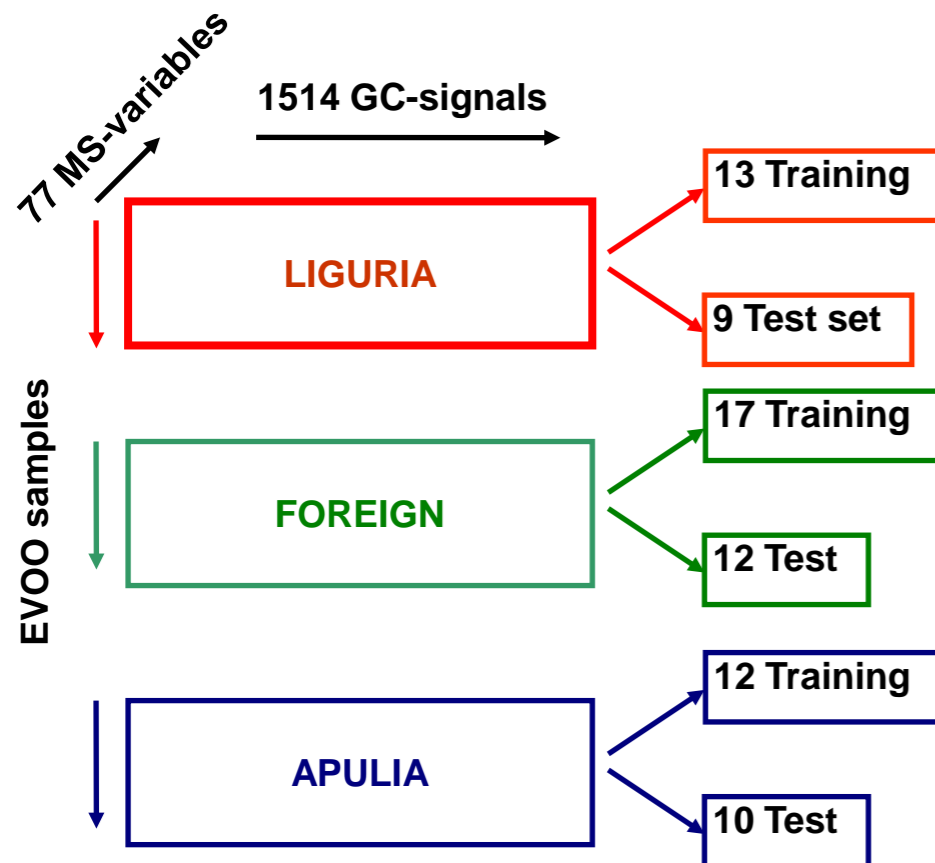
EVOO Data set

Preprocessing

3rd Mode variables reduction*



* For computational reason (too much memory required)

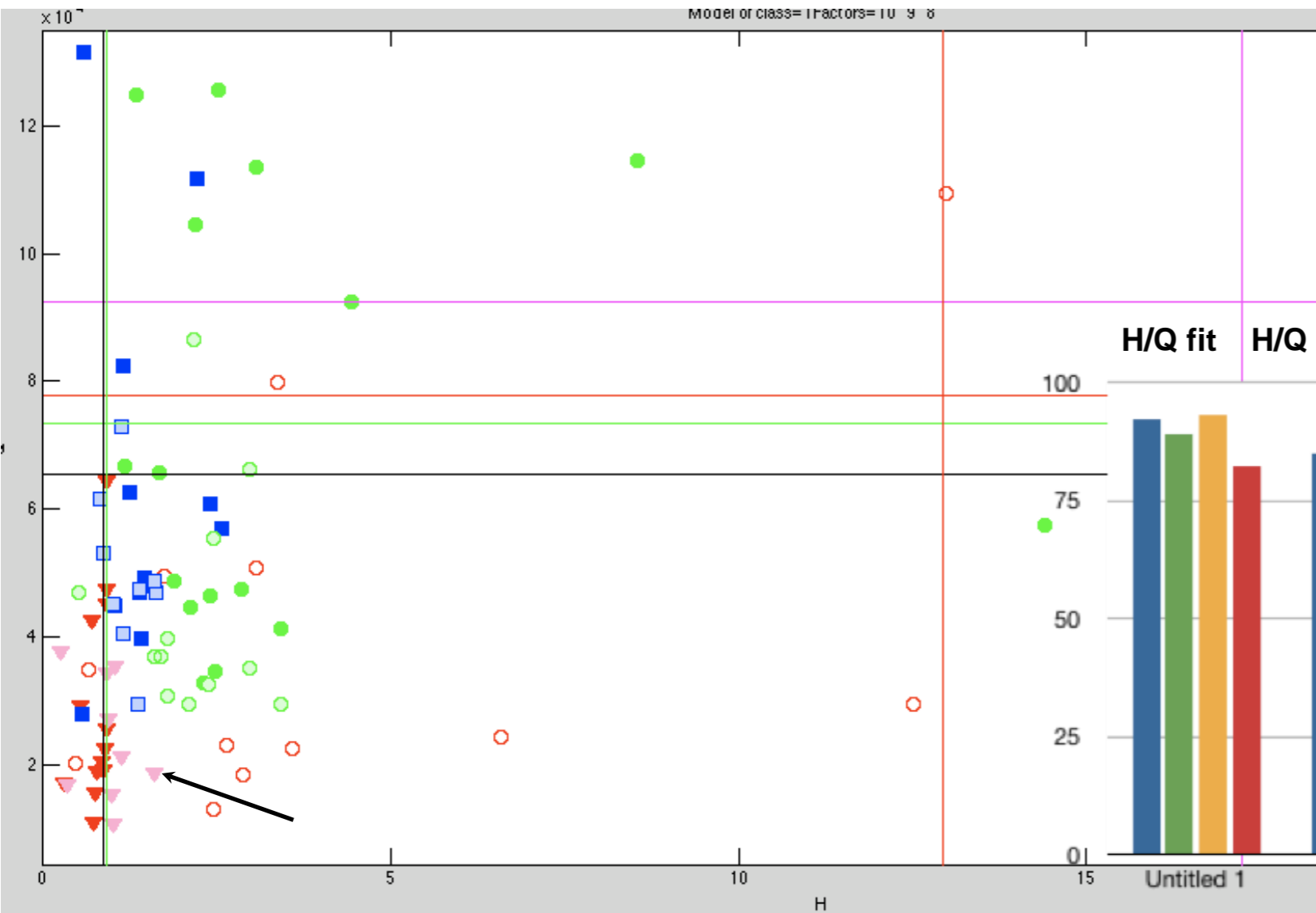


Ⓢ PRETREATMENT :

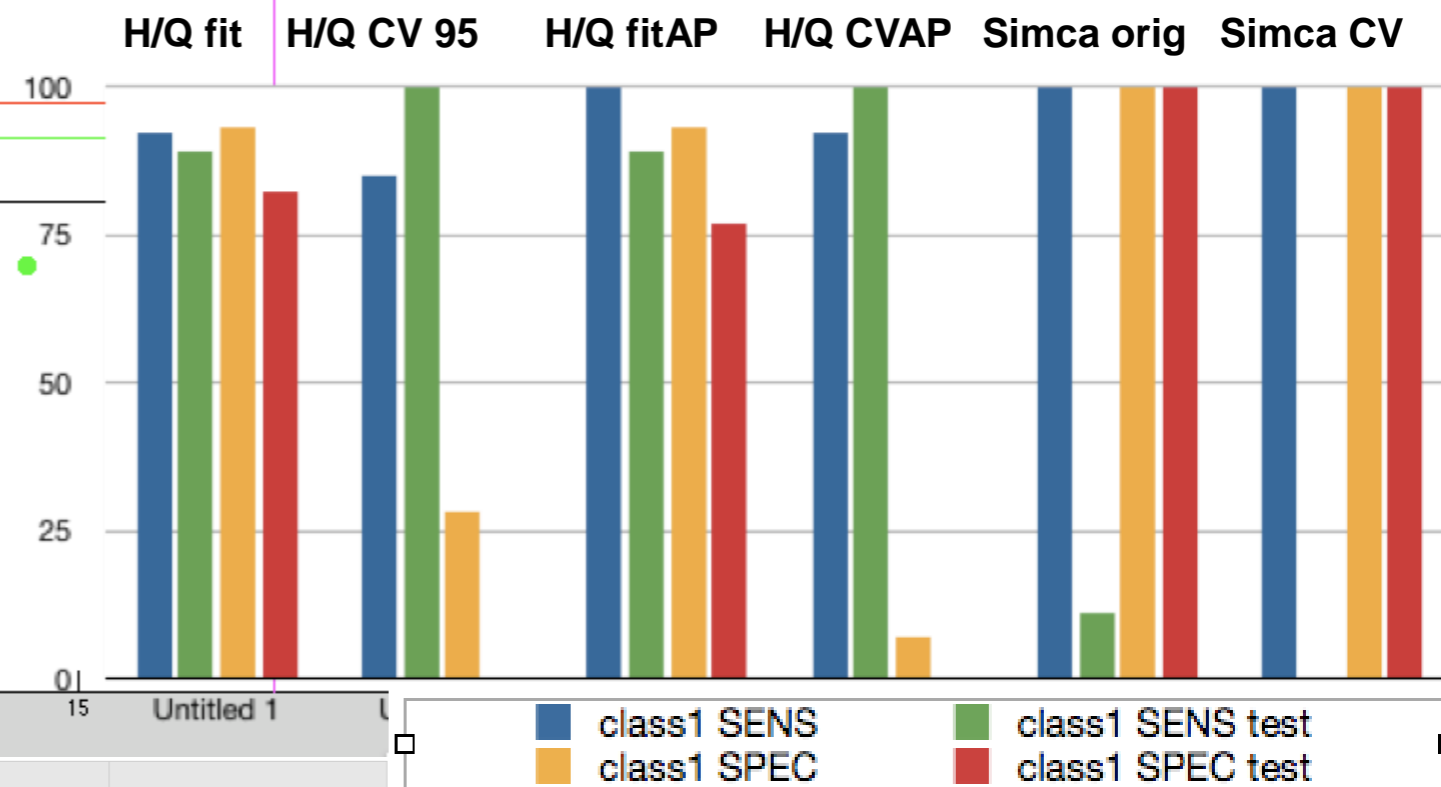
- Ⓢ Autoscaling within the third mode
 - Centering across the first mode
 - Block-scaling within the second mode
-
- Ⓢ random split in training and test sets for each class
 - Ⓢ exploratory single class data analysis (TUCKTEST) for choosing number of factors to explore

EVOO Data set

NSIMCA analysis



▶ in this case D lim were much worst



LIGURIA [10 9 8]	H/Q Fit	H/Q CV 95%	H/Q Fit AP	H/Q CV AP	SIMCA_orig	SIMCA_origCV
SENS	92 (12/13)	85	100	92	100	100
SPEC	93 (27/29)	28	93	7	100	100
SENS test	89 (8/9)	100	89	100	11	0
SPEC test	82 (18/22)	0	77	0	100	100

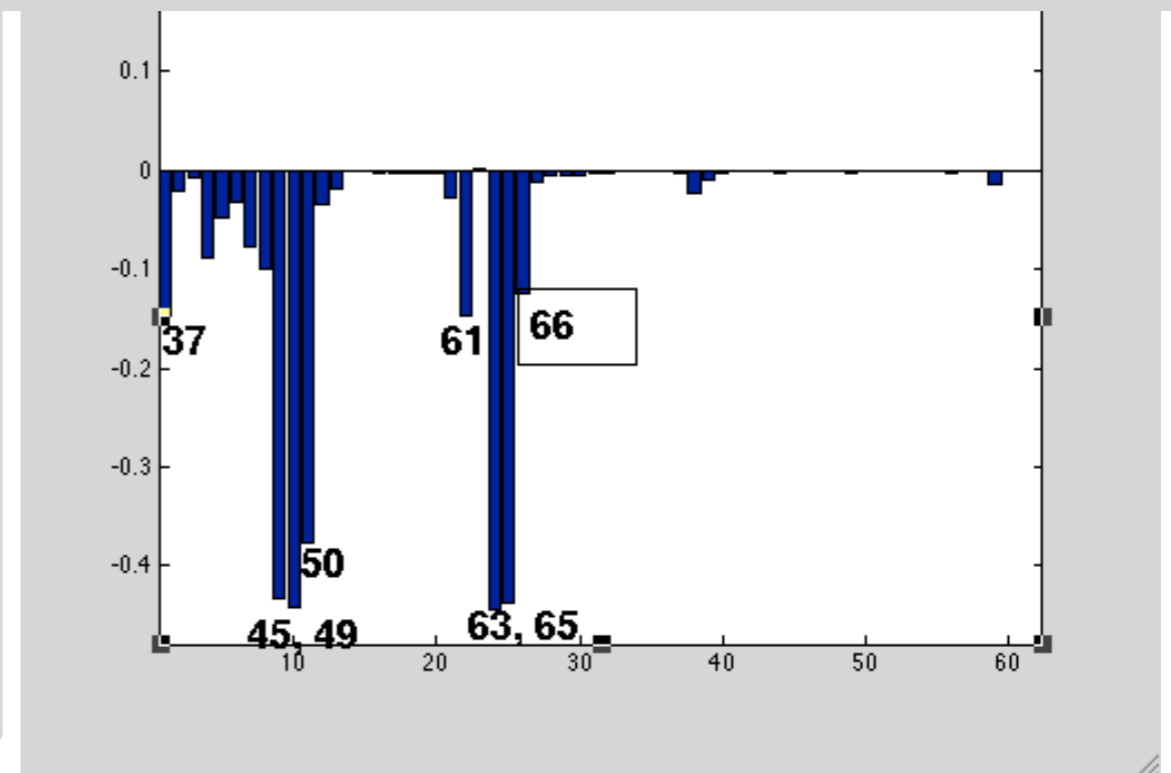
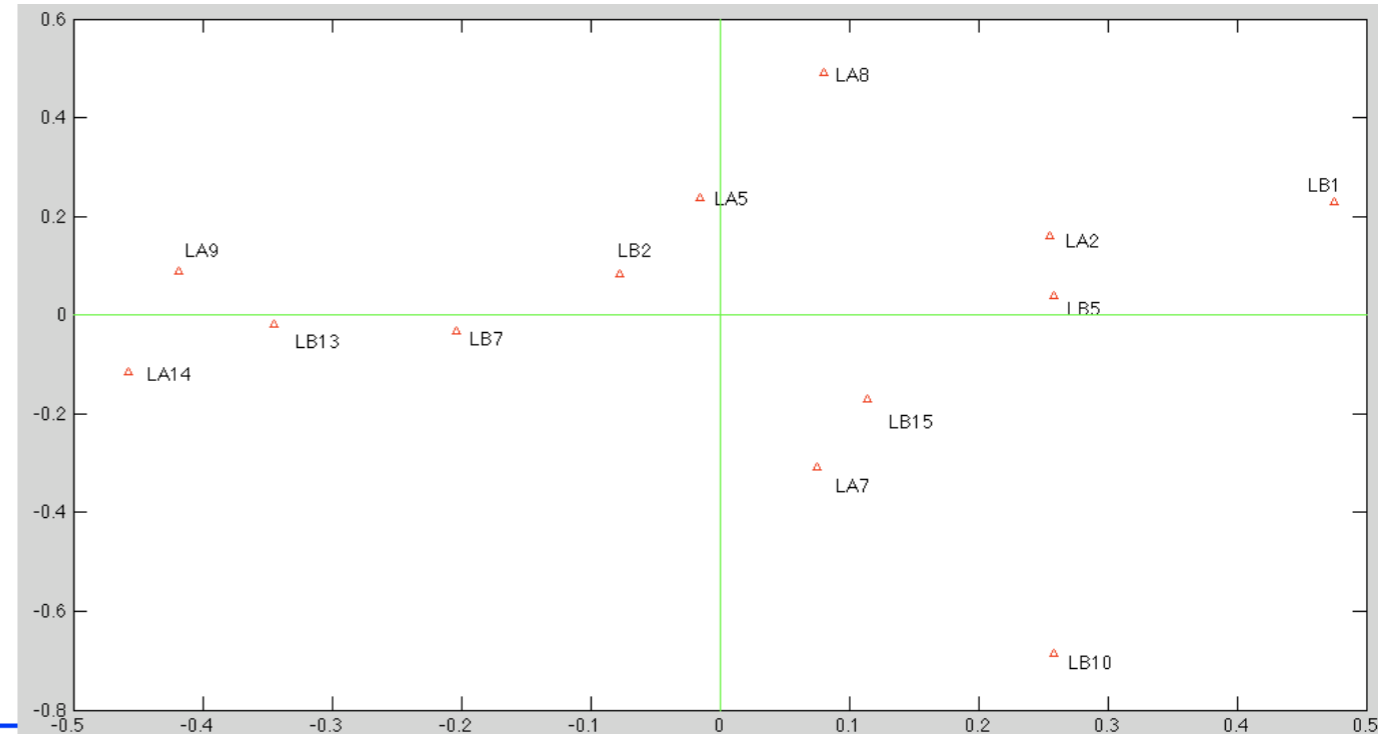
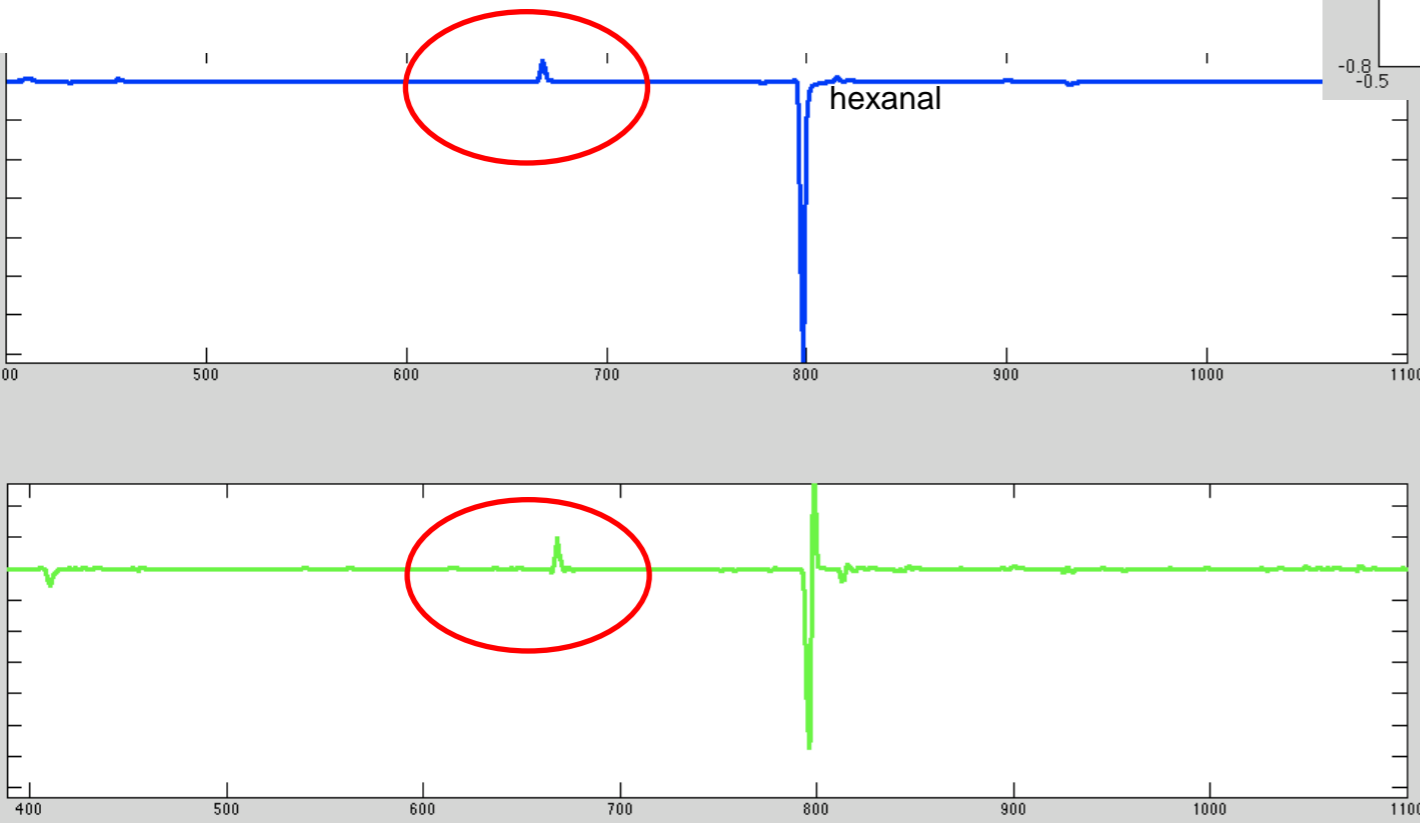
EVOO Data set

NSIMCA analysis

core

1	(1, 1, 1)	21.77464%	-250.92303
2	(2, 2, 1)	9.60127%	166.62087
3	(2, 1, 1)	7.23106%	144.59926

- Hexanal common to oliv oil but typical voc of Ligurian oliv oil appear
- there may be Rt shift



Final remarks

Perspective

- ▶ sampling is a more critical issue than limits criteria
- ▶ choice of dimensionality as low RMSECV suboptimal with respect to SENS/SPEC compromise
- ▶ empirical CV 95% promising considering it is simple
- ▶ there may be cases where fit performs better than CV and viceversa independently of the way limits are estimated

- ▶ test on more data sets and applications needed
- ▶ code needs some refinement, availability at www.life.model.dk
- ▶ implement a parameter analogous to the variable discriminant power, for model interpretation
- ▶ test use of only a reference distribution for combined Q and D
- ▶ robust decomposition methods

THANKS FOR YOUR ATTENTION

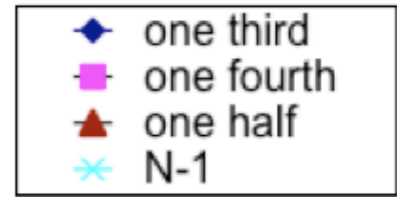
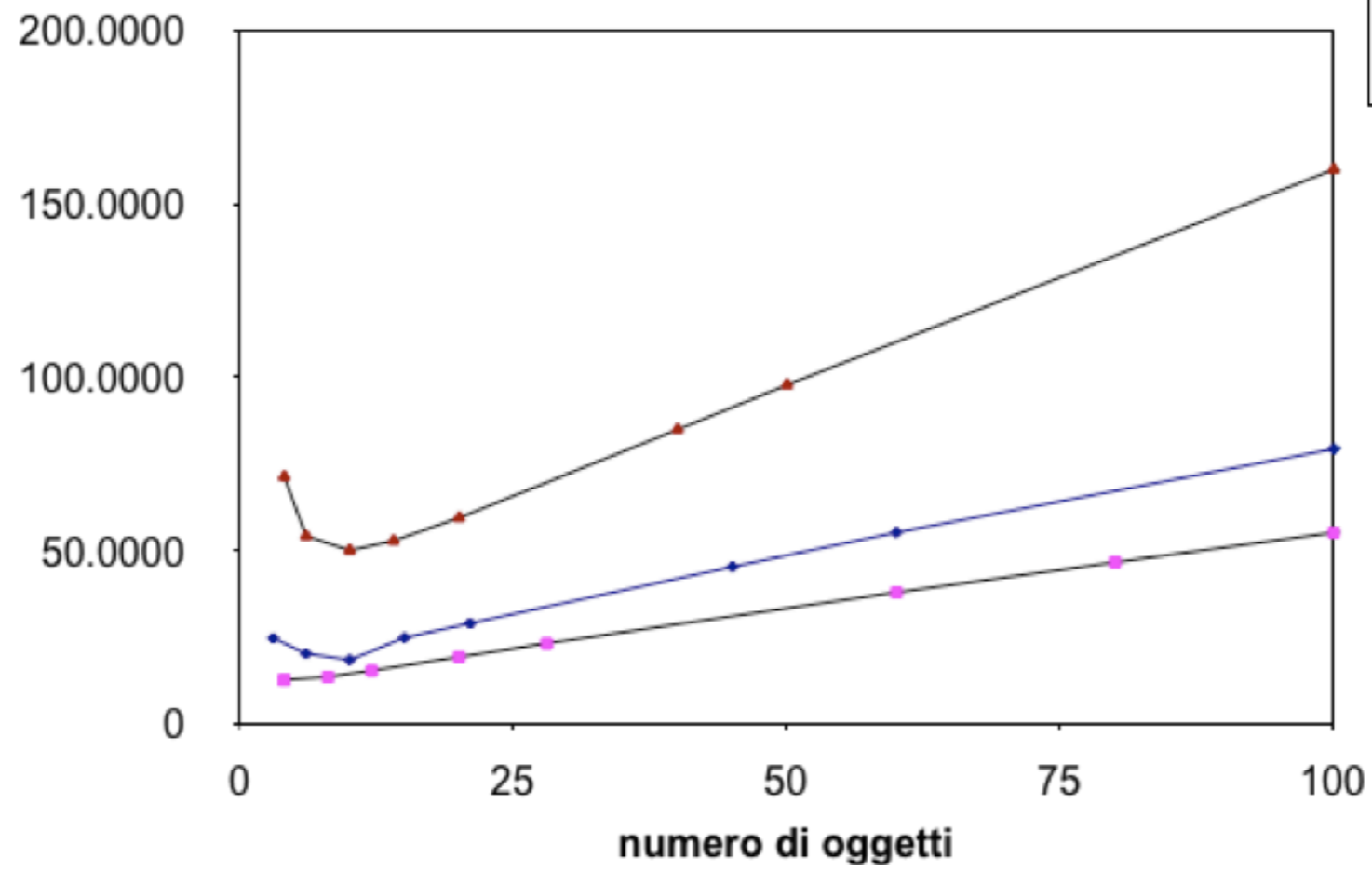
*and your writing interesting papers that
induced me more doubts than certainty*

Dubito ergo indago



Dim

ratio objs/LVs



Block-scaling within the second mode

