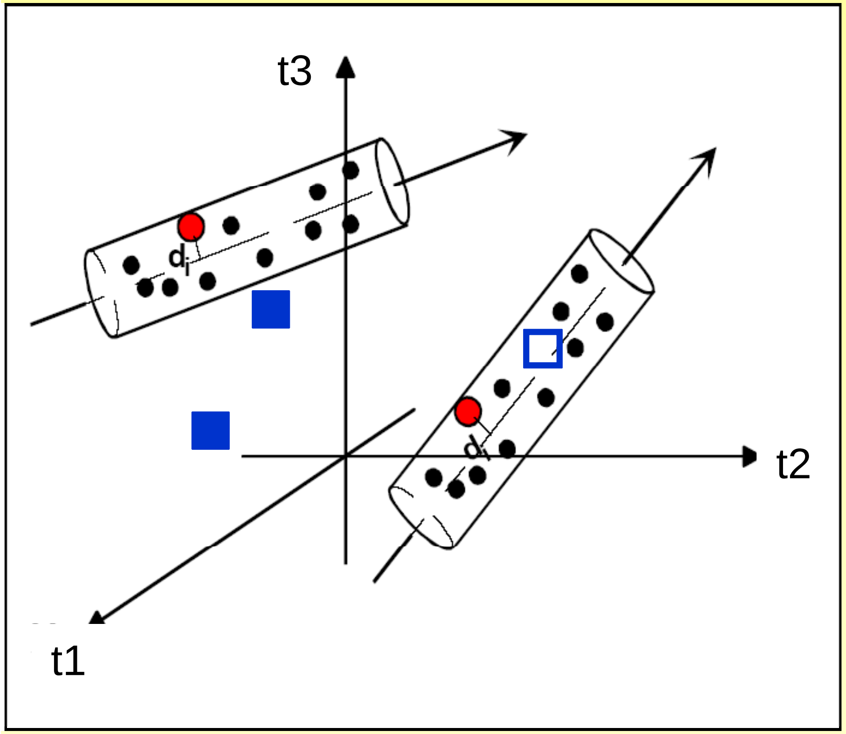
Robust SIMCA bearing on non-robust PCA

O. Rodionova, A. Pomerantsev,

Semenov Institute of Chemical Physics, RAS
Moscow

Russian Chemometric Society

SIMCA (Soft Independent Modeling of Class Analogy)



Classification in either of a number of predefined classes

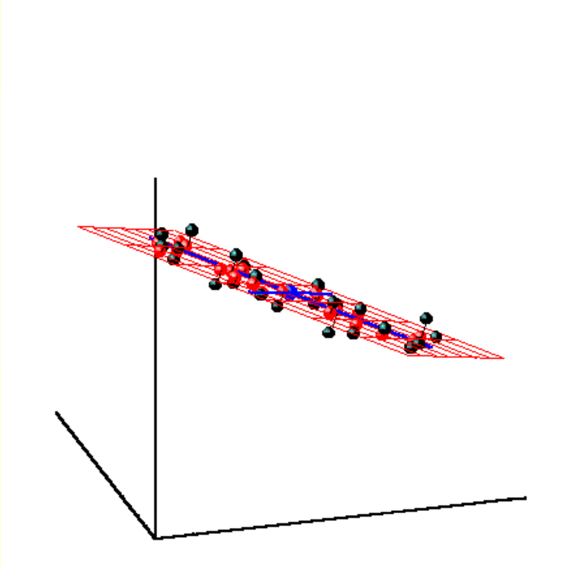
Outlier detection

Disjoint PCA class-modeling

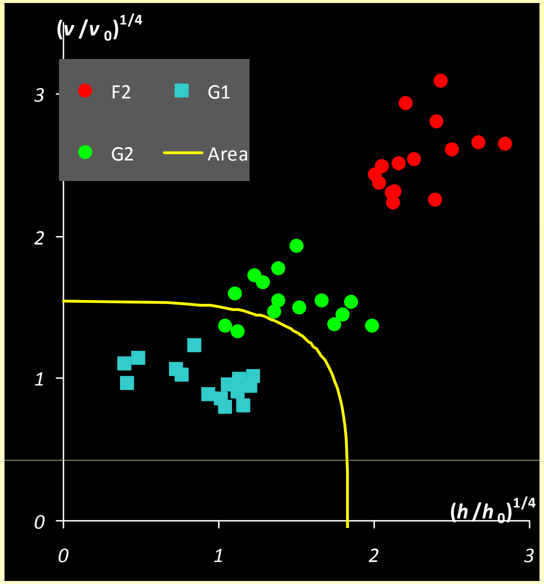
New object is compared with each class

S. Wold 1976

SIMCA: Main Steps

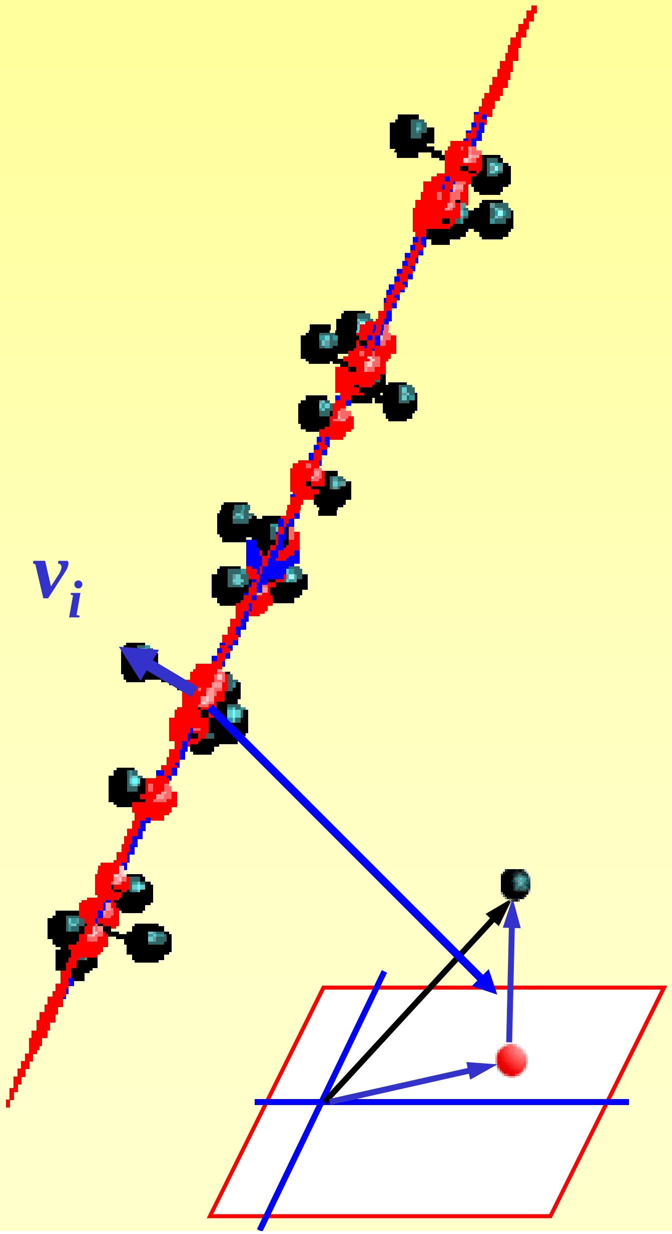


1-st step: Principal Component Analysis



2-nd step: Construction of the Acceptance Area

Orthogonal distance (OD), v_i



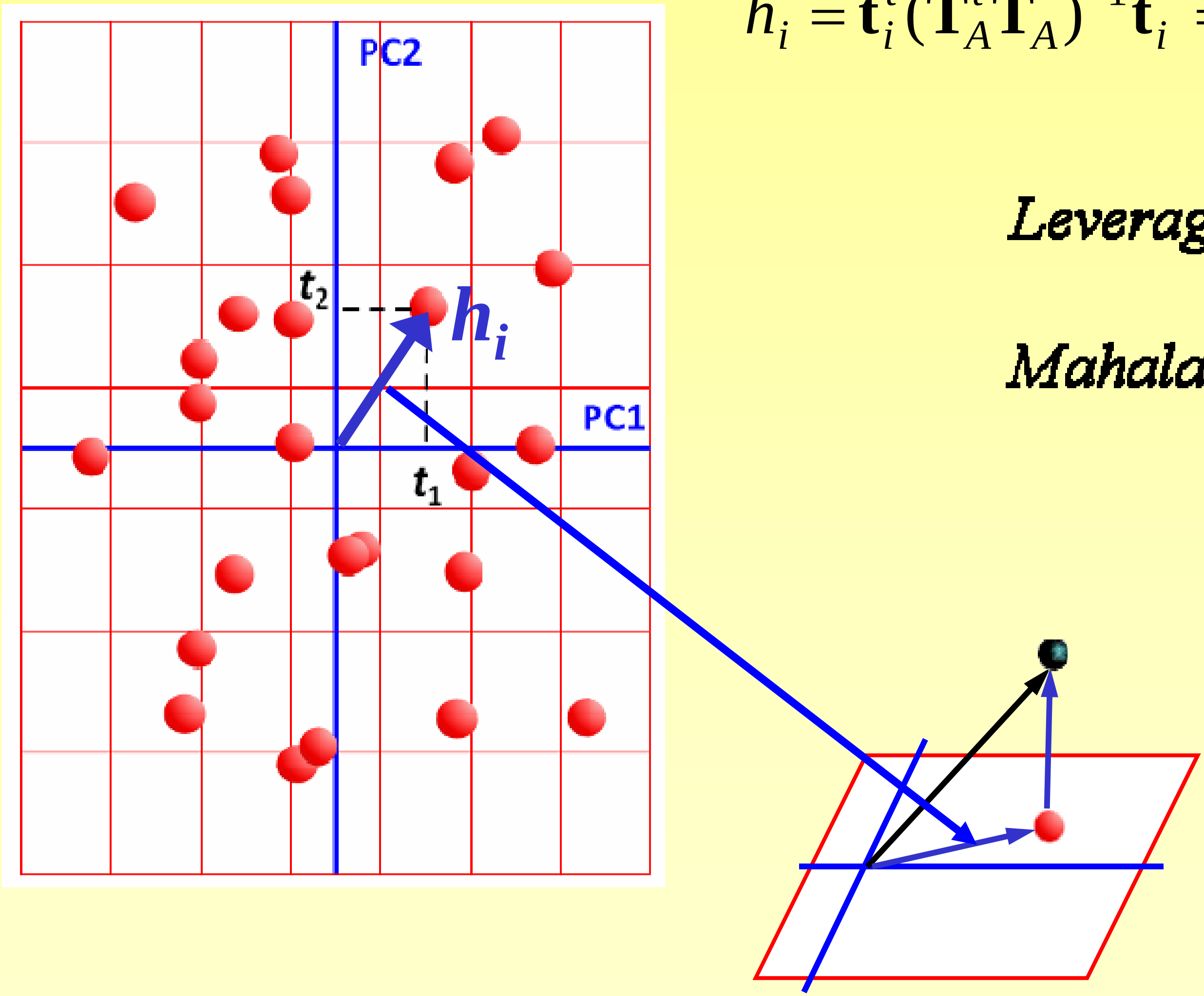
$$v_i = \sum_{j=1}^{J} e_{ij}^2 = \sum_{a=A+1}^{K} t_{ia}^2 = L_0 - \sum_{a=1}^{A} t_{ia}^2$$

Variance per sample $= v_i / J$

Q statistics $= v_i$

$$\mathrm{OD_i} = \sqrt{v_i}$$

Score distance (SD), h_i



$$h_i = \mathbf{t}_i^{\mathrm{t}}(\mathbf{T}_A^{\mathrm{t}}\mathbf{T}_A)^{-1}\mathbf{t}_i = \sum_{a=1}^{A}\frac{t_{ia}^2}{\lambda_a}, \quad i = 1,\ldots,I$$

Leverage = h_i+1/I

Mahalanobis = $(h_\mathrm{i})^{½}$

$$\mathrm{SD}_\mathrm{i} = \sqrt{h_i}$$

Acceptance areas

Calculated by the PCA decomposition

$$v/v_0 \sim \chi^2(N_v)/N_v$$

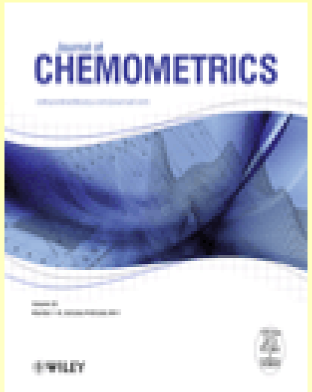
$$h/h_0 \sim \chi^2(N_h)/N_h$$

Estimated DoF

$$N_v\,,\ N_h$$

Set by a researcher

Type I Error = α



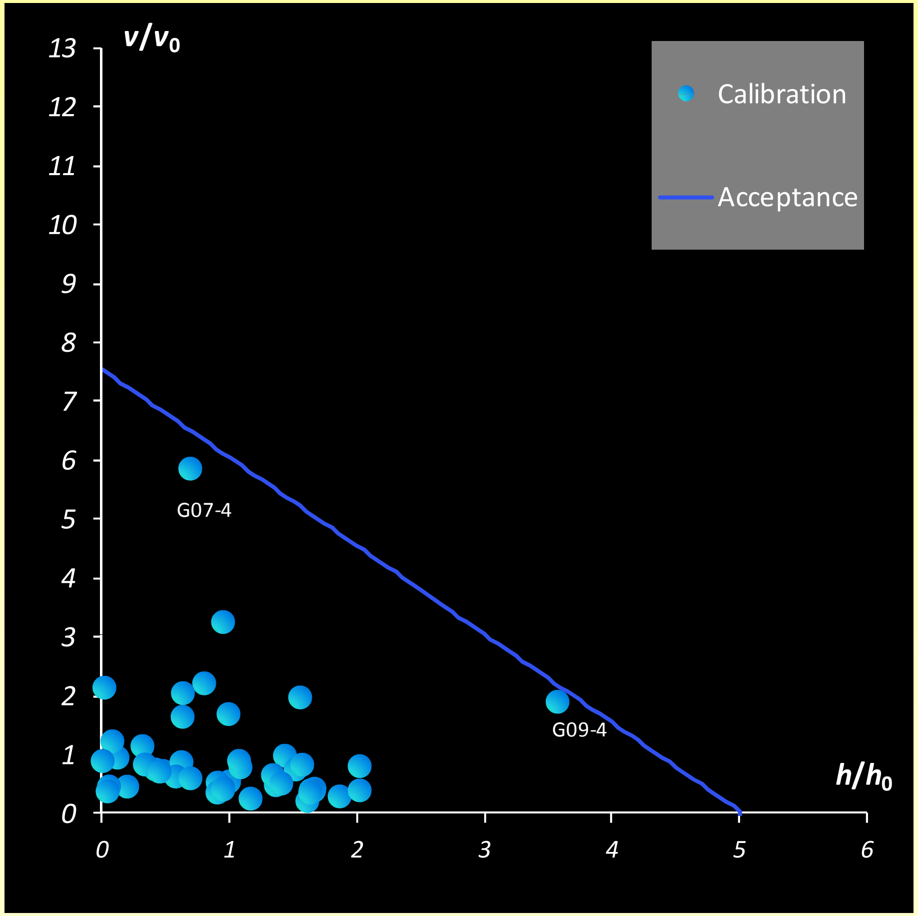
J. Chemometrics 2008; 22; A. Pomerantsev

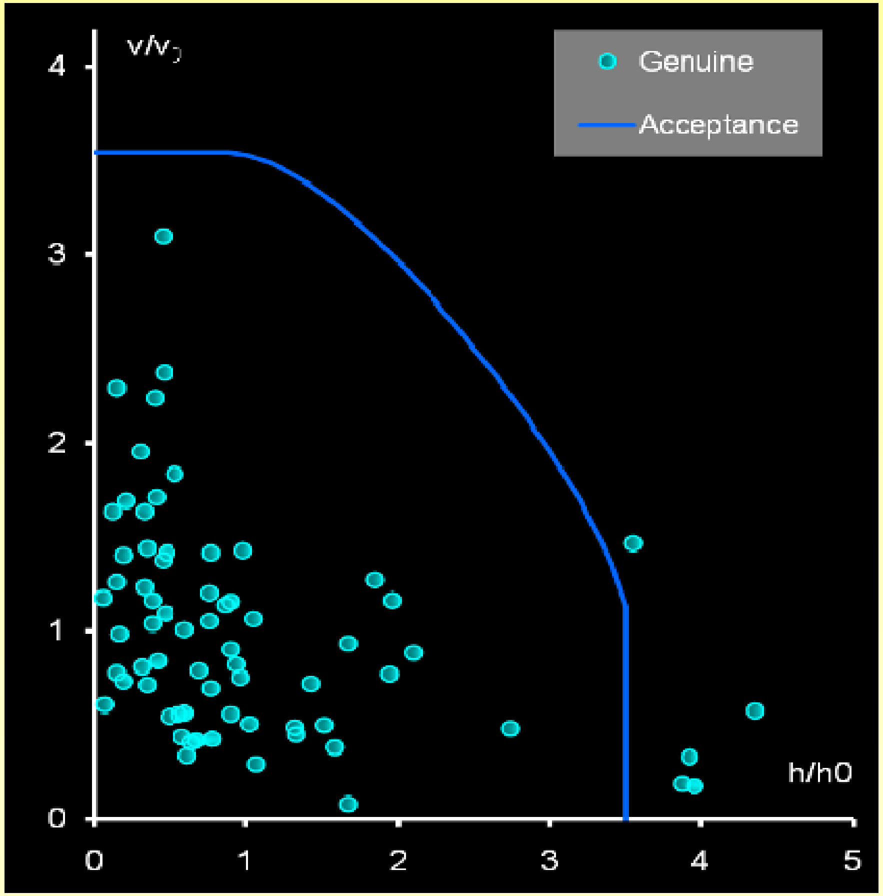
Acceptance areas for multivariate classification derived by projection methods

Acceptance areas

$$N_h \frac{h}{h_0} + N_v \frac{v}{v_0} \sim \chi^2 (N_h + N_v)$$

Modified Wilson-Hilferty approximation for χ^2

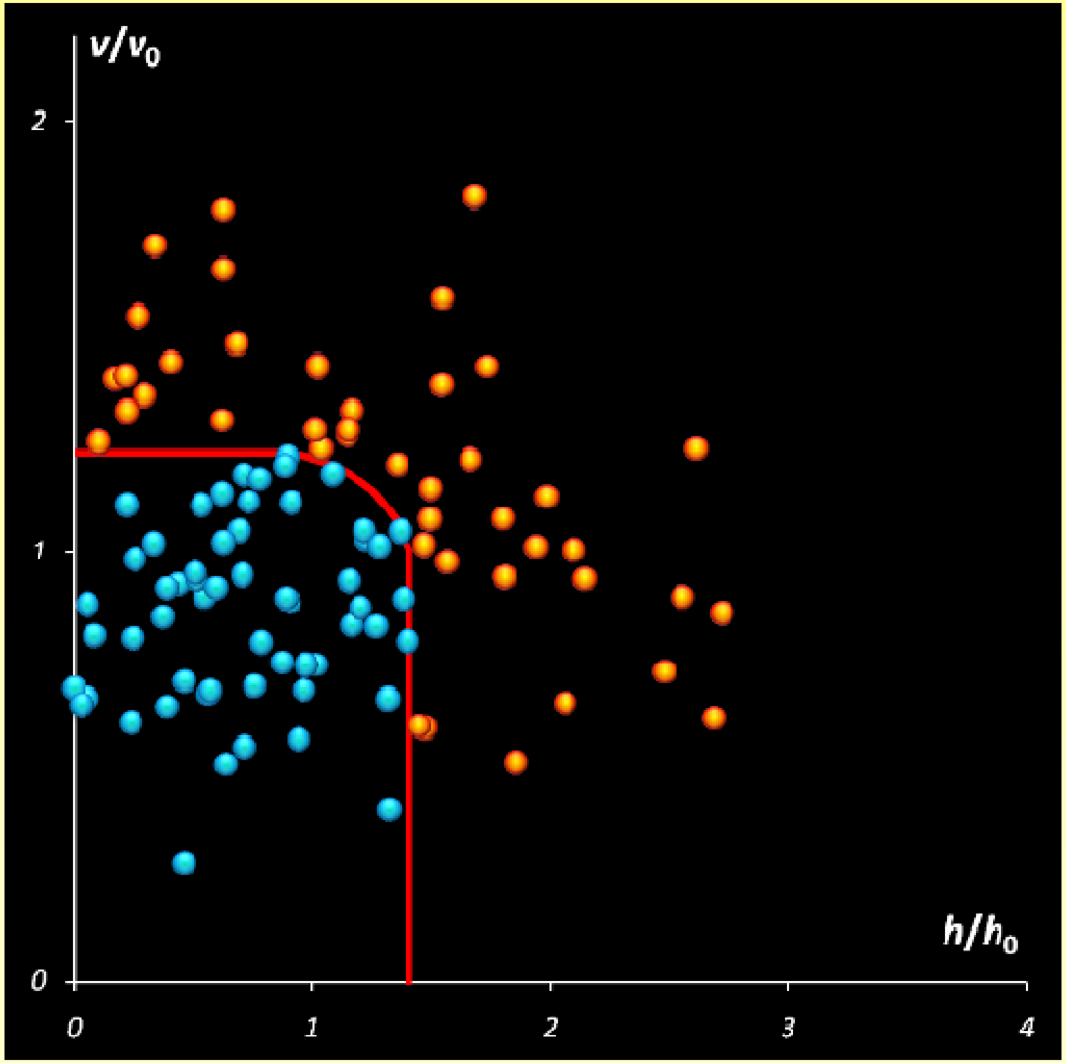




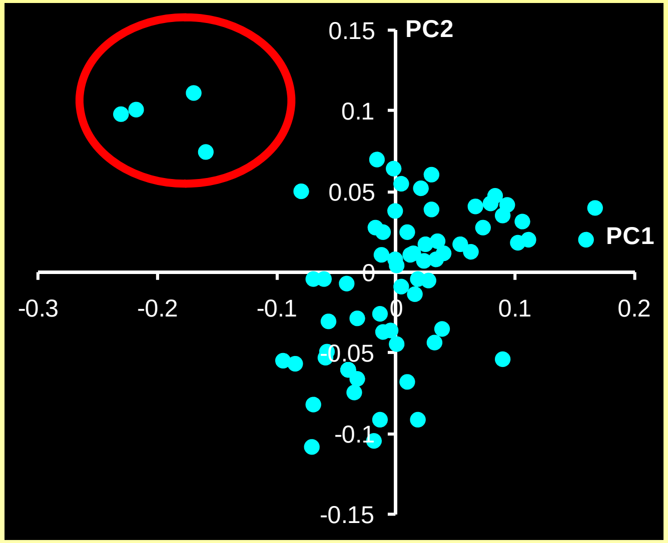
Type I error α. *I*=100

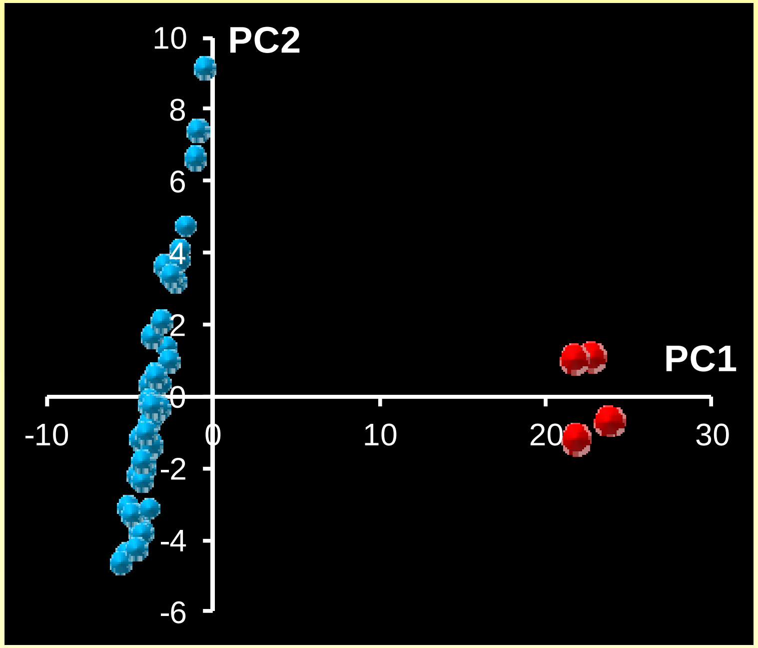
α=0.4

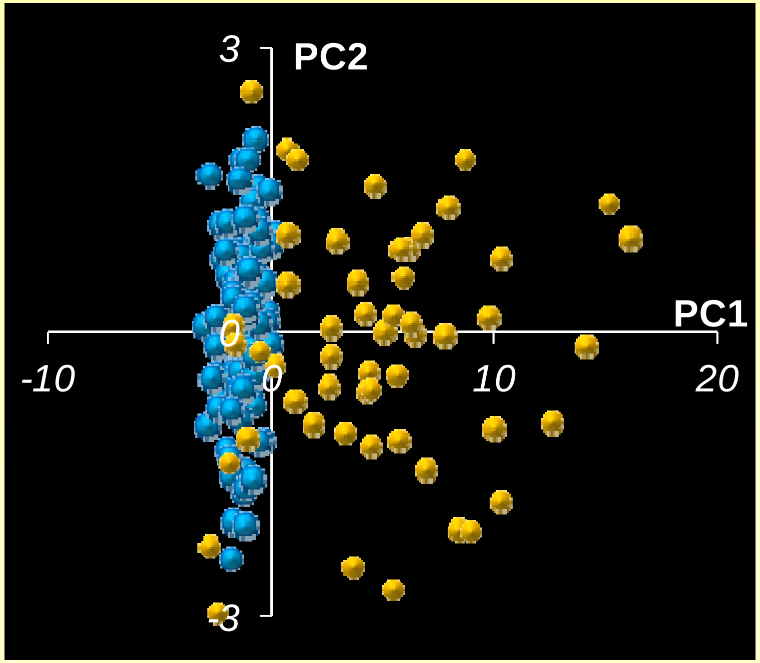
OUT
43 object

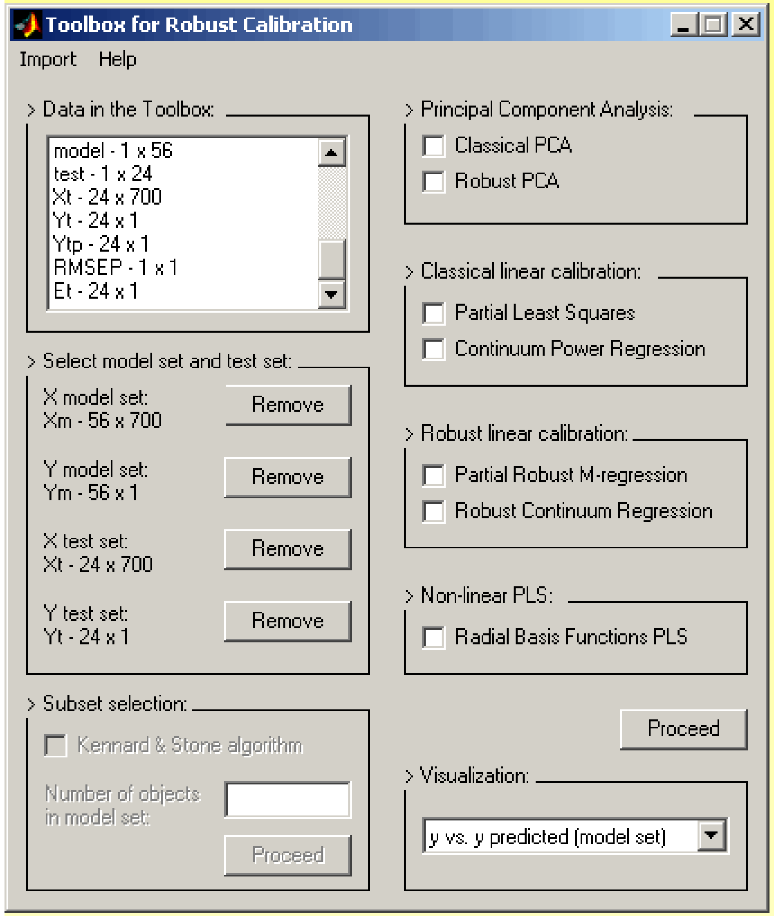


Abnormal observations









TOMCAT ToolBox

http://chemometria.us.edu.pl/RobustToolbox/

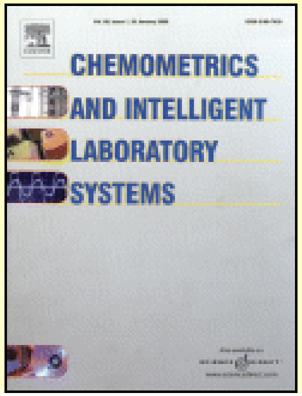
Chemometric Research Group, The University of Silesia

1. Robust PCA
robust PCs, robust singular values

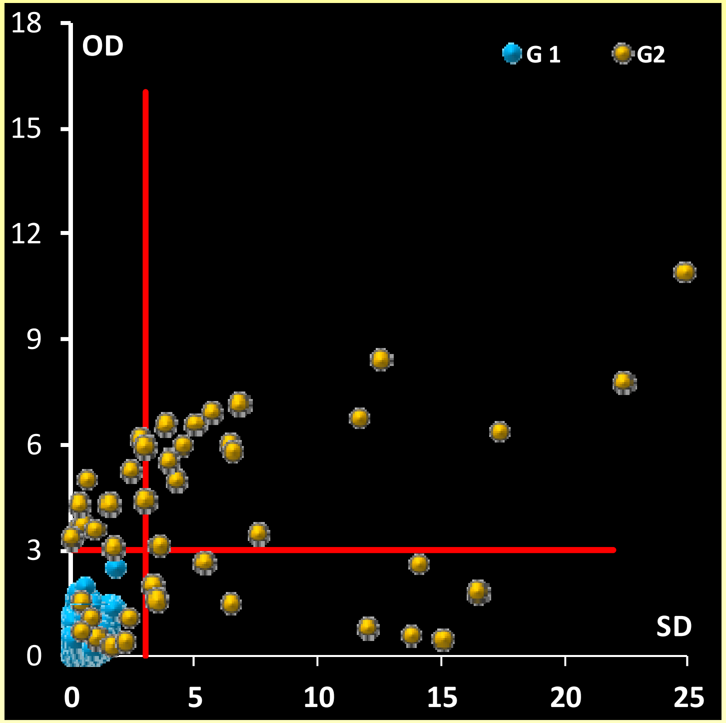
2. Robust classification rules
z-transformed robust OD and SD

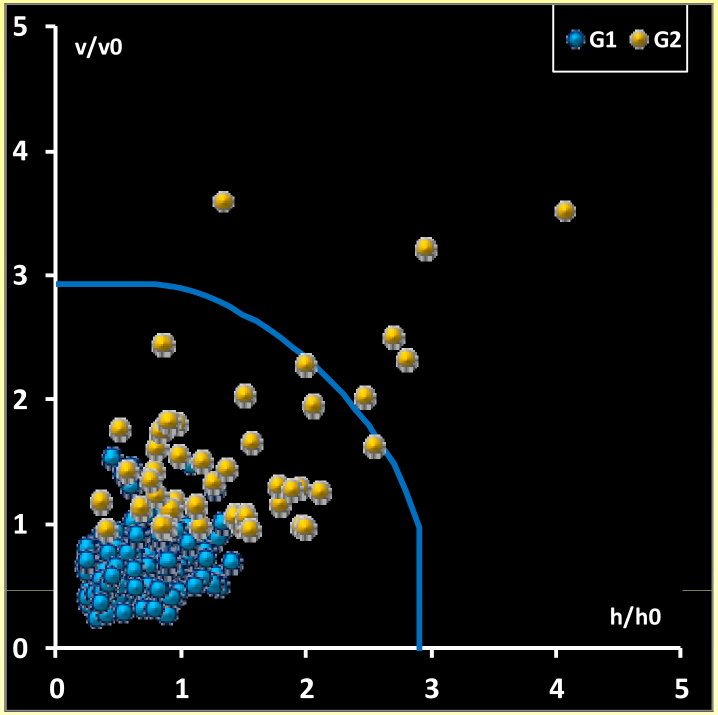
$$z = \frac{\left| x - \mathrm{median}(x) \right|}{\sigma_{Q_n}(x)}$$

M. Daszykowski, S. Serneels, K. Kaczmarek, P. Van Espen, C. Croux, B. Walczak, TOMCAT: a MATLAB toolbox for multivariate calibration techniques :*Chemometrics and Intelligent Laboratory Systems*, 85 (2007) 269-277.



Robust and non-robust classification





Construction of the Classification Rules

$$N_h \frac{h}{h_0} + N_v \frac{v}{v_0} \sim \chi^2 (N_h + N_v)$$

$$x = \begin{cases} = h \\ = v \end{cases} \qquad N_x \frac{x}{x_0} \sim \chi^2 (N_x) \quad \Rightarrow \quad \begin{matrix} N_x = ? \\ x_0 = ? \end{matrix}$$

Regular case

$$h_0 = \frac{1}{I} \sum_{i=1}^{I} h_i \equiv \frac{A}{I}$$

Method of Moments

$$\hat{N} = \frac{2}{S^2}$$

$$v_0 = \frac{1}{I} \sum_{i=1}^{I} v_i \equiv \frac{L_0}{I} \left(1 - R(A)\right)$$

$$S^2 = \frac{1}{I} \sum_{i=1}^{I} (x_i - 1)^2$$

Construction of the Classification Rules

Robust Estimators

Median *M*

$$M = \frac{x_0}{N_x}\chi^{-2}(0.5, N_x)$$

Interquartile *R*

$$R = \frac{x_0}{N_x}\left[\chi^{-2}(0.75, N_x) - \chi^{-2}(0.25, N_x)\right]$$

Empirical formula, *a* ,*b*, *d* -constants

$$\hat{N}_x = \exp\left[\left(\frac{1}{a}\ln\frac{bR}{M}\right)^{\frac{1}{d}}\right]$$

$N_x = ?$

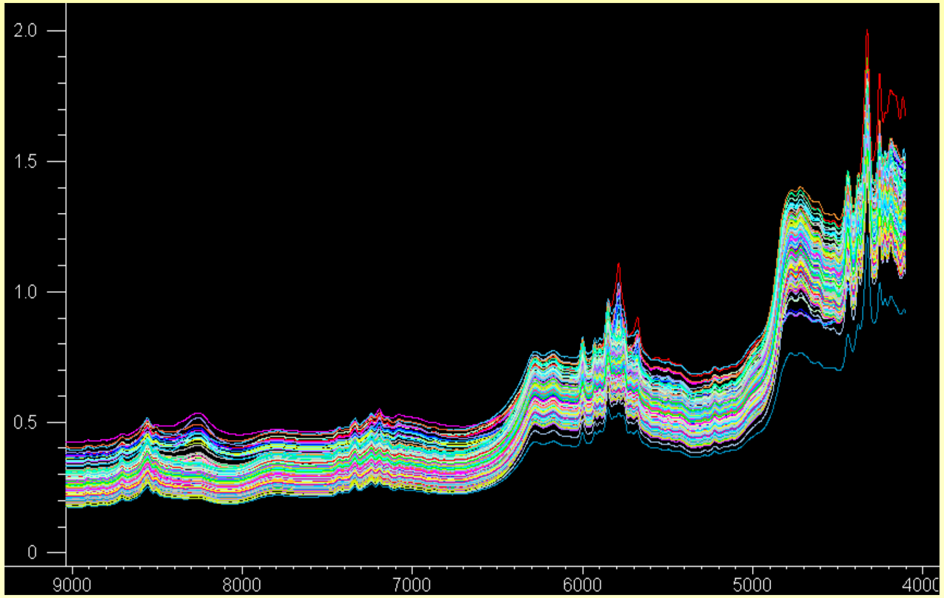
$x_0 = ?$

$$\hat{x}_0 = 0.5\hat{N}_x\left(\frac{M}{\chi^{-2}(0.5, \hat{N}_x)} + \frac{R}{\chi^{-2}(0.75, \hat{N}_x) - \chi^{-2}(0.25, \hat{N}_x)}\right)$$

Case Study

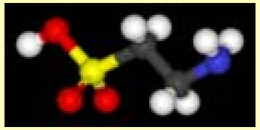
Data acquisition with fiber-probe: NIR spectra in 4100 –10000 cm^{-1} region

Data set: Substance in the closed PE bags, 82 drums, each bag measured 3 times, totally: 246 spectra +4 drums with other substance



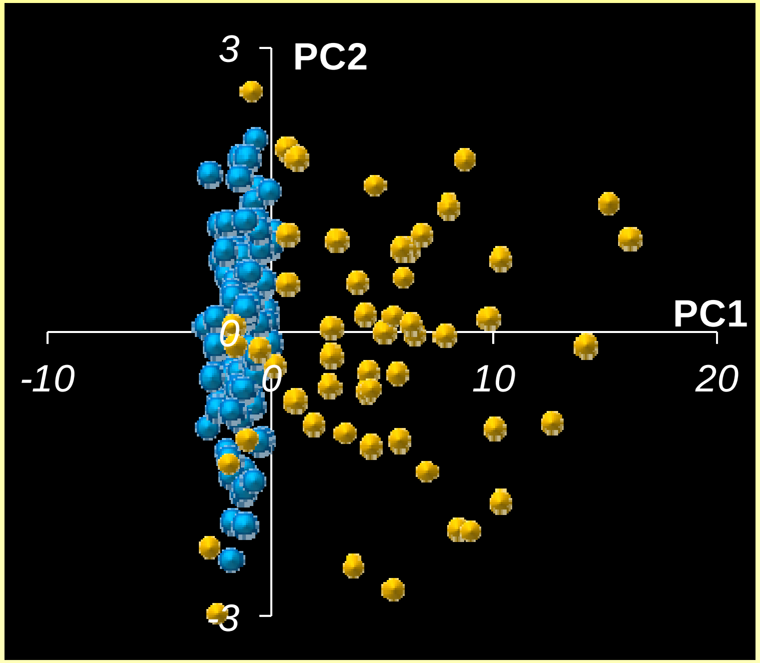
Substance in closed PE bags

Taurine,



2-aminoethanesulfonic acid.

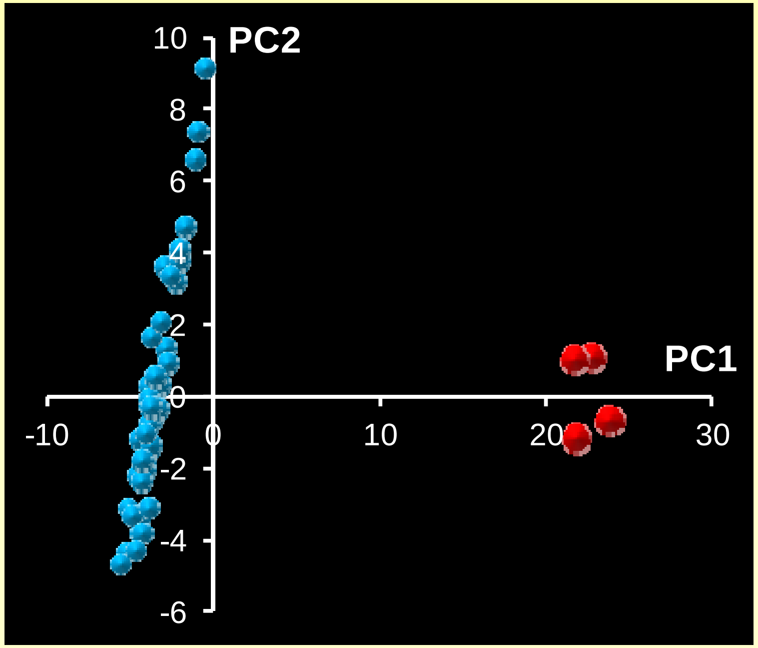
Data Sets' Description



Group G1: 170 objects
Group G2: 46 objects

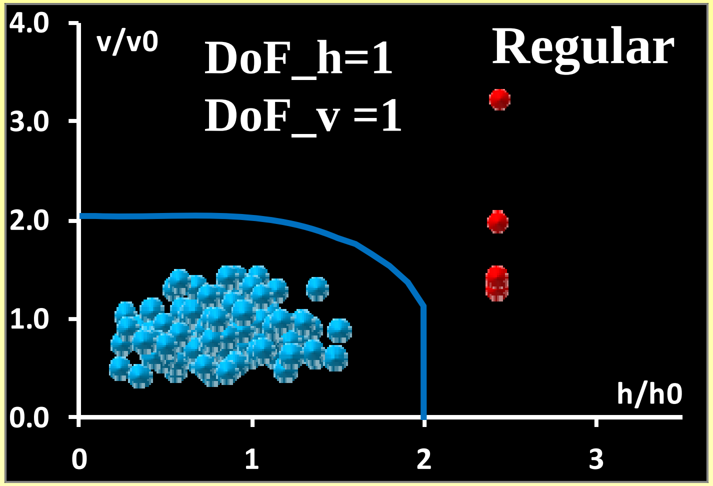
<u>Test Set</u> : 30 objects
G1: 24 objects
G2: 6 objects

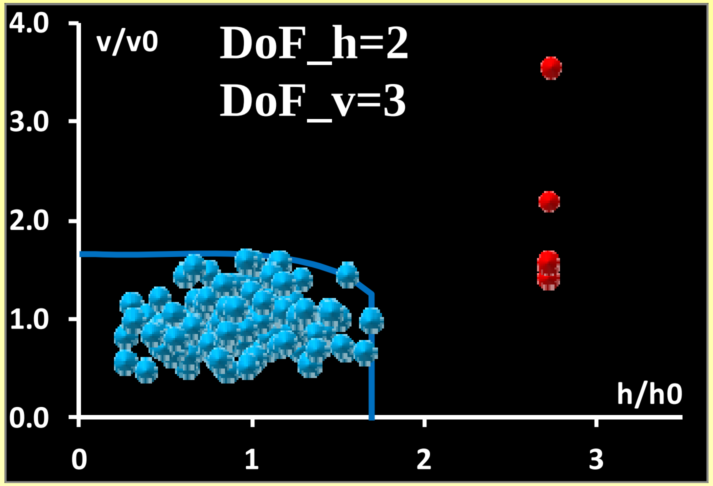
Group G3: 4 objects



Model with Evident Outliers

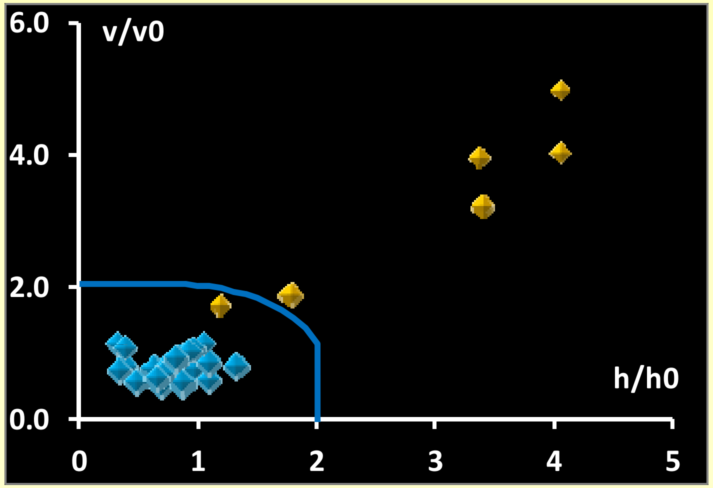
Training set: G1+G3

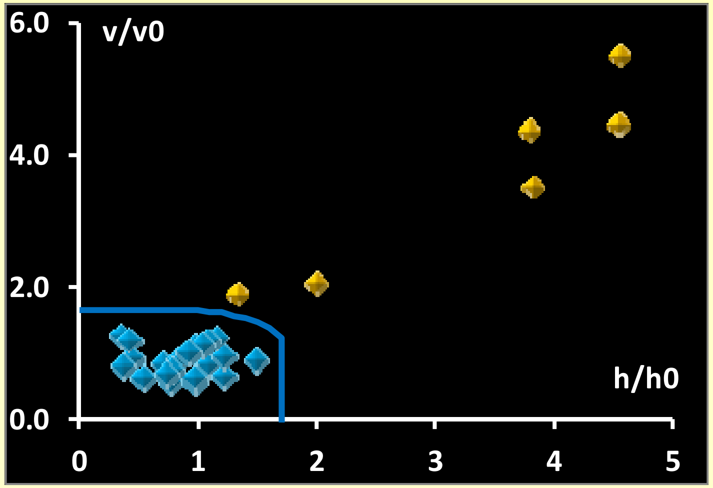




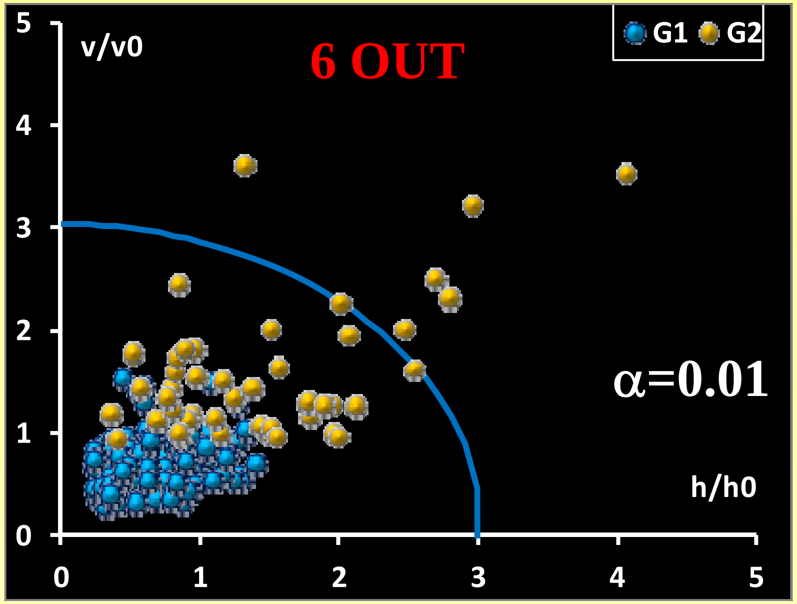
Robust

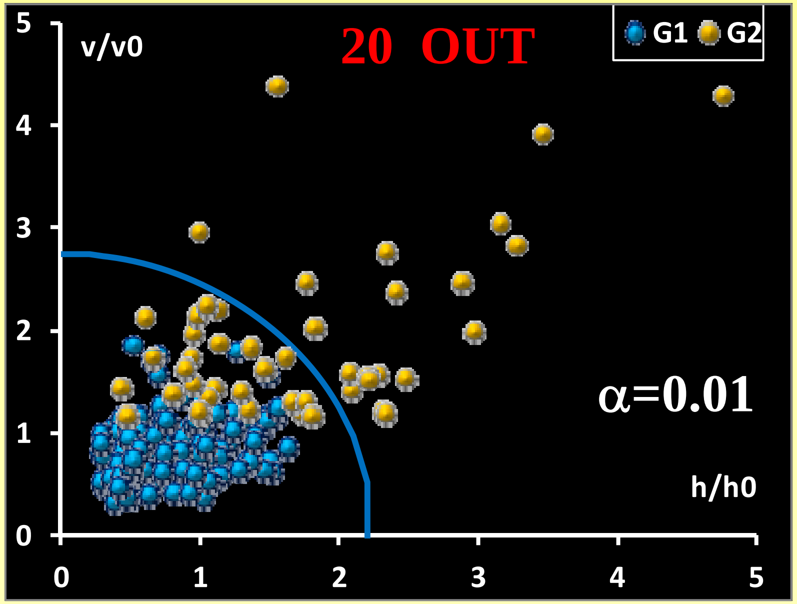
Test set: 30 objects

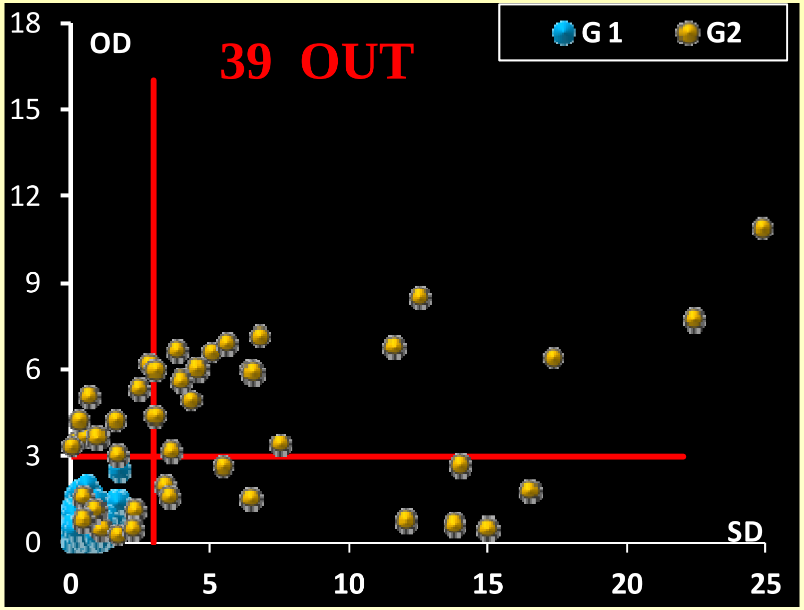


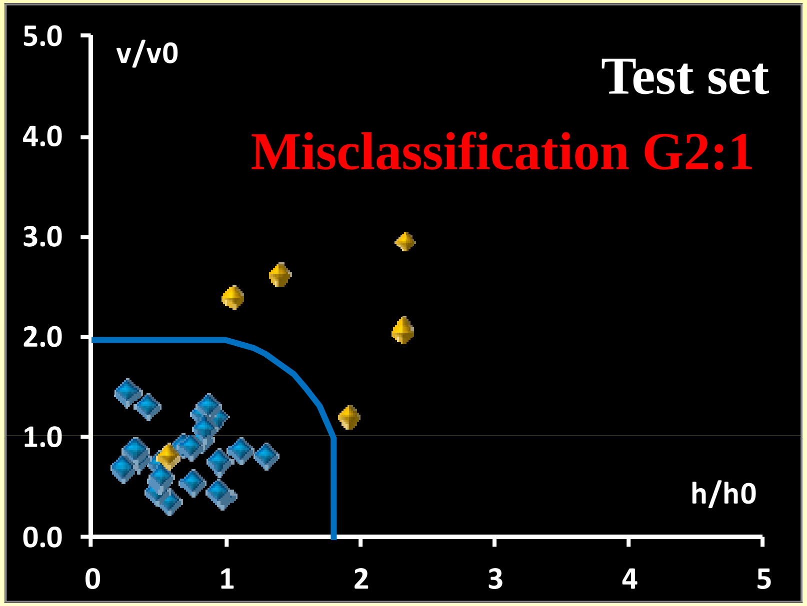


Training set : G1+G2



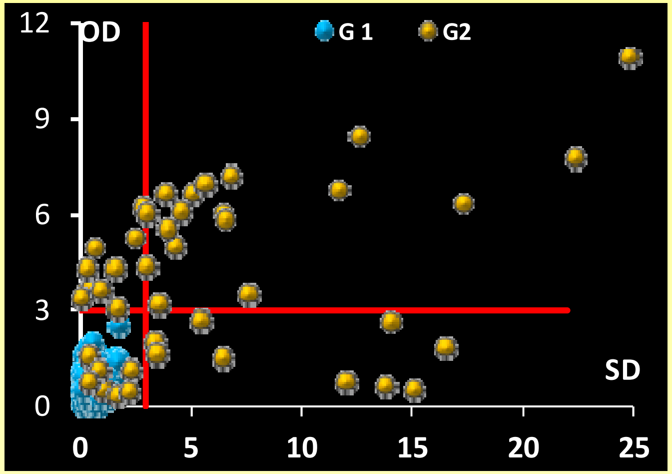




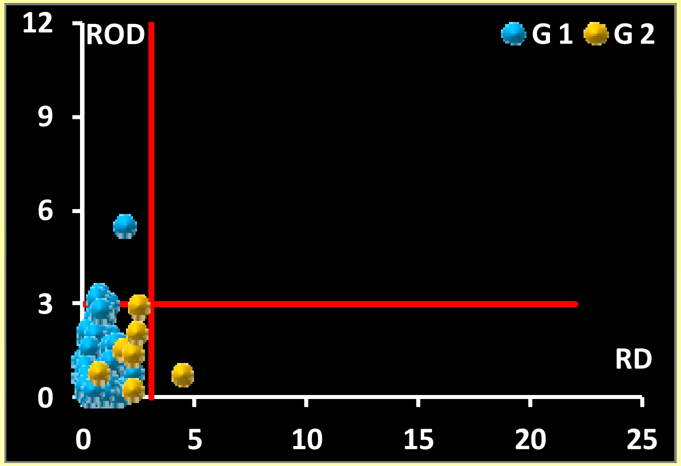


Robust SIMCA Results

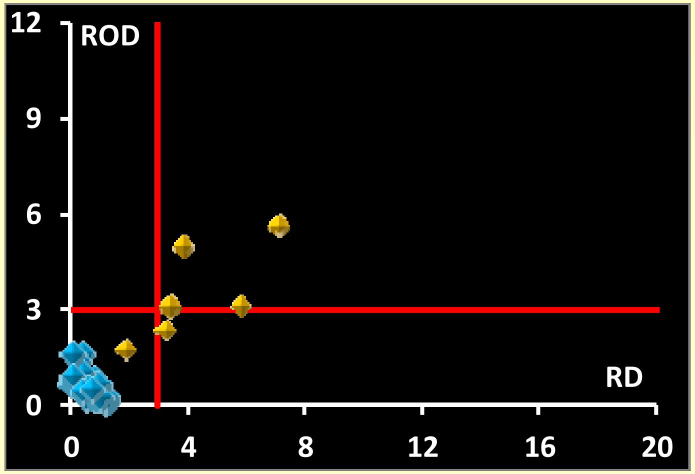
Training set 216 objects (G1: 170 +G2: 46)



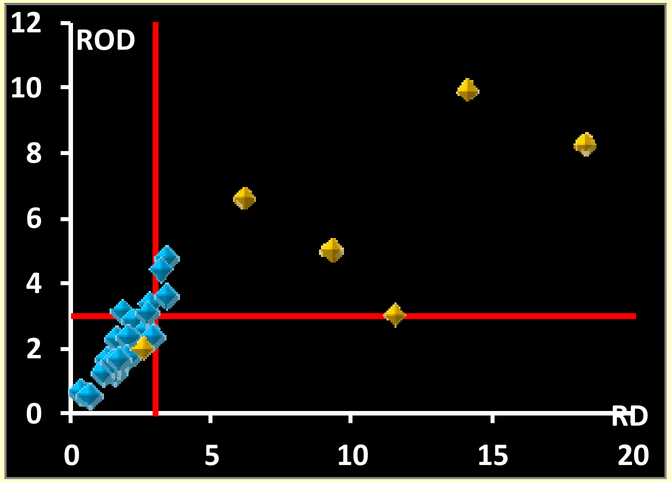
Training set 177 objects (G1: 170 +G2: 7)



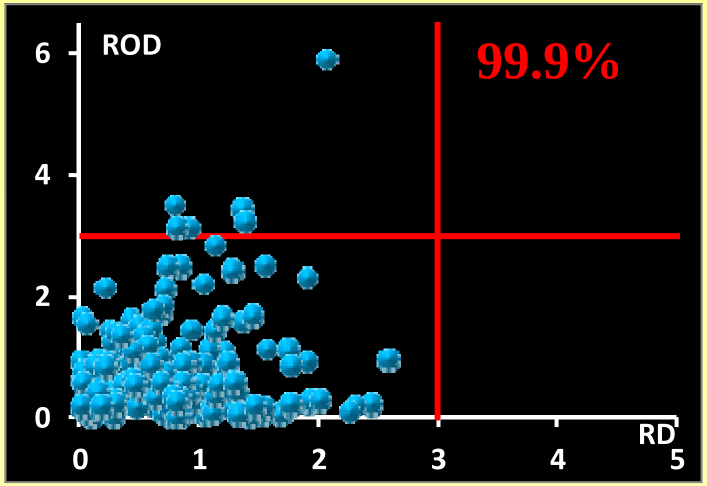
Test set 30 objects (G1: 24 +G2: 6)
Misclassification G2:1



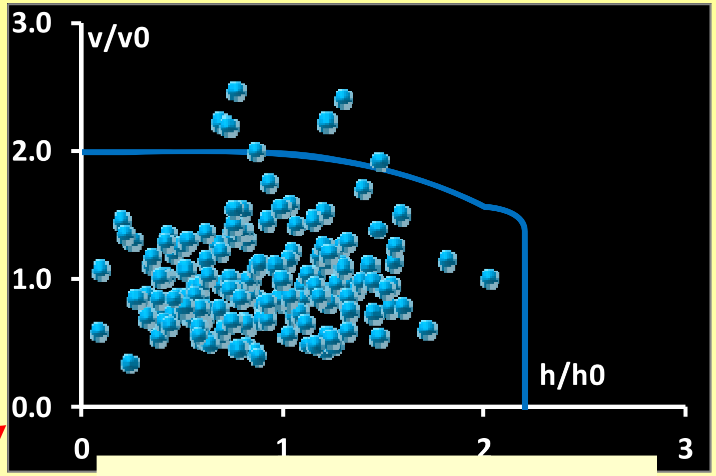
Test set 30 objects (G1: 24 +G2: 6)
Misclassification G1:7; G2:1



Training set G1: No Outliers



OUT: 6 objects

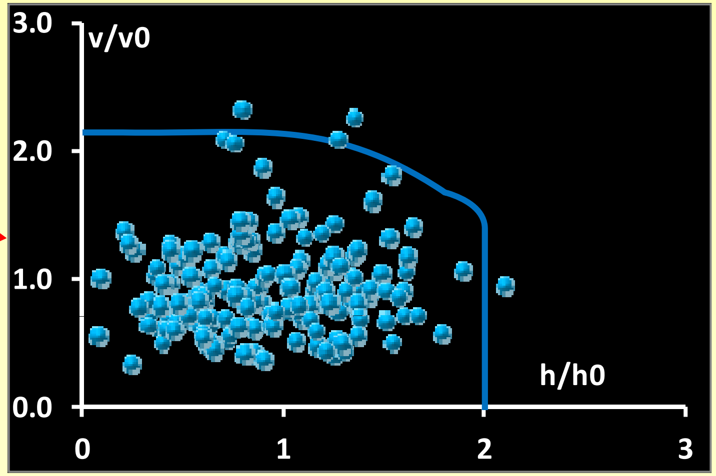


OUT: 7 objects

Robust estimates

DoF_h= 2
DoF_v= 4

α =0.01



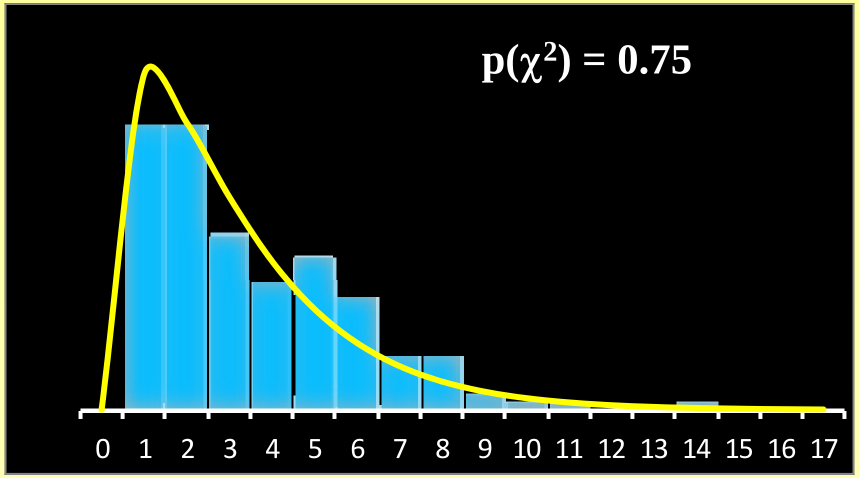
Regular estimates

DoF_h= 3
DoF_v= 3

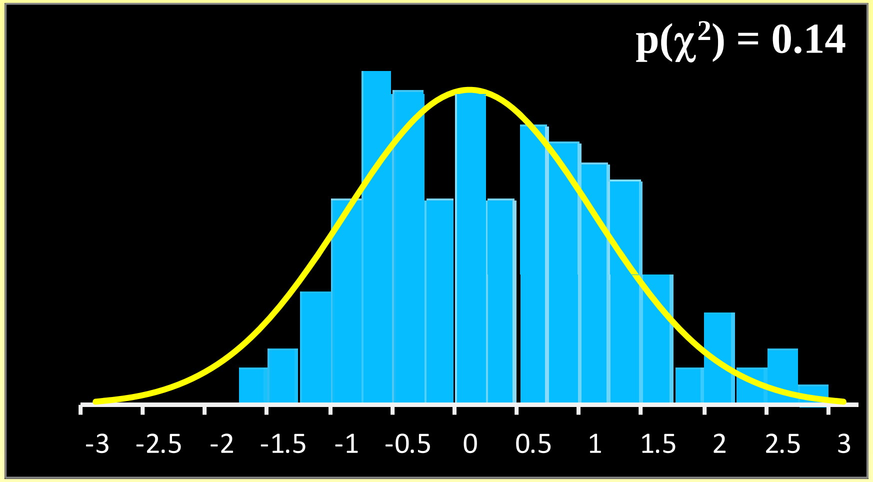
OUT: 3 objects

Residuals' Distributions

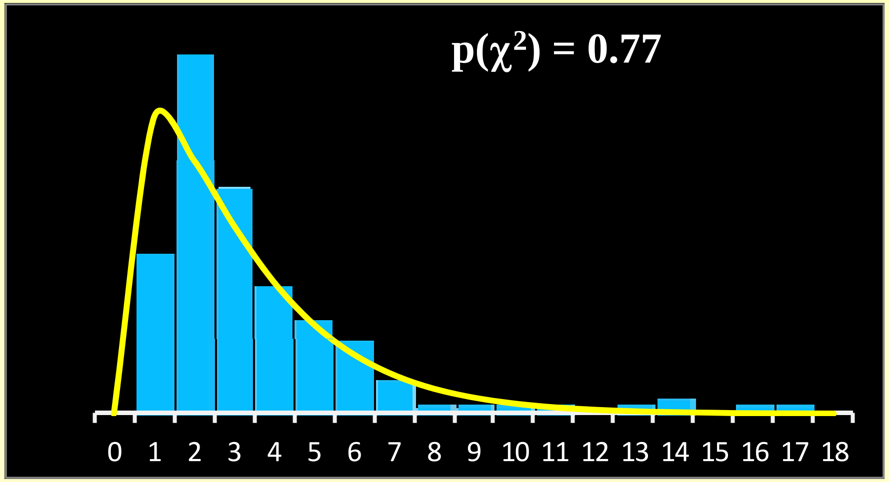
$N_h\ h/h_0$



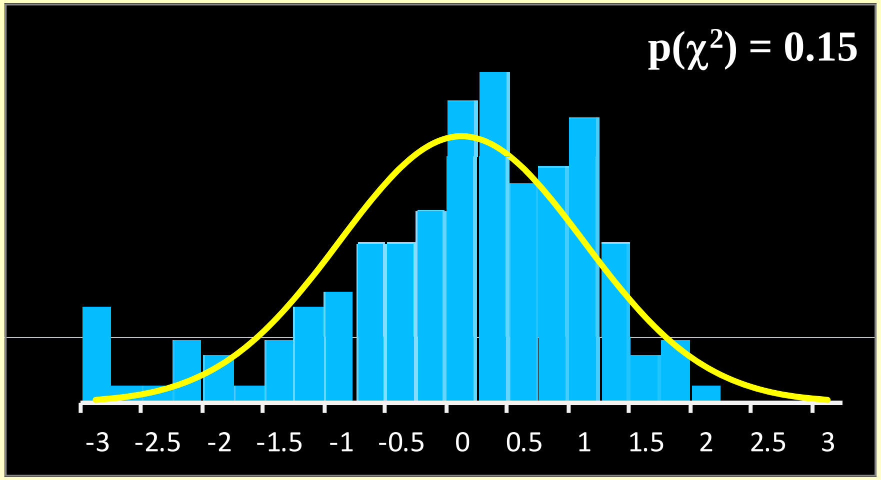
z(SD)



$N_v\ v/v_0$



z(OD)



Conclusions

Any classification problem should be solved with respect to a given type I error.

Application of the robust procedure for the construction of the classification rules provides reliable outlier detection

It is important to have a possibility for switching between robust and non-robust classification methods.