Use and abuse of robust PCA

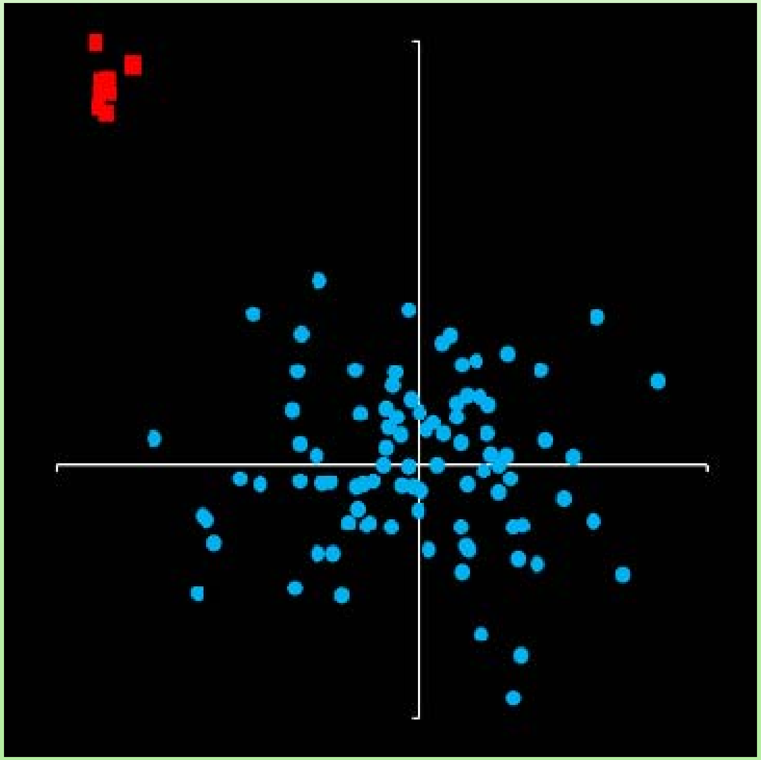
Alexey Pomerantsev & Oxana Rodionova

Semenov Institute of Chemical Physics, Moscow

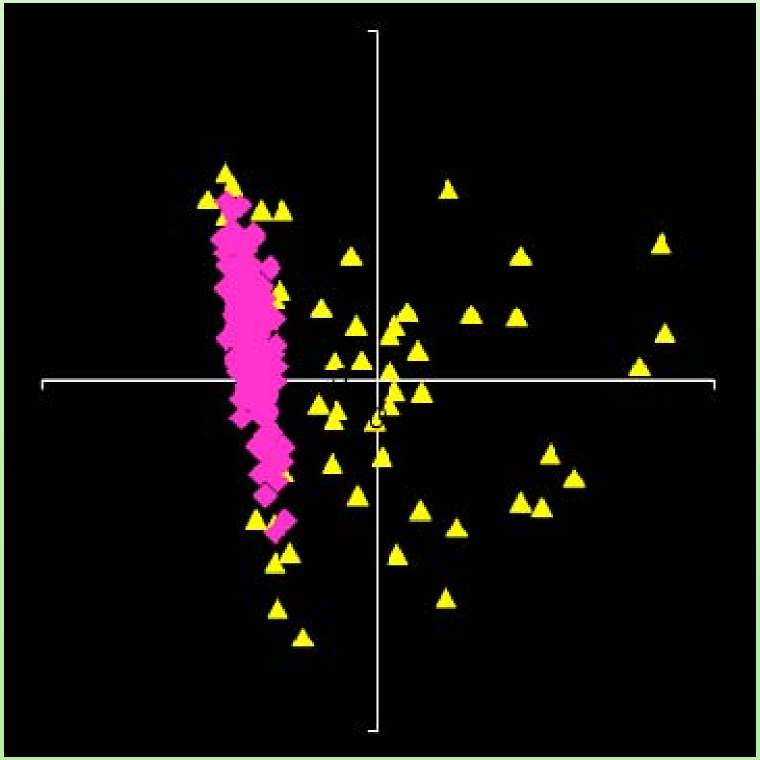
Contaminated data

$$R_\gamma = (1 - \gamma)R + \gamma D$$

Few outliers

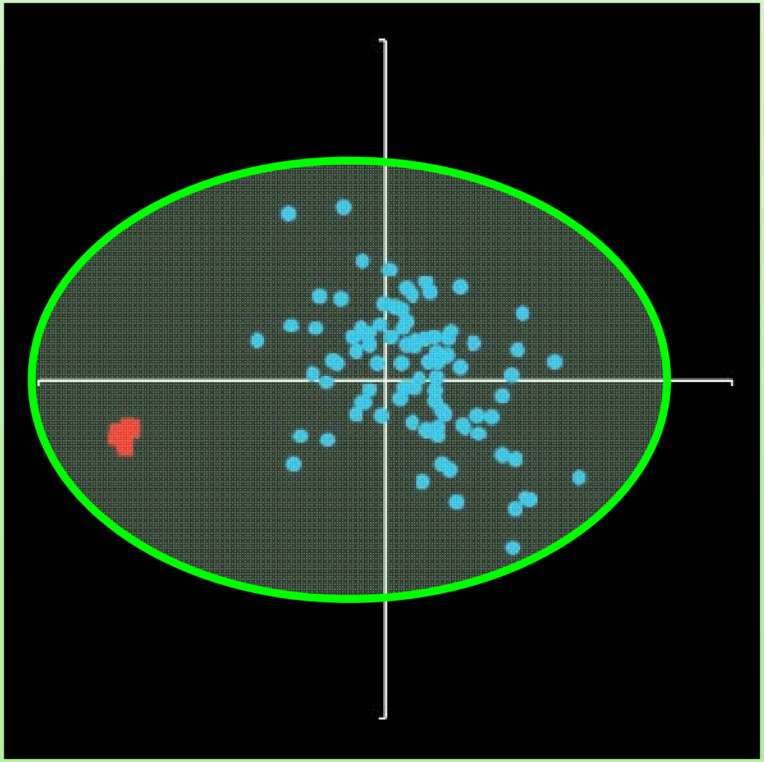


Two groups

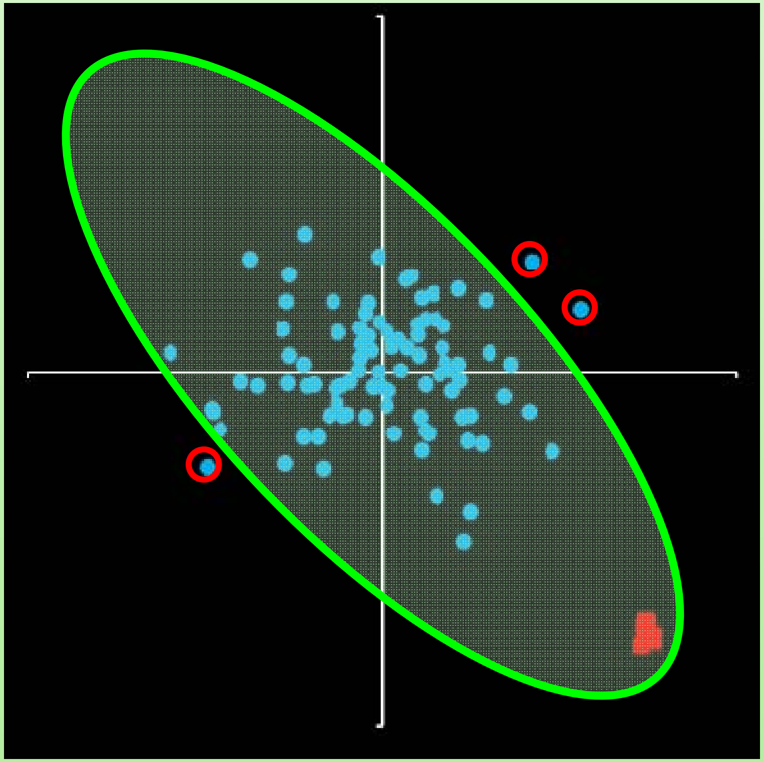


Classical methods for contaminated data

Masking effect

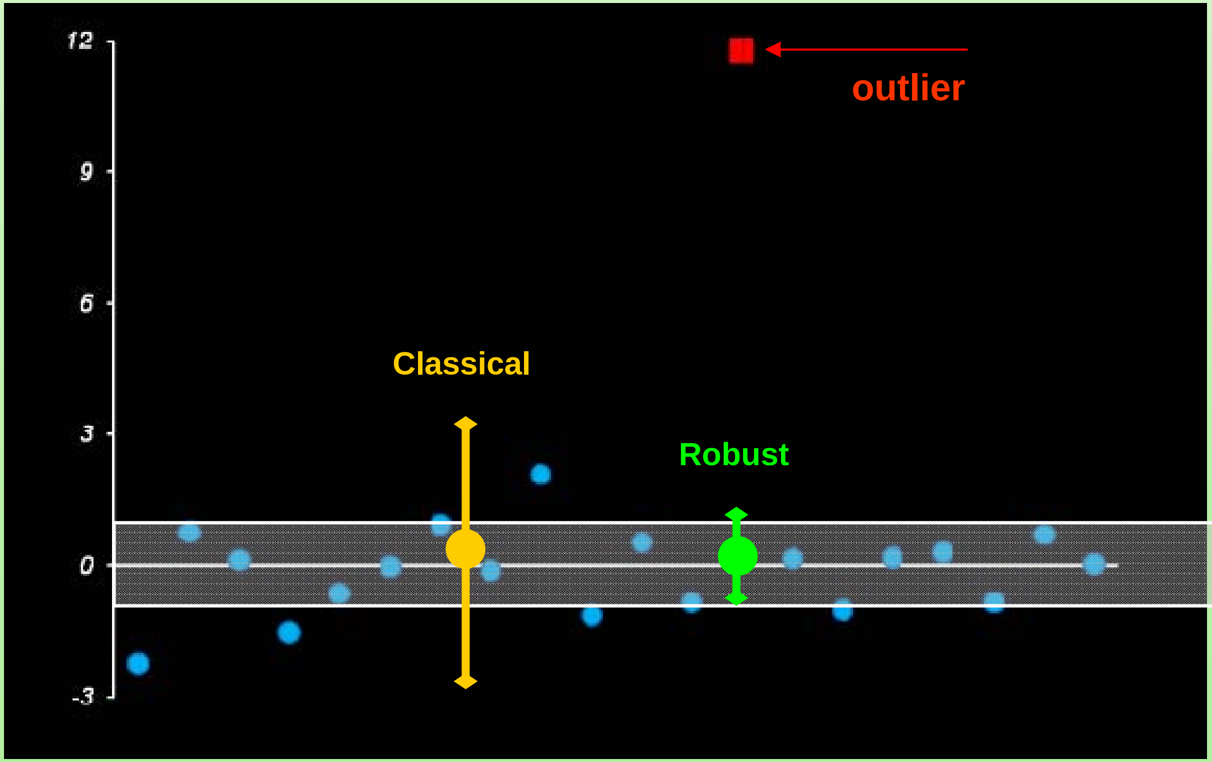


Swamping effect



Classical and robust statistics

Classical	Robust
$\bar{x} = \frac{1}{I}\sum_{i=1}^{I} x_i$	$\tilde{x} = \mathrm{median}(\mathbf{x})$
$s^2 = \frac{1}{I-1}\sum_{i=1}^{I}(x_i - \bar{x})^2$	$s_{\mathrm{MAD}} = 1.4826\,\mathrm{median}(\lvert\mathbf{x} - \tilde{x}\rvert)$



Extremes and Outliers

$$\begin{pmatrix} x \\ y \end{pmatrix} \propto N(\mathbf{0},\mathbf{I})$$

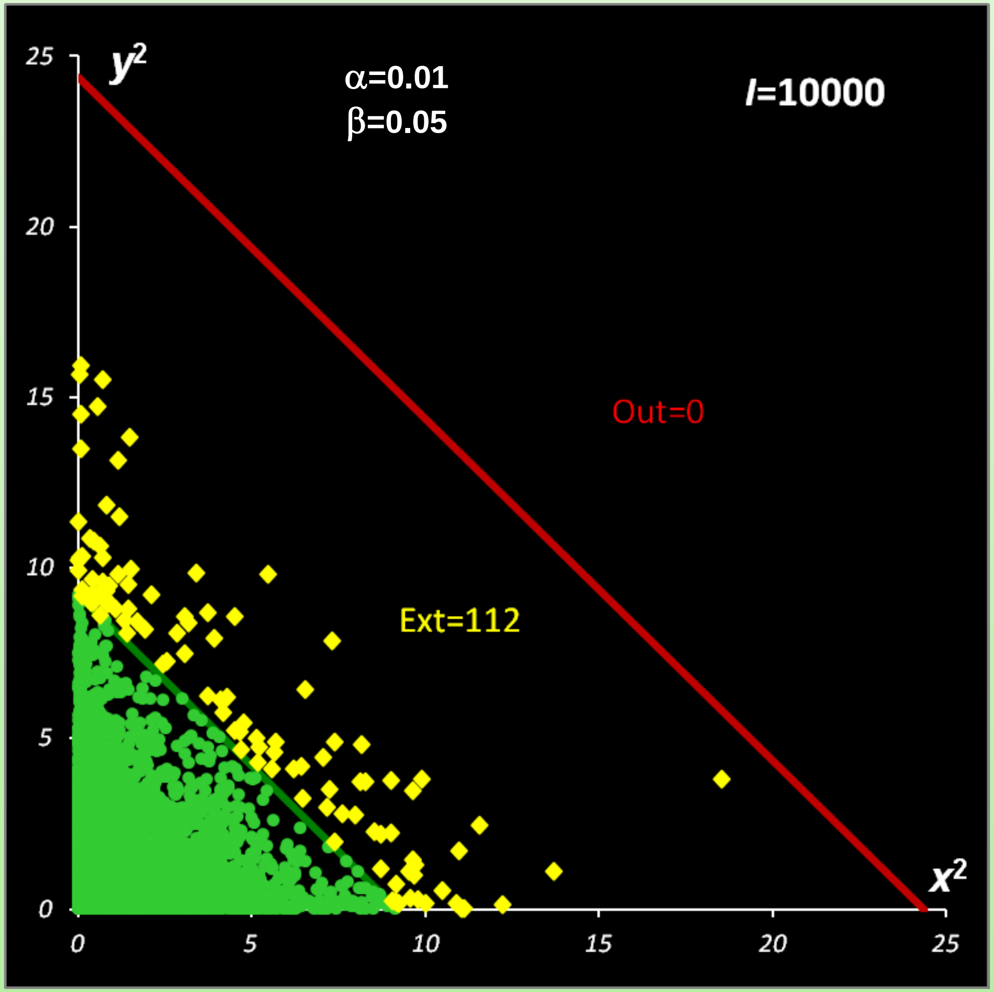
$$x^2 + y^2 \propto \chi^2(2)$$

α is Extreme significance

$$x^2 + y^2 \le \chi^{-2}(2 \mid 1-\alpha)$$

β is Outlier significance

$$x^2 + y^2 \le \chi^{-2}\left(2 \mid (1-\beta)^{1/I}\right)$$

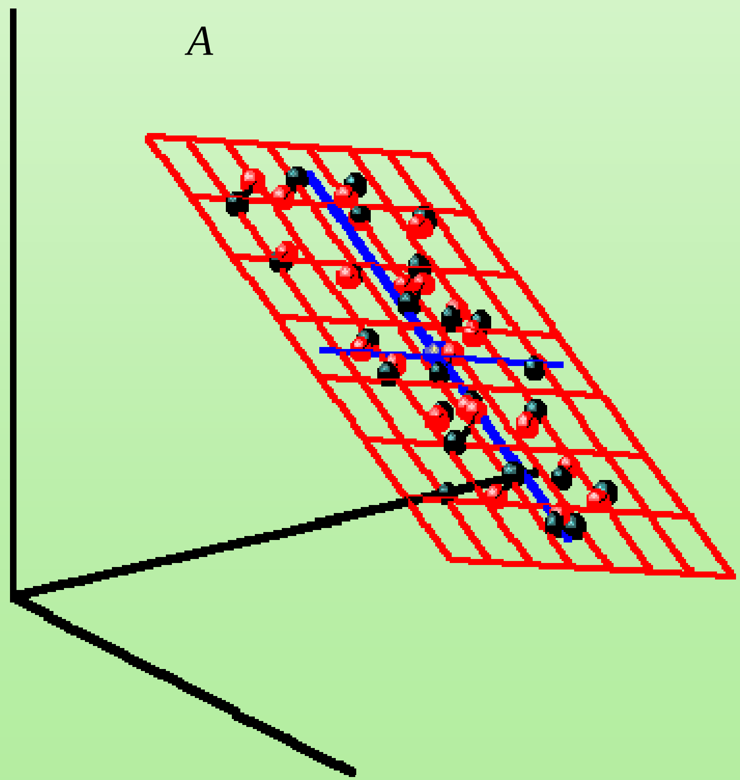


Problem

	Regular data	Contaminated data
Classical methods	OK	BAD
Robust methods	?	OK

Principal Component Analysis

$$I\ \underset{J}{\boxed{\mathbf{X}}} = I\ \underset{A}{\boxed{\mathbf{T}_A}} \times A\ \underset{J}{\boxed{\mathbf{P}_A^t}} + I\ \underset{J}{\boxed{\mathbf{E}_A}}$$



PCA robustification

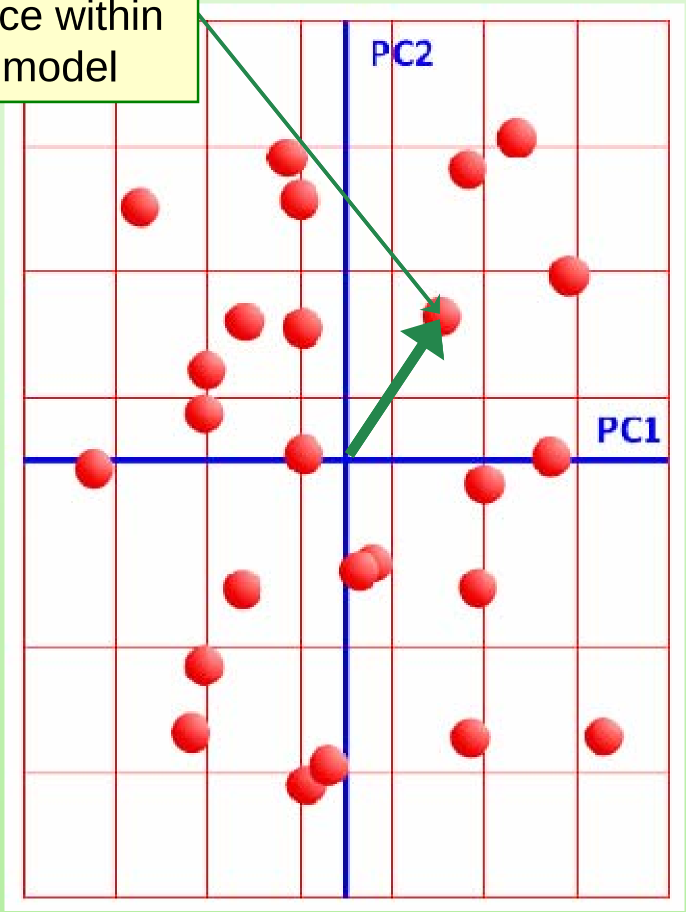
(1) Data pre-processing;

(2) Decomposition;

(3) Calculation of thresholds.

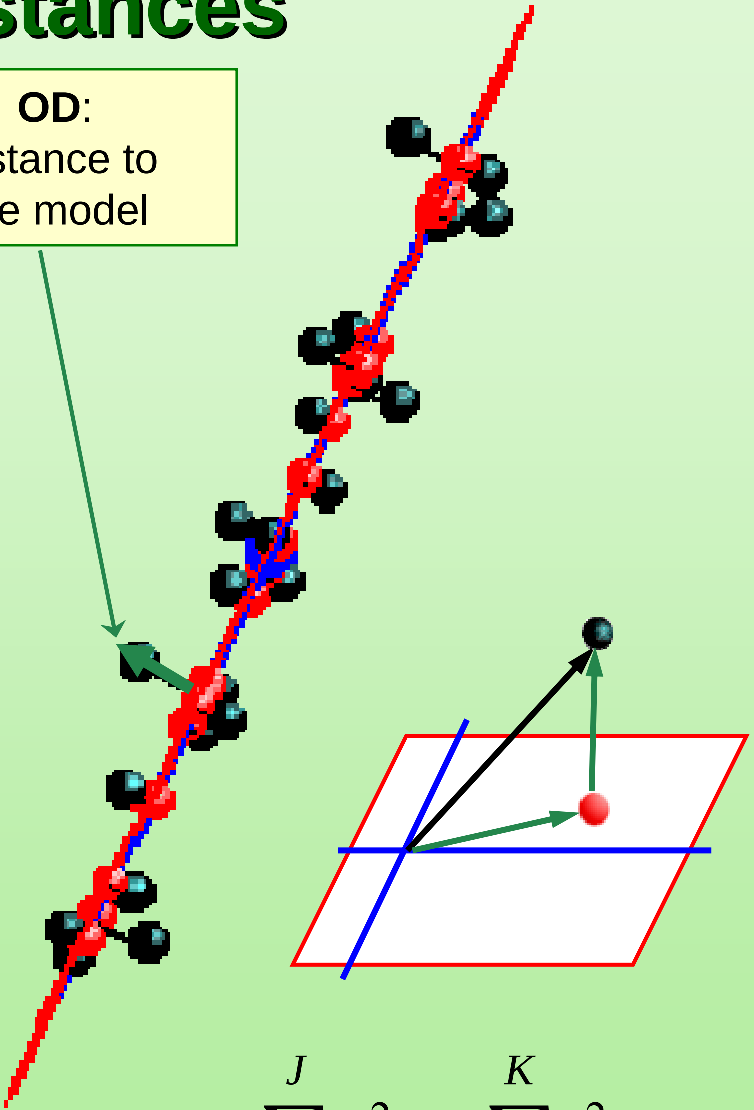
Scores & Orthogonal Distances

SD:
distance within the model



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

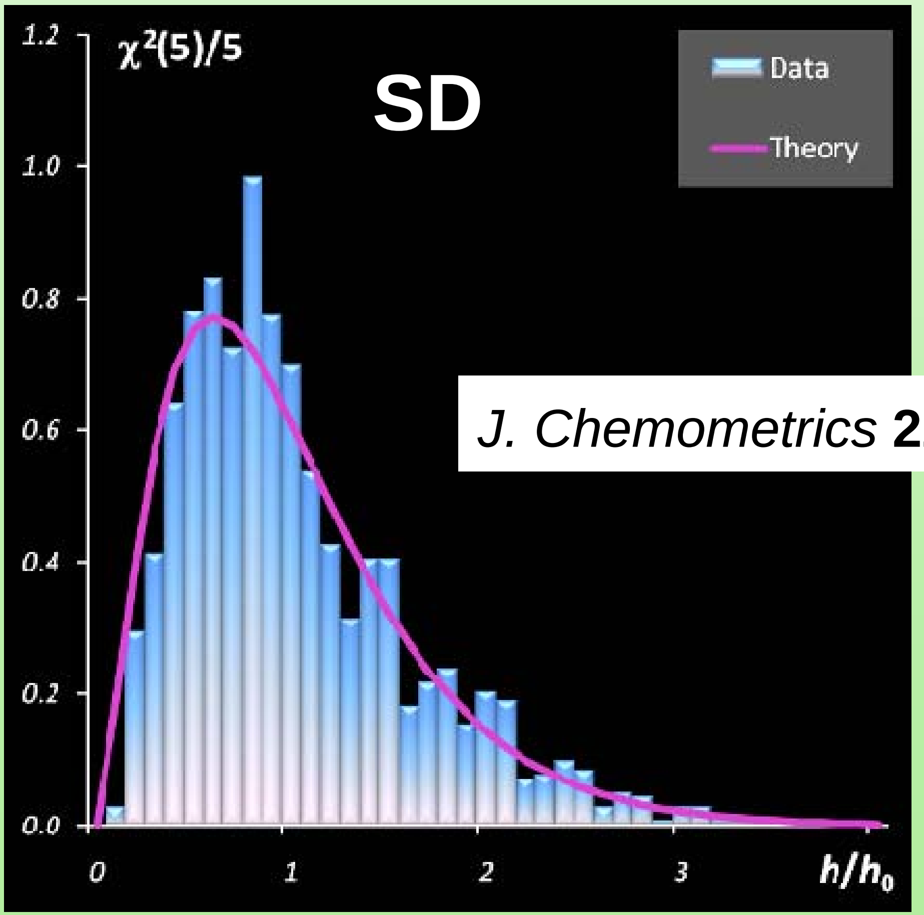
OD:
distance to the model

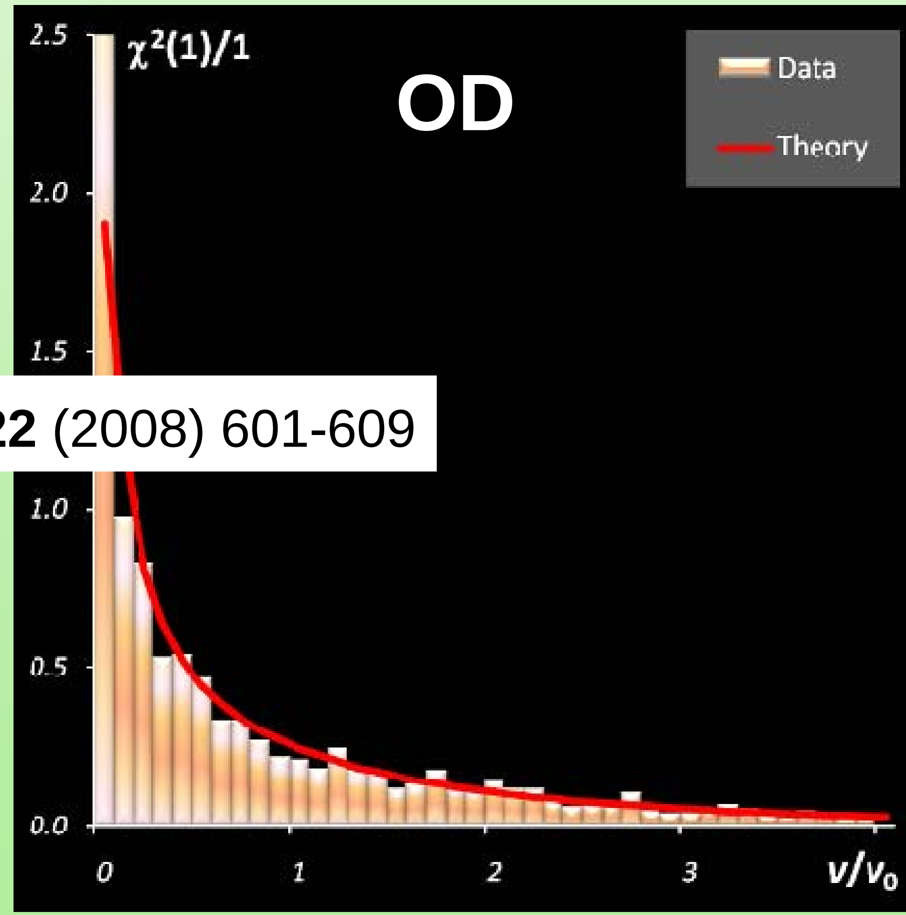


$$v_i = \sum_{j=1}^{J} e_{ij}^2 = \sum_{a=A+1}^{K} t_{ia}^2$$

Data Driven SIMCA

$$u = \begin{cases} h \\ v \end{cases} \qquad (u_1, \ldots, u_I) \propto \left(u_0/N\right) \chi^2(N) \Longrightarrow \begin{cases} u_0 = ? \\ N = ? \end{cases}$$





J. Chemometrics **22** (2008) 601-609

Tolerance Areas

$$z = N_h \frac{h}{h_0} + N_v \frac{v}{v_0}$$

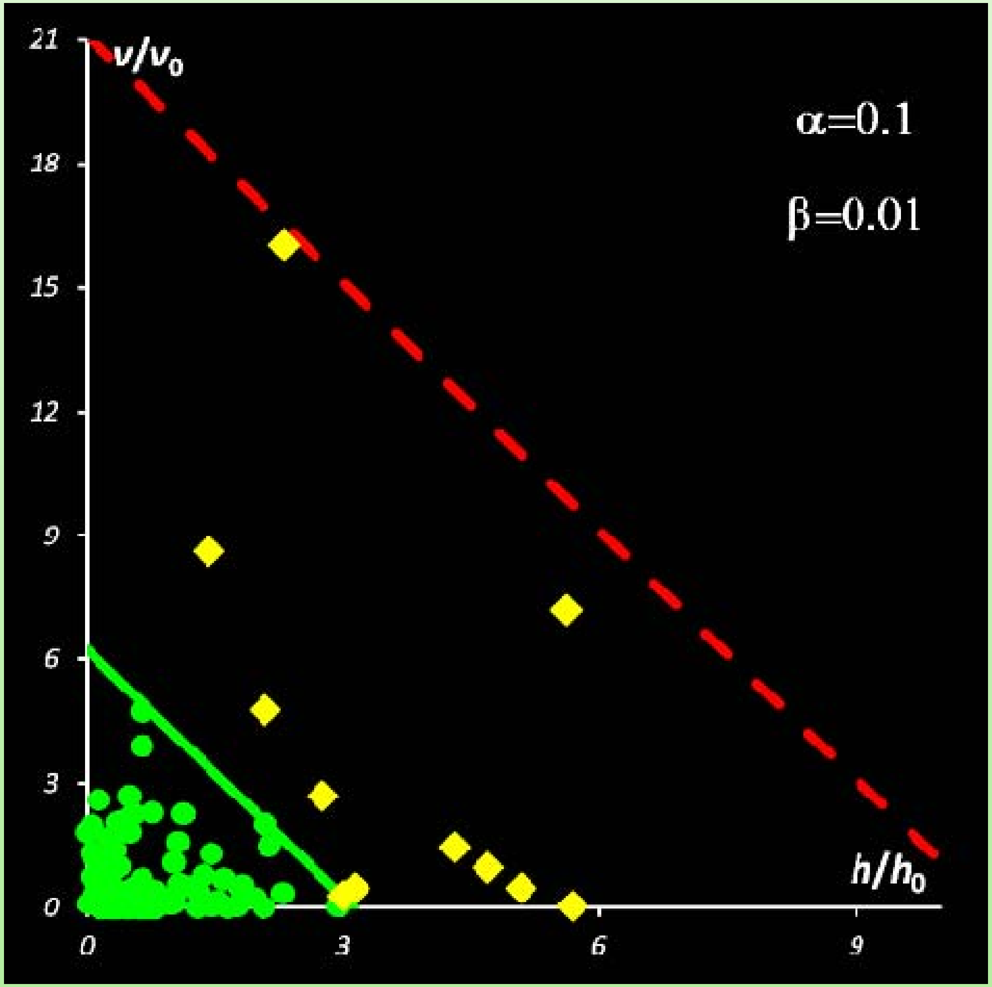
$$z \propto \chi^2 (N_h + N_v)$$

α is Extreme significance

$$z \le \chi^{-2}(N_h + N_v \,|\, 1-\alpha)$$

β is Outlier significance

$$z \le \chi^{-2}\left(N_h + N_v \,|\, (1-\beta)^{1/I}\right)$$



Classical Data Driven (CDD) SIMCA

Classical Method of Moments

Given

$$(u_1, \ldots., u_I) \propto \left(u_0/N\right) \chi^2(N)$$

Then

$$\hat{u}_0 = \bar{u}, \qquad \hat{N} = \operatorname{int} \frac{2\hat{u}_0^2}{s_u^2}$$

Where

$$\bar{u} = \frac{1}{I}\sum_{i=1}^{I} u_i\,, \quad s_u^2 = \frac{1}{I-1}\sum_{i=1}^{I}(u_i - \bar{u})^2$$

Robust Data Driven (RDD) SIMCA

Robust Method of Moments

Given

$$(u_1,\ldots., u_I) \propto \left(u_0/N\right) \chi^2(N)$$

Then

$$\begin{matrix} \tilde{u}_0 \\ \\ \tilde{N} \end{matrix} \Leftarrow \begin{cases} M = \dfrac{u_0}{N}\chi^{-2}(0.5, N) \\ \\ R = \dfrac{u_0}{N}\left[\chi^{-2}(0.75, N) - \chi^{-2}(0.25, N)\right] \end{cases}$$

Where

M=median($\mathbf{u}$) R=interquartile($\mathbf{u}$)

¼ ¼

$$u_{(1)} \le u_{(2)} \le \ldots. \le u_{(I\text{-}1)} \le u_{(I)}$$

½ ½

Dual Data Driven (3D) SIMCA

Given

$$\mathbf{X}=\mathbf{T}^{t}\mathbf{P}+\mathbf{E}$$

$$\mathbf{h}=(h_1,\ldots., h_I) \qquad \mathbf{v}=(v_1,\ldots., v_I)$$

Then

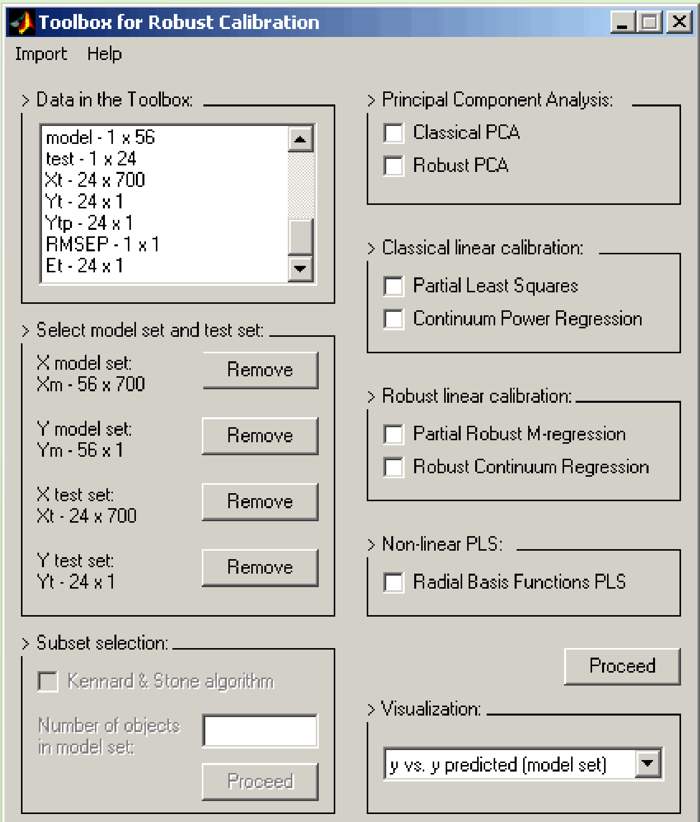
CDD SIMCA	RDD SIMCA
$\left(\hat{h}_0 \ \hat{N}_h\right) \quad \left(\hat{v}_0 \ \hat{N}_v\right)$	$\left(\tilde{h}_0 \ \tilde{N}_h\right) \quad \left(\tilde{v}_0 \ \tilde{N}_v\right)$

$$\left(\tilde{N}_h \approx \hat{N}_h\right) \& \left(\tilde{N}_v \approx \hat{N}_v\right)$$

Yes CDD SIMCA	No RDD SIMCA

TOMCAT ToolBox

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



Robust pre-processing
robust centering & scaling

Robust PCA
robust PCs, robust singular values

Robust classification rules
z-transformed robust OD and SD

$$\mathrm{RD}_i = \frac{\left|\sqrt{\mathrm{SD}_i} - \mathrm{median}(\sqrt{\mathbf{SD}})\right|}{Q_n(\sqrt{\mathbf{SD}})}$$

$$\mathrm{ROD}_i = \frac{\left|\sqrt{\mathrm{OD}_i} - \mathrm{median}(\sqrt{\mathbf{OD}})\right|}{Q_n(\sqrt{\mathbf{OD}})}$$

M. Daszykowski, S. Serneels, K. Kaczmarek, P. Van Espen, C. Croux, B. Walczak, *ChemoLab* **85** (2007) 269-277

Case study I. Simulated regular data

$$\mathbf{x} = \boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\delta} \propto \mathrm{N}(\mathbf{0}, \mathbf{V}) \qquad \boldsymbol{\varepsilon} \propto \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

The numbers of variables, J=3

The numbers of objects, I=100

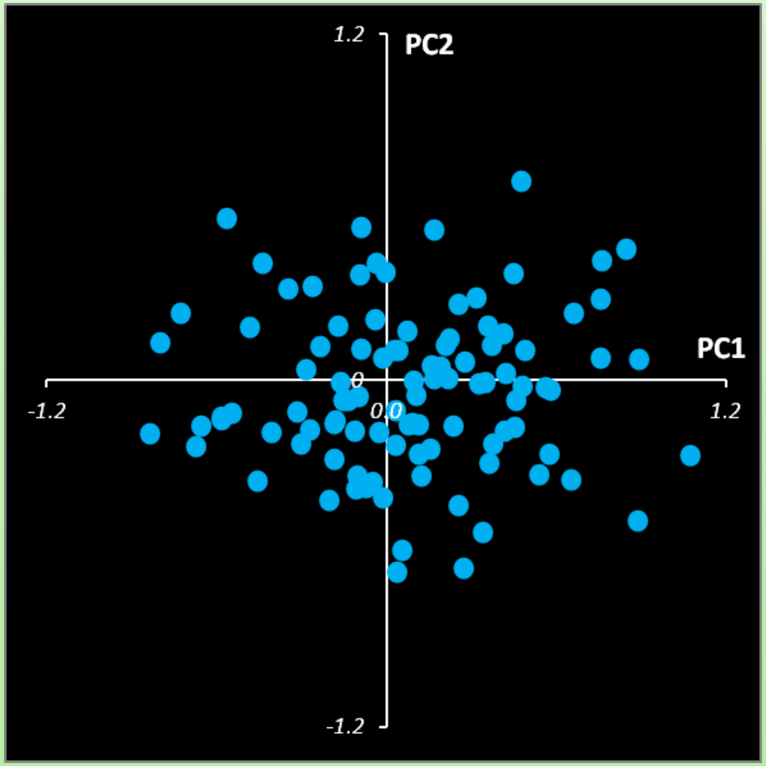
The number of principal components, A=2

The δ properties are:

E(δ) = 0, v_{11}= v_{22} = v_{33} = 0.28, rank(**V**) = 2.

The ε component properties are:

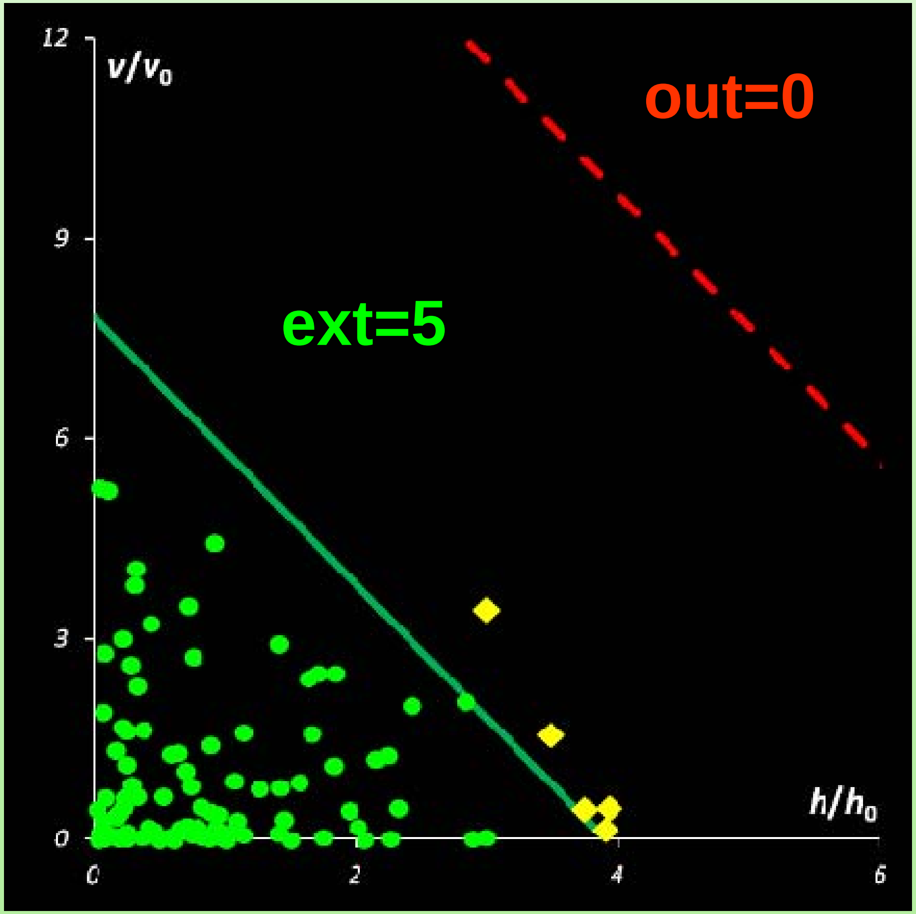
E(ε) = **0**, σ=0.05



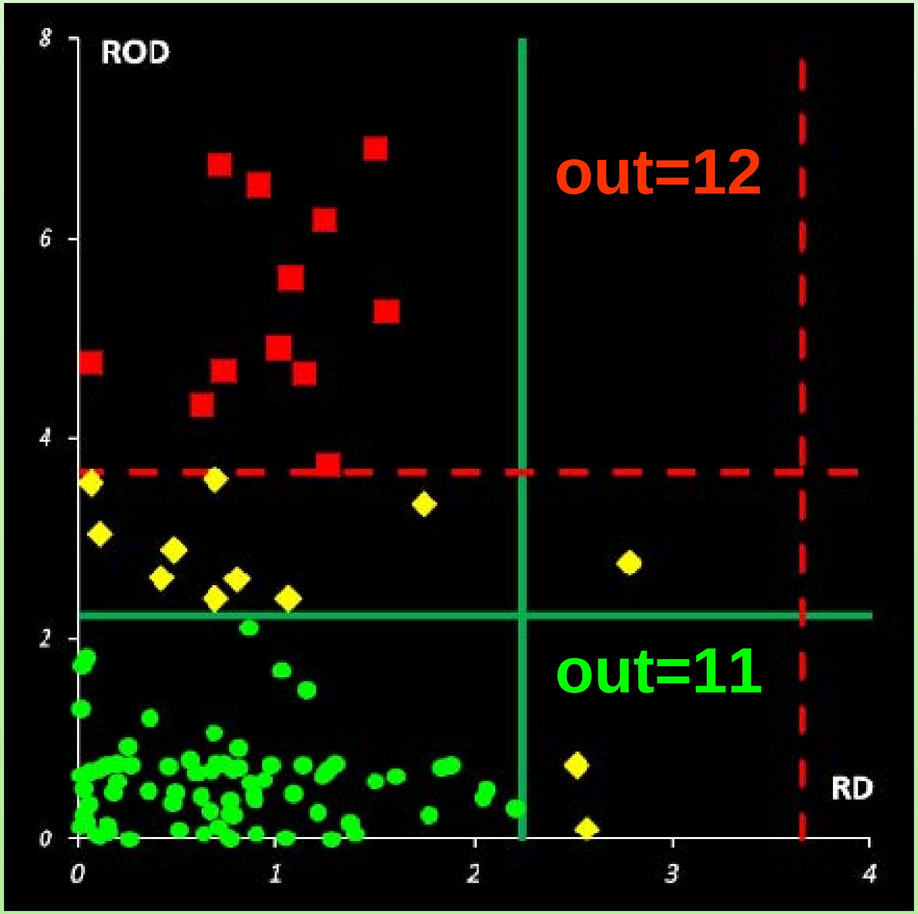
SIMCA plots

extreme area (α=0.05) outlier area (β=0.05)

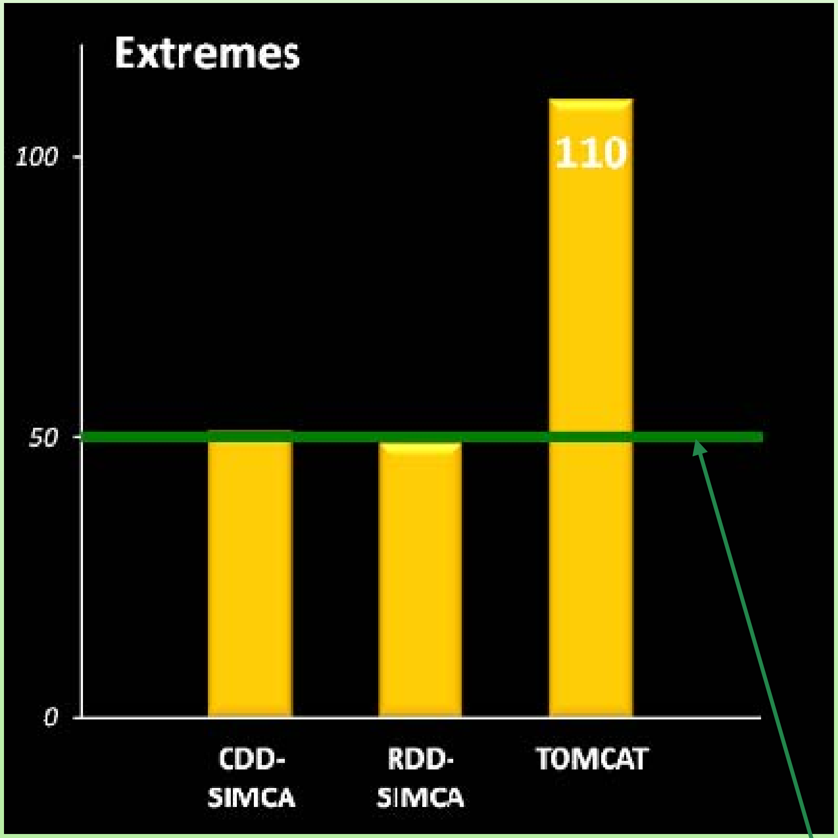
CDD SIMCA

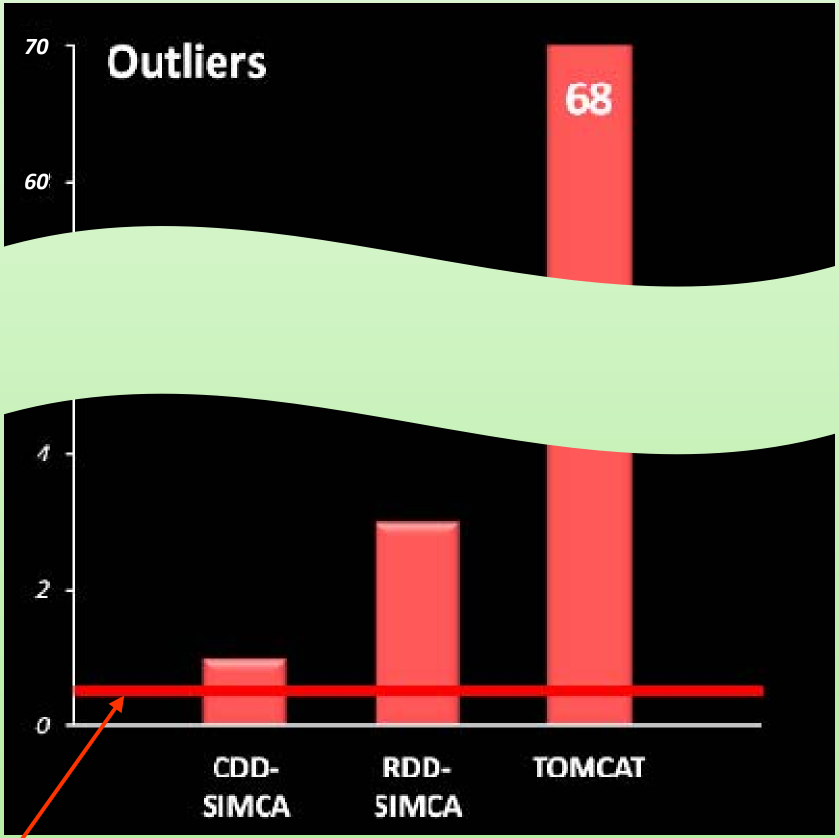


TOMCAT



Totally in 10 regular data sets





Expected

Case study II. Simulated data with outliers

$$\mathbf{x} = \boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\delta} \propto \mathrm{N}(\mathbf{0}, \mathbf{V}) \qquad \boldsymbol{\varepsilon} \propto \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

The numbers of variables, J=3

The numbers of objects, I=100

The number of principal components, A=2

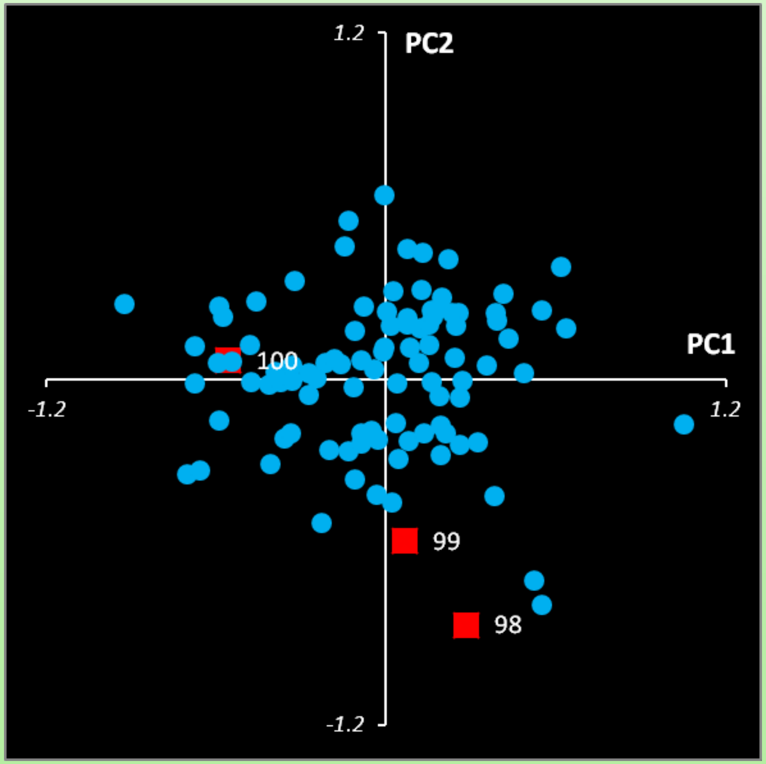
The δ properties are:

E(δ) = 0, $v_{11} = v_{22} = v_{33} = 0.28$, rank(**V**) = 2.

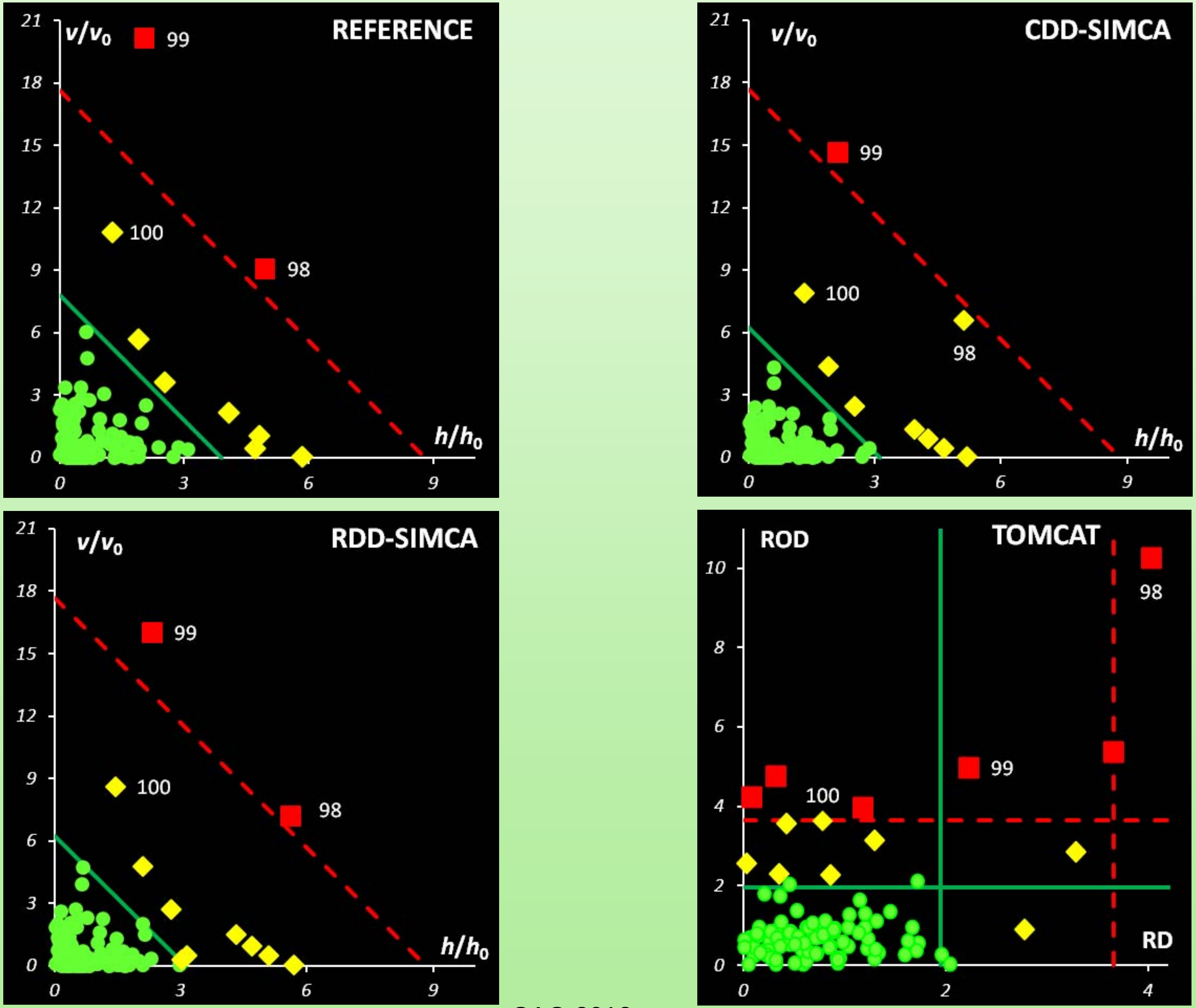
The ε component properties are:

E(ε) = **0**, σ=0.05 (first 97 objects)

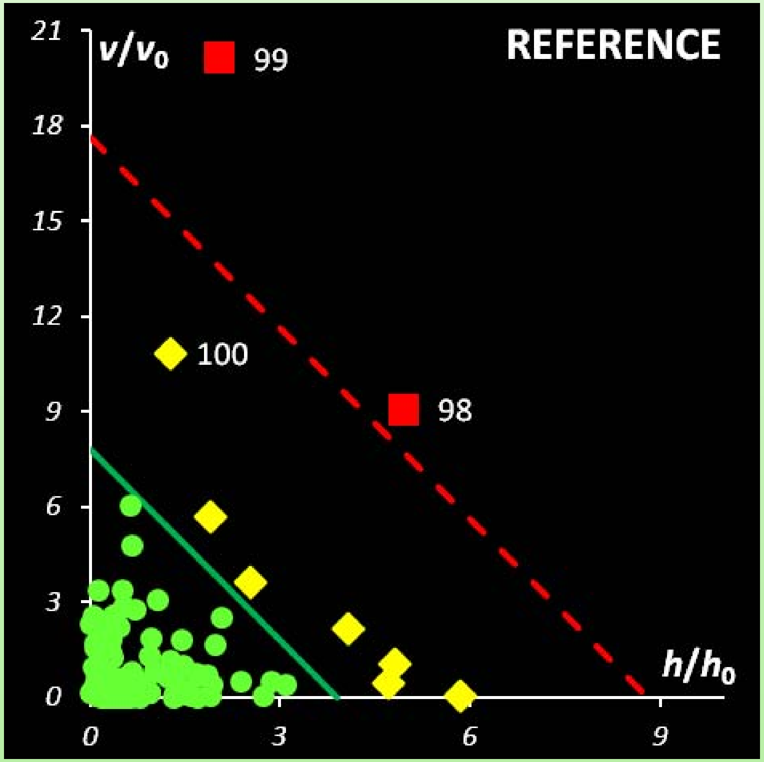
E(ε) = **0**, σ =0.2 (last 3 objects)

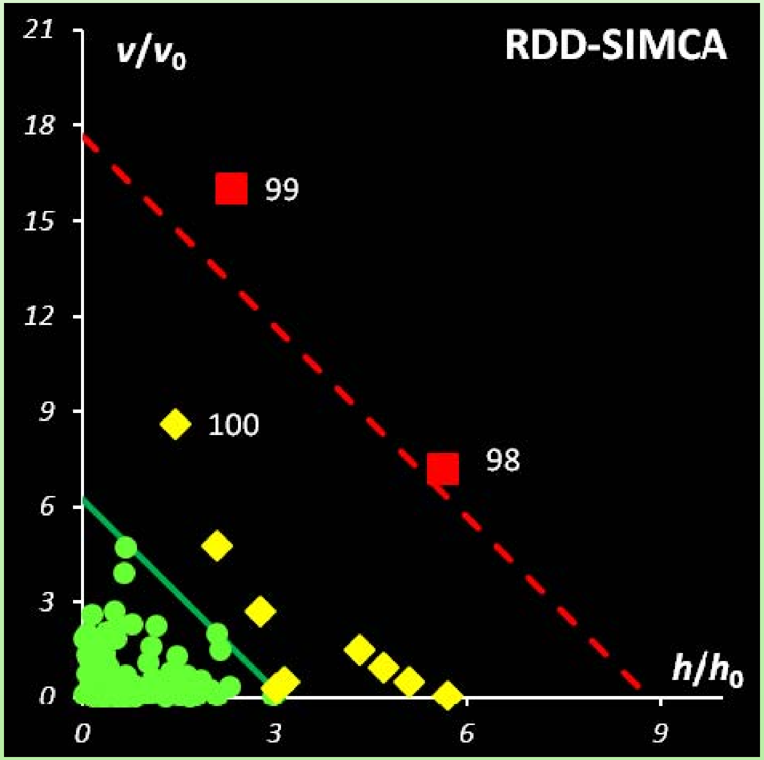


SIMCA plots

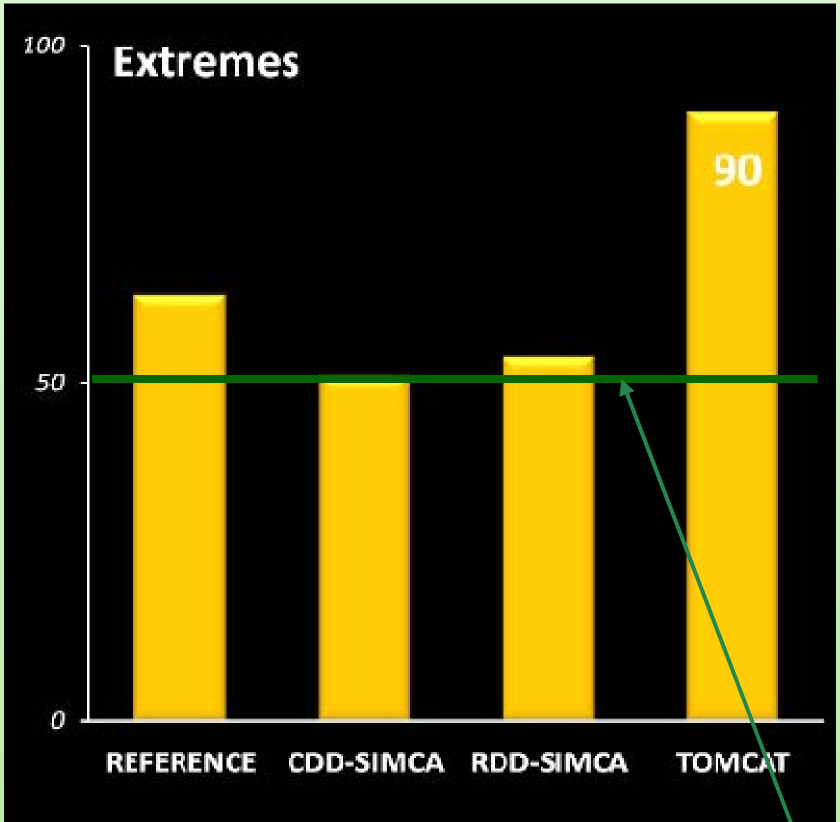

REFERENCE
v/v0
h/h0
99
100
98
CDD-SIMCA
v/v0
h/h0
99
100
98
RDD-SIMCA
v/v0
h/h0
99
100
98
TOMCAT
ROD
RD
98
99
100

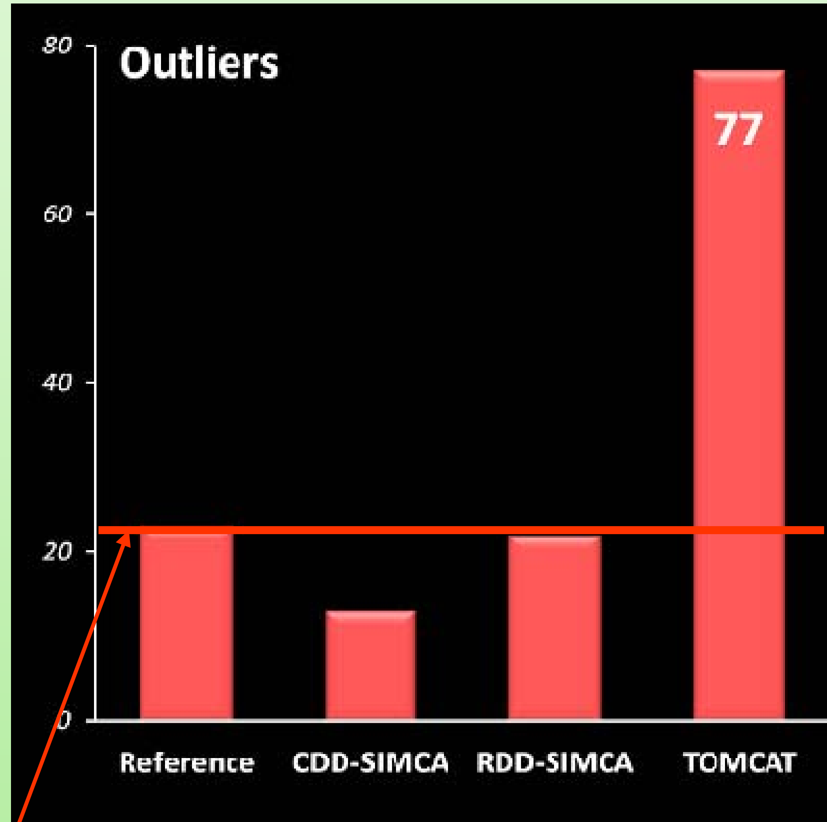
REFERENCE & RDD-SIMCA





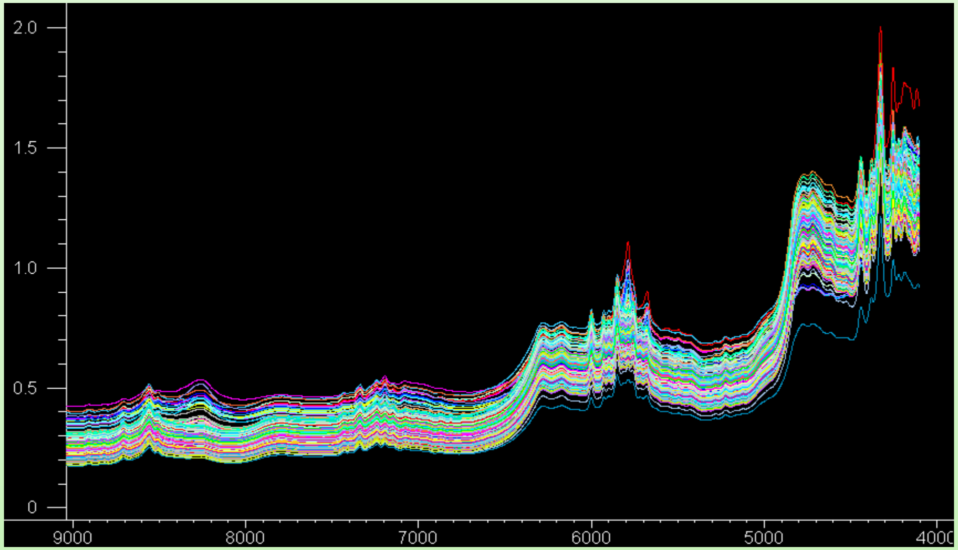
Totally in 10 data sets with outliers

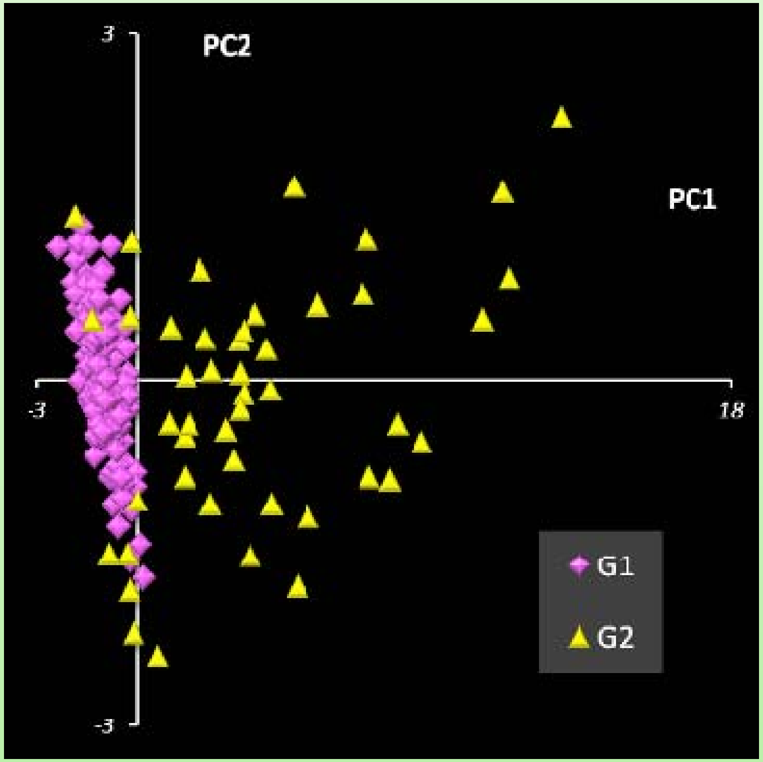




Expected

Case study II. Real world data with 2 groups





Substance in the closed PE bags,

82 drums measured by NIR.

Totally: 246 spectra

Group G1: 196 objects

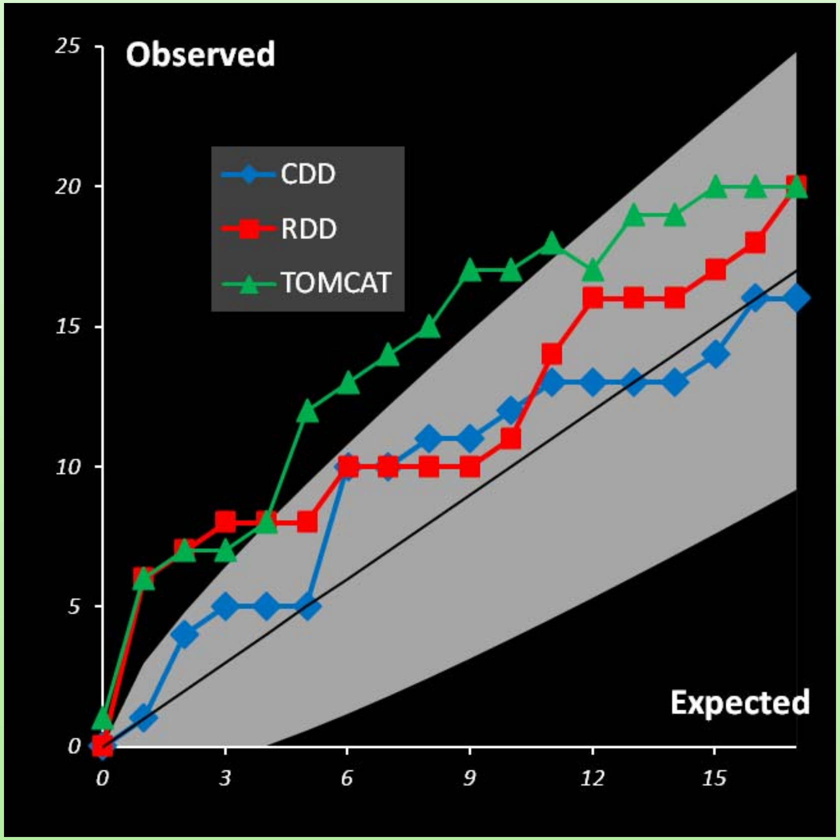
Group G2: 50 objects

ACA 642 (2009) 222-227

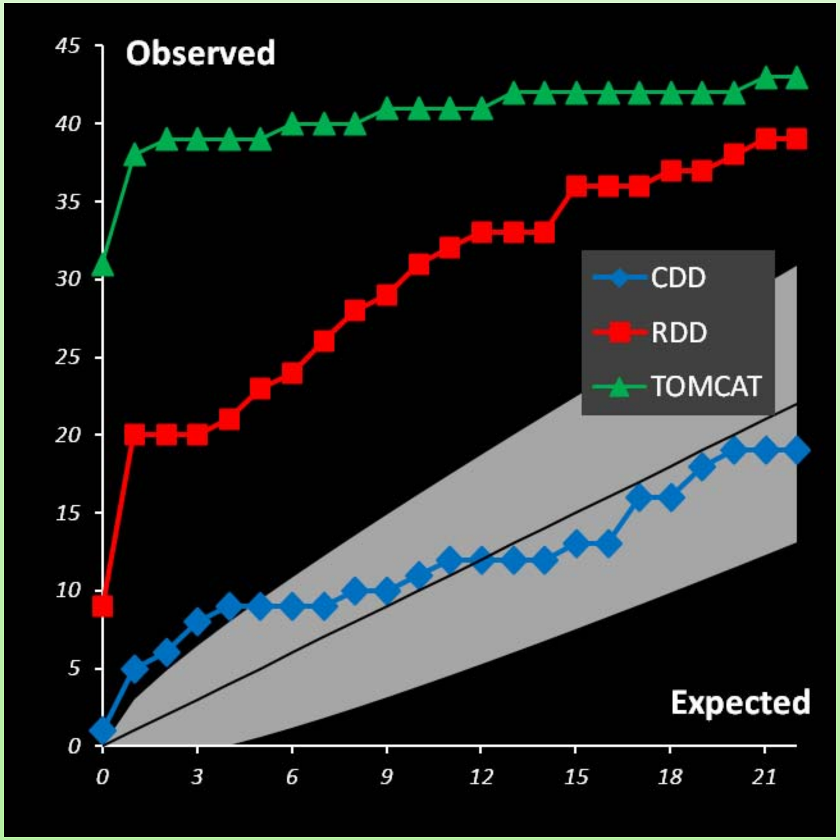
Expected/observed number of extremes

Expected number of extremes $N=\alpha l$

Clean subset G1

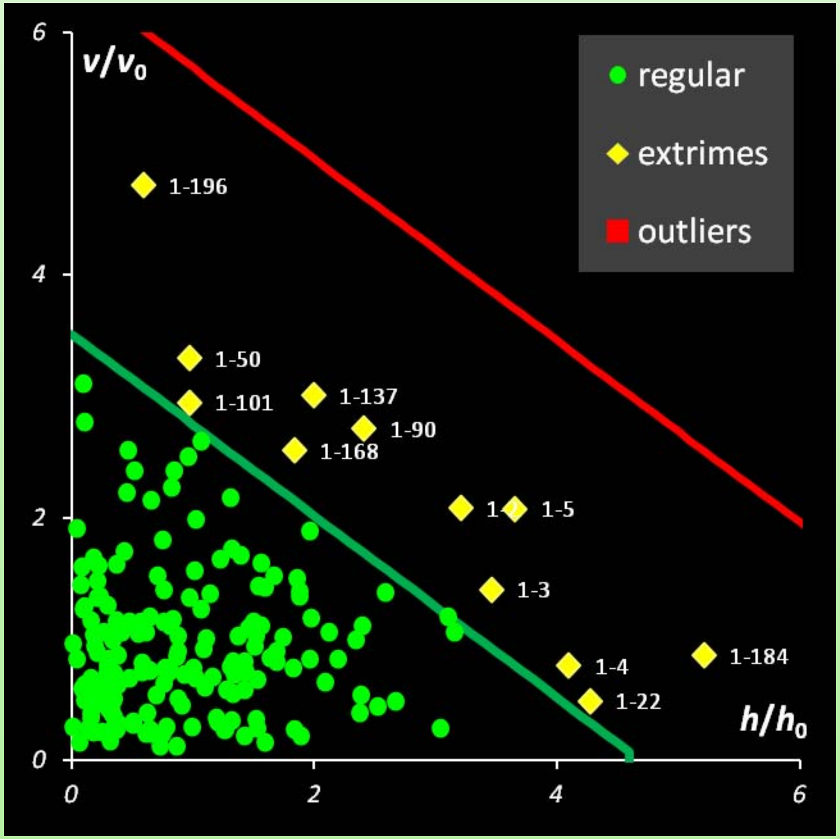


Contaminated dataset G1+G2

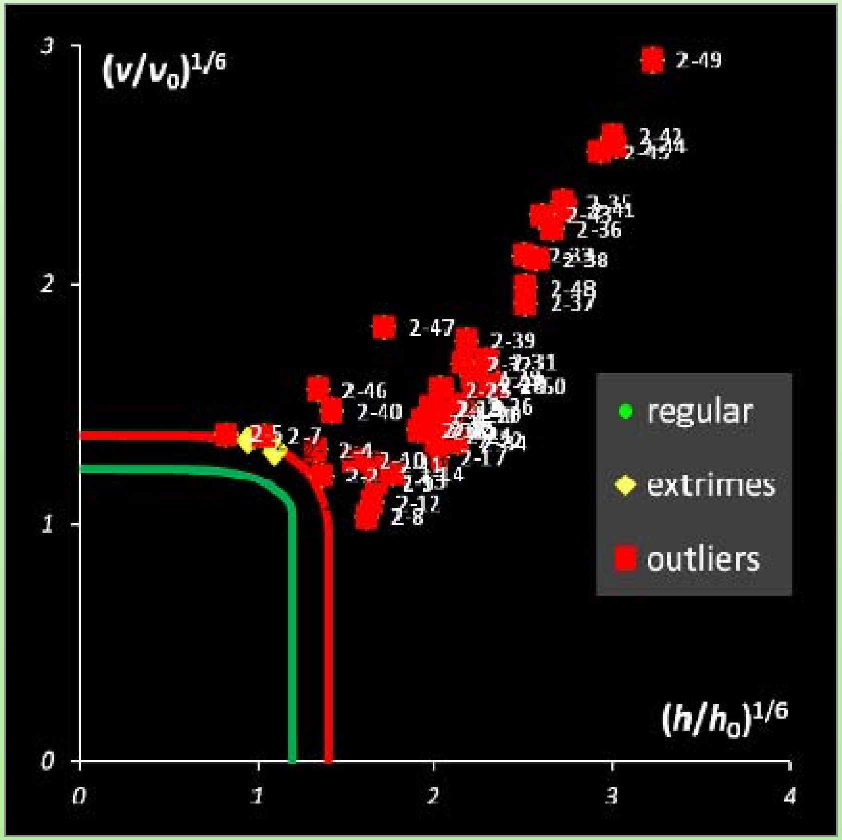


Results of separation

Subset G1 revealed



Subset G2 revealed



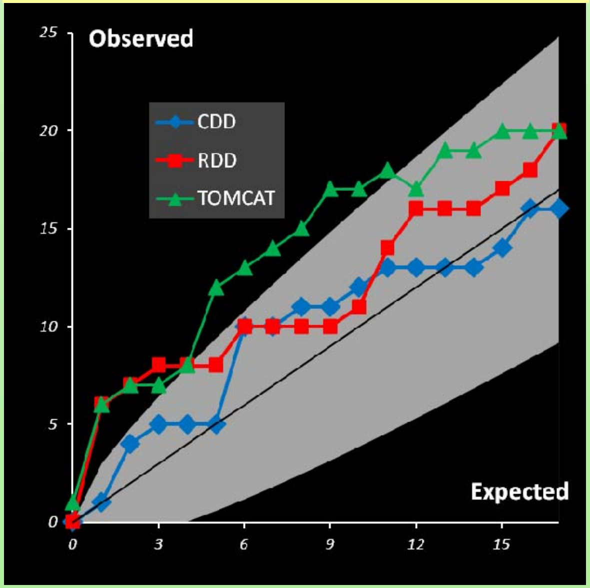
Conclusion 1

Each tool has its purpose: classical methods are for regular data, whereas robust methods should be used for contaminated data. Do not expect that there exists a common tool that yields reasonable results in both cases.

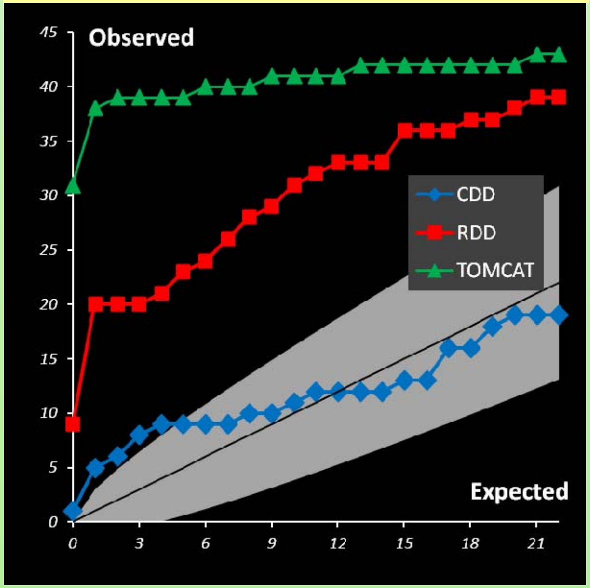
Conclusion 2

Extreme objects play an important role in data analysis. These objects should not be confused with outliers. The number of extremes should be compared to the expected number, coupled with the significance level α.

Clean dataset



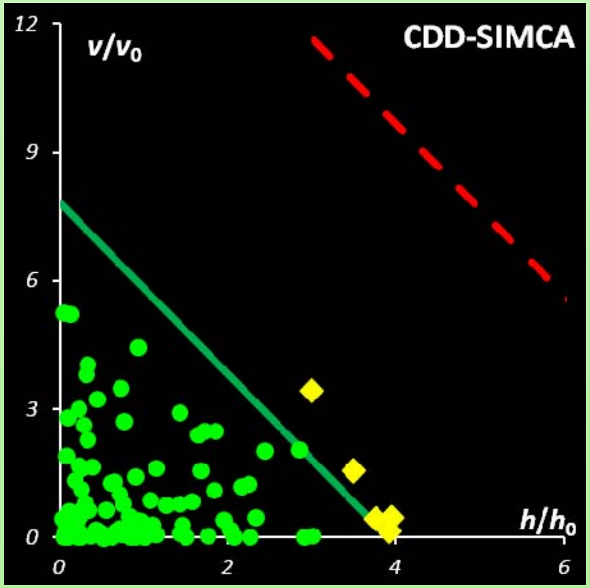
Contaminated dataset



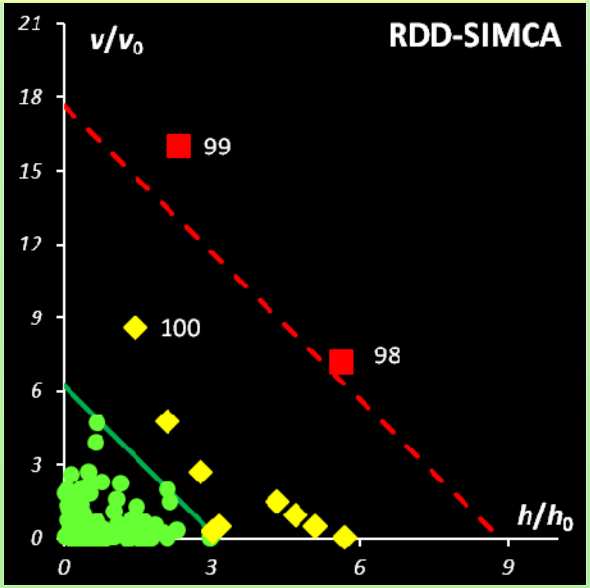
Conclusion 3

The proposed Dual Data Driven PCA/SIMCA approach looks like a fine competitor to the pure classical and to the strictly robust methods. This technique has demonstrated a proper performance in the analysis of both regular and contaminated data sets.

Clean dataset



Contaminated dataset



Thank you for your attention

A Lawyer's Mistake