Feasibility of Dual Data Driven SIMCA application for explorative analysis and classification of drugs with identical API content

O.Ye. Rodionova[1,2], N.G. Izmaylova[1], K.S. Balyklova[1], A.V.Titova[1], A.L. Pomerantsev[2,3]

[1]Information and Methodological Center for Expertise, Stocktaking and Analysis of Circulation of Medical Products, Slavyanskaya sq.,4-1, 109074, Moscow, Russia
[2]Semenov Institute of Chemical Physics RAS, Kosygin 4, 119991, Moscow, Russia, [3] Institute of Natural and Technical Systems RAS, Kurortny pr. 99/18, 354024, Sochi, Russia

Objectives:

To develop a math concept for the NIR based recognition of medicines with

(1) data-driven acceptance rules accounting for α-error

(2) data-driven extended acceptance area for the β-error control.

Samples:

Intact tablets packed in Polyvinyl Chloride (PVC) blisters.

Overall 457 tablets of Branded and Generic drugs are analyzed.

Measurements:

NIR spectra are acquired in 4000 - 12500 cm^{-1} range with resolution of 8 cm^{-1} by FTNIR spectrometer (Bruker Optics) equipped with a handheld fiber-optic probe.

Triplicate readings are averaged.

Dual Data-Driven SIMCA – a one-class classifier*

PCA

Def. Null class is a sample set which is used as the target for a one-class classifier.

PCA of Null class

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

Same **P** → PCA of an alternative class

$$\mathbf{T}' = \mathbf{X}'\mathbf{P} \qquad \mathbf{E}' = \mathbf{X}' - \mathbf{T}'\mathbf{P}$$

Same **T** → SDs and ODs

Score distance (SD)

$$h_i = \mathbf{t}_i^{t}(\mathbf{T}^{t}\mathbf{T})^{-1}\mathbf{t}_i \propto h_0\chi^2(N_h)$$

Chi-squared

$$h_i' = \mathbf{t}_i'^{t}(\mathbf{T}^{t}\mathbf{T})^{-1}\mathbf{t}_i' \qquad v_i' = \sum_{j=1}^{J} e_{ij}'^2$$

Orthogonal distance (OD)

$$v_i = \sum_{j=1}^{J} e_{ij}^2 \propto v_0\chi^2(N_v)$$

Total distance (TD)

Noncentral chi-squared

$$c_i' = N_h\frac{h_i'}{h_0} + N_v\frac{v_i'}{v_0} \propto c_0'\chi'^2(N_h + N_v, s)$$

Total distance (TD)

$$c_i = N_h\frac{h_i}{h_0} + N_v\frac{v_i}{v_0} \propto \chi^2(N_h + N_v)$$

False negative → Type I error **α**

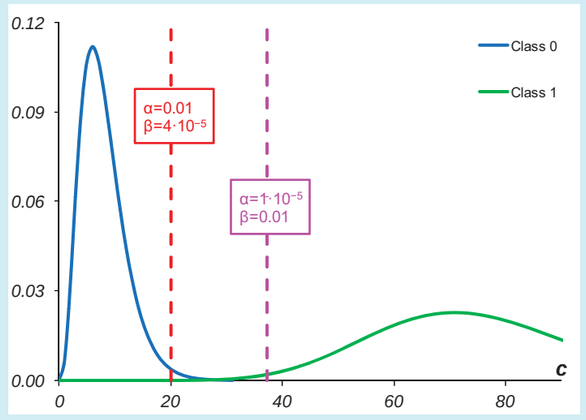
$$c_{\text{cutoff}} = \chi^{-2}(1-\alpha, N_h + N_v)$$

False positive → Type II error **β**

$$\beta = \Pr\left\{\chi'^2(N_h + N_v, s) < \frac{c_{\text{cutoff}}}{c_0'}\right\}$$

*AL Pomerantsev, OYe Rodionova "Concept and role of extreme objects in PCA/SIMCA" *J. Chemometrics*, 2013, (DOI: 10.1002/cem.2506).

Acceptance and extended acceptance areas



The acceptance area is designed by the α–error to ensure a small given rate of false negative decisions w.r.t. the target (null) class.

The extended acceptance area is constructed by the β–error to guarantee a small (given) rate of false positive decisions w.r.t. the nearest alternative class.

Depending on a specific goal, the β–error is established by the α–error, or vice verse.

N_h, N_v are the numbers of degrees of freedom, h_0,v_0,c_0 are the scaling parameters.

All these parameters **are estimated from the training and test set**, but not derived from theory.

Case Study 1 (77 objects). Branded drug

Pancreatin, digestive enzyme, coated tablets, with a high concentration of API, produced by one manufacturer.

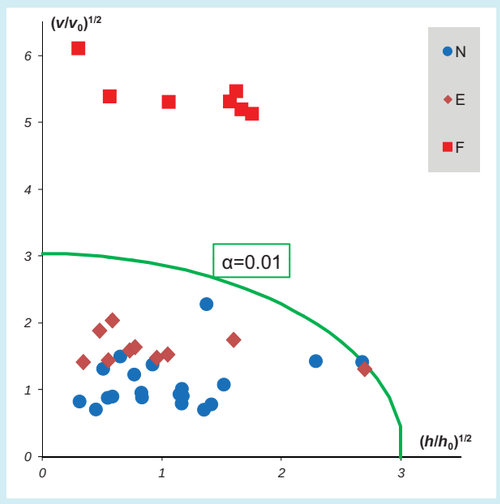
Training set: 40 samples from 4 genuine batches.

Test set: Subset "F" , 7 samples of counterfeited drug.
Subset "E" 10 samples from two genuine batches with expired shelf life.
Subset "N" 20 samples from two genuine batches with actual shelf life.

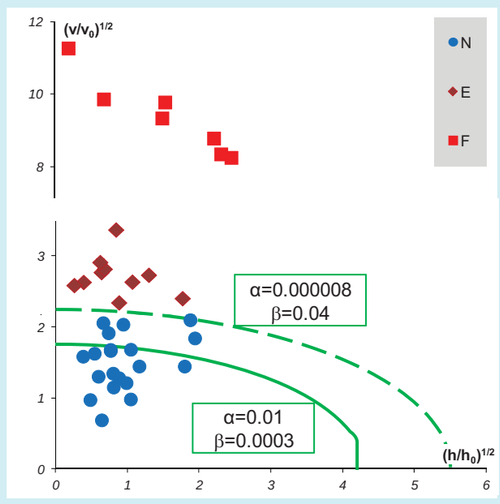
Spectral region: 7286 – 4056 cm^{-1}

Data pre-processing: 2-nd derivative by Savitzky-Golay with a 21 point window and 3-d order polynomial.

Test set recognition. SD-OD plots with the acceptance areas



1 PC (52% of the total variance), N_h=N_v=1



2 PCs (82% of the total variance), N_h=1 , N_v=5

The counterfeited tablets are easily separated from the genuine ones.

For model with 1 PC, the tablets with expired shelf-life, subset **E**, are not discriminated from the genuine ones.

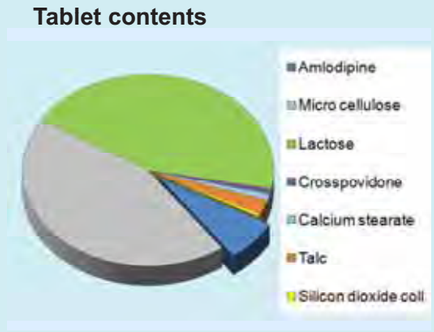
For model with 2 PCs, the **E** tablets are located rather close to the acceptance area.

The expired tablets may be included in the acceptance area, or be considered as aliens depending on the model complexity and the α-error.

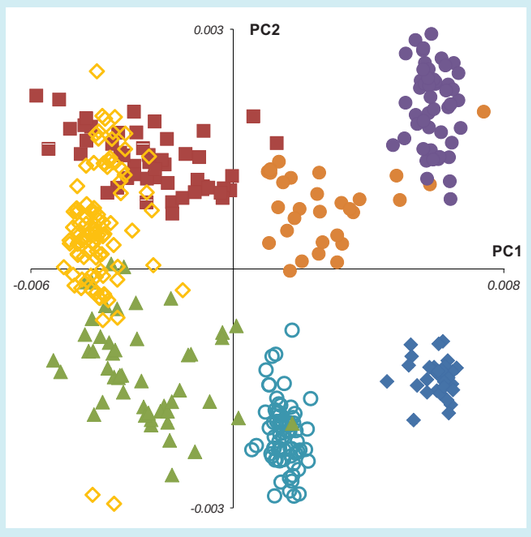
Case Study 2 (380 objects). Generic drug

Amlodipine 10mg, calcium channel blockers. Uncoated tablets with low concentration of API, produced by 7 manufactures in RF.

Tablet contents



Explorative PCA analysis (380 samples)

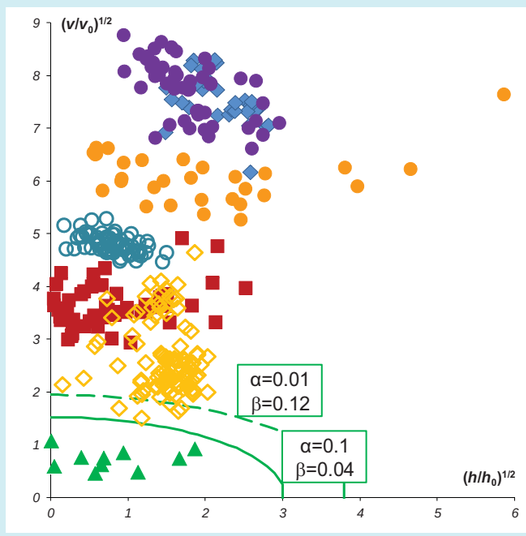


Training set: 40 samples from 4 batches of producer A4.

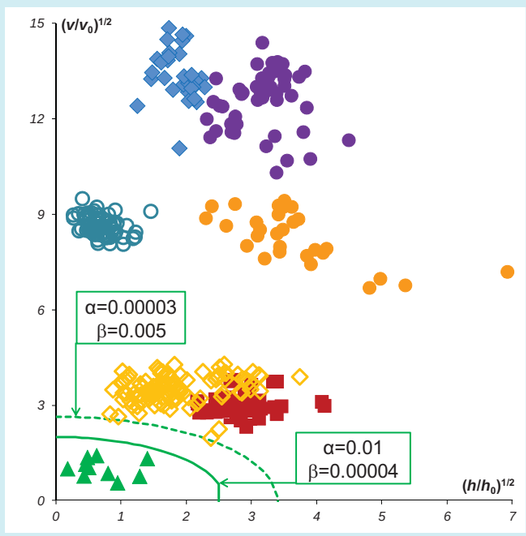
Test set: the rest of A4 (10 samp.) and all tablets from other producers (see Table).

Producer /dataset name	Number of batches	Number of tablets	Marker
A1	3	30	◆
A2	5	50	■
A3	7	70	○
A4	5	50	▲
A5	3	30	●
A6	5	50	●
A7	10	100	◇

Test set recognition. SD-OD plots with acceptance areas



1 PC (76% of the total variance), N_h=1 N_v=4



2 PCs (90% of the total variance), N_h=3 N_v=5

Classification models established with different number of PCs lead to different results.

Application of one-class classifier with data-driven acceptance limits and the ability to assign α-error value helps to solve real world problems for specific drugs.

Special test set consisting of similar quality objects produced by various manufactures enables to establish a Type II error for one-class classifier.

The evaluation of the β-error is done by subjecting the objects belonging to any other class to the model of the class under investigation. This helps to construct extended acceptance area.

Project Challenges

1.It is difficult to collect a really representative training set for the following reasons.

- Central laboratory receives samples directly from the producers. No samples with shelf life of about one/two years are included in the model.
- Manufactures may vary excipients' suppliers or some characteristics of technological process.

So the model and corresponding acceptance area should be possibly general to avoid misclassification of genuine samples.

2. It is incredibly hard to find counterfeited samples for each type of drug and as a result to test the model on its ability to recognize 'high quality fakes'.

In case no fakes are available, the test set can consist of similar drug of various producers. Such drugs should contain identical API, and similar composition of excipients.

The study is conducted within a State project aimed at the medicines quality monitoring and anti-counterfeiting. A part of this project involves establishing and managing a special Near Infra-Red Spectrometers (NIRS) network.