Journal of Pharmaceutical and Biomedical Analysis 98 (2014) 186–192

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

Quantitative risk assessment in classification of drugs with identical API content

O.Ye. Rodionova[a,b,*], K.S. Balyklova[a,c], A.V. Titova[a,d], A.L. Pomerantsev[b,e]

[a] *Information and Methodological Center for Expertise, Stocktaking and Analysis of Circulation of Medical Products, Slavyanskay sq., 4-1, 109074 Moscow, Russia*
[b] *N.N.Semenov Institute of Chemical Physics RAS, Kosygin 4, 119991 Moscow, Russia*
[c] *I.M. Sechenov First Moscow State Medical University, Trubetskaya str., 8. b.2, 119991 Moscow, Russia*
[d] *Pirogov Russian National Research Medical University, Ostrovityanov str., 1, 117997 Moscow, Russia*
[e] *Institute of Natural and Technical Systems RAS, Kurortny pr. 99/18, 354024 Sochi, Russia*

ARTICLE INFO

Article history:
Received 28 March 2014
Received in revised form 30 March 2014
Accepted 21 May 2014
Available online 29 May 2014

Keywords:
Counterfeiting
Near infrared spectrometry
DD-SIMCA
Misclassification errors
Future variability

ABSTRACT

When combating counterfeits it is equally important to recognize fakes and to avoid misclassification of genuine samples. This study presents a general approach to the problem using a newly-developed method called Data Driven Soft Independent Modeling of Class Analogy. The possibility to collect representative data for both training and validation is of great importance in classification modeling. When fakes are not available, we propose to compose the test set using the legitimate drug's analogs, manufactured by various producers. These analogs should have the identical API and a similar composition of excipients. The approach shows satisfactory results both in revealing counterfeits and in accounting for the future variability of the target class drugs. The presented case studies demonstrate that theoretically predicted misclassification errors can be successfully employed for the science-based risk assessment in drug identification.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

More and more falsified drugs are revealed all over the world. According to the European Agency for the Evaluation of Medicinal Products [1] the phenomenon has spread dramatically. Falsification cannot only be applied to branded products but also to generic ones. Different medicines are falsified, and the degree of falsification varies [2]. Among the so called 'low quality' fakes there are pills containing only chalk or starch, remedies without active pharmaceutical ingredient (API), or those containing active substances that are not listed on the packing, samples containing products produced with very low manufacturing quality, i.e. poorly pressed. However, another class of counterfeits exists; the so-called 'high quality' fakes. Ordinarily they are produced using modern equipment and they attempt to repeat the recipe of the genuine drugs.

Various approaches and methods are used to reveal fakes. An efficient approach is the application of the near infra-red (NIR) spectrometry with a subsequent chemometric data analysis [3]. This technique can be carried out not only in laboratories but also in stock houses, drugstores, etc. [4,5]. The NIR spectra are very informative, because they can reveal even non-essential variation in chemical content of an object, especially for organic matters. Moreover, the NIR signal penetrates into a sample and the diffuse reflectance spectra carry information regarding not only chemical but also physical properties, such as particle size and degree of crystallization. Since information is presented in a hidden way, various chemometric methods are used for spectra processing [6,7].

Two types of chemometric methods for qualitative analyses can be underlined. They are correlation techniques and factorization methods. The former include Wavelength Correlation (WC), Euclidian Distance (ED) [8], and similar approaches. These methods are used in European guideline that covers the application of NIR in the pharmaceutical industry [9] and utilized by software such as OPUS [10] and TQAnalyst [11] bundled with commercially available NIR instruments. A major advantage of correlation techniques is their simplicity. The main disadvantage is that processing data without separation of useful signal from noise results in poor discrimination of similar objects from different classes. Techniques based on data factorization, such as principal component analysis (PCA) and soft independent analogy of classes (SIMCA) [6], are more powerful. The NIR-based methods with application of data factorization provide

* Corresponding author at: Kosygin 4, 119991 Moscow, Russia. Tel.: +7 495 9397483; fax: +7 495 9397483.
E-mail address: rcs@chph.ras.ru (O.Ye. Rodionova).

http://dx.doi.org/10.1016/j.jpba.2014.05.033
0731-7085/© 2014 Elsevier B.V. All rights reserved.

This article is protected by the copyright law. You may copy and distribute this article for your personal use only. Other uses are only allowed with written permission by the copyright holder.

Table 1
Description of subsets for Dataset 1, Amlodipine.

Name	Marker	Batches	Tablets all together	Excipients	Tablet mass (mg)
A1	◆	3	30	Lactose, microcrystalline cellulose, magnesium stearate	300
A2	◆	5	50	Lactose, povidone, crospovidone, calcium stearate	200
A3	▲	7	70	Potato starch, lactose, microcrystalline cellulose, magnesium stearate, calcium stearate	200
A4	■	5	50	Corn starch, lactose, microcrystalline cellulose, magnesium stearate	180
A5	△	3	30	Potato starch, lactose, microcrystalline cellulose, magnesium stearate, calcium stearate	150
A6	●	5	50	Potato starch, lactose, microcrystalline cellulose, magnesium stearate, calcium hydrophosphate dihydrate, croscarmellose sodium, silicon dioxide colloidal	500
A7	●	10	100	Lactose, microcrystalline cellulose, magnesium stearate, croscarmellose sodium	200

possibility for rapid and non-destructive [12–14] analysis of a remedy as a whole object without quantitative determination of the API concentration or excipients' composition. As a rule the 'low quality' fakes are easily revealed by any of the abovementioned techniques. Revealing the 'high quality' counterfeits is not that straightforward and requires additional efforts.

A general chemometric approach consists of two important stages. At the first stage a model is developed using the training data set. Ideally this set should account for all possible and even future variations of the target class. At this stage an acceptance area for the target class is constructed and a decision rule is established. The second stage is validation. It is used for testing the performance of the developed model on samples, not used for model construction. For proper model validation the test set should include both genuine samples, being samples from the target class, and the counterfeited samples originating from alternative classes.

It should be recognized that any decision rule 'fake or genuine' cannot be perfect and inadvertent errors should always be expected. These errors are of two kinds [15]. The type I error (α) is the rate of wrong rejection of the target samples, while the type II error, β, is the rate of wrong acceptance of aliens as genuine objects. Both errors are harmful and, generally, for a given sample size, an effort to reduce one type of error results in an increase of the other type of error. An expert should account for the possible errors and develop a rule that is optimal with respect to a specific goal. Naturally, a pertinent tool is paramount to develop such a rule.

We suggest using a well known SIMCA method. It was first proposed in its simplest version in Ref. [16] and underwent several modifications [17–19] later on. Our version is called the data driven SIMCA (DD-SIMCA) [20] and has an ability to characterize classification results in a statistically sound way, i.e. to calculate the errors of misclassification theoretically. The development of various decision rules, both for a regular dataset and in presence of outliers, is described in Ref. [21]. In this paper we describe the application of DD-SIMCA as a part of a national project aimed at medicines quality monitoring and counterfeit combating.

Two important issues are considered here. Firstly, it is incredibly hard to find falsified samples for each type of drug and to test the ability to recognize the 'high quality' fakes. Secondly, it is difficult to collect a fully representative training set that accounts for possible future variations in the genuine drugs or even for their natural aging. Actually, these are two sides of the same problem. How does one develop a decision rule that is as general (i.e. wide) as possible to avoid misclassification of the genuine drug, and, at the same time, is strict enough to reveal falsified products? To resolve this dilemma we suggest using the legitimate analogs of the target class drugs that are manufactured by various producers. Such drugs should contain identical API and have a similar composition of excipients.

2. Materials and methods

All objects are intact tablets packed in Polyvinyl Chloride (PVC) blisters. The samples, except counterfeits from Dataset 2, are provided by the producers. There is no doubt about their authenticity. Thus, no tablets should be removed as outliers, as samples represent real world variations.

2.1. Dataset 1

When compared to a specific genuine class, a range of various producers may be used for assessing the target class acceptance areas, which account for a possible future variation of the target class. Dataset 1 simulates fakes when real counterfeited objects are unavailable.

Dataset 1 consists of uncoated tablets of Amlodipine. These are calcium channel blockers, produced by 7 different manufacturers located in Russia and denoted as A1, A2, . . ., A7. All producers employ the same quantity (10 mg) of the API originated from the same source. Each producer is represented by a set of batches ranging from three to ten. Each batch consists of 10 tablets. Overall there are 380 tablets in Dataset 1. The summary of this dataset is presented in Table 1.

2.2. Dataset 2

Dataset 2 consists of coated tablets of Pancreatin, a digestive enzyme with a high concentration of API, produced by one manufacturer. The data set includes subsets "G", "E" and "F". Subset "G" consists of six batches of genuine medicine (10 objects in each batch) with a valid shelf life and produced over the course of one year. Subset "E" includes two batches of genuine medicine (5 objects in each batch) with an expired shelf life. Subset "F" includes one batch of fakes (7 objects).

There are 77 objects in total in Dataset 2. Tablets in subset "G" can be divided into two groups. The first group of four batches

This article is protected by the copyright law. You may copy and distribute this article for your personal use only. Other uses are only allowed with written permission by the copyright holder.

Table 2
Description of subsets for Dataset 2, Pancreatin.

Name	Marker	Batches	Tablets all together	Comments
G1	🟢	4	4 × 10 = 40	Genuine tablets produced in spring
G2	▲	2	2 × 10 = 20	Genuine tablets produced in autumn
E	🔷	2	2 × 5 = 10	Genuine tablets with an expired shelf life
F	🟥	7	7	Fakes

consists of tablets produced in spring (G1: April and May). The second group of two batches comprises tablets produced in autumn (G2: August and September). These two groups help us to simulate a real-world situation of (routine) testing of samples produced during different periods. A summary of Dataset 2 is given in Table 2.

2.3. NIR measurements

NIR spectra were acquired in the interval 4,000–12,500 cm^{-1} with a resolution of 8 cm^{-1} using the FT-NIR spectrometer (MPA by Bruker Optics) equipped with a handheld fiber-optic probe (FP). Measurements were carried out in a diffuse reflection mode through a PVC blister. Each time triplicate readings were made to control reproducibility. Replicas were averaged for data analysis.

2.4. Chemometrics, DD-SIMCA

Chemometric analysis develops a decision rule that delineates the target class of a specific remedy by exploring its spectral properties. The procedure consists of two steps. The first step is the application of the Principal Component Analysis (PCA) [22]. The PCA model is established using training samples from the target class. The spectral $(I \times J)$ matrix $\mathbf{X}$ (duly preprocessed, e.g. centered) is decomposed by

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{t} + \mathbf{E} \tag{1}$$

where $\mathbf{T} = \{t_{ia}\}$ is the $(I \times A)$ scores matrix; $\mathbf{P} = \{p_{ja}\}$ is the $(J \times A)$ loadings matrix; $\mathbf{E} = \{e_{ij}\}$ is the $(I \times J)$ matrix of residuals; and A is the number of principal components (PC). Matrix $\mathbf{T}^{t}\mathbf{T} = \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_A)$ is a diagonal with elements $\lambda_a = \sum_{i=1}^{I} t_{ia}^2$, which are the eigenvalues of matrix $\mathbf{X}^{t}\mathbf{X}$ ranked in descending order.

At the second step, we employ the PCA results when calculating two relevant distances for each object $i = 1, \ldots, I$ of the training set. They are the score distance (SD), h_i, and the orthogonal distance (OD), v_i:

$$h_i = \mathbf{t}_i^{t}(\mathbf{T}^{t}\mathbf{T})^{-1}\mathbf{t}_i = \sum_{a=1}^{A} \frac{t_{ia}^2}{\lambda_a}, \qquad v_i = \sum_{j=1}^{J} e_{ij}^2 \tag{2}$$

The SD represents the position of a sample within the score space, and the OD characterizes a sample distance to the score space. Similar approach has been used in Ref. [23], where a new multivariate approach for the statistical evaluation of NIR chemical images was developed. The authors calculate Hotelling's T^2 values, which correspond to the score distances.

Our approach has some significant features that set it apart from the one mentioned above. In paper [20], it is shown that distributions of both distances are well approximated by the scaled chi-squared distribution

$$N_h \frac{h}{h_0} \propto \chi^2(N_h), \qquad N_v \frac{v}{v_0} \propto \chi^2(N_v) \tag{3}$$

where v_0 and h_0 are the scaling factors, N_h and N_v are the numbers of the degrees of freedom (DoF). These parameters are considered unknown and estimated using a data-driven method explained in [21].

Statistics c called the *total distance*

$$c = N_h \frac{h}{h_0} + N_v \frac{v}{v_0} \propto \chi^2(N_h + N_v) \tag{4}$$

is used to generate the decision rules. Any decision rule (i.e. an acceptance area) is determined by an inequality

$$c \le c_{\mathrm{crit}}. \tag{5}$$

The first decision rule is developed for the given type I error α,

$$c_{\mathrm{crit}} = \chi^{-2}(1 - \alpha, N_h + N_v) \tag{6}$$

where χ^{-2} is the quantile of the chi-squared distribution with $N_v + N_h$ DoF. To calculate the type II error β, we should assume that an alternative class is available. Then

$$\beta = Pr\left\{ \chi'^2(N_h + N_v, s) < \frac{c_{\mathrm{crit}}}{c'_0} \right\}, \tag{7}$$

where c_{crit} is defined in Eq. (5) and χ'^2 is the noncentral chi-squared distribution. Parameters c'_0 and s are found by the method explained in [24].

Using this approach, every sample and the acceptance areas can be plotted in the coordinates of SD against OD. See Fig. 2 as an example of this distance plot. Applying theory (Eqs. (6) and (7)) to practice, we can yield two acceptance areas. Firstly, we can develop a *regular* decision rule defined by a given α, and then calculate a subsequent β error. On the other hand, we can employ the *extended* rule (area) defined by β, and obtain a sequent α. The first rule is stronger, because all samples accepted by the regular rule are simultaneously accepted by the extended rule, but the converse is not true.

3. Results and discussion

3.1. General

Here, two stages of data processing are conducted in line with a general chemometric approach. A PCA model with the pertinent number of PCs is established. When PCA is used for further prediction purposes, the number of PCs is selected in order to minimize the prediction error. On the contrary, in the SIMCA method the goal of the PCA modeling is to describe the most common features of the investigated class without specifying individual properties of objects or batches included in the training set. Hence, parsimonious models are preferable. After establishing a PCA model, a regular acceptance area is developed by estimating the degrees of freedom for SD and OD, and calculated in line with Eq. (6). Varying the value of α-error allows evaluating the characteristics of the established decision rule.

Validation stage is conducted with the help of a complex test set. This set includes objects of the target class that are not used for model development. In our particular case it is desirable to use the genuine tablets from the batches that are not used in the training set. Such objects help to evaluate model behavior of new samples from the target class. Moreover, it is very important to include alien objects of different types into the test set. A specially selected β-error helps to avoid misclassification of alien objects that are close to the target class and simulate 'high quality' fakes. If the alien

This article is protected by the copyright law. You may copy and distribute this article for your personal use only. Other uses are only allowed with written permission by the copyright holder.

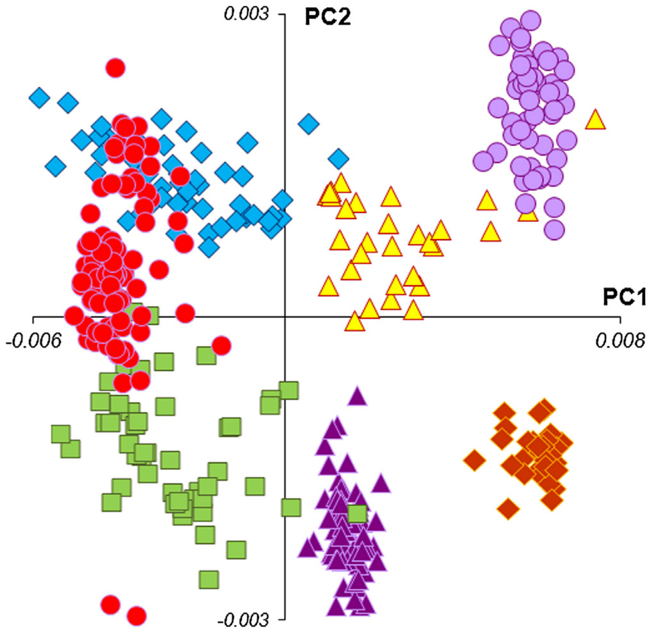

Fig. 1. Explorative analysis. Amlodipine data. PCA using the whole Dataset 1. Scores plot PC1 vs PC2. Markers are shown in Table 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

objects are located far from the regular acceptance area, it is possible to establish an extended area. This area is developed for a given β-error, the sequent α-error is calculated thereafter. Thus, the more versatile the test set the more reliable and informative the decision rule.

In the present study all spectra were pre-processed by the second order Savitzky–Golay differentiation with a 21 point window and a third order polynomial [6]. This transformation was used to remove most artifacts caused by the presence of a PVC blister and a handheld FP application [25]. It is worth mentioning that such transformation decreases the PCA model complexity as it removes low-frequency noise and sharpens spectra peaks.

3.2. *Case study 1. Amlodipine*

The application of DD-SIMCA to Dataset 1 demonstrates the method's performance in absence of falsified samples. For this purposes we verify the model established for batches produced by the same manufacturer with the help of a specially designed test set. Firstly, the test set includes samples produced by the same manufacturer, but not included in the training set previously. Secondly, we add samples produced by other manufacturers. They help to verify the model's performance regarding alien objects. Tablets that are essentially dissimilar from the target class simulate 'low quality' fakes. Tablets close to the target class imitate 'high quality' fakes. The latter help to analyze the most challenging case and to assess the value of type II error.

Preliminary PCA performed on the whole data revels both big similarities and differences in tablets produced by various manufacturers (Fig. 1). This can be explained by different sets of excipients, variability in pelletizing technology, employment of different equipment. It is also important to note that tablets produced by various manufacturers have different mass (Table 1). Though all tablets contain the same API quantity, the mass fraction of API varies from 2% (w/w) (manufacturer A6) to 6.6% (w/w) (A5). These differences also lead to spectra discrepancies, as the spectral absorbance reflects components' concentrations.

Individual decision rules are developed for each producer using the DD-SIMCA method. For this purpose several batches of a target producer are collected in the training set. The test set is formed by one or two batches of the target objects and by the tablets of the other producers. To illustrate, let us choose producer A4 as a target class and analyze all other samples against a model developed for this class. Four batches (40 spectra) are collected in the training set. The test set contains one batch (10 objects) from class A4 and all tablets from the other producers, 340 objects in total.

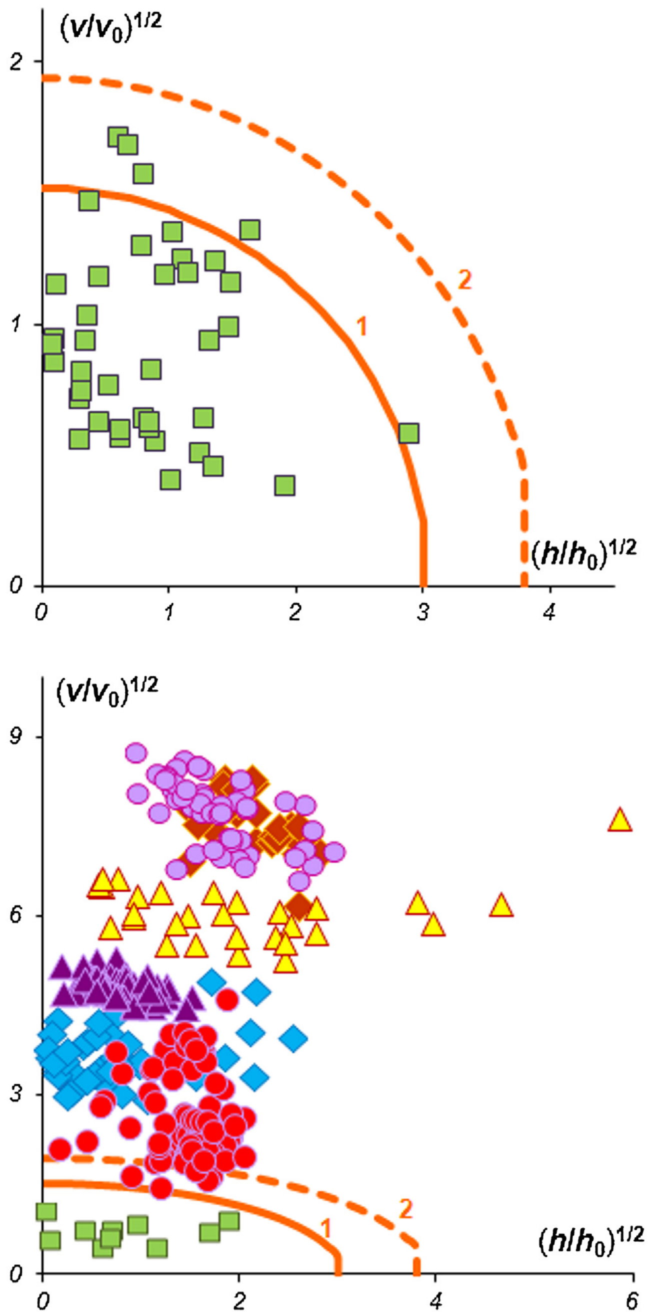

Fig. 2. Amlodipine, producer A4 is employed as a target class. PCA model with one PC. Acceptance areas: $\alpha = 0.1$ (1), $\alpha = 0.01$. (2) Top panel: training set; bottom panel: test set.

The PCA model with one PC explains 76% of the total variance. The decision rules are constructed for the chi-squared distributions (Eq. (3)) with $N_h = 1$, $N_v = 4$ DoFs. For $\alpha = 0.01$ (Fig. 2, the top panel, curve 2) all training objects are located inside the acceptance area. As for the test set (Fig. 2, the bottom panel), the objects originated from all producers except A7 are located far from the acceptance area (curve 2) and can easily be classified as aliens. All test objects

This article is protected by the copyright law. You may copy and distribute this article for your personal use only. Other uses are only allowed with written permission by the copyright holder.

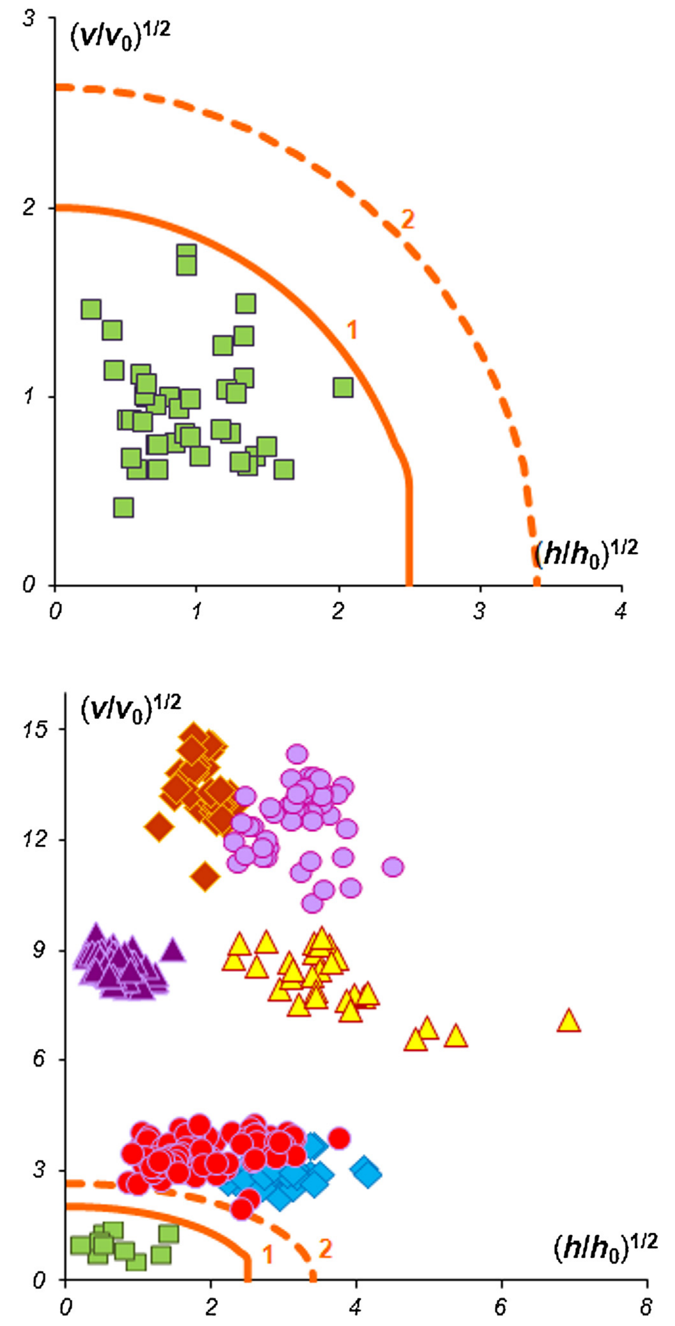

Fig. 3. Amlodipine, producer A4 is used as the target class. PCA model with two PCs. Acceptance areas: regular at $\alpha = 0.01$ (1); extended at $\beta = 0.005$ (2). Top panel: training set; bottom panel: test set.

from target class A4 are classified properly, but 4 objects originated from subset A7 are wrongly attributed to the target class, that is a false acceptance. Considering objects from two classes A4 and A7, it is possible to calculate the empirically observed α and β errors, and then compare them with the tentatively predicted counterparts.

Table 3
Dataset 1. Relationship between the α- and β-errors for models with various number of PCs.

Model complexity (number of PCs)	α-Error	β-Error
1	0.1	0.04
1	0.01	0.12
2	0.01	0.000005
2	0.00003	0.005

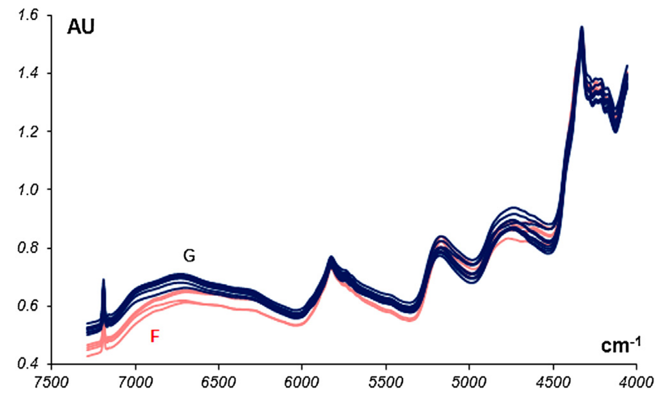

Fig. 4. Dataset 2. Raw NIR spectra of Pancreatin. Blue 'G' spectra, red 'F' spectra. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

For $\alpha = 0.01$, the theoretical sequent β equals 0.035. This means that for subset A7 we can expect 3–4 objects to be wrongly attributed to the target class. In the example, we have 4 such objects. To separate A4 and A7 completely, we should set $\alpha = 0.1$ (Fig. 2, curve 1). In this case, all objects from the test set are properly attributed. At the same time, 5 objects from the training set are located beyond the acceptance area (see Fig. 2, the top panel). This is a statistically sound result, as for a set of 40 objects and at $\alpha = 0.1$, one can expect four wrongly rejected objects.

For a more reliable object classification we establish a PCA model with 2 PCs, which explains 90% of the total variance. The decision rules are constructed for the chi-squared distributions (Eq. (3)) with $N_h = 3$, $N_v = 5$ DoFs. In the test set, the discrimination between the A4 objects and all other producers is much better. At $\alpha = 0.01$ (Fig. 3, curve 1), the objects are clearly separated, and all A4 test objects are located inside the acceptance area. Again, class A7 is the closest one, though it is located at some distance.

Taking into account that the training set may not be fully representative, it is possible to extend the acceptance area to avoid false rejection decisions in the future when new genuine tablets are tested. Considering class A7 as a natural border for new genuine samples from class A4, we can set an appropriate β-value and compute the corresponding α-value. For example, at $\beta = 0.005$, the extended acceptance area corresponds to an α error of 0.00003 (Fig. 3, curve 2). The relationship between the α- and β-errors for the PCA models of different complexity is presented in Table 3.

In this case study, we demonstrated that the PCA model with two PCs reliably separates the target class from all other classes. Samples from class A7 simulate the 'high quality' fakes. The proposed classification method recognized such alien objects successfully. The competing correlation approach failed to discriminate producers A4 and A7.

Additionally, we used a possibility to extend the acceptance area with a goal to account for future variability in class A4.

Similar studies were repeated for cases where each manufacturer was selected as a target class. The results provided individual decision rules with specific values for α and β errors.

3.3. Case study 2. Pancreatin

Dataset 2 is used to illustrate DD-SIMCA ability to reveal real falsified objects and to classify tablets of different grade. A total of 30 tablets from three genuine batches (subset G1) produced in spring were used for training. The test set includes 47 objects that are fakes (subset F); two batches of genuine tablets with expired shelf life (subset E); one genuine batch produced in spring but excluded

This article is protected by the copyright law. You may copy and distribute this article for your personal use only. Other uses are only allowed with written permission by the copyright holder.

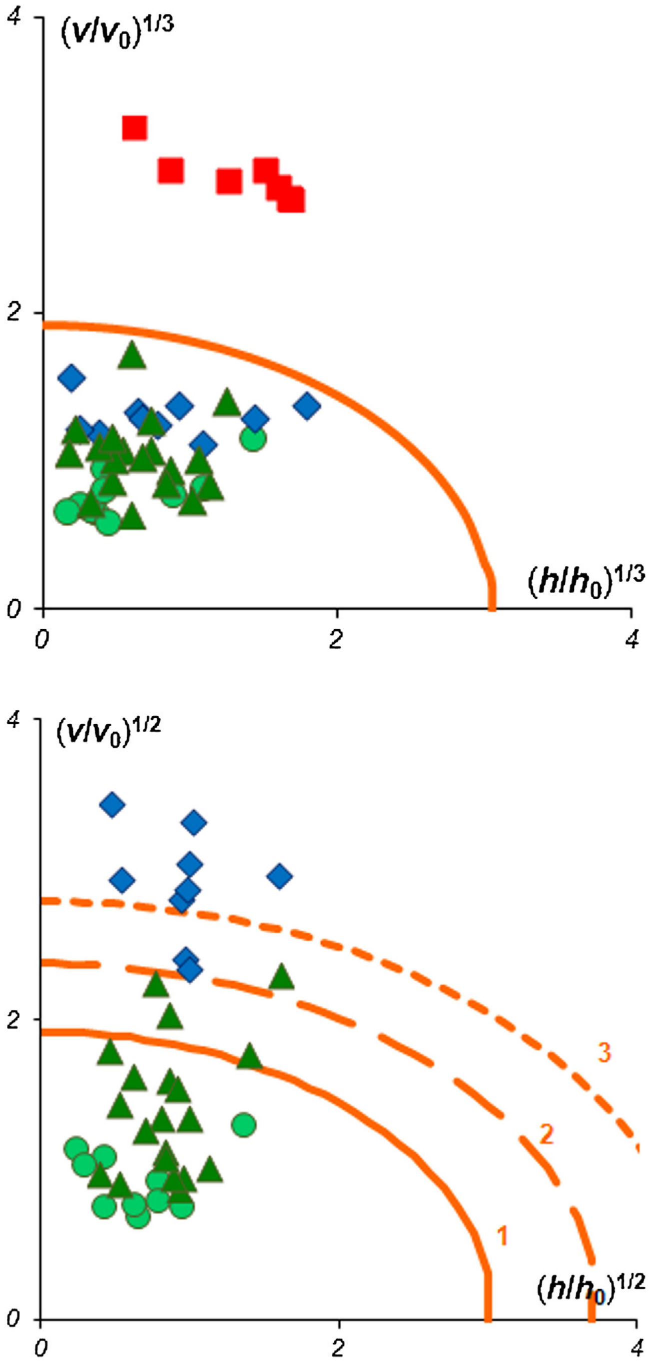

Fig. 5. Distance plots. Pancreatin data. Test set. Markers are shown in Table 2. Top panel: model with 1 PC, acceptance area: $\alpha = 0.01$. Bottom panel: model with 3 PCs, acceptance areas: $\alpha = 0.01$ (1); $\alpha = 0.0001$ (2), $\alpha = 10^{-6}$ (3).

from the training set; and two genuine batches with actual shelf life produced in autumn (subset G2). The spectra (Fig. 4) demonstrate that samples from subset "F" are very similar to the genuine tablets. The correlation method misclassifies four of seven falsified samples and is unable to separate the expired tablets ('E') from the genuine ones ('G').

The most informative spectral region 7286–4056 cm^{-1} is employed for data processing. The PCA shows that if only one PC is used, 42% of the total variance is explained. Even the model with one PC can distinguish fakes from all other samples. The decision rule is constructed for the chi-squared distributions given by Eq. (3) with $N_h = 1$, $N_\nu = 2$ DoFs. For $\alpha = 0.01$ all objects except the counterfeited tablets, are located inside the acceptance area. The 'G' and 'E' tablets are merged and classified as members of the target class (Fig. 5, the top panel). All objects from subset 'F' are located far from the acceptance area and are correctly classified as counterfeited tablets. They are no longer interesting for us, and we will concentrate on the objects from subset 'E' (expired shelf-life) which are not identified as aliens. To establish a more critical model we should increase the model complexity.

The PCA model that can distinguish tablets of subset 'E' from genuine tablets uses three PCs, which explain 90% of the total variance. The decision rules are constructed for the chi-squared distributions given by Eq. (3) with $N_h = 2$, $N_\nu = 5$ DoFs. Unlike the model with one PC, all 'E' objects form a separate group. For $\alpha = 0.01$ (curve 1 in Fig. 5, the bottom panel) all genuine test samples produced in spring are correctly located inside the acceptance area. Here, the β-error calculated with respect to alternative subset 'E' is equal to 0.01, and none of the samples from subset 'E' are misclassified. At the same time, four out of twenty genuine tablets produced in autumn are misclassified as aliens, though they are located rather close to the acceptance area. Extending the acceptance area at $\alpha = 0.0002$ (curve 2 in Fig. 5, the bottom panel) we can see that only one such sample is misclassified. The corresponding β-error equals 0.09. Decreasing the α value we increase the risk of wrong acceptance (type II error). For example, at $\alpha = 2 \times 10^{-6}$ (curve 3 in Fig. 5, the bottom panel), two objects from subset 'E' are wrongly accepted as genuine tablets. The corresponding β-error equals 0.3. Tuning the α-error does not influence the decisions regarding subset 'F' for the current dataset, as these objects are located very far from any reasonable acceptance area.

This case study demonstrates that the suggested procedure is rather critical and works reliably to detect truly falsified drugs. Even though 'F' tablets can be considered 'high quality' counterfeits, the model with one PC can recognize them as aliens. A more complex model which helps to reveal the 'E' subset requires 3 PCs. Note that the application of a more critical model which is able to explain the bulk of variation in the training data, such as the model with three PCs used in the latter case, requires using an extended acceptance area. In case a model with one PC is applied, all genuine tablets produced both in spring and in autumn are reliably classified as target ones.

Our experience shows that in many cases it is difficult to discriminate tablets with expired shelf-life time from those with valid life-time. Expired tablets may be included in the acceptance area, or be considered as aliens depending on the model complexity and the expert-defined α-error.

4. Conclusions

In this paper we have considered a task of revealing counterfeited drugs by the NIR spectroscopy with subsequent chemometric modeling. We have proposed to employ a special test set comprised of legitimate analogs of a target class drugs manufactured by various producers. This test set helps to develop a decision rule as general as possible to avoid misclassification of the genuine drug, and, at the same time, is strict enough to reveal the falsified products of different types. The relevant acceptance areas are established using a newly developed data-driven SIMCA method that allows to compute the misclassification errors theoretically.

The real world examples demonstrate that the proposed technique is flexible enough. The method is capable to recognize 'high quality' fakes successfully. Additionally, with tuning the type I error α, it is possible to extend the acceptance area in order to account for future variability in new genuine samples of the target class. A provisional prediction of the α and β errors can be employed for the science-based risk assessment.

This article is protected by the copyright law. You may copy and distribute this article for your personal use only. Other uses are only allowed with written permission by the copyright holder.

References

[1] Available from: http://www.ema.europa.eu/ema/index.jsp?curl=pages/special_topics/general/general_content_000186.jsp&mid=WC0b01ac058002d4e8

[2] World Health Assembly, Counterfeit Drugs: Threat to Public Health, vol. 55, World Health Assembly, Geneva, 2002.

[3] O.Ye. Rodionova, A.L. Pomerantsev, NIR based approach to counterfeit-drug detection, Trends Anal. Chem. 29 (2010) 781–938.

[4] F.E. Dowell, E.B. Maghirang, F.M. Fernandez, P.N. Newton, M.D. Green, Detecting counterfeit antimalarial tablets by near-infrared spectroscopy, J. Pharm. Biomed. Anal. 48 (3) (2008) 1011–1014.

[5] O.Ye. Rodionova, Ya.V. Sokovikov, A.L. Pomerantsev, Quality control of packed raw materials in pharmaceutical industry, Anal. Chim. Acta 642 (2009) 222–227.

[6] T. Naes, T. Isaksson, T. Fearn, T. Davies, Multivariate Calibration and Classification, Wiley, Christerer, 2002.

[7] Y. Roggo, P. Chalus, L. Lene Maurer, C. Lema-Martinez, A. Edmond, N. Jent, A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies, J. Pharm. Biomed. Anal. 44 (2007) 683–700.

[8] J. Li, D.B. Hibbert, S. Fuller, G. Vaughn, A comparative study of point-to-point algorithms for matching spectra, Chemom. Intell. Lab. Syst. 82 (2006) 50–58.

[9] Note for guidance on the use of Near infrared spectroscopy by pharmaceutical industry and the data requirements for new submission and variations. CPMP/QWP/3309/01, EMEA, 2003.

[10] OPUS Version 7.0, Bruker Optik GmbH, 2011.

[11] TQ Analyst 8.5.21, Thermo Fisher Scientific, 1996–2011.

[12] L. Alvarengaa, D. Ferreiraa, D. Altekruseb, J.C. Menezesa, D. Lochmann, Tablet identification using near-infrared spectroscopy (NIRS) for pharmaceutical quality control, J. Pharm. Biomed. Anal. 48 (2008) 62–69.

[13] O.Ye. Rodionova, A.L. Pomerantsev, L. Houmuller, A.V. Shpak, O.A. Shpigun, Noninvasive detection of counterfeited ampoules of dexamethasone using NIR with confirmation by HPLC-DAD-MS and CE-UV methods, Anal. Bioanal. Chem. 397 (2010) 1927–1935.

[14] P.-Y. Sacré, E. Deconinck, T. De Beer, P. Courselle, R. Vancauwenberghe, P. Chiap, J. Crommen, J.O. De Beer, Comparison and combination of spectroscopic techniques for the detection of counterfeit medicines, J. Pharm. Biomed. Anal. 53 (2010) 445–453.

[15] C. Hartmann, J. Smeyers-Verbeke, D.L. Massart, R.D. McDowall, Validation of bioanalytical chromatographic methods, J. Pharm. Biomed. Anal. 17 (1998) 193–218.

[16] S. Wold, M. Sjostrom, SIMCA: a method for analyzing chemical data in terms of similarity and analogy, in: B.R. Kowalski (Ed.), Chemometrics Theory and Application, American Chemical Society Symposium Series, vol. 52, American Chemical Society, Washington, DC, 1977, pp. 243–282.

[17] G.R. Flåten, B. Grung, O.M. Kvalheim, A method for validation of reference sets in SIMCA modeling, Chemom. Intell. Lab. Syst. 72 (2004) 101–109.

[18] M. Hubert, P.J. Rousseeuw, K. Vanden Branden, ROBPCA: a new approach to robust principal component analysis, Technometrics 47 (2005) 64–79.

[19] C. Durante, R. Bro, M. Cocchi, A classification tool for N-way array based on SIMCA methodology, Chemom. Intell. Lab. Syst. 106 (1) (2011) 73–85.

[20] A.L. Pomerantsev, Acceptance areas for multivariate classification derived by projection methods, J. Chemom. 22 (2008) 601–609.

[21] A.L. Pomerantsev, O.Ye. Rodionova, Concept and role of extreme objects in PCA/SIMCA, J. Chemom. 28 (2014) 429–438.

[22] H.T. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1998.

[23] T. Puchert, D. Lochmann, J.C. Menezes, G. Reich, A multivariate approach for the statistical evaluation of near-infrared chemical images using Symmetry Parameter Image Analysis (SPIA), Eur. J. Pharm. Biopharm. 78 (2011) 117–124.

[24] A.L. Pomerantsev, O.Ye. Rodionova, On the type II error in SIMCA method, J. Chemom. (2014), http://dx.doi.org/10.1002/cem.2610 (early view).

[25] O.Ye. Rodionova, K.S. Balyklova, A.V. Titova, A.L. Pomerantsev, The influence of fiber-probe accessories application on the results of near-infrared (NIR) measurements, Appl. Spectrosc. 67 (12) (2013) 1401–1407.

This article is protected by the copyright law. You may copy and distribute this article for your personal use only. Other uses are only allowed with written permission by the copyright holder.