Special Issue

JOURNAL OF CHEMOMETRICS

Received: 5 June 2007, Revised: 24 February 2008, Accepted: 26 February 2008, Published online in Wiley InterScience: 10 July 2008

(www.interscience.wiley.com) DOI: 10.1002/cem.1147

Acceptance areas for multivariate classification derived by projection methods

Alexey L. Pomerantsev[a]*

In the projection methods (PCA, PLS) two distance measures are of importance. They are the score distance (SD, *a.k.a.* leverage) and the orthogonal distance (OD, *a.k.a.* the residual variance). This paper shows that both distance measures can be modeled by the χ^2-distribution. Each model includes a scaling factor that can be described by an explicit equation. Moreover, the models depend on an unknown number of degrees of freedom, which have to be estimated using a training dataset. Such modeling is further applied to classification within the SIMCA framework, and various acceptance areas are built for a given significance level. A triangular area, constructed using the sum of the normalized SD and OD, is deemed to be the most practical. This theoretical notion is supported by three examples. The first is based on a simulated dataset, while the other two employ real world data. Copyright © 2008 John Wiley & Sons, Ltd.

Keywords: PCA; SIMCA; leverage distribution; residual variance distribution; type I error; acceptance area; classification; influence plot; outlier

1. INTRODUCTION

Projection methods (principal component analysis (PCA), partial least squares (PLS), etc) are the most popular chemometric tools [1]. A simple geometrical representation of these methods provides a good illustration of the approach. Let matrix **X** consist of *I* rows that stand for the objects and *J* columns that correspond to the objects properties. The objects can either belong to different samples or represent a single sample that evolves during a process. The data matrix **X** can be viewed as a cloud of *I* points in the *J*-dimensional property space. It is important to note that in most chemistry-originated applications the cloud has a specific shape [2]. It is flattened in such a way that the points are located close to a hyperplane (a subspace) of the effective dimensionality $A < J$. The oblateness is mainly due to inter-correlations of the properties. The point of the space origin may be placed into the center of the cloud gravity; therefore the center belongs to the subspace as well. There are various techniques which help to reveal the subspace within the whole property space. The PCA uses only one block of data (matrix **X**). PLS regression additionally employs a response data matrix **Y**. However, the projection concept remains the core of both methods.

Each element of the data cloud can be presented as a sum of two vectors: a vector that lies in the subspace (a projection) and a vector transversal to the hyperplane (a residual). The lengths of these vectors are important indicators that characterize a sample position with respect to the subspace (model). These statistics are often referred to as the leverage and the residual variance. In this paper they will be termed as a score distance (SD) and an orthogonal distance (OD) correspondingly. Often the objects are assumed to be randomly selected members of the general totality (class). In this case, the SDs and ODs obtained for the known class members constitute two samplings that represent the population. By exploring these datasets the critical membership levels can be established. Therefore, when a new candidate object **x** is considered it can be projected onto the model subspace, and its own SD and OD values can be calculated. Further on, they are compared with the known critical levels to make a decision on the membership of the class.

The outlined approach has been applied in numerous applications, which can be divided into three main groups. The first one is, obviously, the SIMCA (soft independent modeling of class analogy) method [3], which is a popular chemometric tool for supervised pattern recognition, i.e. for the qualitative analysis. This is a standard problem of the affiliation with a predefined class [4–6]. A typical example has been explored in reference [7], where NIR spectroscopy was employed for the counterfeit drugs detection.

The second application is multivariate statistical process control [8,9]. In this case the class data **X** is a set of process variables measurements obtained under the normal operating conditions. The pertinent PCA model is built and the critical levels for the SD (termed as *D* statistics) and OD (termed as *Q* statistics) are established. A new process batch **x** is projected onto this model and the corresponding *D* and *Q* statistics immediately indicate whether the batch is in control.

The third area is multivariate calibration, namely PCR and PLS methods. A well-known influence plot [10] serves this. Each point of the plot represents a calibration object in the 'SD versus OD' coordinates. The point position marks the role of a sample in the calibration routine: the most influential objects are located in the peripheral areas with the large SD or OD values. The critical levels help to reveal outliers.

Within these applications several statistical problems are of vital importance. Firstly, the form of the SD and OD distributions should be specified. Moreover, in each specific case, the distribution parameters are to be evaluated using a training

* *Institute of Chemical Physics, Kosygin Street 4, Moscow 119991, Russia. E-mail: forecast@chph.ras.ru*

a *A. L. Pomerantsev*
Institute of Chemical Physics, Kosygin Street 4, 119991 Moscow, Russia

data set. Secondly, it is important to set up the rules that reveal the extremes and outliers in the data. Finally, the acceptance area in the influence plot should be defined. This area includes all probable SD–OD values, which indicate the affiliation of a sample to the class. The acceptance area is built with respect to a given type I error [11], which is an event when an object is erroneously assigned to be a class outsider, while it in fact belongs to the class. The corresponding probability is termed as a significance level.

The posed problems have already been discussed in numerous publications; this paper references some of them. We consider these questions to be still topical, however, as the proposed solutions are often inconsistent and contradict each other, for example, in a very popular software tool [12] the critical level for leverage (SD) does not vary with the significance value. This seems to be a practical implementation of the rule of thumb formulated in references [10,13], which states that the critical level of leverage equals its average value multiplied by 2 or 3. The situation with the residuals measures is more intriguing. For instance, paper [14] suggests two feasible critical levels: one based on *F*-distribution (p. 96) and another one, given by a formula derived from approximation of χ^2-distribution (p. 97). A recent discussion at the ICS list [15] has demonstrated that all such problems are still far from solution.

There is another motivation for this research that could be called a mystery of the influence triangle in the multivariate data analysis. It is well-known that the objects with both high leverage and high residual values are dangerous and thus they can be treated as outliers [10]. In other words, all regular calibration samples should be located within a triangle in the influence plot. However, there seems to be no proof, or substantiation for this evident conclusion—a typical SIMCA plot represents a rectangle, not a triangle. We failed to find any reference to such a fact excepting [12]. On the other hand, applying the SIC method [16] for the data analysis one immediately obtains the triangle of insiders, which are the most trustworthy objects within the calibration set.

This paper aims to trigger an open discussion on the topic that could be formulated as follows. How does one establish a proper acceptance area with respect to a given significance level? The presented study does not give the final solutions of the posed problem but presents and substantiates one of the possible ways of the critical limits calculations.

2. THEORY

2.1. Notation

Small bold characters, i.e. **x**, stand for vectors and capital bold characters, i.e. **X**, denote matrices. Non-bold characters are used for vector and matrix elements. Superscript *t* is used for vector and matrix transposition. The *I* and *J* denote the number of objects and variables, respectively, *K* denotes the rank of **X** matrix and *A* denotes the number of latent variables (principal components). An abbreviation DoF stands for the number of degrees of freedom. Other notations used are as follows. $\mathbf{X} = \{x_{ij}\}$ is the $(I \times J)$ data matrix; $\mathbf{T} = \{t_{ik}\}$ and $\mathbf{T}_A = \{t_{ia}\}$ are the full and truncated score matrices with dimensions $(I \times K)$ and $(I \times A)$, respectively; $\mathbf{P} = \{p_{jk}\}$and $\mathbf{P}_A = \{p_{ja}\}$ are the full and truncated loading matrices with dimensions $(J \times K)$ and $(J \times A)$, respectively; $\mathbf{E}_A = \{e_{ij}\}$ is the $(I \times J)$ matrix of residuals; $\mathbf{\Lambda} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_K)$ is the $(K \times K)$ matrix of eigenvalues; **I** is the unit matrix of a relevant dimension; h_i and v_i denote the SD and the OD, respectively of sample $i = 1, \ldots, I$; $\chi^2(N)$ is the χ^2-distribution with *N* DoF; $\chi^{-2}(N, \alpha)$ is the α quantile of $\chi^2(N)$; $F(N_1, N_2)$ is the *F*-distribution with DoF N_1 and N_2; $F^{-1}(N_1, N_2, \alpha)$ is the α quantile of $F(N_1, N_2)$; operators $E()$ and $V()$ denote the mathematical expectation and variance, correspondingly. Symbols S_h and S_v stand for any estimates of the SD and OD variances obtained with the corresponding training sets, h_i and v_i. Notation $x \sim G$ means that variable x is distributed with distribution function G.

2.2. Principal component analysis

Let matrix **X** have a rank $K \leq \min(I, J)$. Note that *K* is an unknown value, which is rather difficult to estimate. However, it does exist and may be used in theoretical calculations. The PCA decomposition of matrix **X** is

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

where $\mathbf{T} = \{t_{ik}\}$ is the $(I \times K)$ score matrix and **P** is the $(J \times K)$ loading matrix. Equation (1) assumes that matrix **X** is column-centered, i.e.

$$\sum_{i=1}^{I} x_{ij} = 0. \tag{2}$$

If matrix **X** does not satisfy Equation (2) it should be modified. The influence of such a transformation on the statistical properties of the PCA decomposition has been repeatedly discussed, i.e. in reference [17]. In particular, it has been proposed to reduce the effective number of objects by unity, i.e. to replace *I* by *I*–1 [10]. This topic will not be further discussed in this paper as a more general approach to the evaluation of DoF is proposed. However, below it is assumed that matrix **X** agrees with the condition given in Equation (2).

The $(K \times K)$ matrix Λ

$$\Lambda = \mathbf{T}^t\mathbf{T} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K) \tag{3}$$

is diagonal with the elements

$$\lambda_k = \sum_{i=1}^{I} t_{ik}^2 \tag{4}$$

These are the first *K* eigenvalues (others are equal to zero) of matrix $\mathbf{X}^t\mathbf{X}$ ranked in the descending order. Matrix **P** consists of the corresponding orthonormalized eigenvectors $\mathbf{P}^t\mathbf{P} = \mathbf{I}$. Therefore,

$$L_0 = \mathrm{Sp}(\mathbf{X}^t\mathbf{X}) = \mathrm{Sp}(\mathbf{T}^t\mathbf{T}) = \sum_{k=1}^{K} \lambda_k \tag{5}$$

Let us consider the first *A* ($A \leq K$) principal components in the decomposition given by Equation (1).

$$\mathbf{X} = \mathbf{T}_A\mathbf{P}_A^t + \mathbf{E}_A, \tag{6}$$

Matrices $\mathbf{T}_A$ and $\mathbf{P}_A$ include the first *A* columns of matrices **T** and **P**, respectively. The $(I \times J)$ matrix $\mathbf{E}_A = \{e_{ij}\}$is the residual matrix. A value

$$R(A) = \frac{\sum_{a=1}^{A} \lambda_a}{L_0} \tag{7}$$

is referenced to the explained data variation. It varies from 0 (at $A=0$) to 1 (at $A=K$)

2.3. The score distance (SD)

For a given number of principal components, A, the value

$$h_i = \mathbf{t}_i^t(\mathbf{T}_A^t\mathbf{T}_A)^{-1}\mathbf{t}_i = \sum_{a=1}^{A}\frac{t_{ia}^2}{\lambda_a}, i=1,\ldots,I \quad (8)$$

is named the SD [13]. It is equal to the squared Mahalanobis distance from the model center to sample i within the score subspace [18]. From Equation (4) it follows that

$$h_0 = \frac{1}{I}\sum_{i=1}^{I}h_i \equiv \frac{A}{I}, \quad (9)$$

Equation (9) is an identity, which always holds by PCA construction. The same equation is given in many publications (e.g., [10,13]), but it is construed statistically, as $E(h)=h_0$. However there is no variability in Equation (9) since the mean SD will always equal A/I regardless of possible changes in the samples, or even in the whole dataset. In other words, if the number of objects equals I, and the number of PCs used is A, then the average of SD is equal to A/I exactly.

Constructing the SD distribution it is necessary to remember that $\mathbf{X}$ matrix is centered. Therefore, at a fixed a, all random variables t_{ia} have zero expectation and variance $\lambda_a I^{-1}$. Taking Equation (3) into account it can be assumed that

$$N_h\frac{h}{h_0} \sim \chi^2(N_h) \quad (10)$$

where N_h is DoF. DoF equal A if the a-scores of all samples are distributed normally. Such an approach is used in reference [19]. In papers [8,9,20] the authors suggest to use another SD distribution

$$\frac{I-A}{I+1}\frac{h}{h_0} \sim \mathrm{F}(A, I-A) \quad (11)$$

This equation results from the assumption that each column vector $\mathbf{x}_j$ is distributed normally. It is worth mentioning that at $I \gg A$, Equation (11) turns into Equation (10), where $N_h=A$.

The normality of either the initial variables $\mathbf{x}_j$, or the scores $\mathbf{t}_a$, is just an assumption that cannot be verified. In our opinion, as the PCA subspace is formed by the linear combinations of vectors $\mathbf{x}_j$, the scores $\mathbf{t}_a$ could be viewed more 'normal' than the initial variables. Therefore, χ^2-distribution seems to be a preferred alternative to F-distribution. We suggest evaluating the DoF in Equation (10) in order to allow for the non-normality in the scores distribution. The DoF can be estimated using the method of moments (MM).

From Equation (10) it follows that the SD variance (denoted as $V(h)$) is equal to

$$V(h) = \frac{2h_0^2}{N_h} = \frac{2A^2}{N_h I^2}. \quad (12)$$

Therefore, DoF, N_h, can be estimated by equation

$$\hat{N}_h = \frac{2h_0^2}{S_h} = \frac{2A^2}{I^2 S_h} \quad (13)$$

where S_h is an estimate of the SD variance, $V(h)$. This variance can be estimated in many ways. The conventional method $S_h = \sum(h_i - h_0)^2 \Big/ (I-1)$ is rather sensitive to outliers. Therefore, in references [14,19,21] it has been proposed that robust variance estimators should be used, e.g. MAD [22].

We apply another robust method. This is the interquartile approach that leads to the following equation:

$$\frac{1}{N_h}\left[\chi^{-2}(N_h, 0.75) - \chi^{-2}(N_h, 0.25)\right] = \frac{1}{h_0}\mathrm{IQR}(h_1,\ldots,h_I) \quad (14)$$

which should be solved with respect to N_h. IQR stands for the interquartile range statistics calculated using the SD training samples. In practice, Equation (14) may be replaced by an explicit equality

$$\hat{N}_h = \exp\left(4.36\ln\frac{1.24}{\mathrm{IQR}}\right)^{0.72} \quad (15)$$

that may be applied for $1 < I < 100$ with the accuracy no worse than 7%.

2.4. The orthogonal distance (OD)

The OD is another important characteristic of PCA model (6). It is calculated as the sum of the squared residuals presented in matrix $\mathbf{E}_A = \{e_{ij}\}$

$$v_i = \sum_{j=1}^{J} e_{ij}^2. \quad (16)$$

The OD, v_i, is the squared Euclidian distance from object i to the model subspace. Often this value is divided by K–A [14,23], or by J–A [24], and then the square root is extracted [14,23,24]. However, following [9,20] we keep value (16) as it is.

It can be shown (e.g. [23]) that

$$v_i = \sum_{a=A+1}^{K} t_{ia}^2 = L_0 - \sum_{a=1}^{A} t_{ia}^2 \quad (17)$$

where L_0 is defined in Equation (5). Taking this into account one can obtain the following equality:

$$v_0 = \frac{1}{I}\sum_{i=1}^{I} v_i \equiv \frac{L_0}{I}(1 - R(A)) \quad (18)$$

which is always fulfilled by the PCA construction. Here $R(A)$ is defined in Equation (7).

The authors of papers [21,23,24] have proposed to employ the following distribution:

$$\frac{v}{v_0} \sim \mathrm{F}(K-A, (K-A)(I-A)) \quad (19)$$

However, in most of the practical cases, the value of $(K–A)$ $(I–A)$ is so large that Equation (19) can be replaced by

$$(K-A)\frac{v}{v_0} \sim \chi^2(K-A) \quad (20)$$

As mentioned previously, K is usually unknown and is difficult to estimate. It is easy to see that the distribution in Equation (20) could be derived from Equation (17), if each component had the same variance. However, in contrast to the SD, the OD consists of

non-normalized variables each of them having its own variance being equal to $\lambda_a I^{-1}$. Therefore, Equations. (19) and (20) are badly fitted for the OD distribution. Back in 1987, it has been noted [25] that SIMCA method has a tendency to reject too many new objects increasing the Type I error. In order to improve the classification efficiency numerous modifications of Equation (19) have been proposed [23]. The apt solution has been found by Nomikos and MacGregor [8], who employed the idea published in reference [26]. They proposed to apply the distribution $g_1\chi^2(g_2)$. The two unknown parameters g_1 and g_2 are to be estimated with the training set. This approach was repeatedly used in the subsequent papers, e.g. in reference [14].

We employ a similar formula for the OD distribution

$$N_v \frac{v}{v_0} \sim \chi^2(N_v) \tag{21}$$

that, however, depends on a single unknown parameter, N_v, because other parameter, v_0, can be estimated independently from Equation (18). The OD variance is equal to

$$\mathrm{V}(v) = \frac{2v_0^2}{N_v}. \tag{22}$$

Therefore,

$$\hat{N}_v = \frac{2v_0^2}{S_v} \tag{23}$$

where S_v is an estimate of the OD variance. Certainly, the IQR approach presented in Equation (14) may be used in the case of OD as well.

3. APPLICATION

3.1. Data

Three data sets are considered in this paper. The first set contains simulated data for which $I = 100$, $J = 25$. These data were built using historical observations of a real industrial process that has been presented earlier in reference [27,28]. Hundred of the most reliable objects selected from the initial dataset are modeled by PCA with five PCs. The matrix $\mathbf{T}_5\mathbf{P}_5^t$ was further disturbed with a pseudorandom white noise with standard deviation being equal to 0.05 max $|\mathbf{x}_{ij}|$. This dataset is denoted as SIM and it is mainly used for the illustration purposes. The DoF values calculated using the MMs are $N_h = 5.7$, $N_v = 21.6$, and these obtained with the IQR approach are $N_h = 5.0$, $N_v = 20.0$. It is known [22] that the robust estimates are less effective but more reliable with respect to outliers. In our case the corresponding DoF values are similar but the IQR estimates are smaller, hence their acceptance areas are wider. Therefore, the IQR estimates look more trustworthy, and they are used in the subsequent analysis.

The second dataset consists of the real spectra obtained by a NIR instrument ($J = 3501$ wavelengths in the interval 4000–7500 cm^{-1}) in the diffuse reflectance measurements. They were used in the discrimination of the genuine and counterfeit tablets by the method presented in reference [7]. The set includes 75 samples that are divided into a training set (40 genuine tablets, $I = 40$) and a test set (15 authentic and 20 fake tablets). The whole dataset is labeled BMT. The number of PCs is 2 ($A = 2$).

The third example represents a multivariate calibration problem that has earlier been described in reference [29]. The **X** set consists of the NIR spectra obtained for the whole grain samples in the interval 9050–10850 cm^{-1} ($J = 118$). The **Y** block includes the water content values obtained by a conventional analytical method. The dataset consists of 123 samples ($I = 123$) and it is named GRAIN.

3.2. Joint distribution

In the theoretical section of this paper, it was shown that employing the training values of h_i and v_i ($i = 1,\ldots, I$), it is possible to build the object distribution within a class. Vector sets ($\mathbf{t}_1,\ldots, \mathbf{t}_A$) and ($\mathbf{t}_{A+1},\ldots, \mathbf{t}_K$) are orthogonal, therefore the SD and OD values are statistically independent. Let us introduce notation c keeping in mind that c (c_0) could be either h (h_0), or v (v_0), and, respectively, N_c is either N_h, or N_v. The corresponding quantiles $\mathrm{Prob}(c > c_\alpha) = \alpha$ are calculated by Equation (10) and Equation (21) as $c_\alpha = c_0\chi^{-2}(N_c, \alpha)/N_c$.

In Figure 1 it is shown that the sample probability $\alpha_n = \mathrm{number}\{c_i > c_\alpha\}/I$ corresponds to the conjectural probability α for SIM data, in which $A = 5$. It can be seen that, in general, the SD and OD are well fit by the χ^2-distributions as all sample values fall into the confidence intervals $[\alpha-2d,\ \alpha+2d]$, where $d^2 = \alpha(\alpha-1)/I$. These intervals are shown in Figure 1.

3.3. Outlier detection

In the outlier detection procedure it is necessary to account for the size of a training set, I, and to obtain the p-values, which are the occurrence probabilities of each particular SD and OD value. They are calculated as

$$P_{\mathrm{out}}(c_i) = 1 - [1 - \Psi(c_i)]^I \approx 1 - \exp[-I\Psi(c_i)] \tag{24}$$

where Ψ is a cumulative distribution function of each statistics h, or v. If any of the p-values is less than a given critical level (e.g. 0.05), the corresponding object can be considered as an outlier. Alternatively, the p-values based on the χ^2-distribution are calculated using $\Psi(c_i) = \chi^2(N_c c_i/c_0, N_c)$.

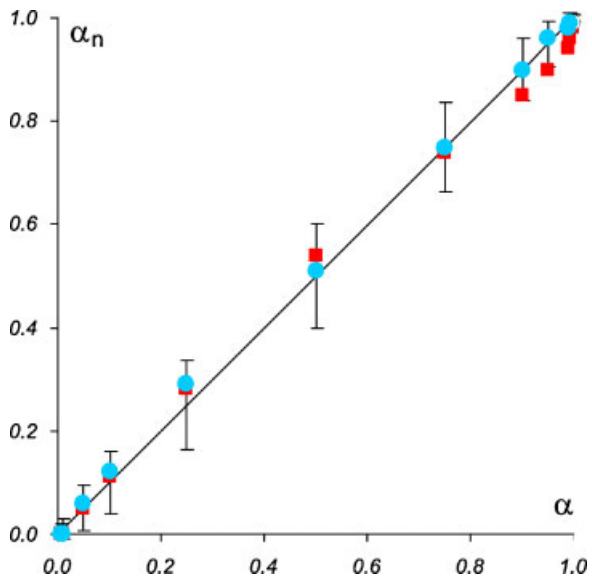

Figure 1. Distribution of the SD (■) and OD (●) in SIM data. Sample probability α_n versus theoretical probability α. Vertical bars represent the tolerance intervals. This figure is available in colour online at www.interscience.wiley.com/journal/cem

The pertinent example of the outlier detection is presented in subsequent section, where the BMT data are explored.

3.4. The acceptance regions

Let **x** be a new sample. It is required to determine whether the sample belongs to the same class *C* as the initial training set **X**. In making the decision two typical errors could occur. The type I error happens in a case when the sample is rejected, while it actually belongs to the class. The type II error is the acceptance of the sample, which in fact does not belong to the class. The type I error is directly connected with the acceptance region of a class. The larger the area is the smaller the type I error. In the composite hypothesis testing (SIMCA is just the case) the type II error cannot be established. Moreover, it is typically equal to 1. A simple example illustrates this unpleasant claim. Let C_ε be a class of samples that does not coincide with *C*. At the same time C_ε converges (in some sense) to *C* at $\varepsilon \to 0$. These two classes may be seen as two parallel planes located at the distance ε It is evident that the type II error tends to 1 as $\varepsilon \to 0$.

Let γ be a given value of the type I error (the significance level). Using two class statistics, *h* and *v*, one can build various acceptance areas, which are dependent on the significance. These areas are demonstrated in the SD–OD plot that is constructed for dataset SIM, at $\gamma = 0.05$. The plot is shown in Figure 2.

Usually (e.g. [19]), the acceptance region H_γ is defined as a direct product of two tolerance intervals for the SD and for the OD values. From Equations. (10) and (21) it follows that

$$H_\gamma = \left[0, \frac{h_0}{N_h}\chi^{-2}(N_h, \alpha)\right] \otimes \left[0, \frac{v_0}{N_v}\chi^{-2}(N_v, \alpha)\right]. \tag{25}$$

Here $\alpha = 1 - \sqrt{1-\gamma}$ is the probability that statistics (*h* or *v*) does not belong to the corresponding interval. This conventional acceptance area is bounded with dashed rectangle **I** in Figure 2. Seven samples of SIM data (7–9, 11–15) are out of the area.

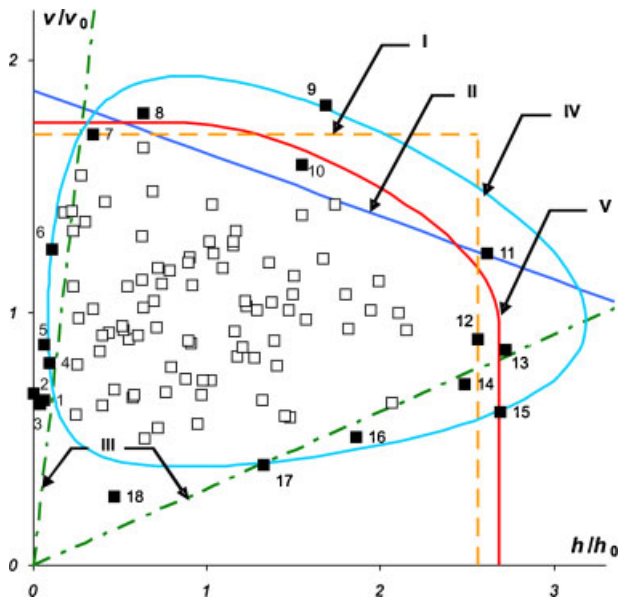

Figure 2. Various acceptance areas for SIM data. Eighteen (of 100) notable samples are marked. This figure is available in colour online at www.interscience.wiley.com/journal/cem

To construct the second area a well-known equation

$$N_h \frac{h}{h_0} + N_v \frac{v}{v_0} \sim \chi^2(N_h + N_v) \tag{26}$$

is used. This gives an area presented by an equation

$$H_\gamma = \left\{(h, v) : N_h \frac{h}{h_0} + N_v \frac{v}{v_0} \le \chi^{-2}(N_h + N_v, \gamma)\right\}. \tag{27}$$

It has a triangular shape bounded with hypotenuse **II** in Figure 2 and the catheti along the axes. Three points 8, 9 and 10 miss the area.

A similar approach has been proposed earlier in reference [19], where classification is made based on a linear combination of the SD and OD

$$\beta\sqrt{\frac{h}{h_0}} + (1-\beta)\sqrt{\frac{v}{v_0}};\ \ \beta\frac{h}{h_0} + (1-\beta)\frac{v}{v_0}. \tag{28}$$

The tuning parameter $\beta \in [0,1]$ is selected to get the sensitivity or the specificity maximized.

The third region follows from another well-known equation

$$\frac{v/v_0}{h/h_0} \sim \frac{\chi^2(N_v)}{N_v}\frac{N_h}{\chi^2(N_v)} \sim F(N_v, N_h) \tag{29}$$

where $F(N_v, N_h)$ is the *F*-distribution. Thus,

$$H_\gamma = \left\{(h, v) : F^{-1}(N_v, N_h, 0.5\gamma) \le \frac{v/v_0}{h/h_0} \le F^{-1}(N_v, N_h, 1-0.5\gamma)\right\}. \tag{30}$$

This unrestricted area is located between two dash–dot lines **III** in Figure 2. Nine samples (1–6 and 14–16) lie outside the area.

The fourth area is constructed using an idea given in reference [8]. It is known [30] that $\chi^2(m)$ distribution can be approximated with the normal distribution as

$$\frac{(\chi^2/m)^{1/3} - (1-s^2)}{s} \sim \mathrm{N}(0,1),\ \ s^2 = \frac{2}{9m} \tag{31}$$

where N(0,1) is the standard normal distribution and *m* is DoF. New variables *z* and *w* are introduced with the transformation, in which the normalized SD and OD values

$$\frac{h}{h_0} \sim \chi^2(N_h)/N_h \to z;\quad \frac{v}{v_0} \sim \chi^2(N_v)/N_v \to w \tag{32}$$

are substituted in Equation (31). SIM data in the new coordinates are shown in Figure 3. The hit probability for circle **1** is equal to $1-\gamma$. The same area presented in the initial coordinates is bounded by curve **IV** in Figure 2. Seven points (1–5 and 17–18) are out.

It can be seen that area **IV** does not include the origin point (0, 0) in Figure 2. This is an evident shortcoming as the small SD and OD values are obviously acceptable. The updated acceptance area follows from the region bounded by curve **2** in Figure 3. Let us calculate the probability of paired statistics (*w*, *z*) to belong to this region. Curve **2** crosses the coordinate axes at distance *r* from the origin. Therefore, the corresponding probabilities for each of

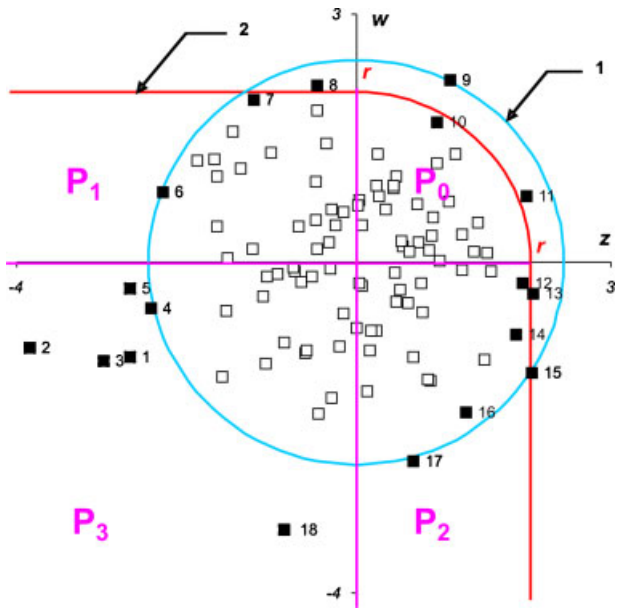

Figure 3. SIM data shown in the transformed variables *z*, *w*. Areas 1 and 2 have the same hit probability $1-\gamma = 0.95$. This figure is available in colour online at www.interscience.wiley.com/journal/cem

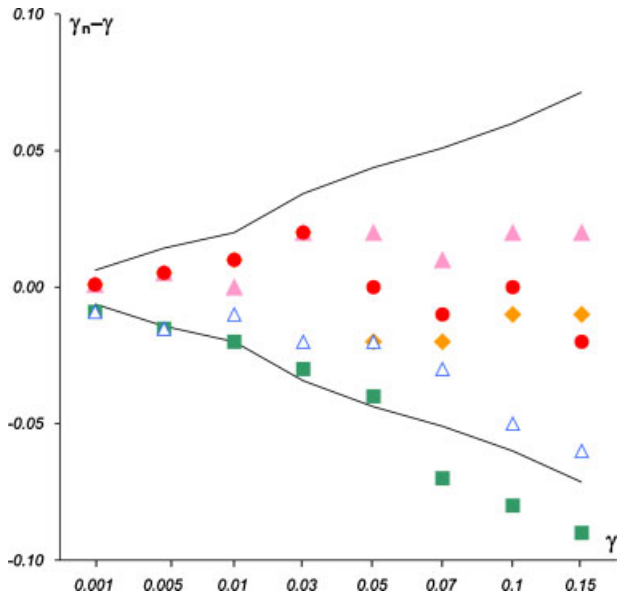

Figure 4. Validation of the acceptance areas at various significance levels γ; γ_n is the rate of samples lying out of the areas: **I** (◆), **II** (▲), **III** (■), **IV** (Δ), **V** (●). Two lines represent the tolerance intervals. This figure is available in colour online at www.interscience.wiley.com/journal/cem

the four composite subregions are

$$P_0 = 0.25\left[1-\exp\left(-0.5r^2\right)\right]; P_1 = P_2 = 0.5(\Phi(r) - 0.5); P_3 = 0.25 \quad (33)$$

where Φ is the cumulative normal distribution function. Summarizing these values an equation for calculation of r is obtained

$$1 - \gamma = \Phi(r) - 0.25\exp\left(-0.5r^2\right) \quad (34)$$

that corresponds to a given significance level γ. Going back to the initial coordinates the area that is bounded by curve **V** in Figure 2 is constructed. Five samples (8, 9, 11, 13 and 15) from SIM dataset lie outside the area.

These acceptance areas can be obtained at various significance levels γ. Figure 4 shows the result of such modeling performed with SIM dataset. Abscissa axis represents the considered values of γ: 0.001, 0.005, etc. Ordinate values are calculated as $\gamma_n-\gamma$, where γ_n is the rate of samples that lie outside the corresponding area (**I**, **II**, etc). In the ideal case $\gamma_n = \gamma$; however, some variability in γ_n should be allowed. This is represented by the tolerance intervals $[\gamma-2d, \gamma+2d]$, where $d^2 = \gamma(\gamma-1)/I$, which are shown by two lines in Figure 4.

The outcomes can be considered as a validation of the proposed areas. It may be concluded that only three of them are of interest. Figure 4 testifies for areas **I**, **II** and **V**. Region **III** is unacceptable as it contains the arbitrarily large values of *h* and *v*. Area **IV** is improper as it does not cover the neighborhood of zero.

3.5. Example 1. BMT dataset

A problem of the medicines counterfeiting is important all over the world [31]. The NIR spectroscopy together with the SIMCA method has earlier been proposed [7] as a promising approach in the rapid recognition of false drugs. The above-mentioned BMT dataset represents a real example that was considered in the method development.

The PCA decomposition with two PCs ($A = 2$) explains 92% of data variation (7). The average values given by Equations. (9) and (18) are: $h_0 = 0.05$ and $v_0 = 5.2 \times 10^{-5}$. The DOFs estimated by the MMs (Equations. (13) and (23)) are equal to $\hat{N}_h = 4.1$ and $\hat{N}_v = 2.8$, and these estimated using the IQR approach given by Equation (14) are $\hat{N}_h = 3.1$ and $\hat{N}_v = 2.3$. The latter estimates are used for the modeling.

Let us begin with the outlier detection in the training set. In Figure 5 the normalized SD (h/h_0) and OD (v/v_0) values are shown together with the relevant critical levels. These levels are calculated applying Equation (24) at significance $P_{out} = 0.05$. The χ^2-distributions (Equations. (10) and (21)) are employed for calculation of levels 1, while critical levels 2 are obtained with *F*-distributions (Equations. (11) and (19)). It can be seen that level 2 is too large regarding the SD, while for the OD it is too small. Therefore, three samples from forty are qualified as the outliers with respect to the OD level 2. They are G11-3, G12-5 and G03-4. The corresponding *p*-values given by Equation (24) are 0.006, 0.004 and 0.01. This illustrates the fact that *F*-distribution is not appropriate for the SD and OD modeling.

Now, let us consider the test set in BMT data. It includes 35 samples: 15 genuine and 20 counterfeit tablets. In Figure 6 the samples are presented in the normalized SD–OD plot, which is similar to Figure 3. The acceptance areas are shown as well: area **I** is the conventional (25), area **II** is the triangular (27) and area **V** is built by approximation (31). The significance level (type I error) is $\gamma = 0.01$.

The counterfeit tablets (squares) are located far outside all acceptance areas. Two genuine samples (filled circles) are worthy of commenting. They are outside of area **I** but belong to both areas **II** and **V**. Sample G10-04 is characterized by a large OD value and a small SD value. In contrast, sample G10-01 has a small OD

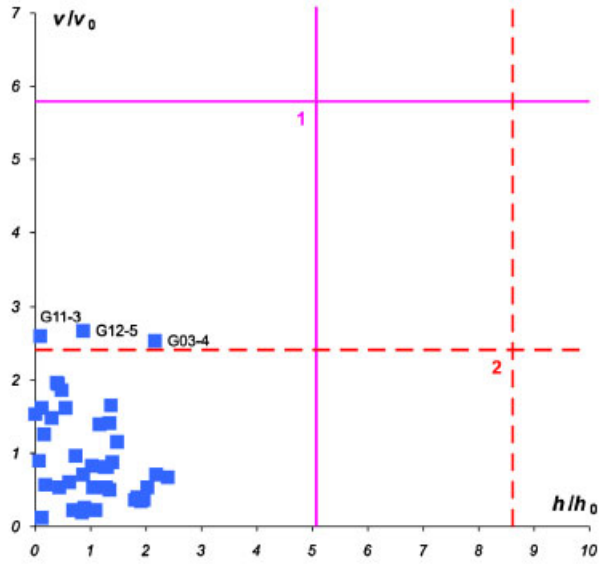

Figure 5. BMT data. Training set. Lines 1 and 2 represent the critical values for outlier detection. Levels 1 are calculated with χ^2-distribution: h by Equation (10), v by Equation (21). Levels 2 are calculated with F-distribution: h by Equation (11), and v by Equation (19). This figure is available in colour online at www.interscience.wiley.com/journal/cem

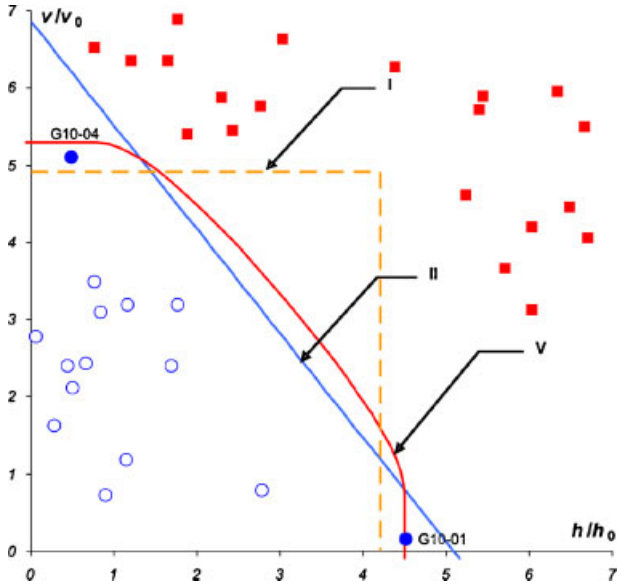

Figure 6. BMT data. Test set. The counterfeit (■) and genuine (○, ●) samples. Three acceptance areas **I**, **II** and **V** are shown. This figure is available in colour online at www.interscience.wiley.com/journal/cem

value and a large SD value. Therefore, areas **II** and **V** seem to be preferable to area **I**.

3.6. Example 2. GRAIN dataset

In this example a calibration problem is considered. This is a well-known task, in which the water content in grain is predicted using NIR spectra. A detailed problem description can be found in reference [29], in which this dataset is designated by 'truncated set'. The PLS model with four PCs explains 99% of the X and 92% of the Y variations, respectively.

The PLS projection composes the space, in which the SDs, h_i, are relevant. Since this is a calibration case, two ODs should be considered. The X residuals give the first OD, v_i, that is obtained using Equation (16). The Y (studentized) residuals present another measure, u_i. The latter is calculated as $u_i = (y_i - \hat{y}_i)^2/(1 - h_i)$, where y_i and $\hat{y}_i$ are the measured and predicted response values, respectively and h_i is the SD value. All these measures, being calculated for the calibration set, constitute three samplings h_i, v_i, and u_i ($i = 1,\ldots, 123$). They are further modeled by the χ^2-distributions in a way that is described in the theoretical section. The corresponding DoF values obtained by the MMs are $N_h = 4.5$, $N_v = 2.4$, $N_u = 0.9$. The IQR approach yields the estimates $N_h = 4.7$, $N_v = 3.0$, $N_u = 1$, which are used later. It can be noted that both N_h values are rather close to $A = 4$, which is the number of PLS components used. It was also expected that N_u would be near 1 as this DoF corresponds to the sum that has only one item.

Figure 7 represents two influence plots, in which the abovementioned acceptance areas ($\gamma = 0.01$) are delineated. The left panel shows the X related measures, i.e. v. The right panel demonstrates the same objects with respect to the Y distance, i.e. u. All objects except sample no. 101 lie inside the areas. This agrees to the given significance level $\gamma = 0.01$ as one outsider may be expected among the 123 calibration objects.

The triangular shape of the sample allocation is clearly seen in both plots. It is natural to suppose that samples that are located

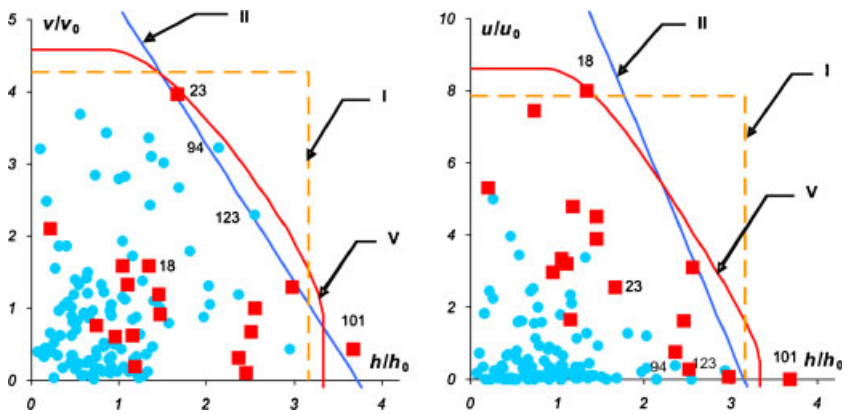

Figure 7. GRAIN dataset. Influence plots for X and Y data. Three acceptance areas **I**, **II** and **V** are shown. Boundary samples (■) are highlighted. This figure is available in colour online at www.interscience.wiley.com/journal/cem

closer to the hypotenuse (**II** or **IV**) are the most influential in calibration. To confirm this claim a concept of boundary samples is applied. This is an essential outcome of the SIC (Simple Interval Calculations) approach [16], which gives a relevant object classification in the calibration problems. The detailed method explanation is presented in paper [32]. The SIC method yields 17 boundary samples highlighted in Figure 7. In the right plot (*Y* related) they are in fact located close to the diagonal. In the left plot this is not the case.

There are some interesting samples in Figure 7. For example, boundary sample 23 is important for *X* relations and not so important with respect to *Y*, and *vice versa*, sample 18 has a strong influence on *Y* and not so important for *X*. Objects 94 and 123 are located near hypotenuse in the left plot but they have very small u/u_0 distances in the right plot. In general, it can be seen that boundary samples mainly relate to the right plot, which represents the response *Y* distances.

4. DISCUSSION AND CONCLUSIONS

It can be concluded that in the modeling of the score and orthogonal distances, χ^2-distribution is preferable to *F*-distribution. The SD and OD distributions are rather similar. Each of them depends on a single unknown parameter, N_h and N_v that are the effective DOFs. The scaling factors (h_0 or v_0) can be estimated in advance. The SD scaling factor (h_0) does not depend on data and thus it can be evaluated by the number of objects, *I*, and by the number of PCs, *A*, as $h_0 = A/I$. It is an astonishing fact that such an evident finding has been still unnoticed in literature.

In our opinion, the estimation of DoF is a key challenge in the projection modeling. In case of SD, DoF should be close to the number of PCs used, i.e. $N_h \approx A$; and, in case of OD, DoF is undoubtedly linked to the unknown rank of the data matrix, $K = \text{rank}(\mathbf{X})$, e.g. $N_v \approx K–A$. However, such evaluations are valid only under an assumption that either data, or scores, are normally distributed, which is always a dubious conjecture. Therefore, we believe that a data-driven estimator of DoF, rather than a theory-driven one should be used. The conventional MMs is sensitive to outliers, therefore other techniques have to be applied. The first approach is the robust estimation that has been studied in papers [14,19,21,22]. The IQR estimator used in this paper is of this kind. The second way is the statistical simulation technique, such as bootstrap and jackknife considered in references [11,17,33].

It is clear that any classification problem within the projection approach should be solved with respect to a given significance level, γ, i.e. the type I error. This is an essential practical demand that has to be complied with a duly care in various applications such as the process control, anti-counterfeiting actions, etc. At the same time, the SD–OD, *a.k.a.* influence plot is a valuable exploratory tool for the identification of the influential, typical, extreme and other interesting objects in data. In this plot different acceptance areas can be constructed. They are the regions where a given share, $1-\gamma$, of the class members belongs to. Five of such areas were presented in the paper. All of them are valid, i.e. they comply with the type I error requirement, but not all of them are practical. The first place should be given to area **II** that is constructed using the sum of the normalized SD and OD. The triangular shape of this area is an evident advantage. For the explanation of this claim let us return to the data cloud image. This cloud could be viewed as an egg-shaped or a box-shaped one. Needless to say, that this choice has nothing in common with the variable correlations. For example, multivariate normal distribution always produces the egg-shaped data cloud, and the correlations only change the cloud orientation. So, the shape problem is a bit deeper. The conventional data interpretation approach is mainly based on the near-normal distributions. This reasonable way of thinking leads to the egg-shaped data cloud, and then to the triangle shaped acceptance areas. Such argumentation explains the above-mentioned mystery of the influence triangle that is widely known but never claimed in literature. It should also be mentioned that area **II** can be applied even for the small values of DoF.

The second place is to be awarded to area **V**, which follows from the normal approximation of the χ^2-distribution. This, however, can be applied when DoF is large enough, namely, more than 30 [30]. This is uncommon case though, as DoF usually does not exceed 10. The conventional rectangle area **I** takes only the third place. Other areas (**III** and **IV**) are unpractical for the reasons explained above.

It is worth mentioning that the designed data should be analyzed with a special care. Usually they do not constitute a representative sampling. Therefore, the data driven methods may lead to dubious DoF values. The HPLC–DAD data are a typical example.

REFERENCES

1. Daszykowski M, Walczak B, Massart DL. Projection methods in chemistry. *Chemom. Intell. Lab. Syst.* 2003; **65**: 97–112.
2. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom. Intell. Lab. Syst.* 1987; **2**: 37–52.
3. Wold S. Pattern recognition by means of disjoint principal components models. *Pattern Recognit.* 1976; **8**: 127–139.
4. Frank IE, Lanteri S. Classification models: discriminant analysis, SIMCA, CART. *Chemom. Intell. Lab. Syst.* 1989; **5**: 241–256.
5. Hall GJ, Kenny JE. Estuarine water classification using EEM spectroscopy and PARAFAC-SIMCA. *Anal. Chim. Acta* 2007; **581**: 118–124.
6. Tominaga Y. Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN. *Chemom. Intell. Lab. Syst.* 1999; **49**: 105–115.
7. Rodionova OYe, Houmøller LP, Pomerantsev AL, Geladi P, Burger J, Dorofeyev VL, Arzamastsev AP. NIR spectrometry for counterfeit drug detection. *Anal. Chim. Acta* 2005; **549**: 151–158.
8. Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. *Technometrics* 1995; **37**: 41–59.
9. Westerhuis JA, Gurden SP, Smilde AK. Generalized contribution plots in multivariate statistical process monitoring. *Chemom. Intell. Lab. Syst.* 2000; **51**: 95–114.
10. Martens H, Naes T. Multivariate Calibration. (2nd edn). Wiley: Chichester, UK, 1998.
11. Flåten GR, Grung B, Kvalheim OM. A method for validation of reference sets in SIMCA modelling. *Chemom. Intell. Lab. Syst.* 2004; **72**: 101–109.
12. The Unscrambler's User's Guide, Version 8.0, Camo, Trondheim.
13. Höskuldsson A. Prediction Methods in Science and Technology, vol. 1. Thor Publishing: Copenhagen, Denmark, 1996.
14. Daszykowski M, Kaczmarek K, Stanimirova I, Vander Heyden Y, Walczak B. Robust SIMCA-bounding influence of outliers. *Chemom. Intell. Lab. Syst.* 2007; **87**: 95–103.
15. F testing in SIMCA. A message posted by Jerry Jin at ICS-L on 5 April 2007 https://listserv.umd.edu/cgi-bin/wa?a2=ind0704& l=ics-l& d=0& p=1265
16. Rodionova OYe, Esbensen KH, Pomerantsev AL. Application of SIC (simple interval calculation) for object status classification and outlier detection–comparison with PLS/PCR. *J. Chemometrics* 2004; **18**: 402–413.
17. Seipel HA, Kalivas JH. Effective rank for multivariate calibration methods. *J. Chemometrics* 2004; **18**: 306–311.

18. De Maesschalck R, Jouan-Rimbaud D, Massart DL. Tutorial. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* 2000; **50**: 1–18.
19. Vanden Branden K, Hubert M. Robust classification in high dimensions based on the SIMCA method. *Chemom. Intell. Lab. Syst.* 2005; **79**: 10–21.
20. Meng X, Morris AJ, Martin EB. On-line monitoring of batch processes using a PARAFAC representation. *J. Chemometrics* 2003; **17**: 65–81.
21. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: A new approach to robust principal component analysis. *Technometrics* 2005; **47**: 64–79.
22. Daszykowski M, Kaczmarek K, Vander Heyden Y, Walczak B. Robust statistics in data analysis—a review: basic concepts. *Chemom. Intell. Lab. Syst.* 2007; **85**: 203–219.
23. De Maesschalck R, Candolfi A, Massart DL, Heuerding S. Decision criteria for soft independent modelling of class analogy applied to near infrared data. *Chemom. Intell. Lab. Syst.* 1999; **47**: 65–77.
24. De Braekeleer K, De Maesschalck R, Hailey PA, Sharp DCA, Massart DL. On-line application of the orthogonal projection approach (OPA) and the soft independent modelling of class analogy approach (SIMCA) for the detection of the end point of a polymorph conversion reaction by near infrared spectroscopy (NIR). *Chemom. Intell. Lab. Syst.* 1999; **46**: 103–116.
25. Droge BM, Van't Klooster HA. An evaluation of SIMCA: part 1. The reliability of the SIMCA pattern recognition method for a varying number of objects and features. *J. Chemometrics.* 1987; **1**: 221–230.
26. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems. *Ann. Math. Stat.* 1954; **25**: 290–302.
27. Pomerantsev AL, Rodionova OYe, Höskuldsson A. Process control and optimization with simple interval calculation method. *Chemom. Intell. Lab.Syst.* 2006; **81**: 165–179.
28. Höskuldsson A, Rodionova OYe, Pomerantsev AL. Path modeling and process control. *Chemom. Intell. Lab.Syst.* 2007; DOI: 10.1002/cem.1103/**88**: 84–99.
29. Pomerantsev AL, Rodionova OYe. Construction of a multivariate calibration by the simple interval calculation method. *Zh. Anal. Khim.* 2006; **61**: 952–966.
30. Abramowitz M, Stegun I. Handbook of Mathematical Functions. National Bureau of Standards: NY, 1964.
31. Deisingh AK. Pharmaceutical counterfeiting. *Analyst* 2005; **130**: 271–279.
32. Rodionova OYe, Pomerantsev AL. Representative and influential subset selection. *J. Chemometrics* 2008.
33. Van Der Voet H. Pseudo-degrees of freedom for complex predictive models: the example of partial least squares. *J. Chemometrics* 1999; **13**: 195–208.