Special Issue Article

JOURNAL OF CHEMOMETRICS

Received: 21 December 2012, Revised: 09 April 2012, Accepted: 17 April 2013, Published online in Wiley Online Library: 23 May 2013

(wileyonlinelibrary.com) DOI: 10.1002/cem.2506

Concept and role of extreme objects in PCA/SIMCA†

Alexey L. Pomerantsev[a,b]* and Oxana Ye Rodionova[a]

For the construction of a reliable decision area in the soft independent modeling by class analogy (SIMCA) method, it is necessary to analyze calibration data revealing the objects of special types such as extremes and outliers. For this purpose, a thorough statistical analysis of the scores and orthogonal distances is necessary. The distance values should be considered as any data acquired in the experiment, and their distributions are estimated by a data-driven method, such as a method of moments or similar. The scaled chi-squared distribution seems to be the first candidate among the others in such an assessment. This provides the possibility of constructing a two-level decision area, with the extreme and outlier thresholds, both in case of regular data set and in the presence of outliers. We suggest the application of classical principal component analysis (PCA) with further use of enhanced robust estimators both for the scaling factor and for the number of degrees of freedom. A special diagnostic tool called extreme plot is proposed for the analyses of calibration objects. Extreme objects play an important role in data analysis. These objects are a mandatory attribute of any data set. The advocated dual data-driven PCA/SIMCA (DD-SIMCA) approach has demonstrated a proper performance in the analysis of simulated and real-world data for both regular and contaminated cases. DD-SIMCA has also been compared with robust principal component analysis, which is a fully robust method.

Keywords: PCA; SIMCA; robust methods; outliers; extreme samples; chi-squared distribution; scores and orthogonal distances; tolerance areas; thresholds

1. INTRODUCTION

The word "robust" is the most popular word in modern chemometrics [1]. The recent trend is employment of robust statistics to any data set regardless of the necessity. This is caused by the growing amount of data and their complexity. We use the word robust in the sense described by Huber [2], that is, robustness signifies insensitivity to small deviations from the assumptions and at the same time larger deviations should not cause a catastrophe.

In this paper, we are discussing the use of robust and "semi-robust" methods on the example of principal component analysis (PCA) and soft independent modeling by class analogy (SIMCA) approach. To avoid any confusion, we should explain the usage of term SIMCA within the paper context. Most often, SIMCA is meant as a one-class classifier. However, it is evident that the acceptance areas derived by the SIMCA classifier can be used in explanatory analysis of data as well, for example, as the decision rules for the outliers' detection [3]. Formally, the regular samples constitute a target class, whereas other data set objects are tested to the class membership. Therefore, in this paper, SIMCA is understood as a procedure consisting of PCA model construction, followed by calculation of the orthogonal and score distances (ODs and SDs) with a subsequent determination of their cutoff levels. The procedure of the cutoff levels' determination is a very important stage. The distance distributions cannot be proposed in advance, because they depend on a particular data.

Chemometrics deals with data that could be viewed as random values, which necessarily vary in replicated measurements. The analysis of data aims at finding estimates for some unknown but not random properties of interest (e.g., concentrations). Any data-driven estimator is a statistic, that is, a function of this data set, and therefore it is a random value, too. Parametric estimators (which are traditionally the most popular ones) are based on distributional assumptions, the majority of which rely on the normal distribution. A violation of these assumptions can potentially lead to wrong estimates that do not meet our expectations. Two typical cases can be considered. The first one is a bi-modality or multi-modality distribution of data, in case the data originate from different populations. A relevant example is presented in Ref. [4] where pharmaceutical substances packed into polyethylene (PE) bags are verified. This data set will be further considered in Section 4.

Another case of a distribution distortion is outliers. These are a few samples, which manifest atypical properties compared with other objects in the data set. An outlier could appear as a blunder in the process of data acquisition, or it could be an intrinsic object in the object population. For example, applying the Near-infrared (NIR) based approach for counterfeit drug detection [5], we revealed a genuine batch that has a higher moisture

* *Correspondence to: Alexey L. Pomerantsev, Semenov Institute of Chemical Physics RAS, Kosygin str. 4, 119991, Moscow, Russia.*
E-mail: forecast@chph.ras.ru

a *A. L. Pomerantsev, O. Y. Rodionova*
Semenov Institute of Chemical Physics RAS, Kosygin str. 4, 119991, Moscow, Russia

b *A. L. Pomerantsev*
Institute of Natural and Technical Systems RAS, Kurortny pr. 99/18, 354024, Sochi, Russia

† *Paper first presented at the Eight Winter Symposium on Chemometrics (WSC8, Russia, 2012).*

contents compared with the other authentic samples. Irrespective of the cause, being spoilage in production or conventional moisture variability, these outliers should be treated with special care. Otherwise, data analysis may lead to misclassification. On the other hand, the outliers should not be confused with extreme samples, which are always present in the data. For example, in a population of 100 normally distributed values, there is a good chance (probability 0.64) to find an extreme sample located beyond 3σ [6]. Excluding such fictitious outliers could significantly deform the data analysis.

It is often difficult to distinguish between the first case and the second case. In the first case, special group of objects should be isolated and analyzed (Section 4.4). The outliers should be detected and removed (Section 4.3), although even a single outlier may indicate the existence of a special population; whereas numerous but non-structured outliers do not form the second mode. The ultimate decision is always made post factum; if an abnormal sample cannot be model, it is an outlier and has to be removed.

All these cases (bi-modality, outliers) are the examples of data that do not follow an assumed regular distribution. To account for these (and other) distortions, a number of robust estimators have been developed [7]. They are based on the assumption that the majority of data samples follow a regular distribution R, and a minor part comes from the other distribution D. The whole data set has therefore a mixed distribution

$$R_\mu = (1-\mu)R + \mu D \tag{1}$$

where $0 \le \mu < 1$. For example, R is the normal distribution disturbed with a few outliers that come from distribution D. Robust estimators are designed to resist any type of distortions: outlier resistant (small μ) and contamination resistant (moderate μ). In any case, distribution R_μ is considered to be close to R. The majority of the robust estimators are developed for the normal distribution R. This, however, is not the only case; the chi-squared distribution may also give a regular basis [8] disturbed by small deviations. Such a case is considered in Section 2.4.

Various estimators (statistics) can be employed for the analysis of the same data set. For instance, both the sample mean $\overline{x}$ and the sample median $\widetilde{x}$ are estimators of the location parameter, that is a data center, but the latter is robust. Obviously, a robust estimator works better for not clean data (outlier contaminated), whereas a classical statistics is preferred when dealing with regular data (which are close to the normal distribution, no outliers). Estimator misuse can lead to various problems. For example, a classical estimator used for contaminated data is often unable to detect outliers (*masking effect*), and, on the contrary, it often identifies regular samples as outliers (*swamping effect*). These issues were often presented in literature [7]. The opposite case, when a robust estimator is used for regular data, is not often discussed by statisticians. However, the gain in robustness is always paid for by the loss in efficiency. For example, the median estimator applied to a regular data is on average 25% farther from the true location value compared with the classical mean estimator [9].

Dealing with a mix of distributions is common in data analysis, chemometrics being no exception. Any data point is influenced by (at least) two random components, which are noted by $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$. The first term, $\boldsymbol{\delta}$, is responsible for the samples (objects) variability, whereas the second term, $\boldsymbol{\varepsilon}$, appears as a result of random distortions, such as measurement errors, violation of experimental conditions, and so on. The difference in these components can be explained in a virtual experiment, where the number of replicated measurements tends to infinity. In this case, the input of the second component can be neglected, whereas the first term still has its influence. Thus, the first term, $\boldsymbol{\delta}$, accounts for the essence of things, whereas the second term, $\boldsymbol{\varepsilon}$, reflects the imperfection of our means to learn about the essence.

The $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ mixture should not be considered as an instance of Equation (1). Moreover, each component could be a mixture of a regular and a distortion fraction. The $\boldsymbol{\delta}$ component is most often contaminated with the second mode distribution, whereas the $\boldsymbol{\varepsilon}$ component can be a source of outliers.

To lessen the influence of outliers, different robust procedures have been developed and applied at various steps of the PCA/SIMCA method. It is possible to classify robust procedures along the following stages:

(1) Data pre-processing,
(2) PCA decomposition, and
(3) Calculation of thresholds.

The first stage includes data scaling and centering using robust estimators, for example, spherical PCA, which firstly projects all data objects onto a hyper-sphere and afterwards uses regular PCA [10] and other similar methods that aim to neglect the influence of atypical samples before data compression. The application of a robust procedure on the second step is the most straightforward approach. Numerous robust PCA algorithms, such as projection pursuit (PP) and PCA algorithms based on robust covariance matrix, are overviewed in Ref. [7,11]. For the threshold development, leverages and residuals are also "robustified" by their normalization with the help of robust estimators. The most popular algorithms, such as robust principal component analysis (ROBPCA) [12] and spherical SIMCA [13], try to apply "robustification" on all, or most, of the stages.

For data exploration and outlier detection, we propose to robustify SIMCA only on the third step. This is performed by conducting a robust data-driven estimation of the ODs and SDs distributions combined with a special design of the acceptance areas (thresholds). Moreover, the robust and non-robust acceptance areas are calculated for various significance levels. As a result, the acceptance area is built for the specified value α of a type I error and taking into account the presence/absence of the outliers.

This paper continues our research started in publication [14].

2. THEORY

2.1. Notation

Small bold characters, for example, $\mathbf{x}$, stand for vectors, whereas capital bold characters, for example, $\mathbf{X}$, denote matrices. Non-bold characters are used for vector and matrix elements. Superscript t is used for vector and matrix transposition. I and J denote the number of objects and variables, respectively; 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; $\mathbf{T} = \{t_{ia}\}$ is the $(I \times A)$ score matrix; $\mathbf{P} = \{p_{ja}\}$ is the $(J \times A)$ loading matrix; $\mathbf{E} = \{e_{ij}\}$ is the $(I \times J)$ matrix of residuals; $\Lambda = \mathrm{diag}\,(\lambda_1,\ldots,\lambda_A)$ is the $(A \times A)$ matrix of eigenvalues; $\mathbf{I}$ is the unit matrix of a relevant dimension; $\mathbf{1}$ is the vector of ones of a relevant dimension. Abbreviations SD and OD are used for the scores and orthogonal distances, correspondingly; h_i and v_i are, respectively, the SD and the OD values for sample $i = 1,\ldots, I$; N($\mathbf{m}$, $\mathbf{V}$) is the

multivariate normal distribution with expectation $\mathbf{m}$ and covariance matrix $\mathbf{V}$; N(0,1) is the univariate standard normal distribution; $\Phi^{-1}(\alpha)$ is the α quantile of N(0,1); $\chi^2(N)$ is the chi-squared distribution with N DoF; $\chi^{-2}(\alpha, N)$ is the α quantile of $\chi^2(N)$; $\chi(N)$ is the chi distribution with N DoF; operators E() and V() denote the mathematical expectation and variance, correspondingly.

2.2. Principal component analysis

The PCA decomposition of matrix $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E} \tag{2}$$

Matrix Λ

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

is diagonal with the elements

$$\lambda_a = \sum_{i=1}^{I} t_{ia}^2 \tag{4}$$

which are the eigenvalues of matrix $\mathbf{X}^t\mathbf{X}$ ranked in the descending order.

Equation (2) assumes that matrix $\mathbf{X}$ is column-wise centered; otherwise, it should be pre-processed using a classical (mean based) or robust (median based) method.

There are two statistics that are important for PCA interpretation. The first one is the SD,

$$h_i = \mathbf{t}_i^t(\mathbf{T}^t\mathbf{T})^{-1}\mathbf{t}_i = \sum_{a=1}^{A} \frac{t_{ia}^2}{\lambda_a}, i = 1, \ldots, I \tag{5}$$

which equals the squared Mahalanobis distance from the model center to sample i within the scores subspace.

The other one is the OD, calculated as the sum of the squared residuals presented in matrix $\mathbf{E} = \{e_{ij}\}$

$$v_i = \sum_{j=1}^{J} e_{ij}^2 = \sum_{a=A+1}^{K} t_{ia}^2, i = 1, \ldots, I \tag{6}$$

where $K \le \min(I, J)$ is the rank of matrix $\mathbf{X}$.

The OD, v_i, is a squared Euclidian distance from sample i to the scores subspace.

2.3. PCA statistics

Let a row vector $\mathbf{x}$ be a point in the J-dimensional variable space. A sample set of such vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_I\}^t$ forms data matrix $\mathbf{X}$. The PCA statistical model defines the distribution of vector $\mathbf{x}$ as a sum of two J-dimensional independent random components

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

The term $\boldsymbol{\delta}$ accounts for the data structure, and $\boldsymbol{\varepsilon}$ is a noise term.

$$\boldsymbol{\delta} \propto D(\mathbf{0}, \mathbf{V}) \qquad \boldsymbol{\varepsilon} \propto F\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \tag{7}$$

In Equation (7), D and F are distributions with zero expectations, $\mathbf{V}$ is the $(J \times J)$ covariance matrix rank A, and $\|\mathbf{V}\| >> \sigma^2$.

In a clean PCA model, both D and F are the normal distributions:

$$\boldsymbol{\delta} \propto \mathrm{N}(\mathbf{0}, \mathbf{V}) \qquad \boldsymbol{\varepsilon} \propto \mathrm{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \tag{8}$$

A non-clean model is more complex and cannot be described by exact equations. A typical case designed in line with Equation (1) is as follows. A contaminated data structure $\boldsymbol{\delta}$ combines two normal distributions

$$D = (1 - \mu_1)\mathrm{N}(\mu_1\mathbf{m}_1, \mathbf{V}) + \mu_1\mathrm{N}(-\mu_1\mathbf{m}_1, \mathbf{V}_1) \tag{9}$$

and the noise variable $\boldsymbol{\varepsilon}$ presumes outliers

$$F = (1 - \mu_2)\mathrm{N}\left(\mu_2\mathbf{m}_2, \sigma^2\mathbf{I}\right) + \mu_2\mathrm{N}\left(-\mu_2\mathbf{m}_2, \sigma_1^2\mathbf{I}\right) \tag{10}$$

It is evident that other models for $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are also possible.

The PCA statistical model is employed for the development of various tolerance areas and, therefore, for the assessment of an object type. The most important are the following areas. At a given significance level α, the *acceptance area* covers $(1 - \alpha)$ 100% of all population. If an object belongs to this area, it is assessed as a regular one; otherwise, it can be classified as an extreme or an outlier. Given I is the data set size, let us consider the acceptance area for a significance level $1 - (1 - \gamma)^{1/I}$. The probability that not all samples from the data set are located in this area (i.e., at least one sample lies outside) is equal to γ, which is therefore the probability of false outlier detection. In this way, the *outlier area* can be defined. The area located between the acceptance and the outlier areas is the *extreme objects area*.

The SIMCA method [15,16] designs the aforementioned areas using the SD and OD statistics. It is evident that small values of both the SD and OD characterize regular samples, whereas large SD or OD values signal the presence of outliers. The pertinent threshold limits can be developed using the SD and OD distributions.

2.4. The SD and OD distributions

Both the SD and OD statistics are quadratic forms of independent random variables. In case of a classical PCA model presented in Equation (8), the SD follows the scaled chi-squared distribution with A DoF [16]. The scaling factor is AI^{-1}. The OD distribution is more complex, even in case of the clean normally distributed data. In contrast to the SD, the OD (Equation (6)) is a sum of heteroscedastic components with variances equal $\lambda_a I^{-1}$. In a realistic case of PCA, given in Equation (9), it is impossible to present a known parametric distribution for both the SD and the OD. To emphasize the similarity between the SD and OD, we will use notation UD, referring to both SD and OD.

The UD distribution problem has been repeatedly discussed in chemometric literature. A detailed review can be found in Ref. [14]. Meanwhile, the solutions to the same problem have being developed in the applied statistics [17–20] since 1941. The following facts are established [20]. The UD distribution can be well approximated by the scaled chi-squared distribution

$$N\frac{u}{u_0} \propto \chi^2(N) \tag{11}$$

where u is the UD variable, u_0 is the scaling factor, and N is the DoF. The parameters u_0 and N are estimated from the training set $\mathbf{u} = \{u_1, \ldots u_I\}$ by the method of moments

$$\hat{u}_0 = \overline{u}, \qquad \hat{N} = \mathrm{int}\frac{2\hat{u}_0^2}{s_u^2} \tag{12}$$

where "int" stands for rounding to the nearest integer greater than 0, and $\overline{u}$ and s_u^2 are the conventional mean and variance estimates

$$\overline{u} = \frac{1}{I}\sum_{i=1}^{I} u_i, \quad s_u^2 = \frac{1}{I-1}\sum_{i=1}^{I}(u_i - \overline{u})^2 \tag{13}$$

This method works well for a clean PCA (Equation (7)), even in case of non-normal distribution of $\boldsymbol{\delta}$, but it fails in a non-clean PCA (Equation (9)), contaminated with outliers (Equation (10)).

Robust estimators for parameters u_0 and N should be developed within the concept presented by Equation (1). In this case, R is the chi-squared distribution distorted by an unknown distribution D. In [14], we proposed a robust estimator of parameter N. In the present paper, we suggest using enhanced robust estimators both for scaling factor u_0 and the DoF N. The method of moments is applied again, but the classical mean and variance statistics are replaced by their robust analogs, being median M and interquartile range S statistics. From Equation (11), it follows that

$$\begin{cases} M = \dfrac{u_0}{N}\chi^{-2}(0.5, N) \\ S = \dfrac{u_0}{N}\left[\chi^{-2}(0.75, N) - \chi^{-2}(0.25, N)\right] \end{cases} \tag{14}$$

The system of equations (14) should be solved with respect to the unknowns u_0 (real number) and N (natural number greater than 0).

Scaling parameter u_0 can be excluded

$$\frac{S}{M} = \frac{\chi^{-2}(0.75, N) - \chi^{-2}(0.25, N)}{\chi^{-2}(0.50, N)} = \mathrm{f}(N) \tag{15}$$

By direct calculations, it can be shown that for $N < 50$, function f(N) is well approximated by a non-linear expression

$$\mathrm{f}(N) = b\ \exp(-az^c), \text{where}\ \ z = \ln(N)$$

with parameters $a = 0.72414$, $b = 2.68631$, and $c = 0.84332$.

By using this approximation, a robust estimator for the DoF can be explicitly obtained

$$\widetilde{N} = int\ \exp\left[\left(\frac{1}{a}\ln\frac{bM}{S}\right)^{\frac{1}{c}}\right] \tag{16}$$

Because of rounding, the solution does not satisfy Equation (15) exactly. Therefore, it is better to use an averaged estimator for the scaling factor u_0

$$\widetilde{u}_0 = 0.5\widetilde{N}\left(\frac{M}{\chi^{-2}\left(0.5, \widetilde{N}\right)} + \frac{S}{\chi^{-2}\left(0.75, \widetilde{N}\right) - \chi^{-2}\left(0.25, \widetilde{N}\right)}\right) \tag{17}$$

Robust estimators presented in Equations (16) and (17) can be used as the alternatives to the conventional estimators given in Equations (12) and (13). In applications, it is always useful to compare the classical and robust estimates of h_0, v_0, N_h, and N_v. If the corresponding values (e.g., $\hat{N}_h$ and $\widetilde{N}_h$) differ considerably, it is indicative of a data set being contaminated with outliers. A special attention should be paid to the cases where the classical DoF estimates $\hat{N}_v = \hat{N}_h = 1$. This often points out a mixed structure of a data set. The examples of that are presented in Section 4.

2.5. The tolerance areas

The fact that both the SD and OD follow the scaled chi-squared distributions provides a possibility for developing of various tolerance areas that were explained in Section 2.3. A weighted sum of the SD variable, h, and the OD variable, v, follows the chi-square distribution with $N_h + N_v$ DoF

$$N_h\frac{h}{h_0} + N_v\frac{v}{v_0} \propto \chi^2(N_h + N_v) \tag{18}$$

Therefore, given a significance level α, the acceptance area $H\alpha$ is calculated as the $(1-\alpha)$ quantile of the chi-squared distribution with $N_v + N_h$ DoF

$$H_\alpha = \left\{(h, v) : N_h\frac{h}{h_0} + N_v\frac{v}{v_0} \le \chi^{-2}(1-\alpha, N_h + N_v)\right\} \tag{19}$$

The α value specifies a type I error in a decision making. This is the share of the false-negative decisions. For example, in a simulated data set (Section 3.3) of 100 objects, theoretically, five extremes can be expected for $\alpha = 0.05$, and there should be only one extreme object for $\alpha = 0.01$. It is worthy of mentioning that the acceptance area does not depend on the data set size I. The number of extremes increases proportionally to I, as it can be seen in the data set of size 10 000 (Section 3.2) and of size 100 (Section 3.3).

The outlier area O_γ has a similar form

$$O_\gamma = \left\{(h, v) : N_h\frac{h}{h_0} + N_v\frac{v}{v_0} > \chi^{-2}\left((1-\gamma)^{1/I}, N_h + N_v\right)\right\} \tag{20}$$

where γ is the outlier significance level. The area of extremes is situated in between. Figure 2 (left panel) illustrates these areas. The statistical meaning of the γ level essentially differs from that of the α level. γ specifies the probability that at least one regular object from the data set will be erroneously considered an outlier. That is why the outlier area depends on the data set size I. For a specific γ value, the greater I, the farther the outlier area. For example, in a case of one large data set (Section 3.2), no outlier was detected among 10 000 objects at $\gamma = 0.5$. At the same time, in a series of 10 data sets of a medium size (100 objects each, Section 3.3), in total, one outlier was singled out for $\gamma = 0.05$. In the latter case, the full probability of outlier detection $(1 - 0.95^{10} \approx 0.4)$ is close to that in the first case.

The whole concept, including the dual (classical and robust) data-driven (no presumed distributions) assessment of the tolerance areas in PCA/SIMCA method, is further validated using simulated and real-world data. For simplicity, this concept will further be called *DD-SIMCA*. The MATLAB code for DD-SIMCA can be downloaded from http://rcs.chph.ras.ru/SIMCA/DDSIMCA.zip

2.6. Fully robust procedure

DD-SIMCA can be applied in a classical variant (CDD-SIMCA), if the conventional estimators (Equations (12) and (13)) are used, or it can be semi-robust (RDD-SIMCA), if the robust estimators (Equations (16) and (17)) are applied. However, it is not fully robust because of the application of classical PCA. Certainly, it would be interesting to combine various robust PCA methods with the DD-SIMCA approach and assess the benefits of this association. This work is being planned, but in this paper, another interesting problem is considered. The goal is to compare the DD-SIMCA with a completely robust method. For comparison, we use the ROBPCA method [12] that employs the PP algorithm [21] realized in the ROBPCA toolbox [22].

In ROBPCA, the SD and OD are calculated similar to Equations (5) and (6) with one distinction of usage of the robust scores, loadings, and eigenvalues. For each distance, a cutoff value is specified. It is supposed that the SD values are approximately χ^2 distributed with DoF equal to A. For the OD values, the Wilson–Hilferty approximation [14,23] is used. This implies that the ODs to the power 1/3 are distributed approximately normally with mean m and variance σ^2. The estimates $\hat{m}$, s^2 for these parameters are obtained using the robust statistics [24]. The original cutoff values then are equal $\chi^{-2}(0.975, A)$ for the SDs and $\left[\hat{m} + s\Phi^{-1}(0.975)\right]^3$ for the ODs.

Unlike the original ROBPCA, we do not use any fixed predefined threshold values in our comparison. We employ the acceptance area that is calculated in a way similar to Equation (19)

$$H_\alpha = \left\{(SD, OD) : SD > \chi^{-2}(\varphi, A) \text{ and } OD > \left[\hat{m} + s\Phi^{-1}(\Phi)\right]^3\right\} \tag{21}$$

where $\varphi = (1-\alpha)^{1/2}$. The outlier area is calculated by Equation (21) with $\varphi = [1-(1-\gamma)^{1/h}]^{1/2}$.

The reasons for the application of Equation (21) are as follows:

1. The cutoff levels determination is an essential part of a specific SIMCA method; therefore, we may not change the procedure suggested by the authors of ROBPCA.
2. At the same time, for impartial methods comparison, the specific cutoff levels for the extreme samples are not fixed but calculated in dependence on the type I error α, in line with the DD-SIMCA method. The same is performed for the outlier cutoff level values, which depend on γ.

Originally, ROBPCA employs the square root of both distances, $OD^{1/2}$ and $SD^{1/2}$. The same transformation is used in all plots related to the ROBPCA application (e.g., Figure 1, right panel). It is also important to emphasize that in ROBPCA, robustification is applied in three steps: firstly, applying a robust preprocessing; secondly, using a robust version of PCA; and thirdly, applying the robust estimates for the cutoff calculations.

3. CASE STUDY I: SIMULATED DATA

3.1. Data design

Simulated data sets are used in case I. The clean normally distributed data are developed in line with Equation (8). They have the following characteristics. The numbers of variables $J = 3$, and the number of principal components $A = 2$. The $\boldsymbol{\delta}$ component properties are $E(\boldsymbol{\delta}) = \mathbf{0}$, $V_{11} = V_{22} = V_{33} = 0.25$, $\text{rank}(\mathbf{V}) = 2$. The $\boldsymbol{\varepsilon}$ component properties are $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\sigma = 0.05$. Two sets of different sizes, $I = 10\,000$ and $I = 100$, were studied. In the third data set (Section 3.4), $\sigma = 0.2$ was used for the outliers generation. Because of the data simulation procedure (with zero expectation and equal variances), no pre-processing is needed.

3.2. Large regular data set

In this case, 10 000 regular samples were simulated and analyzed with DD-SIMCA. The parameters of the SD and the OD distributions were obtained by Equations (12)–(13) and Equations (16)–(17). Classical estimates are

$$\hat{h}_0 = 0.0002000, \hat{v}_0 = 0.000832, \hat{N}_h = 2, \hat{N}_v = 1$$

Robust estimates have very similar values

$$\tilde{h}_0 = 0.0001997, \tilde{v}_0 = 0.000835, \tilde{N}_h = 2, \tilde{N}_v = 1$$

The known theoretical values [14]

$$h_0 = A/I = 0.0002, \qquad N_h = A = 2, N_v = J - K = 1$$

are also very close. Therefore, no practical differences can be noted between the CDD and RDD methods.

The result is shown in Figure 1, left panel. This is a so-called SIMCA plot, in which the scaled ODs (v/v_0) are shown in correspondence to the scaled SDs (h/h_0). A triangular structure of the data points' allocation is evident. The right panel represents the result of ROBPCA. The solid lines (significance level $\alpha = 0.005$) divide the plot into the acceptance area (below) and the extreme area (above). The region above the dashed lines is the outlier area (significance level $\gamma = 0.5$).

Table I shows the number of the extreme objects depending on the significance level α. Theoretical values (column *expected*) are equal to αI. The 0.95 tolerance limits are shown after the sign $\pm$. They are calculated using the binomial distributions. A close alignment between the theory and the experiment (column observed) confirms that DD-SIMCA works well in a classical PCA case. It is also important that both CDD and RDD-SIMCA give almost the same results for a regular data set. ROBPCA provides results that are below the tolerance limits at $\alpha > 0.05$.

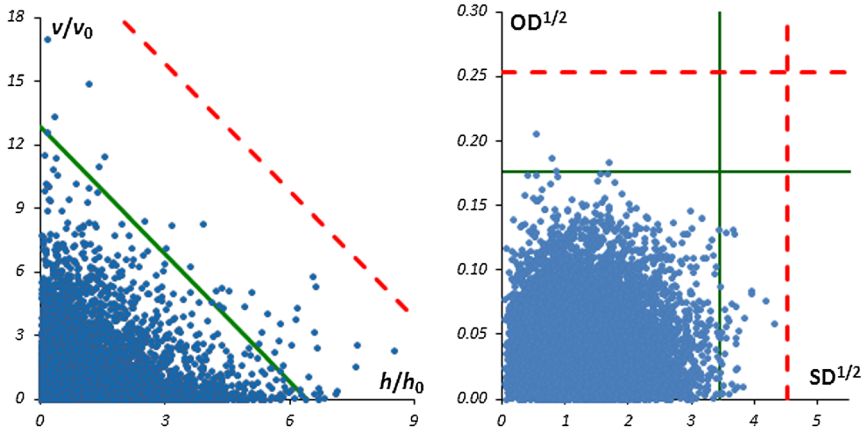

Figure 1. SIMCA plot for a large regular data set. Solid line is the border of acceptance area ($\alpha = 0.005$). Dashed line is the border of outlier area ($\gamma = 0.5$).

Table I. Number of the extreme objects in a regular data set of size 10 000

Significance α	Expected	Observed		
		CDD-SIMCA	RDD-SIMCA	ROBPCA
0.0001	1 ± 2	0	0	0
0.0005	5 ± 4	5	5	2
0.0010	10 ± 6	10	10	4
0.0050	50 ± 14	47	46	32
0.0100	100 ± 20	97	96	75
0.0500	500 ± 43	500	503	423
0.1000	1000 ± 58	1007	1005	878
0.2500	2500 ± 85	2529	2529	2314
0.5000	5000 ± 98	4989	4987	4890

3.3. Many regular data sets of a moderate size

In this experiment, 10 regular data sets of a moderate size ($I = 100$) were simulated and analyzed with DD-SIMCA and ROBPCA. Figure 2 represents the SIMCA plots for the ninth set.

The acceptance area and the outlier area are designed for $\alpha = 0.05$ and $\gamma = 0.05$. The left panel shows the CDD-SIMCA result, whereas the right panel shows the ROBPCA result. CDD-SIMCA revealed five extreme samples and no outliers. The ROBPCA findings are six extreme samples plus one outlier sample.

The results obtained from all data sets are shown in Table II.

They agree well with the expected values that are anticipated for regular data: the total numbers of the extreme and outlier samples are correspondingly $10 \times \alpha I = 50$, $10 \times \gamma = 0.5$. The RDD shows a slight over-robustness with three outliers found.

3.4. Regular data sets with outliers

In this experiment, 10 regular data sets ($I = 100$) were simulated again, but in contrast to the previous section, the ε component in the last three objects was generated with $\sigma = 0.2$. This was performed for simulation of possible outliers. Each data set was analyzed four times, employing different techniques.

Initially, first 97 objects were utilized as a training set, and the last three, which could be outliers, were predicted as a new (test) set. This technique allows us to develop a clear PCA model (a reference model) that is not contaminated with outliers and to compare the results with the other techniques (CDD-SIMCA,

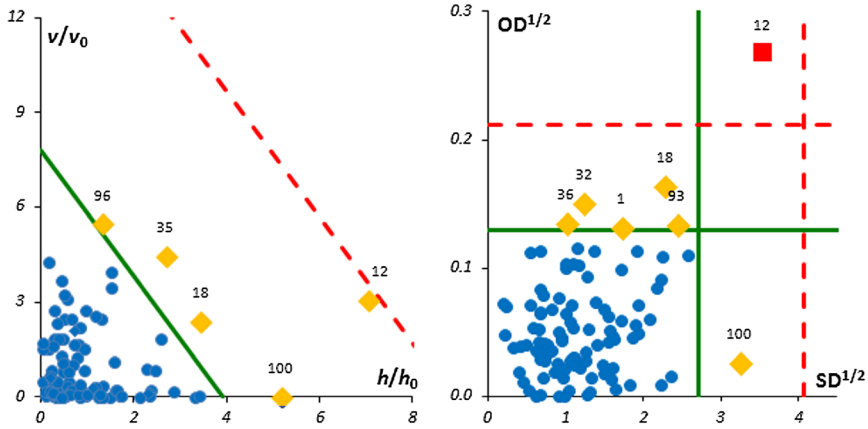

Figure 2. SIMCA plots for a regular data set of a moderate size. The left panel shows the CDD-SIMCA result, and the right one represents the ROBPCA result. Dots are regular samples, and diamonds are extreme samples. Solid line is the border of acceptance area ($\alpha = 0.05$). Dashed line is the border of outlier area ($\gamma = 0.05$).

Table II. The results obtained from 10 simulated regular data sets

Set no.	CDD-SIMCA		RDD-SIMCA		ROBPCA	
	Extremes	Outliers	Extremes	Outliers	Extremes	Outliers
1	5	0	2	0	4	0
2	6	0	5	0	8	0
3	4	0	3	0	4	0
4	3	1	4	1	7	0
5	7	0	8	1	8	0
6	4	0	5	0	5	0
7	6	0	9	0	4	0
8	6	0	6	0	5	0
9	5	0	4	1	6	1
10	5	0	3	0	3	0
Total	51	1	49	3	54	1

Journal of CHEMOMETRICS

RDD-SIMCA, and ROBPCA), which utilize all 100 samples as the training set. Four corresponding SIMCA plots obtained in the analysis of the first data set are presented in Figure 3.

The first reference technique (plot a) revealed two outliers and seven extreme objects. One of the extremes represents an unsuccessful outlier. CDD-SIMCA (plot b) lost one outlier (masking effect) and added it to the extreme objects. The RDD-SIMCA (plot c) results are very similar to the reference model (plot a). At last, ROBPCA (plot d) is close to CDD-SIMCA, having found one outlier and six extreme objects.

In this irregular case, it is difficult to calculate the expected numbers of extremes and outliers; therefore, the reference technique results were considered as the target values: a total of 58 extremes and 23 outliers in 10 data sets. The overall results, obtained from all 10 data sets, are presented in Table III. They confirm the findings from the first set: CDD-

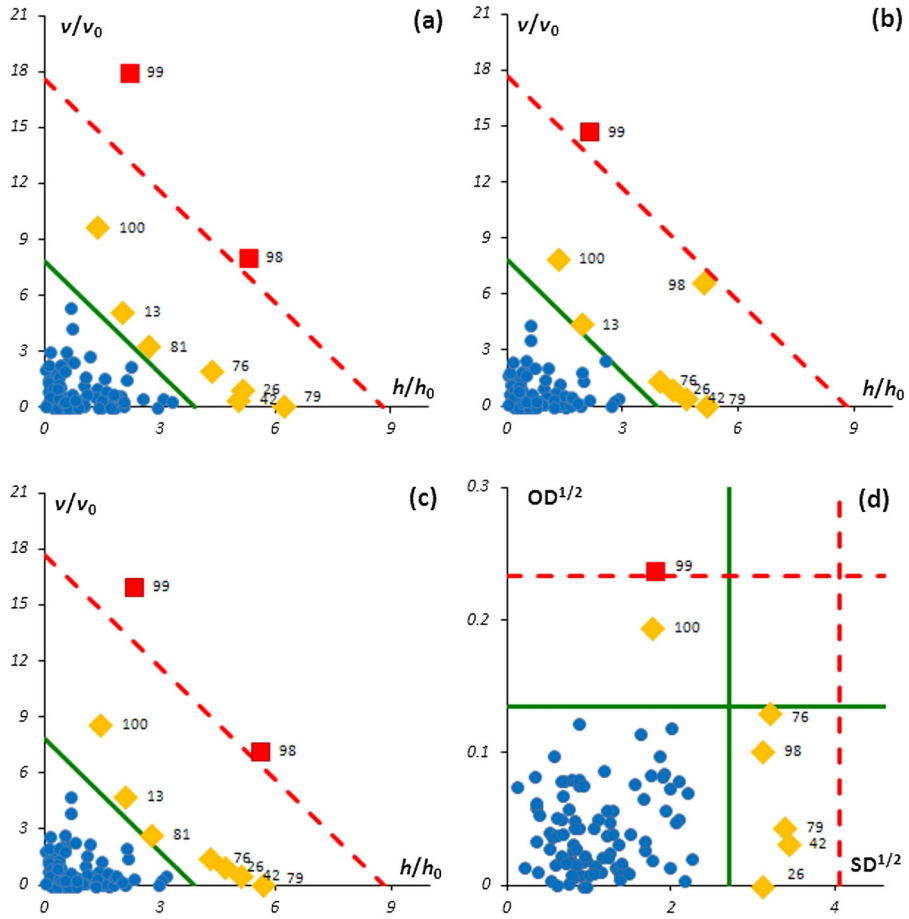

Figure 3. SIMCA plots for a regular data set with outliers. Plot (a) represents a reference technique, plot (b) is for CDD-SIMCA, plot (c) is for RDD-SIMCA, and plot (d) shows ROBPCA. Dots are regular samples, diamonds are extreme samples, and squares are outliers. Solid line is the border of acceptance area ($\alpha = 0.05$). Dashed line is the border of outlier area ($\gamma = 0.05$).

Table III. The results obtained from 10 simulated regular data sets with outliers

Set no.	Reference		CDD-SIMCA		RDD-SIMCA		ROBPCA	
	Ext	Out	Ext	Out	Ext	Out	Ext	Out
1	7	2	8	1	7	2	6	1
2	6	2	4	1	7	2	2	0
3	4	2	4	1	4	2	3	1
4	3	3	4	1	1	3	7	0
5	5	3	2	3	4	3	8	1
6	6	3	7	2	2	3	8	2
7	8	2	8	0	7	2	5	2
8	6	1	6	1	6	1	9	1
9	6	3	5	1	6	2	4	1
10	7	2	3	2	10	2	5	2
Total	58	23	51	13	54	22	57	11

SIMCA and ROBPCA have a tendency to mask the outliers, and RDD-SIMCA works very well almost in parallel with the reference method. Recently developed algorithm [25], which was kindly recommended by an anonymous reviewer, statistically confirms that RDD-SIMCA outperforms other methods presented in Table III.

4. CASE STUDY II: REAL-WORLD DATA

4.1. Packed pharmaceutical substance

This real-world example has been presented in Ref. [4]. The task is a NIR-based incoming inspection of taurine pharmaceutical substance packed in closed PE bags. The NIR spectra were recorded using the Spectrum 100N FT-NIR spectrometer (PerkinElmer, Buckinghamshire, UK) fitted with a handheld diffuse reflectance fiber optic probe with a 2 cm^{-1} spectral resolution. The initial spectral region was 4000–10 000 cm^{-1}, and the final region used for analysis was 4400–7400 cm^{-1}. For a validation purpose, four spectra of the other substance, caffeine, also used at the same pharmaceutical factory, were acquired. Taurine is a non-essential sulfur-containing amino acid, and pure caffeine is a plant-based alkaloid, which is applied to enhance the heart function in a way similar as taurine.

We acquired 246 spectra for 82 bags with taurine. Each bag was measured three times in different places. Additional experiments of measuring the substance in open PE bags directly and the detailed data analysis [4] showed that the whole data set is a mix of two groups. Group G1 consists of 200 spectra of the fine measurements through a single PE layer. Group G2 consists of 46 readings where the main substance peaks were distorted by the varying thickness of PE bags caused by bags' folds. Thus, the data set presents a mix of two populations (Equation (1) with $\mu = 0.25$), or, in other words, the data set is contaminated with abnormal objects, and the share of such objects is about 25%. The third data set (group G3) was obtained in a separate experiment. It includes four caffeine spectra that play a role of gross outliers.

In this case study, groups G1, G2, and G3 are combined in various ways, which illustrates typical problems encountered in data analysis. Before data processing, all spectra are column centered around a (robust) mean value, but not scaled. Data set complexity, that is, the number of PCs, was studded in Ref. [4]. Here, we apply previously yielded results and in all calculations use three PCs/robust PCs.

4.2. Clean data set G1

In this case, CDD-SIMCA provides the most reliable results. By using classical estimators (12), the following DoF were obtained: $\hat{N}_h = 3$, $\hat{N}_v = 5$. As to the robust variant given by Equations (16) and (17), the DoF estimates were $\widetilde{N}_h = 2$, $\widetilde{N}_v = 5$. It is worthy of reminding that the DoFs should be compared in the mating pairs, that is, $\hat{N}_h$ with $\widetilde{N}_h$, and $\hat{N}_v$ with $\widetilde{N}_v$. As a rule, the classical and robust estimates of DoF do not coincide, but do not differ materially either, as it can be in case when a data set is not contaminated with outliers.

For analysis of the quality of the PCA models, we propose to use the extreme plot (Figure 4). This plot demonstrates the dependence of the observed number of the extremes versus theoretically expected values, calculated as $n = \alpha I$. In practice, the plot is obtained by varying $\alpha = n/I$. The gray area represents the tolerance limits calculated as $t_\alpha = n \pm 2\sqrt{\alpha(1-\alpha)I} = n \pm 2\sqrt{n(1-n/I)}$.

The results for data set G1 are shown in the left panel of Figure 4. It confirms that the expected and calculated extremes are sufficiently close for the CDD-SIMCA and ROBPCA methods. RDD-SIMCA is worse, especially at small values of extremes. Both classical and robust estimates of DoF are rather close; the extreme plot shows CDD-SIMCA advantage. Thus, this is a regular data set and CDD-SIMCA should be used.

4.3. Clean data set G1 contaminated with evident outliers G3

In case of set G1 being contaminated with four evident outliers (group G3), a weakness of the classical approach can be seen. Figure 5 presents the acceptance areas ($\alpha = 0.1$, solid curve) and the outliers areas ($\gamma = 0.01$, dashed curve), obtained by CDD-SIMCA (left panel) and RDD-SIMCA (right panel). The curved shape of the areas is explained by the axes transformation $\sqrt{h/h_0}$, $\sqrt{v/v_0}$ that was made for better visualization. This is a useful trick that helps to present the SIMCA plots with large outliers.

All three methods single out all G3 samples as atypical ones. At the same time, the classical approach (CDD-SIMCA) specifies only two objects as outliers, and the other two are considered as extremes. Changing the α level does not influence the CDD-SIMCA results, till $\alpha = 0.1$, when one more extreme object appears. Predictably, ROBPCA as well as RDD-SIMCA shows proper performance, revealing the four outliers. The extreme plot (not shown) demonstrates a proper dependence of the number of extremes on the expected values.

In this case, the DoF estimates were $\hat{N}_h = \hat{N}_v = 1$ (classical) and $\widetilde{N}_h = 2$, $\widetilde{N}_v = 3$ (robust). Strange dependence of the CDD-SIMCA-based acceptance area on the α-level, and an

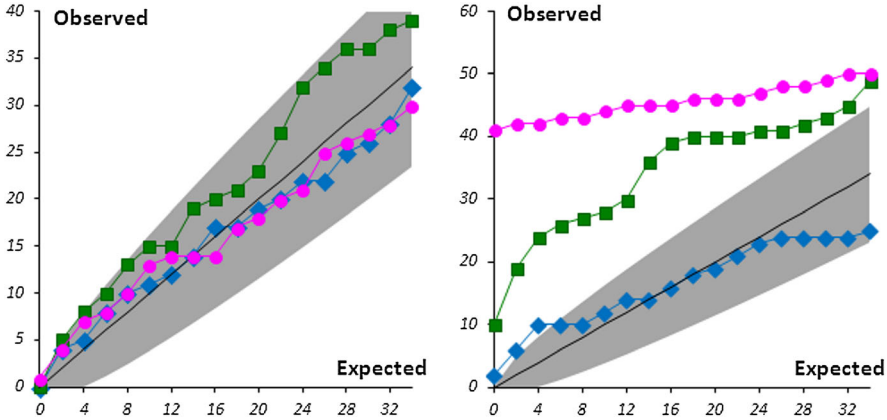

Figure 4. Extreme plots: observed number of extreme objects versus the expected number. Gray area represents the 0.95 tolerance limits. Diamonds show CDD-SIMCA, squares are for RDD-SIMCA, and dots represent ROBPCA.

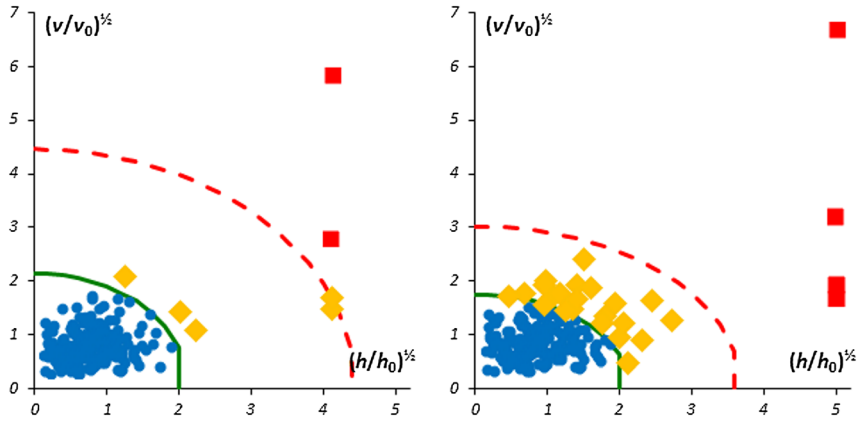

Figure 5. The SIMCA plots for data set G1 + G3. The left panel is for CDD-SIMCA, and the right one is for RDD-SIMCA. Solid curve limits the acceptance area ($\alpha = 0.1$), and dashed curve limits the outlier area ($\gamma = 0.01$).

essential difference between the classical and robust estimators of DoF, indicates the presence of outliers. From Figure 5 (left panel), it can be seen that the CDD-SIMCA acceptance area is distorted, especially in the OD (v) direction, because of a wrong estimation of the DoF. In this case, the masking effect is anticipated.

4.4. Data set consisting of a mixture of two distributions

Let us consider the case of a data set consisting of 200 objects of type G1 and 46 objects of type G2, 246 objects in total. For $\gamma = 0.05$, CDD-SIMCA revealed two outliers, RDD-SIMCA revealed 10 outliers, and ROBPCA revealed 41 outliers. The DoF estimates

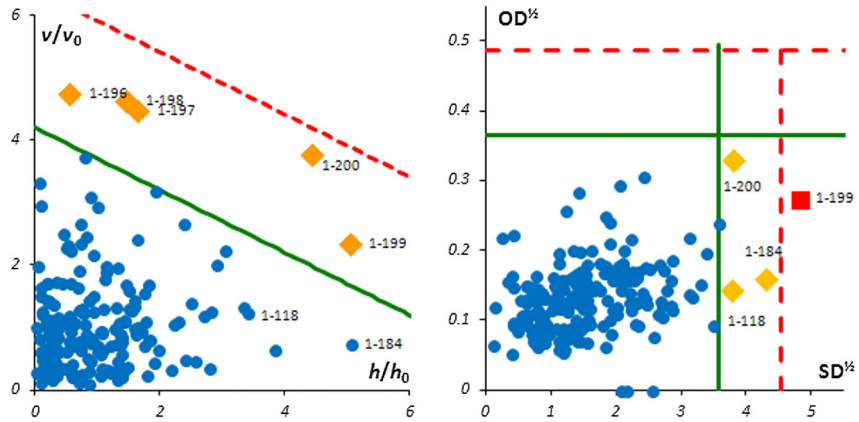

Figure 6. The data set cleaned by RDD-SIMCA. The left panel presents the RDD-SIMCA result, and the right one shows how this result is evaluated by ROBPCA. Solid lines limit the acceptance area ($\alpha = 0.01$), and dashed lines limit the outlier area ($\gamma = 0.05$).

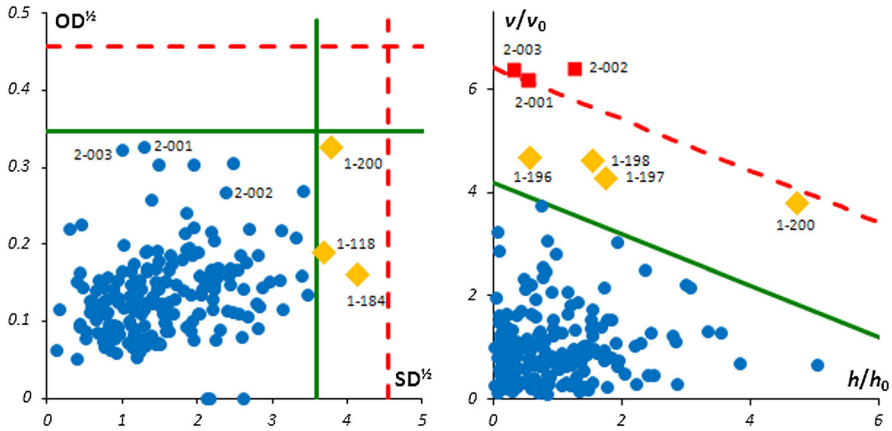

Figure 7. The data set cleaned by ROBPCA. The left panel presents the ROBPCA result, and the right shows how this result is evaluated by RDD-SIMCA. Solid lines limit the acceptance areas ($\alpha = 0.01$), and dashed lines limit the outlier areas ($\gamma = 0.05$).

were $\hat{N}_h = \hat{N}_v = 1$ (classical) and $\tilde{N}_h = 2, \quad \tilde{N}_v = 3$ (robust). The extreme plot (Figure 4, right panel) shows that if the results provided by CDD alone are considered (diamonds), no traces of suspicious objects can be found. CDD-SIMCA cannot reveal the mixture of two distributions and treats the data set as a regular one, without any contamination. This is an evident masking effect. Only a comparison with the robust methods helps to find a significant share of atypical objects. The first symptom can be seen in the extreme plot (Figure 4(b)), the number of extremes obtained by RDD-SIMCA and ROBPCA is always greater than the expected one. The second symptom is the high number of objects revealed as outliers by both robust methods. The most efficient results are provided by ROBPCA, but even this method cannot reveal all objects from G2 group in one step.

To clean the data set, the outliers found during the step should be excluded and the models recalculated. RDD-SIMCA requires nine consequent steps to clean the data set. In Figure 6, the results of cleaning by RDD-SIMCA are presented. Using this approach, all 46 objects from G2 were revealed and excluded. SIMCA plot for the remaining 200 objects (G1) is shown in the left panel. This cleaning can be evaluated by ROBPCA, and the right panel of Figure 6 represents this test. One object from of type G1 (labeled as 1-199) was considered as an outlier.

ROBPCA cleaned the data set in three steps, and the result is presented in the left panel of Figure 7. This method revealed 44 outliers, 43 from group G2 and one object (labeled as 1-199) from group G1. Three objects from G2 (labeled as 2-001, 2-002, and 2-003) have not been revealed. The evaluation of this result by the RDD-SIMCA method is shown in the right panel of Figure 7.

These plots illustrate that both methods lead to the similar results. RDD-SIMCA is stricter to the OD outliers, whereas ROBPCA is more rigorous with the high-leverage objects.

5. CONCLUSIONS

The proposed dual data-driven PCA/SIMCA (DD-SIMCA) technique has demonstrated a proper performance in the analysis of both regular and contaminated data sets. The following issues in the framework of the presented approach should be emphasized.

Extreme objects play an important role in data analysis. These objects are a mandatory attribute of any data set, and they should not be confused with outliers. The number of extremes should be compared with the expected number, coupled with the significance level α. A wrong number of such objects immediately indicates violations in the presumed data structure. Too large a number points out a mixed data structure, whereas too small a number is a sign of the outlier presence. The proposed extreme plot can assist in understanding the problem.

The extreme and outlier significance levels cannot be neglected. If the samples are considered as regular objects, the acceptance probability of $(1 - \alpha)$ should be presented. If an object is detected as an outlier, the probability of false detection (γ) should be given.

The main advantageous of DD-SIMCA are as follows:

- data-driven approach to the evaluation of distances distributions;
- dual method of estimation: classical for regular data and robust for contaminated data; and
- clear association with extreme and outlier significance levels.

It is important to emphasize that for the reliable classification of new objects, DD-SIMCA should be applied in an iterative manner with exclusion of outliers revealed during intermediate steps. Only afterwards, applying PCA to a purified calibration data, the final reliable acceptance area can be built.

The proposed DD-SIMCA approach cannot be viewed as an alternative to fully robust techniques, such as ROBPCA. In our opinion, the maximal benefit can be reaped from the combination of DD-SIMCA with these methods.

REFERENCES

1. Pomerantsev AL, Rodionova OY. Chemometric view on "comprehensive chemometrics". *Chemom. Intell. Lab. Syst.* 2010; **103**: 19–24.
2. Huber PJ. *Robust Statistics*. John Wiley & Sons: Chichester, 1981.
3. Tax D, Duin R. Outlier detection using classifier instability. *Lect. notes Comput. Sci.* 1998; **1451**: 593–601.
4. Rodionova OY, Sokovikov YV, Pomerantsev AL. Quality control of packed raw materials in pharmaceutical industry. *Anal. Chim. Acta* 2009; **642**: 222–227.
5. Rodionova OY, Pomerantsev AL. NIR based approach to counterfeit-drug detection. *Trends Anal. Chem.* 2010; **29**: 781–938.
6. Rodionova OY, Pomerantsev AL, Simple view on simple interval calculation (SIC) method. *Chemom. Intell. Lab. Syst.* 2009; **97**: 64–76.
7. Filzmoser P, Serneels S, Maronna R, Van Espen PJ. Robust multivariate methods in chemometrics. In *Comprehensive Chemometrics*. Chemical and Biochemical Data Analysis. Editors-in-Chief: Brown Stephen D, Tauler Romà, Walczak Beata (eds.). Elsevier: Amsterdam, 2009, 4 volumes.
8. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in one-way classification. *Ann. Math. Statist.* 1954; **25**: 290–302.
9. Maronna RA, Martin RD, Yohai VJ. Robust Statistics. *Theory and Methods*. Wiley: Toronto, 2006.
10. Locantore N, Marron JS, Simpson DG, Tripoli N, Zhang JT, Cohen KL. Principal component analysis for functional data. *Test* 1999; **8**: 1–73.
11. 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.
12. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: a new approach to robust principal component analysis. *Technometrics* 2005; **47**: 64–79.
13. 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.
14. Pomerantsev A. Acceptance areas for multivariate classification derived by projection methods. *J. Chemometrics* 2008; **22**: 601–609.
15. Kowalski BR, Wold S. Pattern recognition in chemistry. In *Handbook of Statistics*, 2. Krishnaiah PR, Kanal LN (eds.). North-Holland Publishing Company 1982, 673–697.
16. De Maesschalck R, Jouan-Rimbaud D, Massart DL. Tutorial. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* 2000; **50**: 1–18.
17. Satterthwaite FE. Synthesis of variance. *Psychometrika* 1941; **6**: 309–316.
18. Ali MM, Obaidullah M. Distribution of linear combinations of exponential variates. *Commun Stat Theor* 1982; **11**: 1453–1463.
19. Wood ATA. An F approximation to the distribution of a linear combination of chi-squared variables. *Commun Stat Simulat* 1989; **18**: 1439–1456.
20. Bentler PM, Xie J. Corrections to test statistics in principal Hessian directions. *Stat. Probabil. Lett* 2000; **47**: 381–389.
21. Croux C, Ruiz-Gazen A. High breakdown estimators for principal components: the projection-pursuit approach revisited. *J Multivariate Anal* 2005; **95**: 206–226.
22. ROBPCA. Robust principal component analysis based on the projection-pursuit approach http://www.econ.kuleuven.be/public/NDBAE06/programs/pca/robpca.txt
23. Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. *Technometrics* 1995; **37**: 41–59.
24. Vanden Branden K, Hubert M. Robust classification in high dimensions based on the SIMCA method. *Chemom. Intell. Lab. Syst* 2005; **79**: 10–21.
25. Héberger K, Kollár-Hunek K. Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. *J. Chemometrics* 2011; **25**: 151–158.

