Chemometrics and Intelligent Laboratory Systems 97 (2009) 64–74

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Simple view on Simple Interval Calculation (SIC) method

Oxana Ye. Rodionova *, Alexey L. Pomerantsev

Semenov Institute of Chemical Physics RAS, Kosygin Street 4, 119991, Moscow, Russia

ARTICLE INFO

Article history:
Received 15 March 2008
Received in revised form 10 December 2008
Accepted 13 December 2008
Available online 25 December 2008

Keywords:
SIC method
Point and interval estimators
Object status classification

ABSTRACT

Various regression methods and new interval approach are two different manners for solution of multivariate calibration problems. Both of these techniques aim to construct a model that associates multiple predictor variables with response and to predict the unknown response value for new predictors. When applied to the same data set both methods provide a researcher with reach information and results that supplement each other, despite the fact that different assumptions regarding the nature of error stand behind the two approaches and different mathematical techniques are applied. The goal of this work is to draw the attention of analysts to the outcomes that may be obtained under the assumption of error finiteness and to compare the results, provided by the interval approach with those yielded by regression modeling.

1. Introduction

Ordinary multivariate calibration (MVC) methods provide prediction result as a point estimate supplemented with some average prediction uncertainty common for all new objects. At the same time prediction with individual uncertainty interval is sometimes more useful for a practical application. This interval should account for all kinds of uncertainties: errors in predictors, errors in response, errors in calibration data, and errors in future prediction data. There are numerous technical methods on how to manage this [1–5], but generally accepted approach still does not exist. The conventional statistical technique is extremely complex [1], while the simulation methods are too time consuming [6]. At the same time, as it is marked in [7] there is more and more conviction that multivariate models should generate quantitative predictions and indicate their level of uncertainty. Application of the fuzzy regression model for the analysis of the uncertainties is presented in [8]. For construction of the robust and reliable models a number of various MVC procedures are designed [9–11].

An alternative to regression analysis is interval approach. In conventional regression analysis the estimates are the values of unknown parameters, which agree with the experimental data *in the best way*. In the interval method any parameter value that *does not contradict* experimental data is accepted as a feasible estimate. Apparently, Kantorovich [12] was the first who proposed this idea, but it was neither accepted nor widely used at that time. The benefits of the interval approach have been demonstrated for the kinetic parameter estimation [13], in the applied spectroscopy [14], for the instrument calibration [15] and signal processing and automatic control [16]. In this paper a Simple Interval Calculation (SIC) method, i.e. the approach within the interval framework, is described. This method returns the results of prediction directly in the interval form [17]. Furthermore, the SIC approach provides wide possibilities for *object status classification*, i.e. assess the relative importance of samples with respect to the constructed calibration model. The advantages of the SIC application are shown for numerous real-world data [18–21]. At the same time it must be emphasized that the SIC method has little in common with the 'interval mathematics' [22]. However, the SIC approach differs from traditional chemometric methods used for MVC [23–25]. It is unusual for analysts to yield the result of prediction/modeling not as a point estimate but directly in the interval form. Moreover, conventional regression approach is based on the traditional assumptions of error normality, errorless of predictors, etc. [26]. These assumptions are rarely held for practical data analysis of technological and natural systems [27]. The SIC approach is based on a postulate that all errors involved in the multivariate calibration problem are limited (sampling errors, measurement errors, modeling errors *etc.*). SIC also uses another mathematical tool (linear programming) in comparison with MVC problems.

These may be the reasons why the SIC approach is not widely-spread among the practitioners. The main goal of the paper is to draw the attention of the analysts to the outcomes that may be obtained under the assumption of error finiteness and to compare the results, provided by the interval approach with those yielded by traditional regression modeling.

The additional goal is to explain the essence of SIC using the simplest example in a step by step manner.

* Corresponding author.
E-mail address: oksana@chph.ras.ru (O.Y. Rodionova).

0169-7439/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2008.12.003

At the same time we do not want to oppose the advocated method to classical MVC approach. The most useful results are yielded when both methods are applied for multivariate data analyses and the regression results are supplemented with the SIC results.

This paper consists of four main sections. In the first section the arguments for the error finiteness are advocated. In the next section the main SIC conceptions are introduced using the univariate simulated example. The calculations are so simple that they can be done even without a computer. In the third section the comparison of the convergence of interval and maximum likelihood estimators is presented. In the last section a well-studied real-world example is described. Basic formulae of the SIC method are briefly reviewed in Appendix A.

2. Error finiteness

In the course of data analysis, the assumption of normal error distribution is a commonplace. Sometimes this is expressed explicitly, but often this is assumed by default. At the same time, chemometricians do not connect the fact of normal error distribution with its unboundedness. On the direct question about how often a researcher takes into account values that are located beyond four standard deviations (4σ), the answer is that if such values occurred they were excluded before data processing. At the same time, the amount of data in the modern multivariate data analysis is often greater than 10^{+6} [28]. Therefore, from the statistical point of view, there should be some 20–30 values that lie beyond 4σ. For example, we can refer to [29] where the authors state that "indeed, in real case studies, the chemist is often able to select, to some degree, the samples, and this will lead to more uniform distribution than normal distribution".

For illustration, let us consider a typical example. One of the traditional MVC problems is the wheat quality analysis that is performed with NIR spectroscopy [1,30]. In the example, NIR measurements were made using InfraLUM FT-10 NIR spectrometer. **X** matrix consists of NIR spectra in the range of 908–1120 nm, recorded at 118 wavelengths; the response ***y*** vector includes the moisture contents of 141 samples as quantified in the laboratory by a standard analytical method (evaporation loss of weight). The initial four-component PLS model based on 141 samples explains 98% of X-variance and 89% of Y-variance.

Fig. 1 shows distribution of the values of water content in wheat kernels (plot a) and PLS score plot PC3–PC4 (plot b). The conventional statistic tests show that response values ***y*** do not contradict a hypothesis about normal distribution. Even the three extreme samples, marked in Fig. 1, look as "admissible" values, as their probabilities are 0.03, 0.21, and 0.38 respectively. Nevertheless, following the ordinary MVC procedure analyst excludes all the samples that are classified as outliers with respect to **X** variables only, or to ***y*** variables only, or to both [30] (marked as filled dots in Fig. 1b) and recalibrate the data set. The results of new calibration with the censored data (124 samples) are shown in Fig. 2. Now, PLS model with 4 components explains 99% variance in **X** and 92% variance in ***y***. Samples are well spread in a score plot (Fig. 2b). At the same time, the distribution of response values becomes the truncated normal distribution cut on $\pm 2.5\sigma$ from the center.

This example demonstrates that traditional methods of data (pre) processing lead to limited errors that conforms truncated normal distribution rather than normal distribution. This could be immediately apprehended as soon as the essence of the MVC concept is recalled. All the conventional regression techniques (OLS, PCR, PLS, or other) employ linear models that are efficacious only in some neighborhood of the data center. Any outstanding data point being a statistical meaningful outlier (Y related), or a space motivated extreme (X related), should be removed in order to save the desired model linearity. The projection technique even aggravates the situation adding more X related samples in the outliers' list.

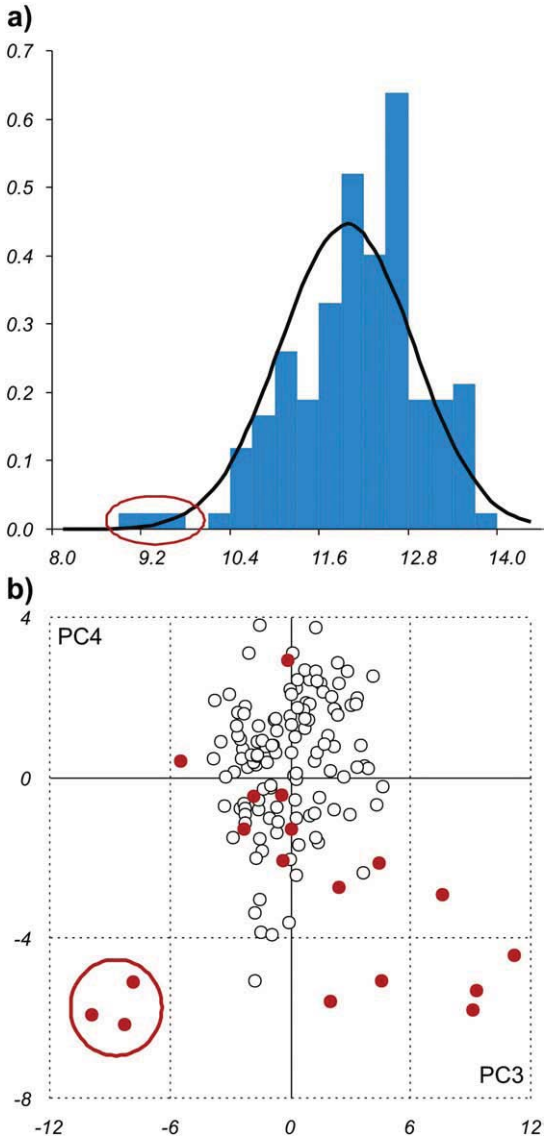

Fig. 1. NIR determination of moisture content in wheat kernels. Initial data set, 141 samples. a) Histogram of moisture content, reference values; b) Score plot for PLS model, PC3 vs. PC4. Filled dots show the "suspicious" objects.

3. Simple explanation of the SIC method. Univariate model

3.1. Simulated example

In this section a simplest univariate regression

$$y = xa + \varepsilon \quad (1)$$

will be used to explain the essence of SIC method. Here y is a response, x is a predictor, a is an unknown parameter, and ε is an error. The main assumption of the method is that the error ε is limited. This may be expressed as follows. The probability that absolute value of ε exceeds some constant β called Maximum Error Deviation (MED) is zero, i.e.

$$\mathrm{Prob}(|\varepsilon| > \beta) = 0. \quad (2)$$

Let us investigate the outcomes that can be drawn from the postulate given by Eq. (2). In Table 1 (columns 1–2) and in Fig. 3 the

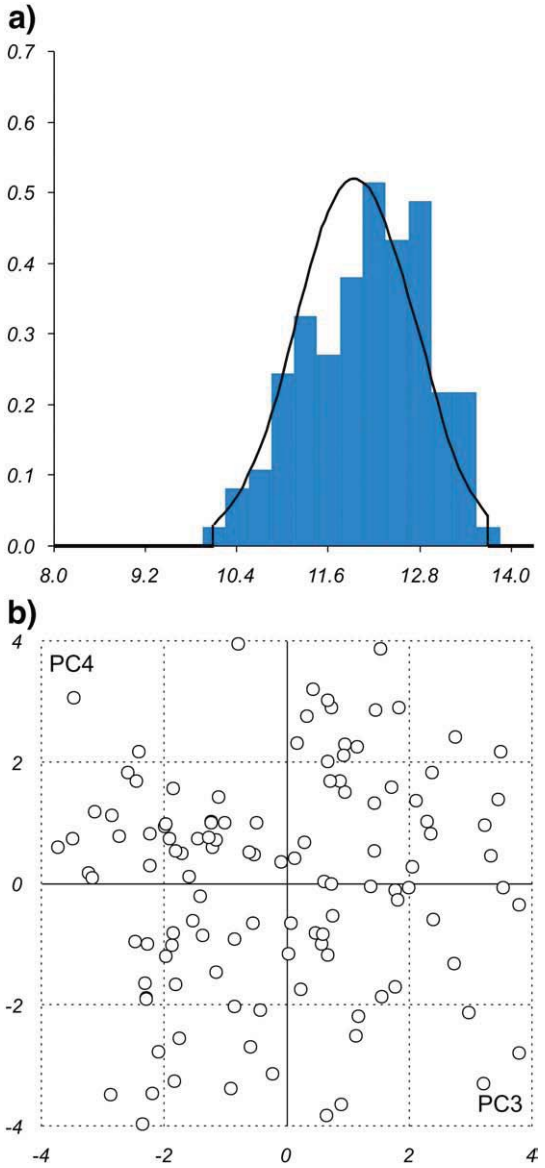

Fig. 2. NIR determination of moisture content in wheat kernels. Truncated data, 124 samples. a) Histogram of moisture content, reference values; b) Score plot for PLS model, PC3 vs. PC4.

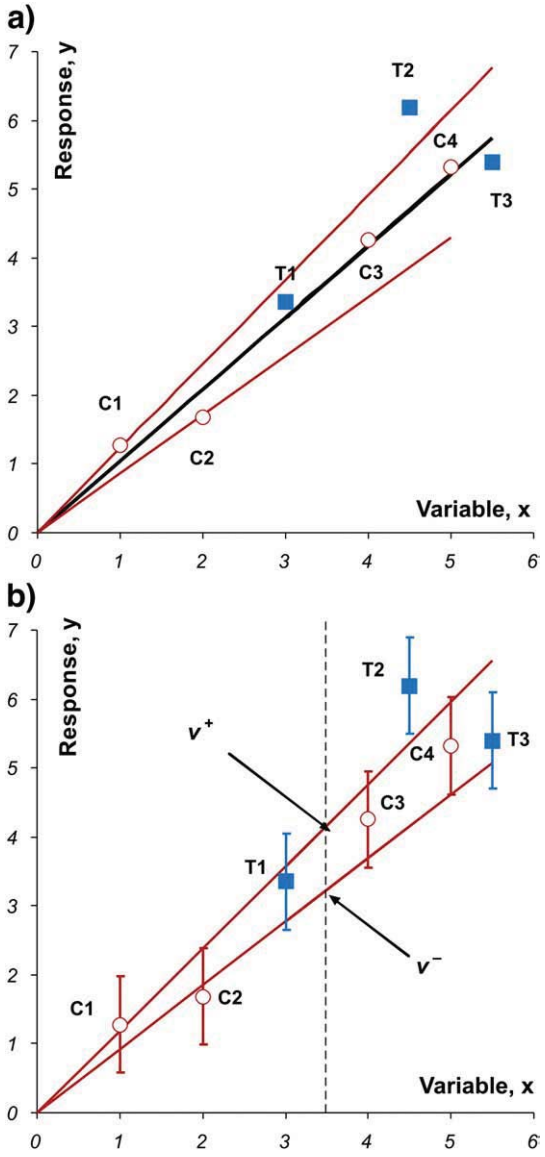

Fig. 3. Univariate simulated example. C–calibration samples, T–test samples. a) OLS estimates: — OLS prediction, — confidence interval limits; b) SIC estimates: I I error-bars, — SIC intervals' limits.

simulated data for regression model (1) with $a=1$ are presented. The error ε is simulated using the uniform distribution with the width that equals 1.4; so in this example $\beta=0.7$. It is worth of mentioning that column 2 in Table 1 presents the "measured" response value but not the "true" y value.

For illustration purposes we use a very small data set, which is divided into two parts. First four objects (C1–C4) are the calibration samples and they are used for modeling. These samples are shown by open dots in Fig. 3. The last three objects (T1–T3) are the test samples, the response values of which are to be predicted and validated. The samples T1–T3 are marked by closed squares in Fig. 3. In spite of the simplicity of the example, it helps to explain all the main properties of SIC method.

Table 1
Simulated data and the results of processing.

Samples	x	y	$\hat{y}$	$\hat{y}^-$	$\hat{y}^+$	$a^{\min}$	$a^{\max}$	v^-	v^+	h_{SIC}	r_{SIC}	$\lvert r\rvert+h$
0	1	2	3	4	5	6	7	8	9	10	11	12
C1	1.0	1.28	1.04	0.86	1.23	0.58	1.98	0.92	1.19	0.19	0.31	0.51
C2	2.0	1.68	2.09	1.72	2.46	0.49	**1.19**	1.85	2.38	0.38	−0.62	1.00
C3	4.0	4.25	4.18	3.43	4.92	0.89	1.24	3.70	4.76	0.76	0.03	0.79
C4	5.0	5.32	5.22	4.29	6.15	**0.92**	1.20	4.62	5.95	0.95	0.05	1.00
T1	3.0	3.35	3.13	2.58	3.69	0.88	1.35	2.77	3.57	0.57	0.26	0.83
T2	4.5	6.19	4.70	3.86	5.53	1.22	1.53	4.16	5.36	0.86	2.05	2.91
T3	5.5	5.40	5.74	4.72	6.76	0.85	1.11	5.08	6.55	1.05	−0.60	1.64

3.2. OLS calibration

Let us begin with the traditional ordinary least square (OLS) method. Using calibration data (x_i, y_i), $i=1,\ldots,4$ (columns 1 and 2 in Table 1, samples C1–C4), the OLS estimate for parameter a can be found

$$\hat{a} = \frac{\overline{y}}{\overline{x}} = 1.044, \quad \text{where } \overline{x} = \frac{1}{4}\sum_1^4 x_i, \quad \overline{y} = \frac{1}{4}\sum_1^4 y_i, \tag{3}$$

and the response value y for any x value, being a calibration, or a test object can be predicted as

$$\hat{y} = \hat{a}x, \tag{4}$$

(column 3 in Table 1, bold line in Fig. 3a). Using a well-known formula, it is also possible to estimate the error variance as

$$s^2 = \frac{1}{3}\sum_{1}^{4}(y_i - \hat{y}_i)^2 = 0.28, \quad (5)$$

and to construct the confidence intervals for y

$$\hat{y}^{\pm} = \hat{y} \pm s\frac{x}{2\bar{x}}t_3(P). \quad (6)$$

Here $t_3(P)$ is the quantile of Student's t-distribution for probability P with 3 degrees of freedom. These confidence limits are presented in Table 1 (columns 4, 5); they are also shown in Fig. 3a by thin lines.

3.3. SIC calibration

Let us see how these data are treated by the SIC method. First, let us assume that value $\beta = 0.7$ is known. (In the most of the real-world data the situation is more complicated and β value is a priori unknown. Later on it will be explained how to deal with such a problem.)

Applying the assumption of the error finiteness (Eq. (2)) to the regression model (Eq. (1)) one can easily see that for every object (x_i, y_i) from the calibration set ($i = 1,\ldots, 4$) the following inequality is fulfilled

$$|y_i - ax_i| \le \beta, \quad (7)$$

or in the equivalent form

$$a_i^{\min} \le a \le a_i^{\max}, \quad (8)$$

where

$$a_i^{\min} = \frac{y_i - \beta}{x_i} \qquad a_i^{\max} = \frac{y_i + \beta}{x_i}. \quad (9)$$

The values given by Eq. (9) are presented in Table 1 (columns 6, 7). Inequalities (Eq. (8)) should be satisfied simultaneously for all calibration samples, i.e. for $i = 1, 2, 3, 4$. Evidently this holds for all values of parameter a, which belong to the interval

$$a^{\min} \le a \le a^{\max}, \quad (10)$$

where

$$a^{min} = \max_{1 \le i \le 4} a_i^{\min}, \qquad a^{\max} = \min_{1 \le i \le 4} a_i^{\max}. \quad (11)$$

The restraining values are marked by the boldface font in the corresponding columns (6 and 7) in Table 1.

The interval given by Eq. (10) defines the *region of possible values* (RPV) for parameter a, i.e. such values of a that do not contradict the calibration data. Obviously, when parameter a is changed within the interval $[a^{\min}, a^{\max}]$, the corresponding response value $y = ax$, for an arbitrary x, is limited by values

$$v^- \le y \le v^+, \quad (12)$$

where

$$v^- = a^{\min}x, \qquad v^+ = a^{\max}x. \quad (13)$$

These limits are presented in Table 1 (columns 8 and 9).

Thus, the interval estimate (Eq. (10)) for parameter a is obtained. Simultaneously, it becomes possible to construct the prediction intervals (Eq. (13)) for response y that are valid both for the calibration samples and for any other (new) sample.

Let us interpret SIC method graphically. Fig. 3b shows the same data as Fig. 3a, but each point is now presented together with its error interval. The half-width of these intervals equals $\beta = 0.7$. When calculating the SIC estimates one should consider each possible line that passes through the origin of the coordinates and crosses/touches the error intervals of every calibration sample. Fig. 3b shows that the low limit is defined by the line that goes through the low bound of sample C4 error interval. The upper limit is determined by the line that goes through the upper bound of sample C2 error interval. Obviously, all the lines inside the two limits satisfy the posed conditions (Eq. (7)), and vice versa, each line that is located outside this angle conflicts with these inequalities. Boundaries are marked in Fig. 3b by bold lines. This is the result of SIC modeling. Now for any new x value (dotted line in Fig. 3b illustrates it) one can calculate the prediction interval $[v^+, v^-]$ for y value.

It is very important that SIC calibration is "based" on the two samples C2 and C4. Only these two samples define the boundaries (Eq. (10)) of the possible values for parameter a. Thus, these samples may be called the *boundary objects*. Other calibration objects, C1 and C3, are inessential in the example. These samples can be removed from the training set and SIC calibration model will not change. This is an important property of the SIC method. Hence, it can be seen that all calibration objects can be divided into two classes: the most important boundary samples, which determine the calibration model, and inessential samples, *insiders*, that may be removed from the calibration data without model deterioration. A more detailed object classification is presented in the next section.

3.4. Object status

Let us consider what happens to the SIC model, or better to say to RPV, if a new sample is added to the calibration set. Obviously, that RPV cannot increase; it can either decrease, or not change. For example, if object T3 is added, then the upper limit (line v^+) moves a little bit lower in such a manner that the line touches the upper bound of the T3 error interval. In this case, $a^{\max}$ changes from 1.19 to 1.11 (see Table 1). This property of the SIC method is called consistency (Eq. (A8) in Appendix A) and it is very important from theoretical point of view as it shows that the more samples are added to the calibration model the more accurate the SIC estimates are. Moreover, if an estimate of MED is selected in a proper way, i.e. it is not less than β, then the true parameter values a are always located inside the RPV (Eq. (10)). This property is called unbiasedness (Eq. (A5) in Appendix A). This is also important for the understanding of the SIC method, as this testifies that SIC estimates tend to the true parameter values a. However, not every new calibration sample improves the model. For example, sample T1 does not change it. This may be seen from Fig. 3b, where the prediction interval (two bold lines) is fully located inside the T1 error interval (the bar). Another case is for sample T2. Its error interval does not intersect with area limited by the boundary lines, i.e. the prediction intervals. Therefore, the addition of T2 to the calibration data destroys the model, as the system of inequalities (Eq. (7)) becomes inconsistent for the given value of β. As it can be easily seen from Table 1, after the addition of T2, the maximum over column 6 (1.22) becomes greater than the minimum over column 7 (1.19).

Thus, new samples can be divided into three groups regarding their influence on the model in case they are included in the calibration set. First of all, all the new samples, which do not change the model after being included in the calibration set, can be classified as *insiders*; those, which do change the model can be classified as *outsiders*. Moreover, among outsiders a group of *outliers* can be distinguished. The outliers cannot be included in the calibration set for the given β value because they destroy the model.

3.5. SIC object status classification

The investigation of an object status in the X–Y plot is inconvenient in the univariate case; moreover, in the multivariate case it is impossible. To make this analysis available in general, two numerical characteristics that reflect the object properties are introduced. They are the SIC residual r

$$r(\boldsymbol{x},y)=\frac{1}{\beta}\left(y-\frac{v^{+}(\boldsymbol{x})+v^{-}(\boldsymbol{x})}{2}\right), \tag{14}$$

which is calculated as the difference between the center of the prediction interval and the reference value (scaled by β); and the SIC leverage h

$$h(\boldsymbol{x})=\frac{1}{\beta}\left(\frac{v^{+}(\boldsymbol{x})-v^{-}(\boldsymbol{x})}{2}\right), \tag{15}$$

which is defined as the width of prediction interval, divided by calibration error. As soon as the prediction interval (Eq. (12)) is calculated, it is very easy to obtain the SIC residual r and SIC leverage h. In our example, the values of r and h are presented in columns 10, 11 of Table 1, and they are also shown in the Object Status Plot (OSP) in the r–h plane (Fig. 4a).

In Fig. 4a the calibration samples are denoted by open dots, and the new (test) samples are denoted as filled squares, in the same way as in Fig. 3. Bold open polygon ABCDE restricts the various object status areas. The form of this zigzag line is defined by two fundamental statements (Eqs. (A13) and (A15), in Appendix A) that establish the relation between h and r.

From Fig. 4a it may be seen that all calibration samples are located inside triangle BCD, so they are *insiders*. For these objects, $|r|+h\leq 1$ (Table 1, column 12). The samples C2 and C4 are located on the border of the triangle and therefore they are called *boundary* samples. For them $|r|+h=1$ (Eq. (A12) in Appendix A). The new object T1 lies inside the triangle, for T1, $|r|+h<1$, hence it may be classified as insider too. The sample T2 is located above line AB; this means that T2 is an *outlier*. For T2, $|r|-h=1.19>1$ (see Eq. (A14) in Appendix A). Object T3 is an *outsider*. From Fig. 4a it may be seen that sample T3 cannot be shifted into the insiders' area with a new value of r (or equivalently with new y value). For T3, $h=1.05>1$. This testifies that such x value contains some essentially new information that has not been presented in the calibration data. Such samples are called *absolute outsiders* (see Eq. (A15) in Appendix A).

Thus, it was shown that the SIC approach introduces an effective method for classification of all MVC objects (calibration samples, as well as new, or test samples). This classification is termed [18] as *Object Status Classification* (OSClas). It is based on definitions (Eqs. (14)–(15)) and statements (Eqs. (A11)–(A15) from Appendix A. OSClas may be applied to a problem with any dimensionality as it is reduced to calculation of values r and h for each object with their subsequent allocation in OSP.

It should be mentioned that a triangular shape of the insider's area in OSP (Fig. 4a) might appear somewhat similar to the conventional influence plot (Fig. 4b), in which the same samples are presented in coordinates OLS leverage (h_{OLS}) vs. OLS residual (r_{OLS}) [24]. The latter ones are calculated as follows

$$h_{OLS}=\boldsymbol{x}^{t}\left(\mathbf{X}^{t}\mathbf{X}\right)^{-1}\boldsymbol{x}=\boldsymbol{x}^{2}/\sum_{i=1}^{4}x_i^2, \qquad r_{OLS}=\frac{1}{\beta}(y-\hat{y}).$$

In reference [24, p. 286] it is written that: "Large leverage alone or large studentized residual alone is not necessary enough for the observation to be influential. At least a moderate contribution from each of these quantities is required for the influence to be large". This finding is very much along the same lines as the one developed here.

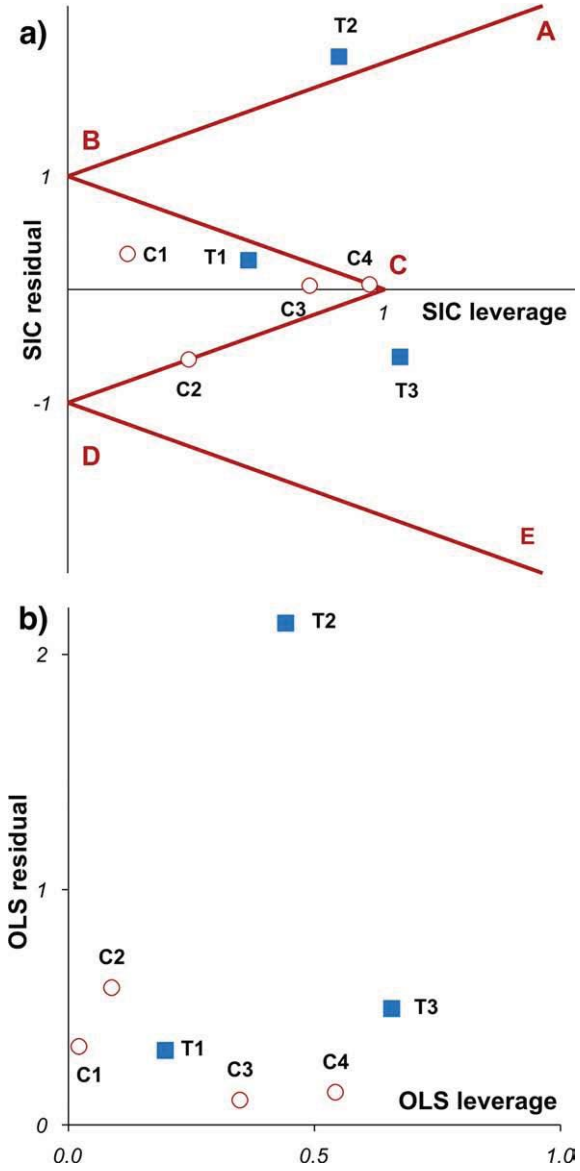

Fig. 4. Definition of the influence of each object for the univariate simulated example. ○—calibration samples, ■—test samples. a) SIC Object Status Plot; b) OLS influence plot.

Certainly, similarity between the influence plot and OSP is not a coincidence. This comes from a well-known basic statistical relationship [24], which relates the modeling *accuracy* (RMSEC), *precision* (SEC), and *bias* (BIAS):

$$\mathrm{RMSEC}^2\approx SEC^2+BIAS^2. \tag{16}$$

In the SIC approach, where MED value, β, is the calibration accuracy, SIC leverage, h, stands for the (normalized) precision, and SIC residual, r, is responsible for the (normalized) bias, this equation may then be represented in a form:

$$\beta^2=\beta^2h^2(\boldsymbol{x})+\beta^2r^2(\boldsymbol{x},y), \tag{17}$$

which actually conforms to Eq. (A12) in Appendix A. At the same time, we should recognize a substantial difference between Eqs. (16) and (17), as the former one makes sense only for the whole data set, i.e., *on average*, while the latter equality is valid for every sample in the data set.

In conclusion of this section we demonstrate the plots that show the relationship between the OLS and SIC characteristics of object

importance. Fig. 5a illustrates a very high correlation ($R^2=0.999$) between the OLS residuals and the SIC residuals. The relationship between the leverages is more complicated (Fig. 5b). Even higher correlation ($R^2=1.000$) is observed between the square root of OLS leverage and SIC leverage. This can be explained by the following. The OLS leverage is proportional to the prediction variance [25] that defines the size of the confidence interval, which is proportional to the square root of the variance. In its turn, the SIC leverage is proportional to the prediction interval width. Of course, for more complicated problems the relationships between OLS and SIC characteristics are not so simple, but the main tendencies are kept. Different aspects of similarities and dissimilarities of OLS and SIC methods as well as comparison of SIC prediction intervals and OLS confidence intervals have been investigated in [19].

3.6. Unknown MED value and how it can be estimated

The total procedure of β estimation is rather complicated and it is briefly described in Appendix A. The more detailed explanation may be found in [17]. For general understanding of the matter it is important to note that an estimate of β (denoted by b) always framed by 2σ–4σ, where σ is the error variance. Evidently, for any truncated error distribution, the β value cannot be less that 2σ; the smallest $\beta=1.71\sigma$ is for the uniform distribution. Then, for an ordinary sample set (say, less than 1000 objects), one cannot expect an extreme samples farther than 3σ. At last, the 4σ limit gives an assurance that any sample will cross this border. In the example under consideration $\beta=2.5\sigma$. This is greater than the value of 1.71, which corresponds to the uniform distribution applied for the error simulation, but this could be expected as only four points were available for calculation of the standard deviation s.

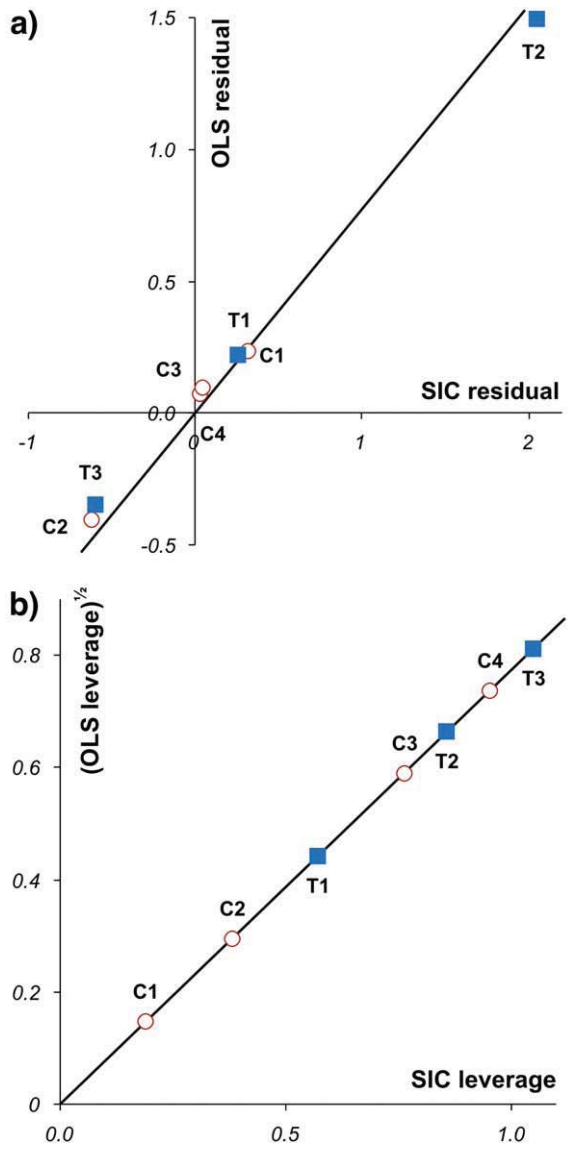

Fig. 5. Object characteristics in the univariate example. ○—calibration objects, ■—test objects. a) Comparison between OLS residuals and SIC residuals; b) Comparison between OLS leverages and SIC leverages.

It is natural to consider whether such variations in MED estimation have an influence on the results and conclusions of OSClas. The answer is negative, inasmuch as the main SIC quality measures, r and h, are defined as the relative ratios; see Eqs. (14) and (15)). SIC prediction intervals are quite another matter. They grow with b, and when $b=\beta$, these intervals have the covering probability of 1 by definition. At the same time, it can be shown, see e.g. [17], that the same intervals constructed with the estimated MED, b_{SIC} (for $P=0.90$), instead of the true β value, display the covering probability, which in any case is not less than 0.9999. This result confirms that not only the proposed OSClas alone, but also the whole SIC theory, in general, can be used in practice [18].

To illustrate this statement let us compare the OLS estimates and the SIC estimates in our univariate example. Comparing plots a) and b) in Fig. 3 one can see that each 95%-confidence interval for OLS prediction is wider than the correspondent SIC interval. If one calculates the probability of SIC interval using Eq. (6), it appears equal to 0.91. It is worth of mentioning that there we used a rather high value $\beta=2.5\sigma$, which corresponds to the normal probability $\mathrm{Prob}[-2.5,+2.5]=0.99$. As a result, our simple example reveals two important issues. First, the application of unlimited (normal) distribution for the confidence estimation leads to unreasonable wide intervals. Secondly, even for a small sample set, the SIC method provides us with the reasonable prediction intervals, which coincide with practical experience. To confirm the latter statement, 100,000 repeated simulations have been performed with our example, and the true value $y=x$ always fell into the SIC prediction intervals.

4. Convergence of interval estimates

One more elementary example is used for comparison the properties of interval (SIC) estimator to traditional maximum likelihood (ML) estimator. Let us consider a sample set $\mathbf{x}=(x_1,\ldots,x_n)$ from normal distribution $N(\alpha,\sigma^2)$, truncated on the interval $[\alpha-\beta,\alpha+\beta]$, $\beta=\kappa\sigma$. Parameter $\kappa=\beta/\sigma$ determines the truncation level (see Fig. 6). The task is to construct an estimator for the unknown parameter α with known values of β and κ and to investigate the estimator convergence, i.e. the dependence of estimate accuracy on the sample size.

Let us start with the traditional approach. It is well known that the ML estimator for parameter α is calculated as

$$a_{ML}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{18}$$

To characterize the a_{ML} accuracy the confidence interval may be used

$$\mathrm{Prob}(|a_{ML}-\alpha|<\beta h_{ML})=P.$$

Here

$$h_{ML}(P)=\frac{x_{0.5(1-P)}}{\sqrt{n}}\varphi(\kappa) \tag{19}$$

is the normalized half width of the confidence interval (ML leverage), and x_γ is the quantile of the normal distribution.

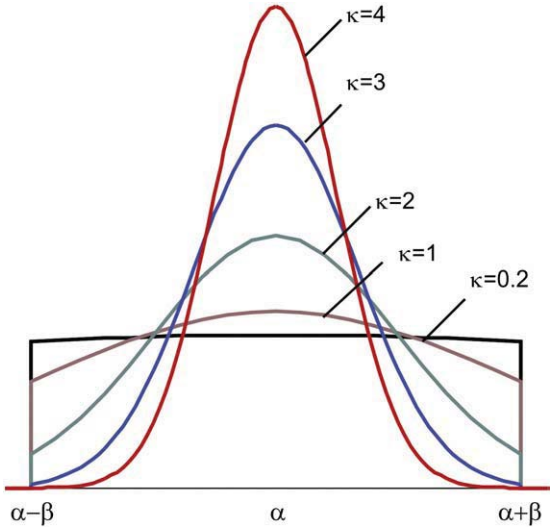

Fig. 6. Several truncated normal distributions with various κ values.

At the same time the SIC estimator for parameter a is calculated as

$$a_{SIC} = [\max(x_i - \beta), \min(x_i + \beta)]. \quad (20)$$

For confidence probability P, the half width (SIC leverage) of interval (Eq. (20)) is calculated as

$$h_{SIC}(P) = -\frac{\ln(1-P)}{n} 2\psi(\kappa). \quad (21)$$

Functions $\varphi(\kappa)$ and $\psi(\kappa)$ depend on parameter κ, which determines the way of distribution truncation ($\kappa = 0.2$, 1, 2, 3, 4). When $\kappa = 0.2$, the distribution under consideration is close to the uniform one, and for $\kappa = 4$ the distribution is hardly distinguished from normal distribution.

Comparing Eqs. (19) and (21) the following conclusions can be drawn.

1. In general, the width of uncertainty interval for the SIC estimator decreases $\sqrt{n}$ times faster than for the ML estimator.
2. For small sample sizes and/or high κ values conclusions are not that straightforward. Direct calculations show that, the SIC estimator is more efficient than the ML estimator beginning from some n_0 value that, in its turn, depends on κ.

Fig. 7 helps to compare the interval width (Eqs. (19), (21)) with the sample size and κ value. For $\kappa \leq 1$, and $n > 10$, h_{SIC} is less than h_{ML}. For $\kappa > 1$ and small sample sizes the width of uncertainty intervals for the ML estimator is less than that for the SIC estimator. However, the situation changes as the sample size increases. For example, when $\kappa = 2$ and confidence probability is $P = 0.9$, the width of the SIC uncertainty interval becomes less than the ML interval when the sample size is greater than 100. As to $\kappa = 3$, the same result is reached only when the sample size is greater than 3500 (Fig. 7b).

Recently, a new property of the SIC estimator has been found [31]. Let $f(x)$ be a symmetric finite density of error ε distribution; such that $f(x) = 0$ at $|x| > \beta$. The condition for effectiveness is $f'(\beta) \neq 0$. This means that in case of uniform distribution, the SIC estimator is more effective than in case of triangular (Simpson's) distribution. For the latter distribution the SIC estimator has the same effectiveness as the ML estimator.

In case of multivariate linear regression expressions given by Eqs. (19) and (21) should become much more complicated; the interval width depends on predictor matrix **X**, but not on n only. In such a case the proof of a similar statement turns into a challenging mathematical task.

5. Real-world example. Multivariate model

5.1. Data

In order to demonstrate the feasibility of SIC approach when applied to the real-world, and even multicollinear data, we use the well-known didactic "Octane Rating" [32] example. The **X** predictors are NIR-measurements (absorbance spectra) over 226 wavelengths in the range of 1100–1550 nm, while the responses y are the corresponding reference measurements (from the laboratory testings) of octane number. There are two sample sets: the calibration and test sets. In both sets the octane numbers vary from 87 to 93. The calibration set consists of $n = 24$ production gasoline samples and it is used for modeling. We also have an access to the genuine test set of 13 new samples, which is used for prediction testing only. In fact, this set includes four samples with alcohol added in them (Nos. 10–13). The sets without, or including these samples, are called the *short test set* (Nos. 1–9) and the *long test set* respectively (Nos. 1–13).

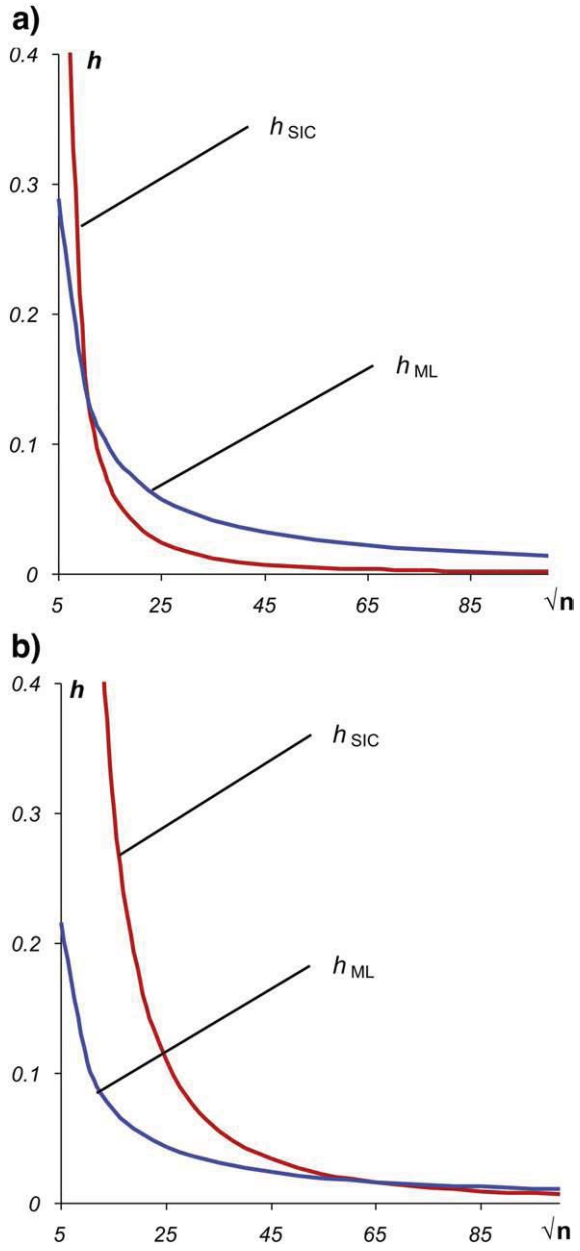

Fig. 7. Half width of the uncertainty intervals vs. sample size. Confidence probability $P = 0.9$. a) $\kappa = 2$; b) $\kappa = 3$.

5.2. Methods

Matrix **X** is rank deficient and as it is mentioned in Appendix A some regularization procedure should be used before application of the SIC method. For this particular NIR octane problem the calibration model carries two PLS-components

$$\boldsymbol{y} = m_0 1 + \mathbf{T}\boldsymbol{a} + \boldsymbol{\varepsilon},$$

where m_0 is the mean value of $\boldsymbol{y}$ and $\mathbf{T}$ is the $n \times 2$ score matrix. This model explains 97.5% of **X**-variance and 98.1% of $\boldsymbol{y}$-variance; RMSEC = 0.28, RMSEP = 0.31.

5.3. Calibration

To apply the SIC method it is necessary to evaluate the value of MED, β. It may be estimated as described in Appendix A. Here one should take into account that the projection methods always increase the total error deviation by the modeling errors, which are ascribable to the fact that bilinear (PLS, PCR) models are merely approximations to the complex systems. Thus β is necessarily always greater than an estimate of the measurement error alone.

Applying formula given by Eq. (A18) in Appendix A we obtain $b_{\min} = 0.484$. This means that for the given data set SIC method cannot be applied when b is less than 0.484, because the RPV will be empty. This is the lower bound of the MED estimate and its upper bound b_{SIC} should be evaluated as well. According to Eq. (A20), $b_{\mathrm{SIC}} = 0.880$. This value is used in all the subsequent calculations in this example. In addition, it can be concluded that for the given calibration set

$$0.48 \le \beta \le 0.88. \tag{22}$$

Calculating $s = \mathrm{RMSEC} = 0.268$, it is possible to compare the accuracy of modeling by PLS and SIC methods. Here $b_{\min}/s = 1.81$ and $b_{\mathrm{SIC}}/s = 3.28$.

Having the β value estimated with $b = 0.880$, one can proceed to the SIC model calibration. In general, it is not necessary to construct the RPV explicitly, and in case when model complexity is greater than two this is a very complex problem. In the octane rating example $p = 2$. Therefore, for illustration purposes, and in order to explain the SIC techniques, the RPV shape is presented in Fig. 8a.

In the same way as in the above univariate case, the RPV is formed not by all the calibration samples, but with the boundary objects only. There are six boundary samples (Nos. 7, 9, 13, 14, 18, 23), which are marked with closed dots in Fig. 8b that presents OSP for the calibration set. All boundary samples fall on the border of the triangle, i.e. for them $|r| + h = 1$. Only these objects form RPV as it is shown in Fig. 8a, where each line corresponds either to equation $\boldsymbol{t}_i^t\boldsymbol{a} = y_i - m_0 + b$ (marked with "+" subscript) or to equation $\boldsymbol{t}_i^t\boldsymbol{a} = y_i - m_0 - b$ (marked with "−" subscript). The numbers near these lines represent the boundary samples numbers.

So, the results of the SIC calibration are the estimate of β, and the set of boundary samples that form the RPV.

5.4. Prediction

Now, let us consider the SIC prediction intervals. For each sample from the test set, one should calculate the values v^- and v^+ (Eq. (A10) in Appendix A) that determine the limits of individual prediction interval. This optimization problem is solved using the linear programming methods that could give the problem solution in general, without explicit RPV presentation.

The problem of linear programming [33] is to minimize/maximize a linear function of continuous real variables, subject to linear constraints. The function that is being optimized is called the objective function. For purposes of describing and analyzing algorithms, the problem is often stated in the restricted normal form that is

$$\min_{\boldsymbol{a}} \left\{ \boldsymbol{c}^t\boldsymbol{a}, \text{ subject to } \mathbf{T}\boldsymbol{a} = \boldsymbol{d} \text{ and } \boldsymbol{a} \ge 0 \right\},$$

where $\boldsymbol{a} \in R^p$ is vector of unknowns, $\boldsymbol{c} \in R^p$ is so-called cost (known) vector, and $\mathbf{T} \in R^{n \times p}$ is the constraint matrix. The feasible region described by the constraints is a polyhedron, and the solution lies at a vertex of this polyhedron (Fig. 8a). Any system of linear inequalities may be translated to a restricted normal form using additional variables: slack variables are added to a problem to eliminate 'less-than' constraints, and surplus variables are added to a problem to eliminate 'greater-than' constraints. Moreover, any maximization problem may be converted to a minimization problem by changing the signs of the $\boldsymbol{c}$ coefficients in the objective function [34].

To solve the minimization problem the well-known Simplex method [33,34] is applied. This method generates a sequence of feasible iterates by repeatedly moving from one vertex of the feasible set to an adjacent vertex with a lower value of the objective function. When it is not possible to find an adjoining vertex with a lower value

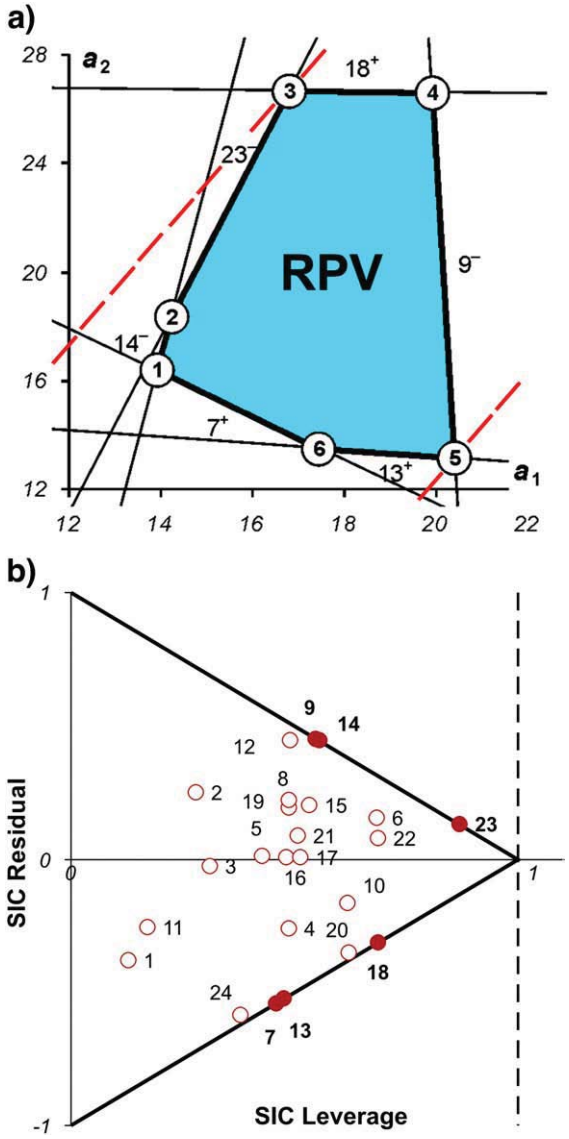

Fig. 8. Octane rating PLS model with 2 components. Training set. a) RPV in parameter space: — boundary line, C–vertex; — — prediction, solution of the optimization problem. b) SIC object status plot. C–insiders, ●–boundary samples.

of $\boldsymbol{c}^t\mathbf{a}$, the current vertex must be optimal, and termination occurs. To aid finding the first feasible solution (any vertex of the polyhedron) the artificial variables are added. The algorithm does not demand constructing the polyhedron explicitly, but calculates vertexes algebraically using the pertinent systems of linear equations. As a result, we get the optimal solution, i.e. vector $\boldsymbol{a}$, which is both feasible (satisfying the constraints) and optimal (obtaining the smallest objective value). The simplex method is well elaborated and it is included in many math packages.

In our example, the polyhedron formed by the linear constrains is RPV, which has six vertexes. For illustration purposes, each vertex in Fig. 8a is numbered. Let us find the prediction interval for the first test sample, which will be later marked with 0 sub-index. Using conventional PLS procedure its score vector, $\boldsymbol{t}_0 = (-0.0689; 0.0343)$, can be calculated. To find the limits of prediction interval that are v^- and v^+, it is necessary to solve two optimization problems

$$v^- = \min_{\boldsymbol{a}} \boldsymbol{t}_0^t\boldsymbol{a}, \qquad v^+ = \max_{\boldsymbol{a}} \boldsymbol{t}_0^t\boldsymbol{a},$$

where parameter vector $\boldsymbol{a}$ satisfies the constrains given by the calibration data ($\boldsymbol{y}$, $\mathbf{T}$),

$$y_i - m_0 - \beta \le \boldsymbol{t}_i^t\boldsymbol{a} \le y_i - m_0 + \beta, \quad i = 1, 2, \ldots, 24.$$

This means that vector $\boldsymbol{a}$ lies within RPV shown in Fig. 8a. Solving the problems, we yield the prediction interval for response $y_0 = m_0 + \boldsymbol{t}_0^t\boldsymbol{a}$, as

$$v^- + m_0 \le y_0 \le v^+ + m_0.$$

These solutions are indicated by the dashed lines in Fig. 8a.

Table 2 presents the related values calculated in the each vertex of RPV. It can be seen that prediction interval $v^- = 88.30$ and $v^+ = 89.01$ ensues from vertex 5 (minimum) and vertex 3 (maximum). Values for corresponding rows are marked by the boldface font.

Actually, in a complex problem, there is no need to examine each vertex, as this is inefficient and very time-consuming method. To find the optimum value the standard Simplex algorithm is used. The first feasible solution found by Simplex method is vertex 1. To find v^- value, the algorithm moves in the following way: vertex 1 → vertex 6 → vertex 5; to find v^+ value, the algorithm moves as vertex 1 → vertex 2 → vertex 3 (Fig. 8a).

5.5. Results

In this section the SIC prediction will be compared with the results calculated by PLS model $\hat{y}_{\text{test}}$ (Fig. 9a). To evaluate the PLS prediction uncertainty the traditional technique is applied, in which the root mean square error of prediction (RMSEP) is the average prediction error estimated at the validation stage. If new test samples are of the same kind and in the same range as the calibration samples, one should expect roughly the same average prediction error. In the example, it is exactly the case for the short test set. RMSEP calculated by leave-one-out cross validation is equal to 0.322. The correspondent

Table 2
Construction of SIC prediction interval.

Vertex #	a_1	a_2	$\boldsymbol{t}_0^t\boldsymbol{a}$	y_0
1	13.91	16.36	−0.398	88.85
2	14.22	18.36	−0.351	88.90
3	**16.79**	**26.66**	**−0.244**	**89.01**
4	19.91	26.61	−0.461	88.79
5	**20.41**	**13.16**	**−0.956**	**88.30**
6	17.43	13.51	−0.739	85.52

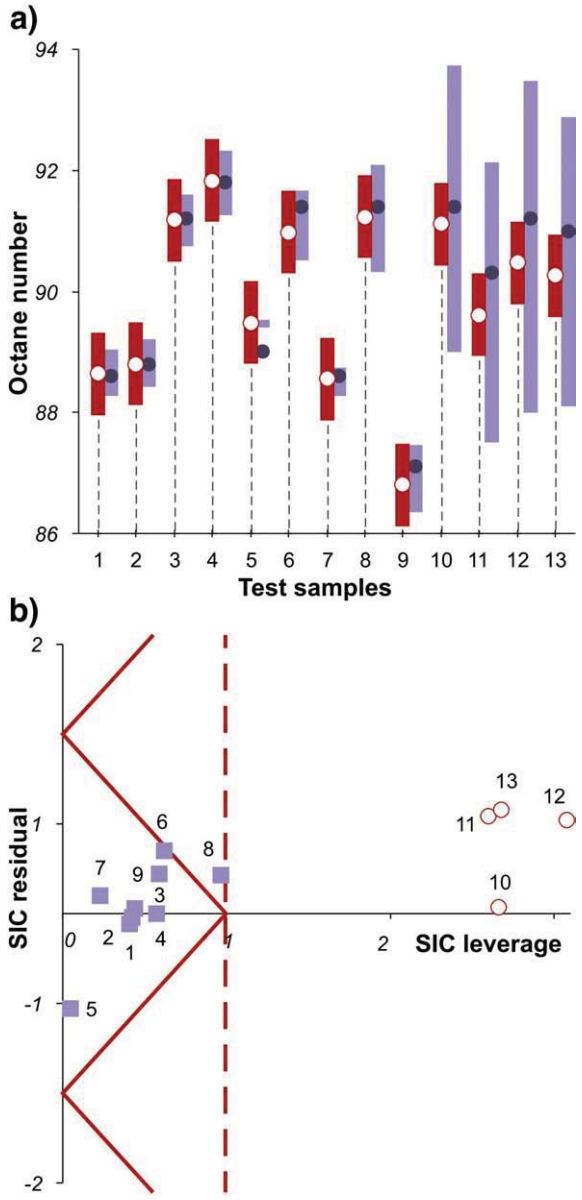

Fig. 9. Octane rating PLS model with 2 components. Test data. a) Prediction: ●—reference values, ▮—SIC prediction intervals, ○—PLS prediction, ▮—uncertainty bars; b) SIC object status plot: ▪—Nos. 1–9, ○—Nos. 10–13.

intervals $[\hat{y}_{\text{test}} \pm 2\text{RMSEP}]$ are shown in Fig. 9a as dark bars. To avoid the wrong results, such a technique obviously cannot be applied to the samples that are treated as outliers in PLS. The external validation [24], i.e. calculation of RMSEP on the test set, leads to the similar result, RMSEP = 0.250. Of course, for this calculation the short test set (without outliers) should be used.

For the last four test samples very large SIC prediction intervals are obtained (Fig. 9a). This is because such samples contain alcohol, and in that way they are different from the calibration set. In the conventional projection approach, such samples are treated as outliers. These samples may be easily determined also in OSP (Fig. 9b). The OSClas treats them as the absolute outsiders, i.e. the samples that are non-similar to the calibration samples.

Studying Fig. 9a, one can see that the reference values (closed dots), as well as the results of PLS prediction (open dots), lie inside the SIC-prediction intervals (light bars) and the uncertainty intervals derived from PLS model (dark bars) agree with SIC intervals for the 'normal

case' (test samples Nos. 1–9). At the same time, the SIC prediction intervals are individual for each new sample and therefore they are more informative in comparison with average value calculated for all samples by PLS. For the absolute outsiders (test samples Nos. 10–13), SIC intervals immediately signal their abnormality. It is also worthy of mentioning that for the 'regular' samples SIC intervals are smaller than the confidence intervals by PLS.

This example demonstrates how the SIC approach answers the main questions that are of great importance for an end-user.

1. Maximum error deviation (MED β) estimate represents the calibration accuracy and in this way characterizes the reproducibility for all samples that are similar to the calibration ones.
2. SIC prediction intervals present the uncertainty of each individual new sample.
3. The position of each sample in the Object Status Plot (OSP) determines whether this or that object is similar to the calibration set samples, i.e. it determines the sensible range for the model application.

If we are only working out a standard method for octane rating by NIR measurements, this could be cautious to state that such a technique may be applied to those samples, which are insiders. Just in this case we can guarantee that prediction uncertainty is not worse than calibration accuracy, which approximately is equal to the precision of the traditional ASTM based measurements. On the other hand, if we are developing a method for research investigations, it would be enough to warn that the method may not be applied for the absolute outsiders, because such samples are seriously different in their predictors' structure.

6. Conclusions

We believe that the presented SIC approach for the prediction interval construction may be useful in many practical applications within the multivariate calibration and data analysis. The validity of the method with respect to the above-mentioned assumptions is beyond all doubts. The main advantages of the SIC method are as follows. The method

- does not depend on the form of error distribution, although the efficiency of the estimate depends on the distribution form, more precisely on the heaviness of its tails;
- presents the results in the interval form with all uncertainties included;
- comprises the internal object classification approach, which can distinguish the reliable 'insiders', the doubtful 'outsiders', the significant 'boundary samples', the irrelevant 'absolute outsiders', and the destructive 'outliers'.
- uses no extra parameters, which cannot be evaluated by the data and have to be set a priori.

We started with the assumption that all errors are limited, and this assumption resulted in RPV, which is a volumetric estimate of unknown model parameter. In its turn, the application of RPV gave the results of prediction directly in the interval form. The specific calculation aspects of the SIC method are rather simple, since they are based on the well-designed procedures for linear programming and do not demand elaboration of new algorithms. Now, the SIC method is implemented in MATLAB script-language [37]. This is a beta release of the program that can be downloaded and used for free.

Appendix A

The SIC method is described in details in [17] and the object status classification is considered in [18]. Only basic features of the SIC method essential for the current paper are presented here.

A.1. Region of possible values

Let us consider a linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{a} + \boldsymbol{\varepsilon}, \tag{A1}$$

where $\boldsymbol{y}$ is the n-dimensional response vector; $\boldsymbol{a}$ is the p-dimensional parameter vector; $\mathbf{X}$ is the $(n\times p)$-predictor matrix, $\boldsymbol{\varepsilon}$ is the error vector. It is presumed that error ε is limited that means that there exists such a value $\beta>0$, called maximum error deviation (MED), that

$$\exists\beta > 0\ \mathrm{Prob}\{|\varepsilon| > \beta\} = 0, \text{ and for any } 0<b<\beta\ \mathrm{Prob}\{|\varepsilon| > b\} > 0, \tag{A2}$$

where Prob{•} denotes the probability that an event occurs. β considered to be common for all objects however this is not critical.

According to Eq. (A2), for each calibration object ($i=1,\ldots, n$) and known β value the following inequalities are fulfilled

$$y_i^- \le \boldsymbol{x}_i^t\boldsymbol{a} \le y_i^+, \qquad y_i^- = y_i - \beta, \qquad y_i^+ = y_i + \beta. \tag{A3}$$

As the true parameter vector $\boldsymbol{\alpha}$, is unknown, it is possible to consider all the vectors $\boldsymbol{a}$, which agree with the inequalities. All such vectors $\boldsymbol{a}$, for a given i, form a *strip* $S(\boldsymbol{x}_i, y_i)$ in the space of parameters R^p. Any vector $\boldsymbol{a}$ satisfies all inequalities given by Eq. (A3) simultaneously, if and only if (later 'if and only if' is abbreviated as 'iff') it belongs to all strips $S(\boldsymbol{x}_i, y_i)$.

A *region of possible values* (RPV) A for parameter $\boldsymbol{a}$ is a set in parameter space determined by the intersection of all strips, i.e.

$$A = \bigcap_{i=1}^{n} S(\boldsymbol{x}_i, y_i). \tag{A4}$$

Region A is a *closed convex polyhedron* [34], delineated by the boundaries of intersecting strips. This is a *random set* because the RPV is constructed using random values $\boldsymbol{y}$.

A.2. The RPV properties for model Eq. (A1)

1. The region A is an unbiased estimator of parameter $\boldsymbol{\alpha}$.

$$\mathrm{Prob}\{\alpha\in A\} = 1. \tag{A5}$$

2. The region A is bounded iff rank$\mathbf{X}=p$ [34]. To apply the SIC method a standard technique [17,24] should be used to *project* the initial data on a lower-dimensional subspace

$$\boldsymbol{y} = \mathbf{T}\mathbf{P}^t\boldsymbol{a} + \boldsymbol{f} = \mathbf{T}\boldsymbol{c} + \boldsymbol{f}, \tag{A6}$$

where the score matrix $\mathbf{T}$ has the full rank $k<p$.

3. The region A is a consistent estimator of $\boldsymbol{\alpha}$, i.e.

$$\mathrm{Prob}\{A\cap\alpha\} = 1 \text{ as } n\to\infty, \tag{A7}$$

under the same traditional *weak* conditions

$$\lambda_p \to \infty \text{ as } n\to\infty, \tag{A8}$$

as for the OLS estimate.

The RPV is formed not by all objects from the calibration set, but by a subset of *boundary objects* only.

A.3. Predicting the response

Consider a response prediction for any new vector $\boldsymbol{x}$ using model in Eq. (A1). If parameter $\boldsymbol{a}$ varies over the RPV A, the predicted value $\boldsymbol{y}=\boldsymbol{x}^t\boldsymbol{a}$ belongs to the interval

$$V = \left[v^-, v^+\right], \tag{A9}$$

where

$$v^- = \min_{\boldsymbol{a}\in A}\left(\boldsymbol{x}^t\boldsymbol{a}\right), \qquad v^+ = \max_{\boldsymbol{a}\in A}\left(\boldsymbol{x}^t\boldsymbol{a}\right). \tag{A10}$$

Interval V is the result of the SIC prediction. The solutions of optimization problems given by Eq. (A10) may be obtained by the linear programming methods [33,34], which are commonly used [35].

A.4. List of statements for SIC object status classification

Statement 1. All *calibration* samples satisfy inequality

$$|r(\boldsymbol{x},y)| \le 1 - h(\boldsymbol{x}). \tag{A11}$$

Statement 2. Calibration object $(\boldsymbol{x}_i, y_i)$ is a boundary object, iff

$$|r(\boldsymbol{x}_i,y_i)| = 1 - h(\boldsymbol{x}_i). \tag{A12}$$

Statement 3. An object $(\boldsymbol{x},y)$ is an insider, iff

$$|r(\boldsymbol{x},y)| \le 1 - h(\boldsymbol{x}). \tag{A13}$$

Statement 4. An object $(\boldsymbol{x},y)$ is an outlier, iff

$$|r(\boldsymbol{x},y)| > 1 + h(\boldsymbol{x}). \tag{A14}$$

Statement 5. An object $(\boldsymbol{x},y)$ is an absolute outsider, iff

$$h(\boldsymbol{x}) > 1. \tag{A15}$$

Applying statements given by Eqs. (A11)–(A15), one can construct an *object status plot* (OSP), which is a two-dimensional plot for *any* dimensionality of the initial data $(\mathbf{X}, \boldsymbol{y})$ and for *any* number of model parameters.

A.5. Estimation of MED

As a rule, the MED value is unknown and some estimate b is used instead of β. In this case RPV A depends on b and $A(b)$ is extended monotonically with increasing of b

$$b_1 > b_2 \Rightarrow A(b_1) \supset A(b_2). \tag{A16}$$

Therefore, it can be claimed that for a sequence of consistent β estimates $b_1 > b_2 > \ldots \ge \beta$, properties (A5)–(A8) are true for $A(b_n)$ as well. Furthermore,

$$A(0) = \varnothing, \quad A(\infty) \neq \varnothing. \tag{A17}$$

From Eqs. (A16)–(A17) it follows that there exists *minimum* b such that $A(b) \neq \varnothing$. This minimum value can be taken as an estimator for the unknown parameter β

$$b_{\min} = \min\{b, \quad A(b) \neq \varnothing\}. \tag{A18}$$

Estimate in Eq. (A18) is a consistent but biased ($b_{\min} \le \beta$), and it is the low limit of all possible β values.

Applying the traditional statistical approach [36], it is possible to find such an estimator b that $\mathrm{Prob}\{b > \beta\} > 0.95$ and b is as close to β as possible. Let us consider $\bar{\boldsymbol{a}}$—some point (regression) estimate of parameter $\boldsymbol{\alpha}$, residuals $\boldsymbol{e} = \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{a}}$, and statistics

$$b_{\mathrm{reg}} = \max(|e_1|, \ldots, |e_n|). \tag{A19}$$

Statistical simulations help to construct the enhanced estimator b_{SIC}

$$b_{\mathrm{SIC}} = b_{\mathrm{reg}} C\left(n, s^2\right) \tag{A20}$$

as the 0.95 upper limit for β. Empirical function C [17] depends on n that is the number of objects in the calibration set, and on the residual *variance* s^2, which characterizes the heaviness of tails of the error distribution.

References

[1] K. Faber, B. Kowalski, Propagation of measurement errors for the validation of prediction obtained by principal component regression and partial least squares, J. Chemom. 11 (1997) 181–238.
[2] M. Hoy, K. Steen, H. Martens, Review of partial least squares regression prediction error in Unscrambler, Chemom. Intell. Lab. Syst. 44 (1998) 123–133.
[3] K. Faber, Comparison of two recently proposed expressions for partial least squares regression prediction error, Chemom. Intell. Lab. Syst. 52 (2000) 123–134.
[4] S. De Vries, C.J. Ter Braak, Prediction error in partial least squares regression: a critique on the deviation used in The Unscrambler, Chemom. Intell. Lab. Syst. 30 (1995) 239–245.
[5] J.A. Fernandez Pierna, L. Jin, F. Wahl, N.M. Faber, D.L. Massart, Estimation of partial least squares regression prediction uncertainty when the reference values carry a sizeable measurement error, Chemom. Intell. Lab. Syst. 65 (2003) 281–291.
[6] A.L. Pomerantsev, Confidence intervals for nonlinear regression extrapolation, Chemom. Intell. Lab. Syst. 49 (1999) 41–48.
[7] N.M. Faber, X.-H. Song, P.K. Hopke, Sample-specific standard error of prediction for partial least squares regression, Trends Anal. Chem. 22 (2003) 330–334.
[8] R. Rajkó, Treatment of model error in calibration by robust and fuzzy procedures, Anal. Lett. 27 (1994) 215–228.
[9] M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, B. Walczak, Robust statistics in data analysis — a review. Basic concepts. Robust statistics in data analysis — a review basic concepts, Chemom. Intell. Lab. Syst. 85 (2007) 203–219.
[10] S. Frosch Moeller, J. von Frese, R. Bro, Robust methods for multivariate data analysis, J. Chemom. 19 (2005) 549–563.
[11] iPLS, web: http://www.models.kvl.dk/source/iToolbox/index.asp).
[12] L.V. Kantorovich, About some new approaches to calculation methods and data processing, Sib. Math. J. 3 (1962) 701–709 (in Russian).
[13] S.I. Spivak, M.G. Slinko, V.I. Timoshenko, V.Y. Mashkin, Interval estimation in determination of parameters of a kinetic-model, React. Kinet. Catal. Lett. 3 (1975) 105–113.
[14] V.M. Anisimov, A.L. Pomerantsev, A.G. Novoradovskii, O.N. Karpukhin, Determining the sensitivity of materials to polychromatic light, J. Appl. Spectrosc. 46 (1987) 97–101.
[15] A.P. Votshinin, A.F. Bochkov, G.R. Sotirov, Method of data analysis at interval non-statistical error, Ind. Lab. 56 (1990) 76–95.
[16] B.M. Ninness, G.C. Graham, Rapprochement between bounded error and stochastic estimation theory, Int. J. Adapt. Control Signal Process. 9 (1995) 107–132.
[17] O.Ye. Rodionova, A.L. Pomerantsev, Principles of Simple Interval Calculations, in: Pomerantsev (Ed.), Progress in Chemometrics Research, NovaScience Publishers, NY, 2005, pp. 43–64.
[18] O.Ye. Rodionova, K.H. Esbensen, A.L. Pomerantsev, Application of SIC (Simple Interval Calculation) for object status classification and outlier detection — comparison with regression approach, J. Chemom. 18 (2004) 402–413.
[19] A.L. Pomerantsev, O. Ye Rodionova, Hard and soft methods for prediction of antioxidants' activity based on the DSC measurements, Chemom. Intell. Lab. Syst. 79 (2005) 73–83.
[20] A.L. Pomerantsev, O. Ye Rodionova, A. Höskuldsson, Process control and optimization with Simple Interval Calculation method, Chemom. Intell. Lab. Syst. 81 (2006) 165–179.
[21] O. Ye Rodionova, A.L. Pomerantsev, Subset selection strategy, J. Chemom. 22 (2008) 674–685.
[22] B.E. Moore, Interval Analysis, Prentice-Hall, N.Y., 1966.
[23] P.J. Brown, Multivariate calibration (with discussion), J. Roy. Stat. Soc. B 44 (1982) 287–321.
[24] H. Martens, T. Næs, Multivariate Calibration, Wiley, New York, 1998.
[25] A. Höskuldsson, Prediction Methods in Science and Technology, vol. 1, Thor Publishing, Copenhagen, Denmark, 1996.
[26] G.A.F. Seber, Linear Regression Analysis, Wiley and sons, NY, 1997.
[27] V.J. Clancey, Statistical methods in chemical analyses, Nature 159 (1947) 339–340.
[28] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multi- and megavariate data analysis, Umetrics, 2001.
[29] D. Jouan-Rimbaud, D.L. Massart, C.A. Saby, C. Puel, Characterization of the representativity of selected sets in multivariate calibration and pattern recognition, Anal. Chim. Acta 350 (1997) 149–161.
[30] E.L. Sulima, K.A. Zharinov, V.A. Zubkov, L.A. Rusinov, Specific features of practical implementation of calibration model transfer from a master instrument to slave NIR analyzers for analysis of main characteristics of wheat, in: Pomerantsev (Ed.), Progress in Chemometrics Research, NovaScience Publishers, NY, 2005, pp. 212–218.
[31] A.L. Bulyanitsa, Conditions of effectiveness of Simple Interval Calculation method, Rev. Ind. Appl. Math. 15 (2) (2008) 254 (in Russian).
[32] K.H. Esbensen, Multivariate Data Analysis — In Practice, 4-th Ed.CAMO, 2000.
[33] G. Dantzig, Linear Programming and Extensions, Princeton University Press, Princeton, N. J., 1963.
[34] S. Gass, Linear Programming, (4-th ed.) McGow-Hill, New York, 1975.
[35] J.-H. Jiang, Y.-Z. Liang, Y. Ozaki, On simplex-based method for self-modeling curve resolution of two-way data, Chemom. Intell. Lab. Syst. 65 (2003) 51–65.
[36] E. Gumbel, Statistics of Extremes, Columbia University Press, N.Y., 1962
[37] Software implementation of SIC method for MATLAB, http://rcs.chph.ras.ru/sic (1 Aug 2008).