ISSN 1061-9348, Journal of Analytical Chemistry, 2006, Vol. 61, No. 10, pp. 952–966. © Pleiades Publishing, Inc., 2006.
Original Russian Text © A.L. Pomerantsev, O.Ye. Rodionova, 2006, published in Zhurnal Analiticheskoi Khimii, 2006, Vol. 61, No. 10, pp. 1032–1047.

Construction of a Multivariate Calibration by the Simple Interval Calculation Method

A. L. Pomerantsev and O. Ye. Rodionova

Semenov Institute of Chemical Physics, Russian Academy of Sciences, ul. Kosygina 4, Moscow, 119991 Russia

Received May 26, 2005

Abstract—Simple interval calculation is a method for linear modeling and for constructing interval estimates of predictions in a multivariate calibration. Simple interval calculation gives results in a convenient interval form with regard to all the existing uncertainties: measurement errors of predictors and responses, discrepancies of bilinear modeling, and so on. In addition, the simple interval calculation method opens new opportunities for constructing an informative classification of object significance. The method is based on only the assumption of the error boundedness; it uses linear programming algorithms for data analysis. This approach differs substantially from the conventional regression methods used in chemometrics and, therefore, is poorly understood by analysts. This paper gives an elementary explanation of the simple interval calculation method illustrated by the simplest model and real examples.

DOI: 10.1134/S1061934806100030

Experimental data description, model construction, and the prediction of new values, which are referred to collectively as calibration, are among the oldest but still urgent problems exploited extensively in analytical chemistry [1]. The problem of multivariate calibration is essentially the following. Let there be some experimental data represented by two matrices, a matrix $\mathbf{X}$ of analytical signals (such as spectra) and a matrix $\mathbf{Y}$ of the corresponding chemical values (such as concentrations). The number of rows in these matrices is equal to the number of the reference objects under study; the number of columns in the $\mathbf{X}$ matrix corresponds to the number of channels (wavelengths) at which the signal is recorded; and, finally, the number of columns in the $\mathbf{Y}$ matrix equals the number of the chemical values, or responses. Based on the set of reference objects $\{\mathbf{X}, \mathbf{Y}\}$, it is required that a mathematical model $\mathbf{Y} = \mathbf{X}\boldsymbol{a}$ be constructed, by which new responses $\boldsymbol{y}$ can be predicted using a given new row of analytical signals $\boldsymbol{x}$. Evaluation of this model is a complicated ill-posed mathematical problem [2]. However, the multivariate model gives a substantial gain in accuracy as compared to simple calibration by several “characteristic channels” [3].

Since Gauss (1794), the regression approach has been used for the analysis of experimental data. The approach is based on the minimization of deviations of the calculated model values $\hat{\boldsymbol{y}}$ from their corresponding experimental values $\boldsymbol{y}$, or the least squares method [4]. Extensions of this approach, such as the method of principal components (1901) [5], method of maximum likelihood (1912) [6], ridge regression (1963) [2], projection to latent structures (1975) [7], etc., made it possible to use it for complicated ill-posed problems, for example, in spectroscopy, where the number of unknown parameters (wavelengths) is much greater than the number of objects under study [8]. However, all these methods give the prediction result as a point estimate, whereas interval estimate taking into account the uncertainty of prediction is often needed in practice. The confidence intervals cannot be constructed by conventional statistical methods, because the problem is too complex [9], and simulation methods can hardly be used, because calculations take too much time [10].

In 1962, Kantorovich [11] proposed a different approach to data analysis: instead of minimizing the sum of squared deviations, simultaneous inequalities should be used, which can be solved by linear programming methods. In this case, the prediction result is immediately obtained as an interval; therefore, this method was referred to as *simple interval calculation*. In its time, this concept did not obtain proper recognition and development, which was probably because of insufficient computer speed. In the 1980s and 1990s, several interesting applied studies [12–19] were performed using this method, including those in the field of analytical chemistry [18]. These studies have been summarized in monograph [20], where the main problem solved by the authors of the above papers has been considered in detail. This is the problem of interval estimation of the model *parameters* and embedding the domain of these parameters in a hypercube, parallelepiped, ellipsoid, etc.

This statement of the problem seems unproductive and showing little promise, which was supported by practice: no new papers in this field were observed in the recent decade. At the same time, we believe that the concept of Kantorovich can give some interesting results if multivariate calibration is considered as a problem of making interval prediction of *response y*. In this case, two equally important practical problems can be solved. First, one can find the range of uncertainty [21] for predicting the required response (chemical value), that is, to evaluate the *accuracy* of the calibration constructed individually for each object. Second, using the simple interval calculation approach, one can construct object *classification* [22], that is, establish individual peculiarities of each object governed by its relation both to the model and to other reference objects. Such concepts as *outlier* (an object notably standing out of the general regularity) or *extreme object* (an object lying in the peripheral region of the model and having a significant effect on its construction) are well-known examples of this classification. In spite of the extensive use of these concepts in various works [23–31], there are neither generally recognized definitions nor methods for their detection. The simple interval calculation method can fill this gap.

Theoretical aspects of the simple interval calculation method are published in [22, 32], and the results of its practical application are discussed in papers [21, 33, 34]. However, this method differs substantially from the conventional regression approach used for multivariate calibration problems. Its philosophy, mathematics, and vocabulary are unusual for analysts. Hence, we propose an elementary explanation of the simple interval calculation method based on the primitive uni- and bivariate examples, with which the most important concepts and results can be explained and demonstrated.

In the first part of the paper, we will give the reasons justifying the basic postulate of the simple interval calculation method, namely, error boundedness. We will give both theoretical and practical arguments supporting this postulate. The second part of this paper will be devoted to the detailed consideration of a primitive model example, in which all the necessary calculations can be performed by a pencil-and-paper method. By this example, basic concepts used in the simple interval calculation method will be introduced and illustrated, and it will be shown what conclusions follow from consistent application of the error boundedness principle. In the third part of this paper, a real example of a classical multivariate calibration problem will be discussed, namely, the prediction of gasoline octane number from IR spectra [35, 36]. The appendix gives a formal and mathematically rigorous description of the simple interval calculation method.

WHY THE ERRORS ARE BOUNDED

The main assumption of the simple interval calculation is the boundedness of the measurement error. This approach to the experimental data interpretation needs some substantiation. In data analysis, the principle of normal error distribution is conventionally assumed either explicitly or implicitly. However, the assumption of the normal error distribution has been repeatedly subjected to criticism from different points of view. Some papers, such as [37, 38], demonstrated that measurement error is usually bounded rather than normally distributed. It is significant that most analysts do not associate error unboundedness with the normal distribution principle. When being asked how often a researcher has to deal with data including values lying beyond four standard deviations (4σ), as a rule, the answer is that if these values ever occur, they are unconditionally removed as early as in the preprocessing (data reduction) stage. At the same time, the amount of data analysts deal with at present often exceeds 10^{+6} [39]. Hence, among them one can expect about 20–30 "normal" values lying beyond this threshold. The opinion of the authors of [30] is noteworthy. They state, "Indeed, in real case studies, the chemist is often able to select, to some degree, the samples, and this will lead to a more uniform distribution than normal distribution."

Let us consider a typical example supporting this point of view. Determining grain quality by NIR (near infrared) spectra is a classical problem of multivariate calibration [40]. In the example considered, the measurements were performed using an InfraLUM FT-10 spectrometer at 8000–14000 cm^{-1}, and water content in grain is the analytical value to be predicted. The spectral data **X** were prepared according to the procedure involving (1) averaging the spectra over three repeated measurements, (2) taking the logarithm, (3) smoothing the spectra according to the second-order three-point Savitzky–Golay algorithm [41], (4) normalization of each spectrum along the spectral lines, and (5) centering and normalization of all the spectra over the samples. Data ***y*** were also averaged over three repeated measurements, centered, and normalized. A multivariate calibration model was constructed using 141 samples at 9000–11000 cm^{-1} by projection to latent structures (see Methods). To construct multivariate calibration, four PLS principal components are sufficient, which explain 99 and 90% of **X** and ***y*** variance, respectively.

Figure 1a shows the distribution histogram of water content in grain, and Fig. 1b shows the PLS score plot in PC4 vs. PC3 coordinates. Applying conventional statistical analysis to this data, one can notice that they do not conflict with the hypothesis of normal distribution of responses. Even three extreme samples marked in the plots seem "tolerable"; their occurrence probabilities are 0.03, 0.21, and 0.38. Nevertheless, acting according to conventional multivariate calibration procedure, we removed all the samples marked in Fig. 1b by filled points as the outliers and performed a new model calibration. The results of processing the censored data (124 samples) are shown in Fig. 2.

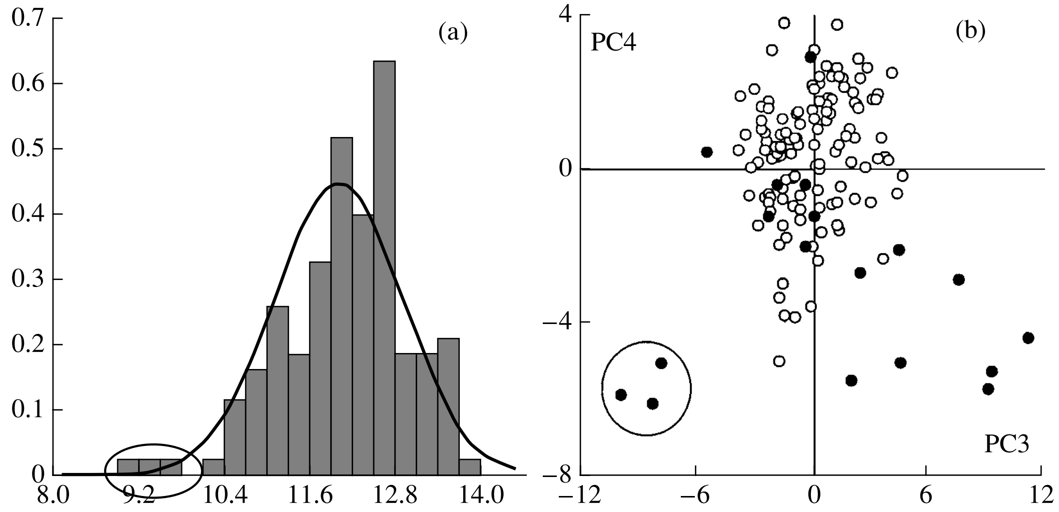

Fig. 1. Multivariate calibration of water content in grain from NIR spectra using an initial set of 141 samples. (a) Distribution histogram of water content in grain and (b) projection to latent structures scores plotted as PC4 vs. PC3. Filled points denote "suspicious" samples.

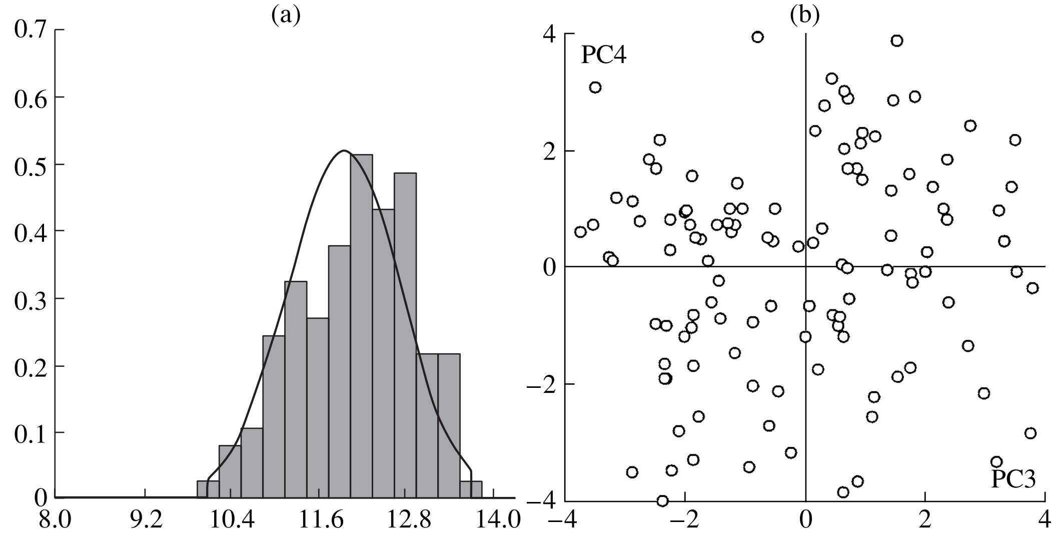

Fig. 2. Multivariate calibration of water content in grain from NIR spectra using a censored set of 124 samples. (a) Distribution histogram of water content in grain and (b) projection to latent structures scores plotted as PC4 vs. PC3.

Now, the PLS model explains 99 and 92% of **X** and ***y*** variance, respectively. The score plot (Fig. 2b) exhibits no suspicious samples. However, in this case, the response distribution corresponds to a *truncated* normal distribution cut off at $\pm 2.5\sigma$ from the center.

The example considered demonstrates that, following conventional procedures for data analysis and processing, we obtain bounded errors that obey a truncated normal distribution rather than a normal one. In the following section, we will see what can be inferred from the error boundedness considered further as a postulate.

EXPLANATION OF THE SIMPLE INTERVAL CALCULATION METHOD: A UNIVARIATE EXAMPLE

Model example. Let us explain how the simple interval calculation method works using the simplest univariate regression:

$$y = xa + \varepsilon. \tag{1}$$

The appendix gives a more rigorous mathematical description, from which only a few simple formulas will be used here.

Table 1. Model data and results of data processing

Objects	x	y	$\hat{y}$	$\hat{y}^-$	$\hat{y}^+$	$a^{\min}$	$a^{\max}$	v^-	v^+	h	r	$\lvert r\rvert + h$
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

The postulate of boundedness of the measurement error ε is the main assumption of the simple interval calculation method. This postulate can be formulated as follows. No error ε can exceed some constant β in absolute value, that is,

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

Let us consider the elementary conclusions following immediately from this postulate. Table 1 (columns 1 and 2) and Fig. 3 show some model data we constructed for regression (1) for $a = 1$. Errors in the response y were simulated by a uniform distribution with a width of 1.4, that is, $\beta = 0.7$.

In this example, a very short set of data (reference objects) is used, which is divided into two parts. The first four objects denoted as C1–C4 make up the training set used for model construction. They are represented in Fig. 3 by open circles. The last three objects denoted as T1–T3 are test objects, for which the prediction is constructed. They are represented in Fig. 3 by filled squares. In spite of the extra simplicity of this example, it helps us to explain the most important properties of the simple interval calculation method.

Least-squares calibration. Let us begin with the conventional least-squares method [6]. Using training data (x_i, y_i), $i = 1–4$ (columns 1 and 2 in Table 1, objects C1–C4), one can find the least-squares estimate of the parameter a,

$$\hat{a} = \frac{\bar{y}}{\bar{x}} = 1.003, \quad \bar{x} = \frac{1}{4}\sum_{1}^{4} x_i, \quad \bar{y} = \frac{1}{4}\sum_{1}^{4} y_i, \tag{3}$$

and predict the response y at all points x, both training and new ones:

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

(column 3 in Table 1 and Fig. 3a, thick line). We can also estimate the error ε by using the well-known formula [4]

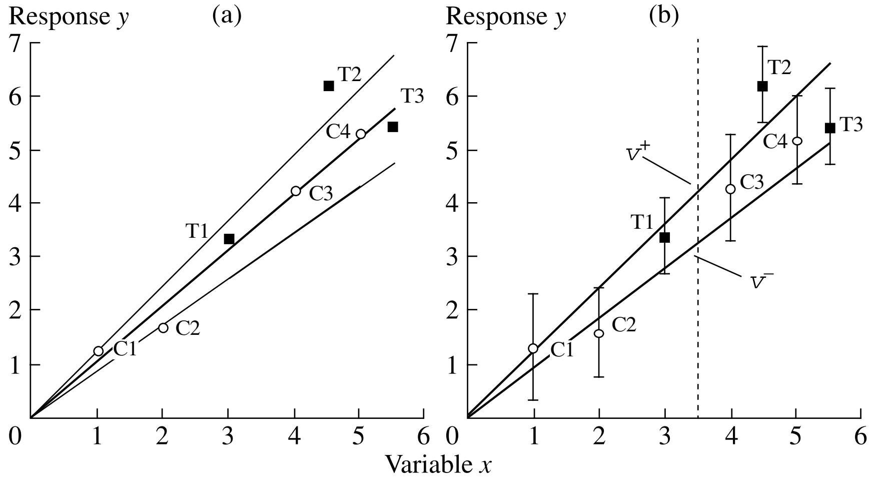

Fig. 3. Univariate model example. Open circles are training objects; filled squares are test objects. (a) Least-squares method: thick lines are least-squares predictions; thin lines are boundaries of the confidence intervals. (b) Simple interval calculation method: bars are error intervals; lines are boundaries of the prediction intervals.

$$s^2 = \frac{1}{3}\sum_{1}^{4}(y_i - \hat{y}_i)^2 = 0.078 \qquad (5)$$

and construct confidence intervals for the response

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

Here, $t_3(P)$ is the inverse Student's distribution with three degrees of freedom for the probability P. Confidence limits for $P = 0.95$ are given in columns 4 and 5, Table 1, and in Fig. 3a (thin lines).

Calibration by simple interval calculation. Now, let us consider how these data are interpreted by the simple interval calculation method. Suppose that we know that $\beta = 0.7$. In most practical applications, the situation is much more difficult, and β is not known a priori. Later, we will see how this problem can be solved.

From the regression equation (Eq. (1)) and the error boundedness principle (Eq. (2)), it follows that, for each object from the training set (x_i, y_i), i = 1–4, the following holds:

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

or, equivalently,

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

where

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

The values (9) are given in the sixth and seventh columns (Table 1). Inequalities (8) should hold simultaneously for all the training objects, that is, for i = 1, 2, 3, and 4. It is clear that this can be the case only for the values of the parameter a lying within the interval

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

where

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

These values are set in bold type in the corresponding columns of Table 1.

Interval (10) defines the *region* of the parameter a, that is, the values that agree with the experimental data. It is clear that, when the parameter a varies within the interval (10), the corresponding response $y = ax$ at an arbitrary point x is bounded by the values

$$v^{-} \le y \le v^{+}, \qquad (12)$$

where

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

These values are given in columns 8 and 9 (Table 1).

Hence, an interval estimate of the parameter a (10) is constructed. This estimate is analogous to the point estimate $\hat{a}$ obtained using the least squares method. In addition, the prediction intervals (13) for the response y are also found. These intervals are valid for either training or any other (new) objects.

Let us consider a graphical interpretation of the simple interval calculation method. Figure 3b shows the same data as Fig. 3a, but now each point is accompanied by the error interval (vertical bars) of half-width $\beta = 0.7$. When constructing simple interval estimates, one should consider all the possible straight lines passing through the origin so that each of them "touch" all the error intervals for all the four training objects. One can see from the plot that the lower boundary is the line passing through the lower point of the interval for the object C4. The upper boundary is the straight line passing through the upper point of the interval for the object C2. All the lines contained by these boundaries will obviously meet conditions (7) and, vice versa, any line lying off this angle will conflict with these conditions. The boundaries are represented in Fig. 3b by two thick lines v^+ and v^-.

Let us mention the obvious fact that, in our example, the construction of the calibration by the simple interval calculation method is based only on two objects, C2 and C4. They govern the boundaries (10) of the region of the parameter a, so that we can call these objects *boundary.* Other training objects C1 and C3 are inessential. We can remove them from the training set, and the result remains the same. This is a very important property of the simple interval calculation method; it finds an application in choosing a representative object set [22].

Thus, we have shown that all the objects from the training set in the simple interval calculation method can be divided into two groups: the most important boundary objects that form the basis for the model and inessential insiders that can be removed from the training set without changing the model.

Object status. What can happen to the simple interval calculation model or, more precisely, to its region of possible values when a new object is added to the training set? It is clear that the RPV can only shrink. Thus, when the object T3 is added, the upper boundary (v^+ line) goes below the old one so as to "touch" the upper boundary of the error interval for T3. In this case, $a^{\max}$ becomes equal to 1.11 instead of 1.16 (Table 1). This property of the simple interval calculation method, called *consistency* (see (24) in the appendix), is important from the theoretical point of view. It shows that, as the number of objects in the training set increases, the uncertainty of the simple interval estimates decreases. In this case, if the maximum error is chosen properly, or, at least, it is no less than β, the true value of the parameter a is always within the possible values (10). This property called *unbiasedness* (see (21) in the

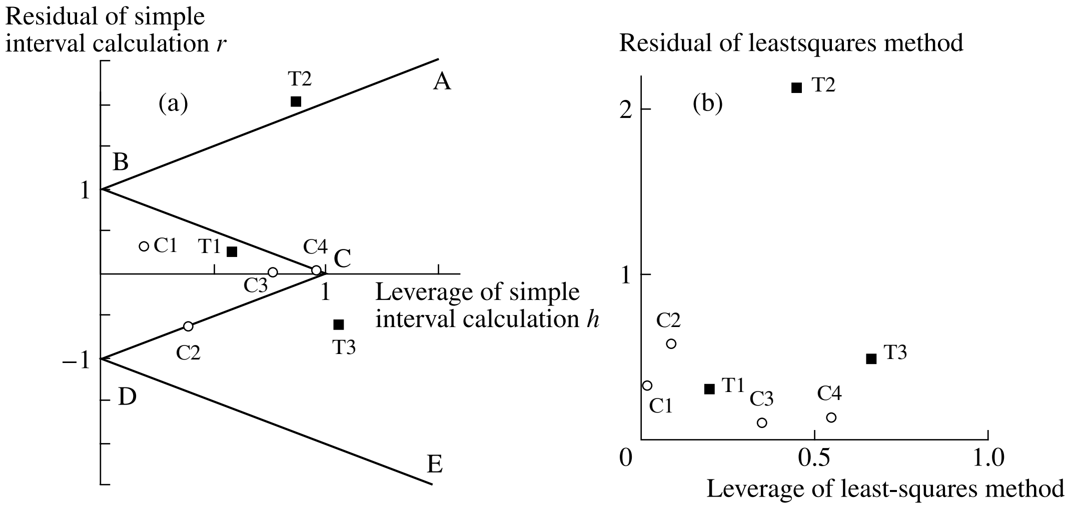

Fig. 4. Determining object status in the univariate model example. Open circles are training objects; filled squares are test objects. (a) Object status plot obtained by simple interval calculation. (b) Object influence plot obtained by the least-squares method.

appendix) is also important for understanding and substantiation of the simple interval calculation method.

However, not any new object included in the training set results in the model refining. Thus, the object T1 does not change the simple interval calculation model. This can be seen from Fig. 3b, where the prediction interval lies entirely within the corresponding error interval of T1, as well as from the sixth and seventh columns of Table 1. Another example is the object T2. Its error interval is disjoint with the prediction interval; therefore, when T2 is added to the training set, the model is destroyed, because simultaneous inequalities (7) become inconsistent. This can also be seen from Table 1: the minimum over column 7 (1.11) becomes less than the maximum over column 6 (1.22). Hence, one can classify new objects into three groups with respect to how they affect the model when added to the training set, that is, to determine the status (influence) of each object. First, one can recognize a class of *insiders* that do not change the model and a class of *outsiders* that do change the model. In addition, in the outsiders, a group of *outliers* can be recognized, that is, objects that cannot be added to the training set (at given β) because they destroy the model.

Object status classification based on simple interval calculation. Studying the object status using the plot of Y vs. X is inconvenient, and it becomes impossible in the multivariate case. To perform this analysis in the general case, two variables reflecting object properties are introduced: residual of SIC r and leverage of SIC h (Eqs. (27) and (28) in the appendix). Having prediction intervals (12), one can easily calculate these values. In our example, r and h are given in columns 10 and 11 (Table 1) and shown in the *object status plot* (OSP) in Fig. 4a in the coordinates (h, r).

In this diagram, training objects are denoted by open circles, and test objects are denoted by filled squares as in Fig. 3. The thick line ABCDE bounds the regions with different object statuses. The shape of this line is governed by the two fundamental inequalities ((29) and (31) in the appendix) relating h and r. One can see from the OSP that all the objects from the training set lie within the BCD triangle (their status is insider; for them, $|r| + h \leq 1$; see column 12 of Table 1) with the objects C2 and C4 lying on its boundaries (their status is boundary; for them, $|r| + h = 1$). The test object T1 also falls in the insider triangle, because $|r| + h = 0.83 < 1$. The object T2 lies below line DE, which indicates that it is an outlier. For it, $|r| - h = 2.91 > 1$ (see inequality (31) in the appendix). The object T3 is an outsider, and one can see from the OSP that at no residue r can it fall within the insider region. For it, $h = 1.05 > 1$. This indicates that its variable x contains some new important information lacking in the calibration model. These objects are called *absolute outsiders*.

Hence, it is shown that the simple interval calculation method makes it possible to introduce a new classification of all the objects of multivariate calibration from the training set as well as the test and new objects. This classification is called *object status classification* [22]. It is based on definitions (27) and (28) and on statements (29)–(32) in the appendix. Practical application of the object status classification consists in calculating r and h, constructing the corresponding plot status diagram, and studying the location of objects in this plot.

One can notice that the triangular shape of the insider region in the object status plot (Fig. 4a) resembles an ordinary influence plot [7] shown in Fig. 4b. In this plot, all the objects of the model example are represented by their least-squares leverage

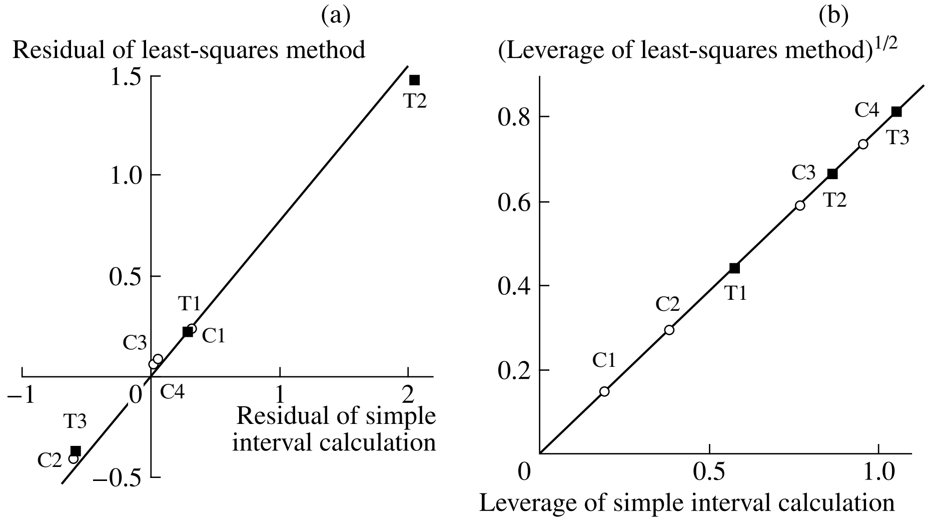

Fig. 5. Object characteristics in the univariate model example. Open circles are training objects; filled squares are test objects. (a) Comparing residues obtained by the least-squares method and by simple interval calculation. (b) Comparing leverages obtained by the least-squares method and by simple interval calculation.

$$h_{\mathrm{ls}} = \mathbf{x}^{\mathrm{t}}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{x} = x_i^2/\sum_{i=1}^{4} x_i^2$$

against their normalized least-squares residual.

$$r_{\mathrm{ls}} = (y - \hat{y})/\beta.$$

The resemblance between the status and influence plots follows from the well-known statistical relationship [8], which relates the *accuracy* (root-mean-square error of calibration, RMSEC), *precision* (standard deviation of error of calibration, SEC) and *bias* (systematic error, BIAS),

$$\mathrm{RMSEC}^2 \approx \mathrm{SEC}^2 + \mathrm{BIAS}^2. \qquad (14)$$

For the simple interval calculation method, in which β is the accuracy, the simple interval calculation leverage h characterizes the normalized precision, and simple interval calculation residual r is responsible for the normalized bias, Eq. (14) can be represented as

$$\beta^2 = \beta^2 h^2(\boldsymbol{x}) + \beta^2 r^2(\boldsymbol{x}, y), \qquad (15)$$

which coincides with Eq. (29) of the appendix. On the other hand, we should acknowledge a substantial difference between Eqs. (14) and (15). It consists in that the latter equation holds for each object, whereas Eq. (14) makes sense only for the entire collection of objects, that is, on average.

In the end of this section, let us show two plots (Figs. 5a and 5b) indicating a close relationship between the object characteristics in the least-squares and simple interval calculation methods. From Fig. 5a, one can see a strong correlation ($R^2 = 0.999$) between the least-squares and simple interval calculation residuals. The relationship between the leverages is more complex (Fig. 5b), but the correlation between the square root of the least-squares leverage and the simple interval calculation leverage is also observed here ($R^2 = 1.000$). This relationship obviously follows from the definitions of these values. The leverage in the simple interval calculation method is proportional to the prediction interval width (12), whereas the leverage in the least-squares method is proportional to the prediction variance [42], which governs the confidence interval width proportional to the square root of variance. It is clear that, in more complex problems, the relationship between the least-squares and simple interval calculation characteristics is not so simple, but the principal trend remains the same. The problem of similarity and difference between the least-squares and simple interval calculation methods and comparison of the simple interval calculation prediction intervals and confidence least-squares intervals are considered in detail in [21].

Estimation of β. Constructing the estimate b of parameter β, or the maximum error, is a rather complex statistical procedure, which is briefly outlined in the appendix and considered in detail in [32]. To understand the matter, one should keep in mind only the fact that, depending on the nature of the distribution of error ε, the estimate b constructed lies, as a rule, within 2σ–4σ, where σ is the standard deviation of this distribution. It is clear that, for any truncated distribution, β cannot be less than 2σ. The extreme case is the uniform distribution, for which $\beta = 1.71\sigma$. For a usual number of objects in the experiment (say, less than 1000), we cannot expect extreme values beyond 3σ. Finally, the limit of 4σ makes sure that new objects will never go beyond this boundary as well. In the example considered, $\beta = 2.5\sigma$, which is somewhat higher than it should be for the uniform distribution, which we used for data simulation. This result is completely explainable, because the estimate of parameter β was calculated using a very small training set (only four objects).

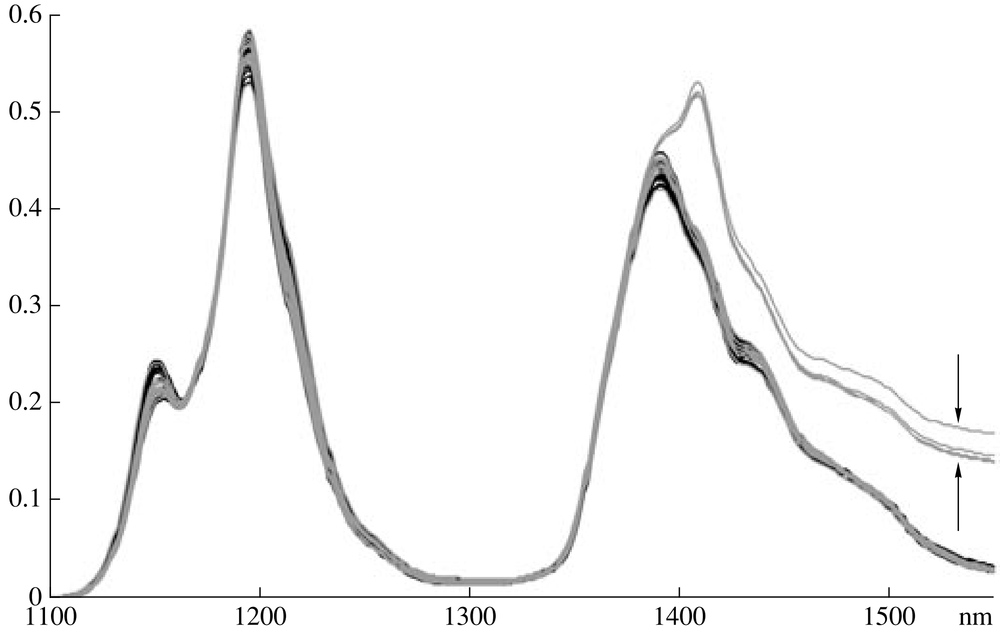

Fig. 6. Absorption spectra of gasolines. Arrows denote ethylated samples.

It is reasonable to ask how these variations in the estimate of the maximum error β can affect the results of the object status classification. Because such object characteristics as the simple interval calculation residual *r* and simple interval calculation leverage *h* are defined in the simple interval calculation method as relative values divided by β (see (27) and (28) in the appendix), overestimated maximum error β does not affect the results of the object status classification. The contrary is the case for the prediction intervals in the simple interval calculation method. They increase with the estimate *b*, and at $b = \beta$ the intervals, by definition, cover the true value with a unitary probability. On the other hand, it was shown in [32] that the prediction intervals constructed for the estimate b_{SIC} ((37) for $P = 0.90$) instead of the true β have a probability of covering no less than 0.9999. This result supports the conclusion that not only the proposed object status classification but the entire simple interval calculation method as well can be used in practice.

To illustrate this statement, let us compare the least squares and simple interval estimates for our example. When comparing the plots in Figs. 3a and 3b, one can see that the confidence interval ($P = 0.95$) for the least-squares prediction is wider than the prediction range of the simple interval calculation method. If one calculates the covering probability of SIC interval using formula (6), it is equal to 0.91. Note that we have taken rather high $\beta = 2.5\sigma$, which corresponds to the normal probability Prob[–2.5, +2.5] = 0.99.

Hence, the considered primitive model example shows two important facts. First, the use of the unbounded (normal) distribution for constructing confidence estimates results in unnecessarily wide intervals. Second, even for a small amount of data, the simple interval calculation method gives reasonably wide intervals that represent the facts well. To support the last statement, we have repeated the simulation of our example 100000 times. Never was the true $y = x$ beyond the predicted intervals obtained by the simple interval calculation.

A REAL EXAMPLE: A MULTIVARIATE MODEL

Data. We used the well-known didactic example of predicting the octane number of gasoline [35, 36] to demonstrate that the simple interval calculation method can successfully be applied to real problems of multivariate calibration, including situations in multicollinearity conditions. In this example, the matrix **X** consists of NIR absorbance spectra obtained for 226 wavelengths at 1100–1550 nm. They are shown in Fig. 6.

Components of the vector ***y*** are the results of the corresponding laboratory measurements of octane numbers [46]. The data are divided into two sets; in each, the octane number varies within 87–93. The first training set consists of 24 samples of commercial gasolines (nos. 1–24); it is used for constructing the multivariate calibration. The second set includes 13 samples that serve for validating the predictive properties of the model. It is important that this test set contains four samples (nos. 10–13) that represent ethylated gasolines lacking in the training set. Let us call the test sets with these samples (nos. 1–13) and without them (nos. 1–9) long and short test sets, respectively.

Methods. When constructing the multivariate calibration, the most important mathematical problem consists in inverting the matrix $\mathbf{X}^t\mathbf{X}$, which, in our case, has the dimensions 226 × 226. If this matrix were nondegenerate (had full rank [6]), one might invert this matrix and find the estimates of the unknown model parameters $\hat{\boldsymbol{a}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{y}$ for the calibration by the least squares method. However, in our example, as with most practical problems, this matrix is degenerate. Therefore, according to property (22) in the appendix, the simple interval calculation method cannot be used. To overcome this difficulty, various data regularization methods are used, such as the principal component method, ridge regression, and so on. We used the *projection to latent structures* method [35]. The method consists in simultaneous decomposition of the matrix $\mathbf{X}$ and the vector $\boldsymbol{y}$ in the form

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E}; \quad \boldsymbol{y} = \mathbf{T}\boldsymbol{q} + \boldsymbol{f}, \tag{16}$$

where $\mathbf{T}$ is the matrix of *scores*, $\mathbf{P}$ and $\boldsymbol{q}$ are the matrix and vector of *loadings*, and $\mathbf{E}$ and $\boldsymbol{f}$ are the matrix and vector of *residuals*. To construct this decomposition, three rules are used. First, columns $\boldsymbol{t}_i$ of the matrix $\mathbf{T}$ are linear combinations of columns $\boldsymbol{x}$ of the matrix $\mathbf{X}$; that is, $\boldsymbol{t}_i = \mathbf{X}\boldsymbol{w}_i$. Second, coefficients w are chosen so as to maximize the correlation between the response $\boldsymbol{y}$ and vector $\boldsymbol{t}_i$. Third, the number of columns in the matrix of scores $\mathbf{T}$ and the matrix of loadings $\mathbf{P}$ equals the effective (chemical) rank of the matrix $\mathbf{X}$. This value k is called the number of *principal components* (PCs). Of course, it is less than the number of columns in the matrix $\mathbf{X}$. The possibility of visual interpretation of data in the score plots is an important advantage of the PLS method, which was used in Section 1 (see Figs. 1 and 2). When predicting a new (test) object $\boldsymbol{x}$, it is projected to the vector of scores $\boldsymbol{t}$, which is later subjected to regression (16).

Conventionally, a multivariate calibration problem is set in the homogeneous form $\boldsymbol{y} = \mathbf{X}\boldsymbol{a}$, so that $\boldsymbol{y} = 0$ at $\boldsymbol{x} = 0$. To match the raw data $(\mathbf{X}_{\text{raw}}, \boldsymbol{y}_{\text{raw}})$ with this model, they are *centered*:

$$\boldsymbol{y} = \boldsymbol{y}_{\text{raw}} - m_0\boldsymbol{1}, \quad \mathbf{X} = \mathbf{X}_{\text{raw}} - (m_1\boldsymbol{1}, m_2\boldsymbol{1}, \dots, m_p\boldsymbol{1}).$$

Here, m_0 is the average of the vector of responses y and m_i is the average calculated for all the columns of the matrix $\mathbf{X}_{\text{raw}}$. In addition, it is often necessary to *normalize* data as well. This is done to average the contributions from different variables. If the data are not normalized, the result can depend on some variables that exhibit large variance but small regression significance. To normalize data $(\mathbf{X}_{\text{raw}}, \boldsymbol{y}_{\text{raw}})$ is to multiply them by the diagonal matrices $\mathbf{X} = \mathbf{X}_{\text{raw}}\mathbf{S}_X$ and $\boldsymbol{y} = \boldsymbol{y}_{\text{raw}}\mathbf{S}_y$. Diagonal elements of matrices $\mathbf{S}$ are usually chosen equal to inverse standard deviations s_{ii} calculated for the corresponding columns $\mathbf{X}_{\text{raw}}$ and $\boldsymbol{y}_{\text{raw}}$, that is, $\mathbf{S}_{ii} = (s_{ii})^{-1}$.

One can learn more about the method of projection to latent structures from numerous monographs, such as [7, 8], which are, unfortunately, hardly available in Russia. Recently, this method was presented correctly but briefly in manual [1]. In Russian, detailed presentations of projection to latent structures as well as other methods of multivariate data analysis are given in book [35].

Using the s method of projection to latent structures in our example, one can project the initial multivariate calibration problem onto a two-dimensional subspace, where the new problem is nondegenerate:

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

Here, m_0 is the average $\boldsymbol{y}$ and $\mathbf{T}$ is the $n \times 2$-dimensional matrix of scores. This number of principal components ($k = 2$) explains 97% of $\mathbf{X}$ variance and 98% of $\boldsymbol{y}$ variance.

Calibration. To use the simple interval calculation method, one should determine the maximum error β as described in the appendix. Here, one should take into account that projection methods necessarily increase the overall error due to inaccuracy of modeling. This is because bilinear models (such as projection to latent structures and PCR)[1] are only approximations of complex systems. Therefore, the maximum error β is always greater than any individual measurement error of a response.

Using formula (35) from the appendix, we obtain $b_{\min} = 0.484$. This means that the simple interval calculation method with $b < 0.484$ cannot be used for the data under consideration, because the region of possible values becomes empty. The value of $b_{\min}$ gives a lower boundary of the maximum error β, but we need a corresponding upper estimate as well. Using Eq. (37) from the appendix at $P = 0.90$, we obtain the estimate $b_{\text{SIC}} = 0.880$. This b is used as an estimate of the maximum error β in all the calculations from here on. Having estimated the root-mean-square error of calibration (RMSEC)

$$s_{\text{cal}} = \sqrt{\frac{1}{n_{\text{cal}}}(\boldsymbol{y}_{\text{cal}} - \hat{\boldsymbol{y}}_{\text{cal}})^2} = 0.268,$$

one can compare the accuracy of modeling by the PLS and SIC methods: $b_{\min}/s_{\text{cal}} = 1.81$ and $b_{\text{SIC}}/s_{\text{cal}} = 3.28$.

To apply the simple interval calculation method, one does not need to construct the admitted region explicitly, especially because this is a very complicated problem for more than two variables [20]. However, in the example considered, $p = k = 2$; therefore, to illustrate and explain the simple interval calculation technique, we show the RPV in Fig. 7a.

Just as in the univariate example, the admitted region is generated only by the boundary samples rather than by all 24 samples from the training set. In

[1] PCR is principal component regression.

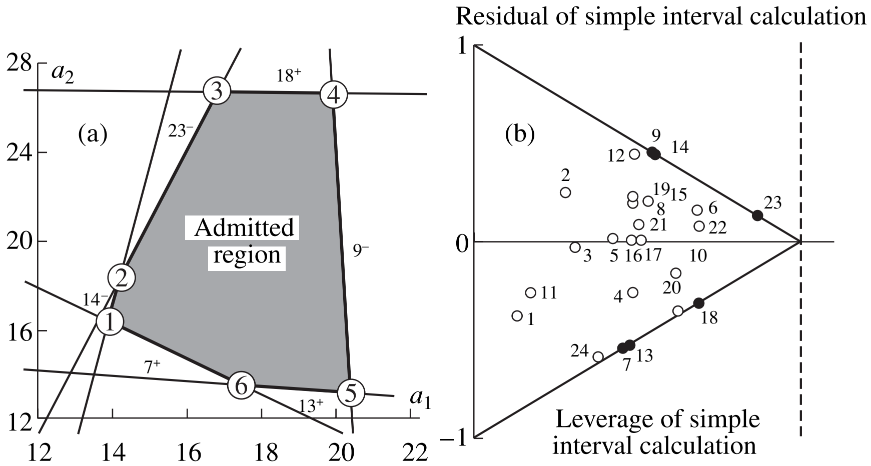

Fig. 7. Prediction of octane number. Projection to latent structures model with two principal components. Training set. (a) Admitted region: lines are boundaries; open circles are vertices. (b) Status plot of the training samples: open circles are insiders; filled circles are boundary objects.

our case, there are six boundary samples (nos. 7, 9, 13, 14, 18, 23). They are denoted by filled points in the status plot of the training samples (Fig. 7b). All the boundary samples lie on the boundary triangle, because $\mid |r| + h = 1$ for them. These samples generate the RPV as shown in Fig. 7a, where each line corresponds to either the equation $\boldsymbol{t}_i^t\boldsymbol{a} = y_i - m_0 + b$ (denoted by "+"), or the equation $\boldsymbol{t}_i^t\boldsymbol{a} = y_i - m_0 - b$ (denoted by "–"). Numbers near the lines correspond to the numbers of the boundary samples, and the RPV is outlined by the thick broken line.

Hence, the calibration by simple interval calculation results in the estimate b of the maximum error β and a set of boundary samples constituting the admitted region.

Prediction. Let us calculate the prediction intervals by the SIC method. To do this, v^- and v^+ values should be calculated for each test object (Eq. (26) in the appendix). These values govern the boundaries of individual prediction intervals. This optimization problem is solved using the *linear programming* technique, which makes it possible to obtain the solution in the general case without explicitly presenting the region of possible values.

The linear programming [43] problem consists in maximization or minimization of a linear function under linear constraints. In the general case, this problem can be presented in so-called canonical form as

$$\min_{a}\{\boldsymbol{c}^{\mathrm{t}}\boldsymbol{a} \text{ a provided that } \mathbf{T}\boldsymbol{a} = \boldsymbol{d} \text{ and } \boldsymbol{a} \geq 0\},$$

where $\boldsymbol{a} \in R^p$ is a vector of unknown parameters and $\boldsymbol{c} \in R^p$ are coefficients of the target function. The matrix $\mathbf{T} \in R^{n \times p}$ is called the constraint matrix, and the vector $\boldsymbol{d} \in R^n$ is called the constraint vector. Any system of linear inequalities can be transformed into the canonical form by introducing additional variables. Slack variables are added to the system to exclude "less than"-type constraints, and surplus variables are added to exclude "greater than"-type conditions. In addition, any maximization problem can be transformed into the minimization problem by changing the signs of the coefficients of the target function [45].

The canonical problem of linear programming can be solved by a simplex method, which consists in the linear search for the vertex points in such a way that the value of the target function $\boldsymbol{c}^t\boldsymbol{a}$ decreases from iteration to iteration. As a result, the solution, point $\boldsymbol{a}$ that is simultaneously admissible (meeting all the constraints) and optimal (giving minimum value), is found. The simplex method [43–45] is a well-known algorithm included into many software packages, such as [46]. It calculates the admissible vertices algebraically by using corresponding simultaneous linear equations rather than explicitly constructing the polyhedron.

In our example, the region of possible values generated by the linear constraints is a polygon with six vertices, numbered for clarity's sake (see Fig. 7a). To illustrate the essence of the simplex method, let us find the prediction interval for the first sample from the test set. We denote it by the index *test*. Using the conventional PLS algorithm, one can find the projection of this 226-dimensional vector $\boldsymbol{x}_{\text{test}}$ on the principal component plane, that is, calculate the score vector $\boldsymbol{t}_{\text{test}} = (-0.0689, 0.0343)$. To determine the boundary of the prediction interval $[v^-, v^+]$, one should solve two linear programming problems:

$$v^- = \min_{\boldsymbol{a}} \boldsymbol{t}_{\text{test}}^{\mathrm{t}}\boldsymbol{a}, \quad v^+ = \max_{\boldsymbol{a}} \boldsymbol{t}_{\text{test}}^{\mathrm{t}}\boldsymbol{a},$$

where the vector of parameters $\boldsymbol{a}$ meets the constraints

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

Table 2. Constructing the prediction interval by simple interval calculation method

Vertex	a_1	a_2	$t^t_{\text{test}}a$	y_{test}
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	88.52

that is, $\boldsymbol{a}$ lies within the RPV shown in Fig. 7b. After solving these problems, one obtains the interval prediction for the desired response $y_{\text{test}} = m_0 + \boldsymbol{t}^{\text{t}}_{\text{test}}\boldsymbol{a}$,

$$v^- + m_0 \le y_{\text{test}} \le v^+ + m_0.$$

Table 2 gives the coordinates of all the six vertices of the RPV (a_1 and a_2) and the corresponding responses. One can find from this table the prediction interval $88.30 < y_{\text{test}} < 89.01$ corresponding to vertices 5 (minimum) and 3 (maximum). In fact, while solving the complex problem, one does not need to act in such an inefficient way and check each vertex individually. It will suffice to use the conventional simplex algorithm. The first admissible vertex found by the simplex algorithm is vertex 1. To find the minimum (v^-), the algorithm travels $1 \longrightarrow 6 \longrightarrow 5$, and to find maximum ($v^+$), $1 \longrightarrow 2 \longrightarrow 3$ (Fig. 7a).

Results. Let us compare the prediction intervals for the test samples found by the simple interval calculation method with the estimates of $\hat{\boldsymbol{y}}_{\text{test}}$ obtained by projection to latent structures method. To do this, we will use a conventional estimate of the prediction accuracy for the PLS, namely, the root-mean-square error of prediction (RMSEP)

$$s_{\text{test}} = \sqrt{\frac{1}{n_{\text{test}}}(\boldsymbol{y}_{\text{test}} - \hat{\boldsymbol{y}}_{\text{test}})^2}.$$

The value of s_{test} found by a leave-one-out cross validation technique (LOO) [7] equals 0.322. If new samples belong to the same type (both qualitatively and quantitatively) as the training samples, one may expect approximately the same accuracy prediction for them. In the example considered, this is the case for the short test set. The corresponding uncertainty intervals $[\hat{\boldsymbol{y}}_{\text{test}} \pm 2s_{\text{test}}]$ are shown in Fig. 8a by black rectangles, and the intervals obtained by simple interval calculation are denoted by gray rectangles. Obviously, using this approach for outlying test samples (nos. 10–13) would be incorrect.

The SIC intervals for the four outlying samples are very large. In the conventional projection approach, they are considered as outliers [35]. These samples can also be easily identified in the object status plot (Fig. 8b). In using the object status diagram, samples nos. 10–13 are characterized as absolute outsiders, that is, the samples that bear absolutely no resemblance to the samples from the training set. Examining the plot in Fig. 8a, one can notice that the test values (open points) as well as the values predicted by the PLS (filled points) lie within the intervals obtained by SIC (gray rectangles) and that the uncertainty intervals found by the projection to latent structures (black intervals) agree with the intervals obtained by simple interval calculation for the short test set of samples (nos. 1–9). At the same time, the length of the interval obtained by simple interval calculation is individual for each new sample and, therefore, is more informative than the average

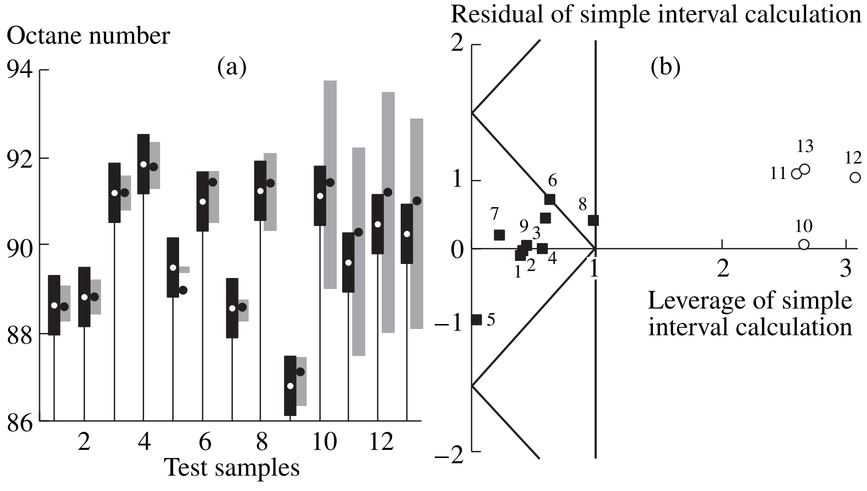

Fig. 8. Prediction of octane number. Projection to latent structures model with two principal components. Test data. (a) Filled circles are test values, gray area is the interval obtained by simple interval calculation, open circles are projection to latent structures estimates, and black area is the uncertainty interval. (b) Status plot of the test samples: filled squares are nos. 1–9; open circles are nos. 10–13.

uncertainty calculated by the projection to latent structures. As for the outlying samples (nos. 10–13), the intervals obtained by simple interval calculation immediately indicate their deviance. Note also that, for "normal" samples, the intervals obtained by simple interval calculation are less than the corresponding uncertainty intervals obtained by the projection to latent structures.

This example demonstrates how the simple interval calculation method can answer questions important for analysts.

1. Estimate of the maximum error β governs the accuracy of calibration and sets the boundary of precision for all objects that are similar to objects from the training set.

2. Prediction intervals obtained by simple interval calculation determine an individual prediction uncertainty of response for each new object.

3. The position of each object in the status plot makes it possible to determine whether this subject is similar to the objects from the training set and thus sets reasonable limits of applicability of the constructed calibration.

As applied to the considered example of the calibration of octane number from NIR spectra, it would be reasonable to state that, in everyday practice, this technique can be applied only to the new samples that are *insiders* with respect to the constructed multivariate calibration. Only in this case is the prediction accuracy surely no worse than the accuracy of calibration, which approximately equals the accuracy of the conventional laboratory technique for fuel rating. On the other hand, if this technique were intended for research purposes, it would be enough to say that it cannot be applied to *absolute outsiders*, because the structure of their predictors differs substantially from the standard training samples.

We believe that the proposed simple interval calculation method can be of practical use for multivariate data analysis. The simple interval calculation method has some advantages as compared to the conventional regression approach.

First, it does not use any assumptions concerning the error form but for its boundedness. Hence, the method can be considered as distribution-free.

Second, it gives the result in the convenient interval form taking into account the prediction uncertainty of the response.

Third, it naturally outlines the limits within which the constructed model can be used. This is achieved by the object status classification discriminating reliable "insiders," significant "boundary objects," suspicious "outsiders," outlying "absolute outsiders," and destructive "outliers."

In our view, the most important (and the only) assumption of the error boundedness is an advantage rather than a drawback of the method because, practically, it seems more sound than the conventional hypothesis of error normality.

The computational procedure of the method is based on known linear programming algorithms and can be easily implemented. Application of the simple interval calculation method to real problems gave results that agree well with the practical experience.

Software. For modeling by the projection to latent structures, we used "The Unscrambler" software package [52]. Simple interval calculation was performed using the software implemented as an add-in for Microsoft Excel. We used the NIPALS algorithm [35] for bilinear modeling, standard SIMPLEX algorithm [43] for optimization, and all the necessary set of special procedures for the preprocessing (data reduction), transformations, and so on. This program is being betatested at present.

APPENDIX

RIGOROUS DESCRIPTION OF THE SIMPLE INTERVAL CALCULATION METHOD

Region of possible values. Let us consider a linear multivariate calibration model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{a} + \boldsymbol{\varepsilon}, \tag{17}$$

where $\boldsymbol{y}$ is an n-dimensional vector of responses, $\boldsymbol{a}$ is a p-dimensional vector of parameters, $\mathbf{X}$ is a $(n \times p)$-dimensional matrix of predictors (independent variables), and $\boldsymbol{\varepsilon}$ is a vector of errors. Let us take error ε as bounded; that is, there is some $\beta > 0$, called maximum error, so that

$$\begin{gathered}\text{Prob}\{|\varepsilon| > \beta\} = 0, \\ \text{and that for any } \ 0 < b < \beta \quad \text{Prob}\{|\varepsilon| > b\} > 0,\end{gathered} \tag{18}$$

where Prob{·} denotes the probability of an event. The symmetry and homoscedasticity of error ε as well as the assumption of the absence of errors in the matrix **X** are not important and can be rejected hereinafter. First, assume that β is known.

Let us call a pair $(\boldsymbol{x}_i, y_i)$, $(i = 1, \ldots, n)$ a training object. Here, vector $\boldsymbol{x}_i^t$ is the ith row of the matrix **X** corresponding to the response y_i in Eq. (17). According to the condition (18), for each $i = 1, \ldots, n$, the following inequalities hold:

$$y_i^- \le \boldsymbol{x}_i^t\boldsymbol{a} \le y_i^+, \quad y_i^- = y_i - \beta, \quad y_i^+ = y_i + \beta. \tag{19}$$

It is reasonable that the true vector of parameters denoted below as $\boldsymbol{\alpha}$ is unknown. However, one can consider all vectors $\boldsymbol{a}$ that meet these inequalities. The values $\boldsymbol{a}$ that meet the condition (19) for a given object i form a strip $S(\boldsymbol{x}_i, y_i)$ in the parameter space R^p. The position and width of this strip is governed by the values $(\boldsymbol{x}_i, y_i)$. Let us consider all the objects from the training set and their corresponding strips. It is clear that the vector of parameters $\boldsymbol{a}$ meets all inequalities (19) simultaneously if and only if it belongs to all the strips.

The *region of possible values A* for parameters $\boldsymbol{a}$ of the system (17) is a set in the parameter space generated as a result of the intersection of all the strips:

$$A = \bigcap_{i=1}^{n} S(\boldsymbol{x}_i, y_i). \qquad (20)$$

Region A is a closed convex polyhedron [48, 49] generated by the boundaries of the intersecting strips. A is a random set, because it is constructed using random values $\boldsymbol{y}$.

Properties of the region of possible values. The region of possible values A exhibits the following properties for any model defined by Eq. (17).

Region A is an *unbiased* estimate of parameter $\boldsymbol{\alpha}$. It follows immediately from the RPV definition that the true value of $\boldsymbol{\alpha}$ always belongs to A:

$$\text{Prob}\{\boldsymbol{\alpha} \in A\} = 1. \qquad (21)$$

Region A is *bounded* if and only if [48, 49] the matrix $\mathbf{X}$ has full rank:

$$\text{rank}\,\mathbf{X} = p. \qquad (22)$$

It follows therefore that, if system (17) is multicollinear, some regularization procedure should be applied before using the simple interval calculation method. Thus, one can use the conventional approach [7, 32] and project the raw data onto a subspace of smaller dimension

$$y = \mathbf{TP}^t\boldsymbol{a} + \boldsymbol{f} = \mathbf{T}\boldsymbol{q} + \boldsymbol{f}, \qquad (23)$$

where the matrix of scores $\mathbf{T}$ has full rank $k < p$, and, next, apply the simple interval calculation method to this system.

Region A is a *consistent* estimate of parameter $\boldsymbol{\alpha}$, that is,

$$\text{Prob}\{A \cap \boldsymbol{\alpha}\} = 1 \quad \text{at} \quad n \longrightarrow \infty \qquad (24)$$

under the same weak conditions [50]:

$$\lambda_p \longrightarrow \infty \quad \text{at} \quad n \longrightarrow \infty,$$

as in the least squares method. This property means that, as the number of the training objects increases, A collapses to the true $\boldsymbol{\alpha}$.

Region A consists of only certain training objects, called *boundary*, rather than all of them. Therefore, all the objects except for the boundary ones can be removed from the training set without changing the region of possible values.

Prediction of response. Let us consider the problem of predicting the response y for a given new vector $\boldsymbol{x}$ from the model (17). If the parameter $\boldsymbol{a}$ varies within the RPV A, it is clear that the value $\boldsymbol{y} = \boldsymbol{x}^t\boldsymbol{a}$ to be predicted belongs to the interval

$$V = [v^-, v^+], \qquad (25)$$

where

$$v^- = \min_{a \in A}(\boldsymbol{x}^t\boldsymbol{a}), \quad v^+ = \max_{a \in A}(\boldsymbol{x}^t\boldsymbol{a}). \qquad (26)$$

Interval V is the result of predicting by simple interval calculation. To obtain it, one does not need to construct region A explicitly, because problem (26) can be solved using standard linear programming methods [43, 44].

Object status classification. To characterize the quality of prediction by simple interval calculation numerically, the following values are introduced.

Value

$$r(\boldsymbol{x}, y) = \frac{1}{\beta}\left(y - \frac{v^+(\boldsymbol{x}) + v^-(\boldsymbol{x})}{2}\right) \qquad (27)$$

is called the *residual of simple interval calculation*. Value r is the difference between the center of the prediction interval and value y (normalized to β); therefore, r characterizes *bias*.

Value

$$h(\boldsymbol{x}) = \frac{1}{\beta}\left(\frac{v^+(\boldsymbol{x}) - v^-(\boldsymbol{x})}{2}\right) \qquad (28)$$

is called *leverage of simple interval calculation*. Value h is calculated as a half-width of the prediction interval divided by the maximum error; therefore, h characterizes β normalized *precision*.

It is clear that when some new object $(\boldsymbol{x}, y)$ is added to the training set, the admitted region A can undergo one of the following events: (1) the RPV does not change; that is, $A_{n+1} = A_n$; (2) the RPV shrinks; that is, $A_{n+1} \subset A_n$; (3) the RPV disappears; that is, $A_{n+1} = \varnothing$. Here, A_n denotes the RPV constructed using the training set of n objects. The first case corresponds to an object that is called an *insider*. These objects completely agree with the model; therefore, one can fully rely on them in prediction. The second case means that the object lies outside the model available; therefore, it may be called an *outsider*. Outsiders do not conflict with the model and improve the accuracy of modeling when added to the training set. However, until these objects are included in the training set, they are unreliable for prediction. This can arise from two factors. First, the width of the prediction interval, that is, the leverage of simple interval calculation, can be greater than the accuracy of calibration. Alternatively, there is a systematic error, which is characterized by the residual of simple interval calculation. Finally, a third event occurs when the new object is completely contrary to the model constructed. These objects, obviously, are *outliers* in all senses and cannot be used for prediction.

It is shown [22, 32] that this classification can easily be performed without explicitly constructing the RPV. Instead, it is performed on the basis of the following statements relating the values of r and h.

Object $(\boldsymbol{x}, y)$ is an *insider* if and only if

$$|r(\boldsymbol{x}, y)| \le 1 - h(\boldsymbol{x}). \quad (29)$$

Training object $(\boldsymbol{x}_i, y_i)$ is *boundary* if and only if

$$|r(\boldsymbol{x}_i, y_i)| = 1 - h(\boldsymbol{x}_i). \quad (30)$$

Object $(\boldsymbol{x}, y)$ is an *outlier* if and only if

$$|r(\boldsymbol{x}, y)| > 1 + h(\boldsymbol{x}). \quad (31)$$

Object $(\boldsymbol{x}, y)$ is an *absolute outsider*, if and only if

$$h(\boldsymbol{x}) > 1. \quad (32)$$

Using definitions (27), (28) and statements (29)–(32), one can construct an object status plot, a prototype of which is shown in Fig. 9. For any dimension of input data $(\mathbf{X}, \boldsymbol{y})$ and any number of parameters, the object status plot is a two-dimensional plot. This makes it a very powerful tool for multivariate calibration. Statements (29)–(32) divide the plane of SIC residuals vs. simple interval calculation leverages into three regions; each of them corresponding to one of three categories of objects: insiders (region (i) in Fig. 9), outsiders (region (ii)), and outliers (region (iii)).

Usually, when a multivariate calibration model is applied to new objects, the corresponding y values are unknown. Therefore, one cannot calculate simple interval calculation residue r (27), but one can always determine simple interval calculation leverage h (28). One can easily see that, if the leverage of the new object is greater than unity ($h > 1$), region (ii)), this object can be classified as insider at no y. These objects form a special class called *absolute outsiders* (statement (32)).

Estimation of β. To apply the simple interval calculation method, one should know the maximum error β. It is usually unknown, and some estimate b is used instead of β. It is clear that in this case the admitted region (A) depends on b and that $A(b)$ monotonically expands as b increases:

$$b_1 > b_2 \Rightarrow A(b_1) \supset A(b_2). \quad (33)$$

It can be shown that, when there is a sequence of estimates $b_1 > b_2 > \ldots \ge \beta$ converging to β, properties (21)–(24) hold for $A(b_n)$ as well. In addition, it is clear that

$$A(0) = \varnothing, \quad A(\infty) \ne \varnothing. \quad (34)$$

It follows from (33)–(34) that a *minimum* value b exists for which $A(b) \ne \varnothing$. This value can be taken as the estimate of β

$$b_{\min} = \min\{b, A(b) \ne \varnothing\}. \quad (35)$$

The estimate (35) is consistent but biased because $b_{\min} \le \beta$. It gives the lower boundary of the possible β values. This is undoubtedly a useful characteristic of the training set and the model, but, in addition to $b_{\min}$, one should estimate the upper boundary of the maximum error as well.

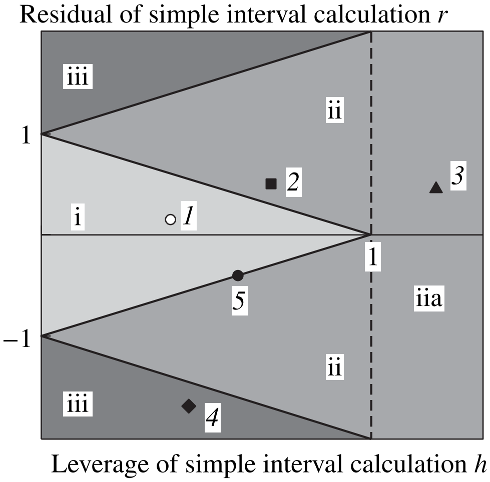
Residual of simple interval calculation r
iii
ii
1
2
3
i
1
1
5
iia
−1
ii
iii
4
Leverage of simple interval calculation h

Fig. 9. Object status diagram: (i) insiders (open circles), (ii) outsiders (filled squares), (iia) absolute outsiders (filled triangles), and (iii) outliers (filled diamonds).

It is clear that any consistent estimate b should depend on two factors:

(1) The number of objects in the training set. The more the objects, the closer b to β.

(2) The heariness of the tail areas of the error distribution. The lighter the tail areas, the worse is this estimate.

Using a conventional statistical approach [51], one can construct an estimate b such that Prob$\{b > \beta\} > P$ and the estimate b is as close to β as possible. Let us consider some point (regression) estimate $\hat{\boldsymbol{y}}$ of the vector $\boldsymbol{y}$, residuals $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$, and a statistics

$$b_{\text{reg}} = \max(|e_1|, \ldots, |e_n|). \quad (36)$$

Statistical modeling performed for different numbers of objects in the training set using various bounded error distributions shows that the estimate

$$b_{\text{SIC}} = b_{\text{reg}} C(n, s^2, P) \quad (37)$$

is the desired upper boundary β with the probability P. The empirical function C [32] depends on the number of objects in the training set n and on the variance of residuals s^2, which characterizes the heariness of the error distribution tail areas.

REFERENCES

1. *Analytical Chemistry. The Approved Text to FECS Curriculum Analytical Chemistry*, Kellner, R., Otto, M., and Widmer, M., Eds., Weinheim: Wiley-VCH, 1998.

2. Tikhonov, A.N., *Dokl. Akad. Nauk SSSR,* 1963, vol. 4, p. 1035.

3. Mar'yanov, B.M., *Izbrannye glavy khemometriki* (Selected Chapters of Chemometrics), Tomsk: Tomsk. Gos. Univ., 2004.

4. Bard, Y., *Nonlinear Parameter Estimation,* New York: Academic, 1974.
5. Pearson, K., *Phillip. Mag*, 1901, vol. 2, no. 6, p. 559.
6. Demidenko, E.Z., *Lineinaya i nelineinaya regressii* (Linear and Nonlinear Regressions), Moscow: Finansy i Statistika, 1981.
7. Martens, H. and Noes, T., *Multivariate Calibration,* New York: Wiley, 1998.
8. Naes, T., Isaksson, T., Fearn, T., and Davies, T., *Multivariate Calibration and Classification*, NIR Publications, 2002.
9. Faber, K., *Chemom. Intell. Lab. Syst.,* 2000, vol. 52, p. 123.
10. Pomerantsev, A.L., *Chemom. Intell. Lab. Syst.*, 1999, vol. 49, p. 41.
11. Kantorovich, L.V., *Sib. Mat. Zh.*, 1962, vol. 3, no. 5, p. 701.
12. Voshchinin, A.P., Bochkov, A.F., and Sotirov, G.R., *Zavod. Lab.*, 1990, vol. 56, no. 7, p. 76.
13. Anisimov, V.M., Pomerantsev, A.L., Novoradovskii, A.G., and Karpukhin, O.N., *Zh. Prikl. Spektrosk.,* 1987, vol. 46, p. 117.
14. Akhunov, I.R., Akhmadishin, Z.Sh., and Spivak, S.I., *Khim. Fiz.,* 1982, vol. 12, p. 1660.
15. Bakhitova, R.Kh and Spivak, S.I., *Khim. Khim. Tekhnol.*, 1999, vol. 42, p. 92.
16. Slin'ko, M.G., Spivak, S.I., and Timoshenko, V.I., *Kinet. Katal.*, 1972, vol. 13, p. 1570.
17. Spivak, S.I., Timoshenko, V.I, and Slin'ko, M.G., *Dokl. Akad. Nauk SSSR,* 1970, vol. 192, p. 580.
18. Belov, V.M., Karbainov, Yu.A., Sukhanov, V.A., et al., *Zh. Anal. Khim.*, 1994, vol. 49, p. 370.
19. Khlebnikov, A.I., *Zh. Anal. Khim.,* 1996, vol. 51, no. 3, p. 347 [*J. Anal. Chem.* (Engl. Transl.), vol. 51, no. 3, p. 321].
20. Belov, V.M., Sukhanov, V.A., and Unger, F.G., *Teoreticheskie i prikladnye aspekty metoda tsentra neopredelennosti* (Theoretical and Applied Aspects of the Uncertainty Center Method), Novosibirsk: Nauka, 1995.
21. Pomerantsev, A.L. and Rodionova, O.Ye., *Chemom. Intell. Lab. Syst.*, 2005, vol. 79.
22. Rodionova, O.Ye., Esbensen, K.H., and Pomerantsev, A.L., *J. Chemom.*, 2004, vol. 18, p. 402.
23. Westad, F. and Martens, H., *J. Near Infrared Spectrosc.*, 2000, vol. 8, p. 117.
24. Cook, R.D., *Technometrics*, 1977, vol. 19, p. 15.
25. Cook, R.D., *J. Am. Stat. Assoc.*, 1979, vol. 74, p. 169.
26. Andrews, D.F. and Pregibon, D., *J. Royal Stat. Soc.*, 1978, vol. 40, p. 84.
27. Draper, N.R. and John, J.A., *Technometrics*, 1981, vol. 23, p. 21.
28. Ns, T., *J. Chemom.*, 1989, vol. 1, p. 121.
29. Höskuldsson, A., *Chemom. Intell. Lab. Syst.*, 2001, vol. 55, p. 23.
30. Jouan-Rimbaud, D., Massart, D.L., Saby, C.A., and Puel, C., *Anal. Chim. Acta*, 1997, vol. 350, p. 149.
31. Fernandez, Pierna, J.A., Wahl, F., de Noord, O.E., and Massart, D.L., *Chemom. Intell. Lab. Syst.,* 2002, vol. 63, p. 27.
32. Rodionova, O.Ye. and Pomerantsev, A.L., *Progress in Chemometrics Research*, New York: Nova Science, 2005, p. 43.
33. Pomerantsev, A.L. and Rodionova, O.Ye., *Progress in Chemometrics Research*, New York: Nova Science, 2005, p. 209.
34. Pomerantsev, A.L. and Rodionova, O.Ye., *Aging of Polymers, Polymer Blends, and Polymer Composites*, New York: Nova Science, 2002, vol. 2, p. 19.
35. Esbensen, K.H., *Multivariate Data Analysis in Practice*, 4th ed., CAMO, 2000.
36. Lang, A., *Neftegaz. Tekhnol.*, 1994, no. 9, p. 10.
37. Clancey, V.J., *Nature,* 1947, vol. 159, p. 339.
38. Rajkó, R., *Anal. Lett.,* 1994, vol. 27, p. 215.
39. Eriksson, L., Johansson, E., Kettaneh-Wold, N., and Wold, S., *Multi- and Megavariate Data Analysis*, Umeå: Umetrics, 2001.
40. Sulima, E.L., Zubkov, V.A., and Rusinov, L.A., *Progress in Chemometrics Research*, New York: Nova Science, 2005, p. 196.
41. Savitzky, A. and Golay, M.J.E., *Anal. Chem.*, 1964, vol. 36, p. 1627.
42. Höskuldsson, A., *Prediction Methods in Science and Technology*, Copenhagen: Thor, 1996, vol. 1.
43. Dantzing, D., *Linear Programming, Its Generalizations, and Applications*, Princeton, N.J.: Princeton Univ. Press, 1963.
44. Taha, H., *Operations Research: An Introduction*, 3 ed., New York: MacMillan, 1982, vol. 1.
45. Karmanov, V.G., *Matematicheskoe programmirovanie* (Mathematical Programming), 5th ed., Moscow: Fizmatlit, 2001.
46. *Linear Programming Packages*, available at http://www.ici.ro/camo/hlp.htm (May 1, 2005).
47. *GOST* (State Standard) *8226-82: Engine Fuel. Research Method for Determining the Octane Number*, 1982.
48. Gass, S., *Linear Programming*, 4th ed., New York: McGow-Hill, 1975.
49. Kuhn, H.W. and Tucker, A.W., *Ann. Math. Studies* (Princeton), 1956, vol. 38.
50. Eicker, F., *Ann. Math. Stat*, 1963, vol. 34, p. 447.
51. Gumbel, E., *Statistics of Extremes*, New York: Columbia Univ. Press, 1962.
52. *The Unscramber*, available at httr://www.samo.no/ (May 1, 2005).