JOURNAL OF CHEMOMETRICS
J. Chemometrics 2004; **18**: 402–413
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.885

Application of SIC (simple interval calculation) for object status classification and outlier detection—comparison with regression approach

Oxana Ye. Rodionova[1]*, Kim Esbensen[2] and Alexey Pomerantsev[1]

[1]Institute of Chemical Physics, 4 Kosygin Street, 119991, Moscow, Russia
[2]ACABS, Aalborg University Esbjerg, Niels Bohrs Vej 8, DK-6700, Esbjerg, Denmark

Received 19 May 2004; Revised 26 October 2004; Accepted 8 November 2004

We introduce a novel approach termed simple interval calculation (SIC) for classification of object status in linear multivariate calibration (MVC) and other data analytical contexts. SIC is a method that directly constructs an interval estimator for the predicted response. SIC is based on the single assumption that all errors involved in MVC are *limited*. We present the theory of the SIC method and explain its realization by linear programming techniques. The primary SIC consequence is a radically new object classification that can be interpreted using a two-dimensional object status plot (OSP), 'SIC residual vs SIC leverage'. These two new measures of prediction quality are introduced in the traditional chemometric MVC context. Simple straight demarcations divide the OSP into areas which quantitatively discriminate all objects involved in modeling and prediction into four different types: *boundary samples*, which are the significant objects (for generating the entire data structure) within the training subset; *insiders*, which are samples that comply with the model; *outsiders*, which are samples that have large prediction errors; and finally *outliers*, which are those samples that cannot be predicted at all with respect to a given model. We also present detailed comparisons of the new SIC approach with traditional chemometric methods applied for MVC, classification and outlier detection. These comparisons employ four real-world data sets, selected for their particular complexities, which serve as showcases of SIC application on intricate training and test set data structures.

KEYWORDS: projection methods; SIC method; object status classification; representative subset selection; outlier detection

1. INTRODUCTION

Projection methods are based on the concept of latent variables (basis vectors spanning a subspace). The best-known example in chemometrics without doubt concerns bilinear projection methods, e.g. PCR and PLS (See References [1,2] and references cited therein). One of the main advantages of these model-forming approaches is the possibility to explore hidden data structures visually with the aid of simple two- and three-dimensional projection windows in both the object and variable spaces. These multivariate calibration (MVC) methods are widely used today within chemometrics and also in increasing fashion outside this field [3,4].

However, there is at least one important aspect of bilinear modeling, and most likely several others, which is still out in the open: *objective recognition and status of outliers*. Consider for example how to form decisions regarding the *importance* and *quantitative role* of a particular sample in a typical data set (training set, test set, etc.), i.e. how to *classify* qualitatively the specific type of outlier and to quantify its 'influence' on e.g. calibration modeling, predictions, etc. Traditionally, the concept of influential objects is mainly discussed in the light of outlier detection. In ordinary regression analysis, different measures are used, i.e. the Cook distance [5,6], the AP distance [7] or a combination of these statistics [8]. A discussion of influence measure in bilinear modeling is given in Reference [9]. Detailed descriptions of different types of errors—errors in predictors, errors in response, errors in calibration data and errors in future prediction data—as well as the different ways of treating these errors are presented in Reference [1]. The most commonly used tool in MVC is the influence plot [1,3], which helps to reveal the most significant and most dangerous outliers and also to find the most informative and therefore important sample in calibration. Although strategies for more or less automatic

*Correspondence to: O. Ye. Rodionova, Institute of Chemical Physics, 4 Kosygin Street, 119991 Moscow, Russia.
E-mail: rcs@chph.ras.ru

outlier elimination are presented in References [10,11], in many cases the ultimate decision can be made only by the user, who must have sufficient contextual background knowledge to distinguish between the different types of outliers; therefore such designations are made *informally* [1–3,12,13]. Another MVC problem closely connected with these issues is that of prediction reliability. There are numerous technical methods on how to control this, but a generally accepted approach does not exist [14–16].

We present here a complete object status classification theory corresponding with most data analytical objectives. It is based on a novel approach termed simple interval calculation (SIC). This is a method for linear modeling [17] which is here applied to the MVC realm, where it is shown to bring about a new systematic insight into *MVC object leverage and outlier analysis*, as well as providing for a new object discrimination *analogy* to various variable selection approaches which have seen significant activity in the last 5–8 years [18–21]. Below we briefly present the SIC theory behind the new *object status classification* (OSClas), which allows us to distinguish quantitatively the most important object types for modeling, termed *boundary objects*, among all calibration samples (training data set), as well as to differentiate three status categories for prediction objects, as *insiders*, *outsiders* and *outliers* respectively. This classification is made automatically with the help of simple and explicit formulae, which can also be geometrically presented using the new object status plot introduced in the paper. This is a two-dimensional plot for any model complexity. The position of an object in this plane fully characterizes the object's status. Such an approach has the evident advantage of being an unambiguous status classification method that draws strict borders between the different classes of model samples. As the SIC method has in its background a postulate that differs from the traditional regression concept, the SIC object status classification brings a new insight into data set structure.

2. SIC BASIC PRINCIPLES

The SIC approach is based on the single assumption that all errors involved in the MVC problem are *limited* (sampling errors, measurement errors, modeling errors, etc.), which would appear to be a reasonable supposition in many practical applications [22,23].

2.1. Region of possible values

Let us consider the linear MVC model

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y}$ is the n-dimensional response vector, $\mathbf{a}$ is the p-dimensional parameter vector, $\mathbf{X}$ is the $n \times p$ predictor matrix and $\boldsymbol{\varepsilon}$ is the error vector.

Definition 1

Error finiteness means that there exists a maximum error deviation (MED) of the error ε, which equals β, i.e.

$$\exists \beta > 0 \quad \text{Prob}\{|\varepsilon| > \beta\} = 0$$

and for any

$$0 < b < \beta \quad \text{Prob}\{|\varepsilon| > b\} > 0 \tag{2}$$

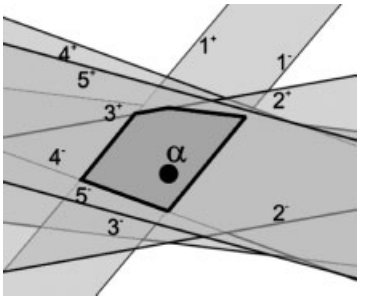

Figure 1. Strips in parameter space, $p = 2$. Typical shape of RPV A (polyhedron), $\boldsymbol{\alpha}$ (●) is the 'true' value.

where Prob{•} denotes the probability that an event occurs.

Symmetry of $\boldsymbol{\varepsilon}$ is used here for simplicity, but this assumption is not essential for the method. We consider β to be *common* for all objects, i.e. we assume error homoscedasticity (which, however, is also not critical).

First we suppose that β *is known*. We call a pair $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, a calibration object. Here vector $\mathbf{x}_i^{\mathrm{T}}$ is the ith row in the $\mathbf{X}$ matrix, which has a corresponding response value y_i. In this work the data analytical term 'object' will be used synonymously with 'sample', both terms signifying one individual row in a conventional data matrix ($\mathbf{X}$, $\mathbf{y}$).

According to our main assumption (2), we can write for each calibration object $i = 1, \ldots, n$

$$y_i^- \le \mathbf{x}_i^{\mathrm{T}}\mathbf{a} \le y_i^+, \qquad y_i^- = y_i - \beta, \qquad y_i^+ = y_i + \beta \tag{3}$$

Naturally we do not know the true parameter vector, which will be denoted here by $\boldsymbol{\alpha}$, but we can consider all vectors $\mathbf{a}$ which agree with Equation (3). It is obvious that such vectors $\mathbf{a}$ for a given i form a *strip* $S(\mathbf{x}_i, y_i)$ in the space of parameters R^p. The position and width of the strip depend on the calibration object values $(\mathbf{x}_i, y_i)$. An example of such strips can be seen in Figure 1, where we have two unknown parameters, i.e. $p = 2$, and five samples in the system, i.e. $n = 5$.

Let us consider all calibration samples and their corresponding strips. It is obvious that any vector $\mathbf{a}$ satisfies all inequalities (3) simultaneously if and only if it belongs to all strips $S(\mathbf{x}_i, y_i)$ (polyhedron in Figure 1).

Definition 2

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

$$A = \bigcap_{i=1}^{n} S(\mathbf{x}_i, y_i) \tag{4}$$

Region A is a *closed convex polyhedron* [24,25] delineated by the boundaries of intersecting strips. This is a *random set*, because the RPV is constructed using random values $\mathbf{y}$. Naturally, randomness of A means that it depends on vector $\mathbf{y}$ and varies when the response values are changed. Formally, randomness of $A(\mathbf{y})$ means that for any $\mathbf{a} \in R^p$ we can calculate the measure $\{\mathbf{y} \in R^n : \mathbf{a} \in A(\mathbf{y})\}$ and thus calculate the probability Prob$\{\mathbf{a} \in A\}$. This is of course just a general definition [26] for the probability that confidence area A covers some point $\mathbf{a}$.

2.2. RPV properties

The RPV A has the following properties for any linear MVC model (1).

Property 1

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

In confidence interval theory [26] this means that the probability of covering a false value of the unknown parameter is not greater than the probability of covering the true value.

From the RPV definition it follows that the true value $\boldsymbol{\alpha}$ always belongs to the RPV:

$$\text{Prob}\{\boldsymbol{\alpha} \in A\} = 1 \tag{5}$$

In particular, this means that if A consists of only one element, $A = \{\mathbf{a}\}$, this element is the true parameter value $\boldsymbol{\alpha}$. It is interesting that in the SIC analysis such a situation is possible even when the number of samples is finite $(n < \infty)$—in contrast with the traditional statistical approach.

Property 2

The region A is bounded if and only if rank $\mathbf{X} = p$ [24,25].

This means that if we work with a multicollinear system, where rank $\mathbf{X} < p$, we have to apply some regularization procedure. To overcome these difficulties, we may apply a standard technique [1,2] and *project* the initial data (Equation (1)) on a lower-dimensional subspace as

$$\mathbf{y} = \mathbf{T}\mathbf{P}^{\mathrm{T}}\mathbf{a} + \mathbf{f} = \mathbf{T}\mathbf{q} + \mathbf{f} \tag{6}$$

where the score matrix $\mathbf{T}$ has full rank $k < p$, and afterwards apply the SIC method to Equation (6). This approach may be used with any particular projection method (PCR, PLS, etc.) or a ridge regression method. In this paper we will neither discuss the choice of a method nor focus on the issue of choosing the correct number of bilinear projection components, etc., as this is widely known within chemometrics.

Property 3

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

$$\text{Prob}\{A \cap \boldsymbol{\alpha}\} = 1 \quad \text{as} \quad n \to \infty \tag{7}$$

under traditional *weak* conditions [27]

$$\lambda_p \to \infty \quad \text{as} \quad n \to \infty \tag{8}$$

as for the OLS estimate. This means that if the number of calibration samples increases, A shrinks towards the true value $\boldsymbol{\alpha}$.

Property 4

The RPV is formed not by all objects from the calibration set, but only by the subset of so-called *boundary objects*. Therefore, if we exclude all objects from the calibration set except these boundary objects, the RPV will not change. In Figure 1, all objects except sample 5 are boundary objects in this sense.

2.3. Predicting the response

Consider a response prediction for any new vector $\mathbf{x}$ using the model in Equation (1). If parameter $\mathbf{a}$ varies over the RPV A, it is clear that the predicted value $\mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{a}$ belongs to the interval

$$V = [v^-, v^+] \tag{9}$$

where

$$v^- = \min_{\mathbf{a} \in A}(\mathbf{x}^{\mathrm{T}}\mathbf{a}), \qquad v^+ = \max_{\mathbf{a} \in A}(\mathbf{x}^{\mathrm{T}}\mathbf{a}) \tag{10}$$

The interval V (Equation (9)) is the result of an SIC prediction. To find this interval, we need not present the RPV explicitly, as the solutions of Equation (10) may be obtained by linear programming methods [24,28,29], which are commonly used [30] to find the optima of linear functions in a convex set. It is known that the optimum is achieved in a vertex of the set, and the *Simplex algorithm* [28] makes this optimization by moving from one vertex to another. Being a standard numerical analysis technique, it is not considered further here.

The SIC interval stands in contrast to the more traditional confidence interval estimators based upon theoretical error distributional model assumptions, which certainly do not always hold for practical data analysis of real-world technological and natural systems anyway [23].

3. SIC OBJECT STATUS CLASSIFICATION

To understand this new approach, one has to keep in one's mind two key issues.

(1) *SIC provides a calibration error measure*, which is the maximum error deviation β. This error is shown in Figure 2(a) by a black error bar associated with the reference value. We call these the calibration intervals.

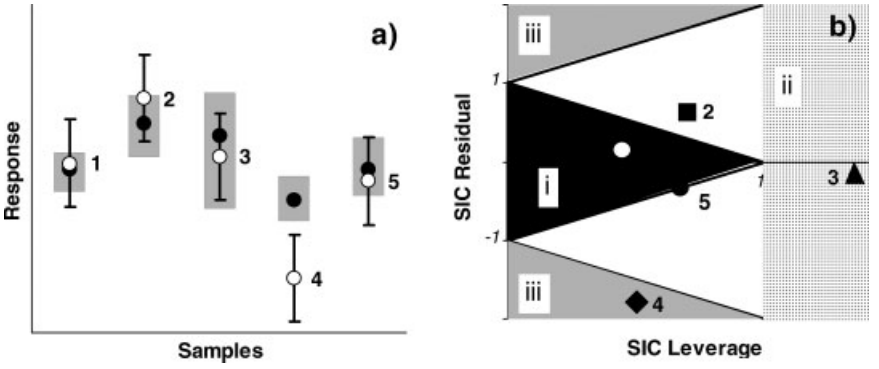

Figure 2. (a) Calibration intervals (black error bars) and prediction intervals (wide gray bars). (b) Object status plot: i, insiders (○); ii, absolute outsiders (▲); iii, outliers (◆).

(2) *The SIC prediction interval V (Equation (9)) is a prediction error measure*. This is presented as wide gray bars in Figure 2(a). It is necessary to mention that in this plot we do not specify the type of the samples, which can be from the calibration set *or* from the test set. Inspecting this plot, one can see different relationships between all calibration and prediction intervals. The gray bars are sometimes wholly inside the black bars, such as for sample 1; this is a good case, because the individual prediction error is here less than the calibration error. This can be a calibration sample or a test sample, which is rather similar to the calibration ones. Sample 3 demonstrates the inverse case, where the gray bar is wider than the black interval; this would reciprocally be a bad case, as the prediction error is worse than the calibration error. This is of course not a calibration sample, but one of the test samples. Sample 2 represents a case in which a small prediction interval is *biased* against the reference value. This is also a test sample which could have a wrongly measured reference value y. Sample 4 displays the worst case, i.e. when the calibration and prediction intervals do not intersect at all. This could for example be a test sample which has totally another structure in predictor vector $\mathbf{x}$. The last sample 5 is of special interest, because its prediction interval touches the calibration interval. If this sample belongs to the calibration set, it is a boundary object (see Property 4).

Thus in this type of SIC plot we can simultaneously observe both the position of interval V Equation (9), regarding the known reference value y as well as the interval's magnitude with respect to value β, which together characterize the 'quality' of the prediction. To quantify this new characteristic ('quality'), we introduce the following two SIC measures.

Definition 3

SIC residual is defined as

$$r(\mathbf{x}, y) = \frac{1}{\beta}\left(y - \frac{v^+(\mathbf{x}) + v^-(\mathbf{x})}{2}\right) \tag{11}$$

The SIC residual is seen to be the difference between the center of the prediction interval and the reference value (scaled by β), so it is a characteristic of *bias*.

Definition 4

SIC leverage is defined as

$$h(\mathbf{x}) = \frac{1}{\beta}\left(\frac{v^+(\mathbf{x}) - v^-(\mathbf{x})}{2}\right) \tag{12}$$

The SIC leverage is calculated as the width of the prediction interval divided by the calibration error, so it has the character of β-normalized *precision*.

Using Equation (3), it can be shown that all *calibration* samples satisfy the inequality

$$|r(\mathbf{x}, y)| \leq 1 - h(\mathbf{x}) \tag{13}$$

Calibration samples for which the equality in (13) is achieved are *boundary* samples (see Property 4).

This approach is seen to be helpful when establishing an explicit classification of *new* objects (new test set samples or new X-data alone) in relation to a specific calibration model, which is represented by its pertinent RPV A. It is evident that adding a new sample $(\mathbf{x}, y)$ to the calibration set could modify A in only one of the following ways:

(1) A does not change, i.e. $A_{n+1} = A_n$;
(2) A shrinks, i.e. $A_{n+1} \subset = A_n$;
(3) A disappears, i.e. $A_{n+1} = \emptyset$.

Here A_n stands for the RPV that is constructed with the help of a calibration set consisting of n objects.

The first case corresponds to samples which are to be termed *insiders* (sample 1 in Figure 2(a)). They agree completely with the model; thus insiders can be trusted absolutely in prediction. The second case means that such objects are located *outside* the existing model, and they are therefore termed *outsiders* (samples 2 and 3 in Figure 2(a)). Outsiders do not contradict the model, but, when added to the calibration set, they actually *improve* the calibration (modeling) accuracy. However, while they are not in the calibration set, outsiders are less than perfect with respect to prediction. There may be two reasons: the width of the prediction interval (i.e. the *SIC leverage*) is greater than the calibration error, or there is a bias (characterized by the *SIC residual*). In the third case, such samples totally conflict with the established model (sample 4 in Figure 2(a)). They are clearly *outliers* in every sense of the term; they cannot be used in prediction at all.

It was shown [31] that such a classification of new objects could easily be performed without explicit construction of the complex RPV in parameter space. It is instead based on the following statements.

Statement 1

An object $(\mathbf{x},y)$ is an insider iff $|r(\mathbf{x}, y)| \leq 1 - h(\mathbf{x})$.

Statement 2

Calibration object $(\mathbf{x}_i, y_i)$ is a boundary object iff $|r(\mathbf{x}_i, y_i)| = 1 - h(\mathbf{x}_i)$.

Statement 3

An object $(\mathbf{x}, y)$ is an outlier iff $|r(\mathbf{x}, y)| > 1 + h(\mathbf{x})$.

Statement 4

An object $(\mathbf{x}, y)$ is an absolute outsider (explained below) for any y iff $h(\mathbf{x}) > 1$.

Using Definitions 3 and 4, one can construct a new *object status plot* (OSP), the archetype of which is shown in Figure 2(b). This OSP has the same appearance for *any* dimensionality of the initial data (**X, y**) and for *any* number of model parameters, which makes it a very powerful MVC tool. Statements 1–4 divide the SIC residual (r) vs SIC leverage (h) plane into three areas, each corresponding to one of the three new object categories: insiders (area i in Figure 2b), outsiders (out of area i) and outliers (area iii). Figures 2(a) and 2(b) show in fact two representations of the same data set.

It should be mentioned that the triangular shape of the insider's area in the OSP (Figure 2(b)) may appear somewhat similar to the conventional influence plot (See Figure 4(b)). In Reference [1] (p. 286) it was written: '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 developed here. Certainly, the similarity between the influence plot and the OSP is not a coincidence. This comes from a well-known basic statistical relationship [2] which relates modeling *accuracy* (RMSEC), *precision* (SEC) and *bias* (BIAS):

$$\mathrm{RMSEC}^2 \approx \mathrm{SEC}^2 + \mathrm{BIAS}^2 \tag{14}$$

In the SIC approach, in which the MED value β is the calibration accuracy, the SIC leverage h stands for the (normalized) precision and the SIC residual r is responsible for the (normalized) bias, Equation (14) may then be represented in the following form:

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

which actually conforms to Equation (14).

On the other hand, we should recognize a substantial difference between Equations (14) and (15), as Equation (14) has sense only for the whole data set, i.e. *on average*, while Equation (15) is *valid for every sample in the data set*.

Usually, when working with a new sample in MVC, we do not know its reference value y. In this case it is of course impossible to calculate the SIC residual r (Equation (11), but we can still calculate the SIC leverage h (Equation (12)). From Figure 2 it is clear that such a new sample for which the leverage is greater than one ($h > 1$, area ii) *cannot* be classified as an insider (area i) for *any* response value. Such samples form a special class of objects which are called *absolute outsiders* (Statement 4). Thus, even when having no information about their reference values, we can nevertheless state that the prediction error of such samples will be greater than the calibration error. Using Equation (12), we can establish the equality

$$h(\lambda\mathbf{x}) = |\lambda| h(\mathbf{x})$$

from which we see that for any calibration predictor (or score) $\mathbf{x}_i$, vector $\lambda\mathbf{x}_i$ is an absolute outsider iff $|\lambda| > 1/h(\mathbf{x}_i)$. As a result, for any given calibration data we can construct the region in predictor space occupied by these absolute outsiders. The following statement defines this area.

Statement 5

Let D be a set in X- (or T-) space, defined as a linear combination of weighted calibration predictors (or scores) $\mathbf{x}_i$:

$$\mathbf{x} = \sum_{i=1}^{n} \frac{\lambda_i}{h(\mathbf{x}_i)} \mathbf{x}_i, \qquad \sum_{i=1}^{n} |\lambda_i| \le 1$$

Then all *absolute outsiders* are to be found exclusively outside this region D. An example of this area is presented in Section 7.1.

Therefore we have shown that the SIC approach lets us introduce a novel method for classification of all MVC objects (calibration samples as well as new or test samples). This classification is termed here *object status classification* (OSClas). It is based on Definitions 3 and 4 and on the consequential Statements 1–5, which follow from them.

To apply OSClas, one has to know the value of MED defined in Equation (2). Ordinarily it is unknown and some estimate b is used instead of β. It is clear that in this case the RPV A depends on b and that $A(b)$ is extended monotonically with increasing b:

$$b_1 > b_2 \quad \Rightarrow \quad A(b_1) \supset A(b_2) \tag{16}$$

Therefore we can claim that if we have a sequence of consistent β estimates $b_1 > b_2 > \ldots \ge \beta$, then Properties 1–4 are true for $A(b_n)$ as well.

Furthermore, it is evident that

$$A(0) = \emptyset, \qquad A(\infty) \neq \emptyset \tag{17}$$

From Equations (16) and (17) it follows that there exists a *minimum* b such that $A(b) \neq \emptyset$. This minimum value can be taken as an estimator for the unknown parameter β:

$$b_{\min} = \min\{b, A(b) \neq \emptyset\} \tag{18}$$

The estimate in Equation (18) is consistent but biased ($b_{\min} \le \beta$), and it is the lower limit of all possible β values. $b_{\min}$ is a useful characteristic of a calibration set and a calibration model, but we need to estimate the upper limit too. Applying the traditional statistical approach [32] to the regression residuals $(\hat{y} - y)$, it is possible to find an estimator b_{SIC} such that, for probability P close to one, $\mathrm{Prob}\{b_{\mathrm{SIC}} > \beta\} > P$ and b_{SIC} is as close to β as possible [31].

The calculation of different β estimators is rather comprehensive and is outside the scope of this paper. However, we can present a rule of thumb that helps to evaluate the estimators roughly. This could be termed the '1-2-3-4 sigma rule'. If we accept that $\mathrm{RMSEC} \approx 1\sigma$, then $b_{\min} \approx 2\sigma$, $b_{\max} \approx 3\sigma$ and $b_{\mathrm{SIC}} \approx 4\sigma$. Certainly, this rule represents just a tendency, which also depends on the number of samples in the calibration set. Nevertheless, our experience in application to numerous examples shows that this rule appropriately characterizes the situation. To confirm this claim, we present Table I. In this table we have collected relevant information from all examples that are examined below.

The arguments for this rule are very simple. For any error distribution, MED should be more than 2σ; the marginal case is the uniform distribution where $\beta = 1.71\sigma$[32]. Then, for the usual number of data samples (say, less then 1000), we could not expect an outlier farther than 3σ. Finally, the 4σ limit gives us the assurance that new samples will never lie outside this border. Therefore we can state that all possible values of unknown MED are located within the interval [3σ, 4σ]. It is natural to ask whether such variations of MED could influence the results and conclusions of OSClas. The answer is negative, inasmuch as the main SIC quality measures r and h are defined as relative ratios; see Equations (11) and (12). Another issue concerns the SIC prediction intervals. These

Table I. Illustration to the '1-2-3-4 sigma rule'

Data	n	RMSEC/σ	$b_{\min}/\sigma$	$b_{\max}/\sigma$	b_{SIC}/σ
Oil	40	1.0	2.3	3.0	4.1
Ship	27	1.0	1.9	2.7	4.0
DSC	11	1.0	1.8	2.0	3.7
Wheat	139	1.0	2.8	3.0	3.7

increase with b, and when $b = \beta$ these intervals have a covering probability equal to one, by definition. On the other hand, it can be shown [31] that the same intervals constructed with the estimated MED b_{SIC} (for $P = 0.90$), instead of the true β value, display a covering probability that in any case is not less than 0.9999. This result confirms that not only the proposed OSClas but also the whole SIC theory in general can be used in practice.

4. SIC IN DATA ANALYSIS

Below we demonstrate how the SIC method can be applied to four real-world data sets, each with its specific data analytical objects, and show the additional SIC information obtainable. Our analysis is based on bilinear projection models that have been established and studied before. Two data sets have been published; the other data can be acquired on demand from the corresponding author. When applying the SIC method to these problems, for each data set we use the following procedure.

1. Establish a pertinent PLS or PCR model for the data set (or use the known model).
2. Validate, in order to find the optimal number of PCs or PLS components (or take this dimensionality from the published literature).
3. Apply SIC modeling for this model complexity, i.e. calculate b_{SIC}, find prediction intervals V and obtain SIC leverages h and residuals r.
4. Analyze and compare SIC and PLS/PCR results.
5. Discuss and conclude on the value of the added SIC information.

Strictly speaking, the only relevant part of these four analytical data sets/problems for the present purpose is that they present (very) different internal data structures, which allows us to demonstrate the SIC approach's new features in a variety of settings—but we do give a brief overall description of their basic data analytical backgrounds nevertheless. The ability of bilinear projection to display the hidden data structures comes to the fore in the *t–t score plots* (for PCA and PCR) and the *t–u cross-score plots* (for PLS). In particular, we shall make use of *dual-space illustrations* combined with a *brushing* facility, i.e. highlighting of objects which have been delineated in one space in (all or selected) other spaces—especially incorporating SIC classifications in the more conventional chemometric plots.

5. ANALYSIS OF CALIBRATION SETS

5.1. Acoustic determination of trace oil concentrations in water

This data set demonstrates the application of acoustic chemometrics for quantitative determination of trace oil concentrations in water [3]. The **X** matrix consists of 1024 acoustic frequency variables (after FFT). The response vector **y** represents reference concentrations of oil in the calibration samples that were specially prepared in the test laboratory. The training data set consists of 40 observations (objects), and there is a test set also consisting of 40 observations.

The original (raw data) PLS model shows a non-linearity in the first-component t–u plot, signifying a non-linear relationship between the oil concentration and its influence on the effective surface tension of water. Therefore the raw y values were transformed by $y = \log(1 + y_{raw})$, which was sufficient to linearize the data set. The final model consists of two PLS components only, since these explain a total of 60% of X-variance and 99.9% of Y-variance, with RMSEC = 0.051, RMSEP = 0.092 for the external validation (test set). The pertinent t–u plots are shown in Figure 3. Circles (open and full) represent training objects.

Using this PLS model, we construct the SIC model and corresponding OSP (presented in Figure 4(a)). The calibration objects in Figure 3 have been annotated with their SIC designations (from Figure 4(a)): full circles represent *boundary samples*. These boundary objects are marked (brushed) in the t–u plots. It is easy to appreciate the peripheral positions of the boundary objects. The gray polygon in Figure 3 is used here to delineate the model backbone formed exclusively by the SIC boundary samples. There are also to be found a few calibration samples outside this boundary, which is quite understandable, since the t–u plots are only projections of a complicated model onto this plane. The interesting issue is: what *additional* SIC information can be gleaned from Figure 4?

By comparing the SIC object status plot (Figure 4(a)) and the influence plot [1,3] (Figure 4(b)), we can state that all the most influential samples found in the influence plot (nos 37, 38 and 40) and the sample with the highest residual variance

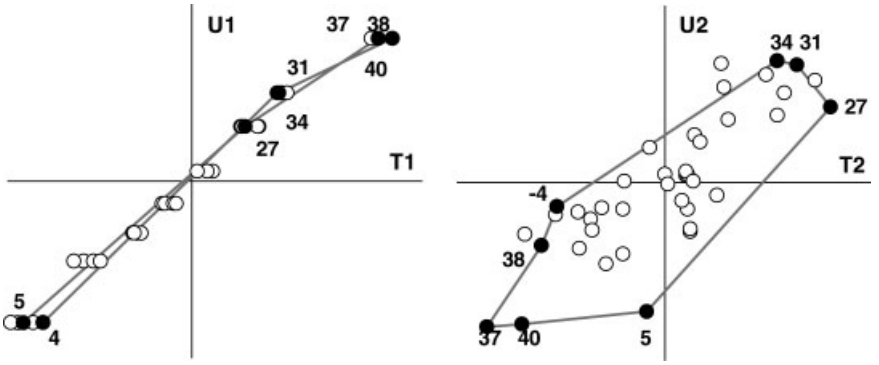

Figure 3. Trace oil-in-water data set. t–u plots for PLS model with two components. Training set: ○, insiders; ●, boundary samples; gray line connects boundary samples.

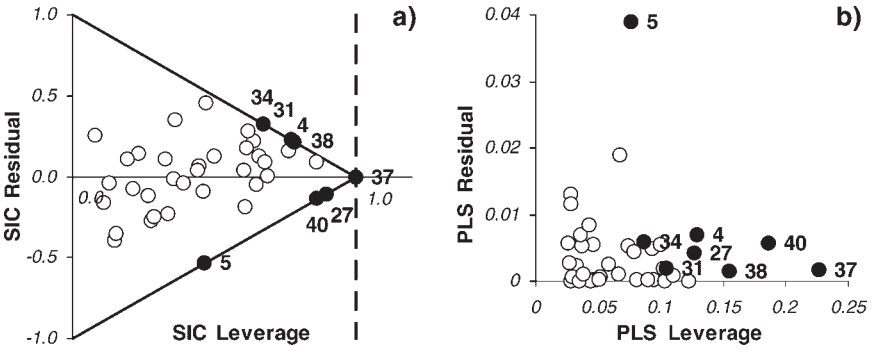

Figure 4. Trace oil-in-water data set. (a) SIC object status plot. (b) influence plot for *y*. PLS model with two components. Training data: ○, insiders; ●, boundary samples.

(no. 5) at the same time are the boundary samples by OSClas. Figures 3 and 4 demonstrate that the SIC object status classification helps to reveal all important samples in the training data set. To identify such samples, we have strict and simple rules (Statements 1 and 2). This example shows that the concept of boundary samples not only makes sense within the SIC approach but also characterizes the existing data set structure optimally. The next subsection and one more example confirm this idea.

5.2. Representative subset selection

Often, e.g. in calibration transfer [33] but also for other data analytical objectives, it is necessary to select a special subset from a large(r) calibration set that will bear the burden of *representing* the entire relevant data model. In general, such a subset will satisfy two opposing requirements: (1) it should be of maximal representativity with respect to the entire set, but (2) it should simultaneously be noticeably smaller than the total set. There have been presented several suggested solutions to this dilemma [34,35], to which we will now append that of the SIC method.

Thus SIC's concept of boundary objects (Statement 2) is applied for this selection. We demonstrate this by using a data set representing wheat calibration [36]. The **X** matrix consists of NIR spectra in the range of 908–1120 nm, recorded at 118 wavelengths; the reference **y** vector includes moisture contents of 139 calibration samples as quantified in the laboratory by a standard analytical method (evaporation loss of weight).

We originally conducted this work at the request of an NIR instrument manufacturer who preferred to use LOO cross-validation in the PLS modeling. The mode1 carries four components, which explain 98% of X-variance and 89% of Y-variance; RMSEC = 0.30, RMSEP = 0.33. On the basis of this PLS model we construct the pertinent SIC model and detect 23 boundary samples. These samples are then the most important objects for modeling, and all other samples may be treated as redundant. Therefore we fix the complexity of the model to these four PLS components and undertake the following procedure.

1. Use the 23 boundary samples as a new calibration set and designate all the remaining objects as a new test set.
2. Construct a new PLS model with four components, based on this new calibration set.
3. Predict all these test set samples using the SIC model with the previously calculated b_{SIC}.

From this we obtain RMSEP = 0.29 via a new *test set validation*. The SIC object status plot (Figure 5) shows that all samples from the new test set are indeed *insiders*. This means that we have in fact reached our goal: (1) the model constructed with the help of the selected subset can predict all other samples with an accuracy that is no worse than the error of calibration evaluated on the whole data set; (2) this subset is indeed significantly smaller, 23 out of 139.

This example demonstrates that boundary samples are not only significant objects for SIC models, but they constitute relevant objects also for bilinear projection models. Not only calibration transfer is in need of such subsets, of course. It bears mentioning that the decision regarding selection of a representative subset may not be so evident as above, since the size of the subset can be constrained by practical necessity and may have to be less than the total number of boundary samples. Another problem concerns multiresponse data. If we have to construct different PLS1 models for each of the responses, i.e. protein, moisture, etc., each boundary sample subset may be different. Nevertheless, the concept of boundary samples is useful as a starting point for such more comprehensive subset selection.

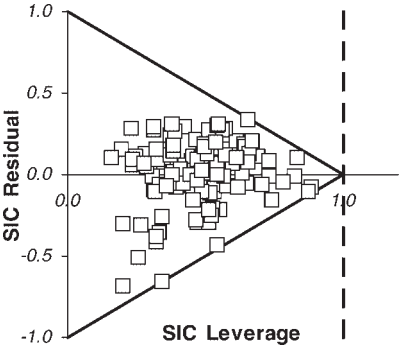

Figure 5. Object status plot for wheat data set: □, 'test samples' (all insiders).

6. SIC ANALYSIS OF TEST SET OBJECTS

The most useful aspect of SIC's facility of object status classification concerns test sets. This is where the new SIC prediction interval estimates *for each test set sample* contrast most with the traditional *ensemble* RMSE estimates based on a suitable validation data procedure [1]. SIC object status classification helps to analyze the role of each sample not only in the calibration set but also in the test set.

6.1. Norwegian cruise ship data set 'Hurtigruten'

The following example is based on an MSc (Eng) optimization study using PLS, aimed at elucidating the complex relationships between the specific loading and ship's engine settings, the objective weather conditions encountered *en route* and the resulting fuel consumption for a particular Norwegian coastal cruise ship.

The data set used here encompasses seven characteristics of weather conditions and officer-determined ship behavior (e.g. wind and current, engine RPM, propeller pitch, etc.), which make up the **X** matrix, while response **y** records the actual bunker fuel consumption ($1\,h^{-1}$). The training set consists of 27 observations (objects), while the test set includes 18 observations, both carefully laid out in a strongly problem-dependent experimental design while the ship traveled the entire length of the Norwegian coastline (autumn 1998). These observations managed to cover a rather wide range of the ship's running parameters under quite different (hence well-spanning) weather conditions and could therefore be considered as fairly representative at large. Test set validation resulted in two PLS components explaining 45% of X-variance and 99% of Y-variance, with RMSEC = 15.2, RMSEP = 42.5. The data set has been under proprietary confidentiality for 5 years but has recently been released; the basic publication is currently being written up by one of us (K.E.).

On the basis of the final PLS model we have constructed the pertinent SIC model and found eight boundary samples. From Figure 6 it is again easy to see that not all SIC boundary samples are tantamount to objects that have a high 'influence' in the traditional PLS model influence plot; in fact, only three objects stands out here (nos 17, 6 and 20). The disposition of the full boundary set gets more interesting when brushed into the pertinent PLS *t–u* plots, displayed in Figure 7. The particular layout of the boundary samples in these *t–u* plots clearly shows their peripheral positions and again confirms our assumption of their special role in model construction; this is especially predominant in the first-component *t–u* plot.

Now we shall analyze the role of each *test sample* regarding the constructed model for the Norwegian cruise ship test set. Using Statement 1, we can find 11 insiders and seven outsiders (Figure 8(a)). Among the outsiders we can also make a more detailed distinction. Samples 4 and 10 are outsiders, but indeed rather close to the model. There may be two reasons why these samples are not insiders: they could have some large errors in the response values, or their X–Y relation differs from that of the model (it is easy to look closer to evaluate the particulars of such samples as soon as they have been pointed to in the OSP). Samples 1, 2 and 5 are *absolute outsiders*, since their X-data structure is not at all compliant with that of the calibration samples. The SIC prediction intervals are accordingly greater than MED β for these samples. Finally, samples 17 and 18 are *outliers*. We cannot trust in the prediction results for these samples and they should be excluded from further consideration.

The specifics of each sample are reflected in their SIC prediction intervals delineated in Figure 8(b). Let us compare the SIC intervals with the prediction uncertainties for individual samples (Figure 8(b), error bars) calculated by the approaches laid out for PLS prediction in The Unscrambler 8.0 software [37], where the variance in an individual predicted response value is calculated as

$$R(\mathbf{y}_{\text{val}})\left(1-\frac{k+1}{n}\right)\left(h+\frac{R(\mathbf{x})}{R(\mathbf{X}_{\text{val}})}+\frac{1}{n}\right)$$

Here $R(\mathbf{y}_{\text{val}})$ is the mean square residual of responses of the validation set, h is the PLS leverage, $R(\mathbf{x})$ is the mean square residual of predictors of the object, $R(\mathbf{X}_{\text{val}})$ is the mean square residual of predictors of the validation set and k is the rank of the PLS model.

As might be expected, these uncertainties reveal the absolute outsiders well (nos 1, 2 and 5), since these samples differ in their X-structure, and here we see a coincidence with the SIC diagnostics. However, there are neither definite diagnostics for outliers (nos 17 and 18) nor for outsiders from these standard prediction uncertainties. It will be particularly illuminating to compare this test set prediction OSClas on the basis of the model *t–u* relationships with Figure 9.

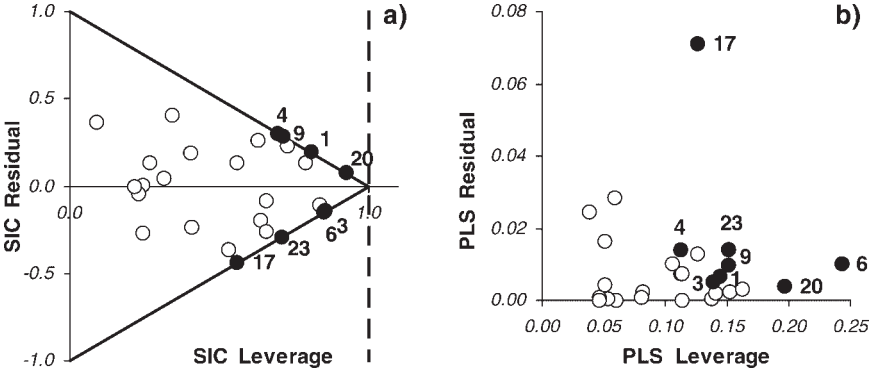

Figure 6. Norwegian cruise ship. (a) SIC object status plot. (b) Traditional 'influence plot'. Training set: ○, insiders; ●, boundary samples.

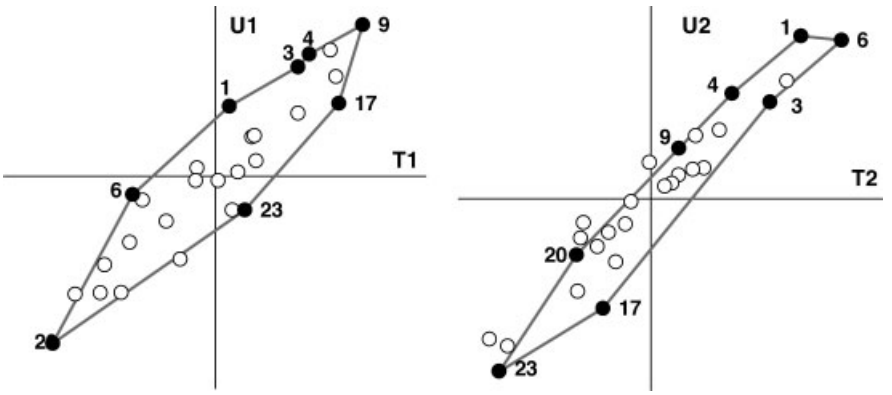

Figure 7. Norwegian cruise ship calibration data set. t–u plots for PLS model with two PCs. Training set: ○, insiders; ●, boundary samples; gray line connects boundary samples.

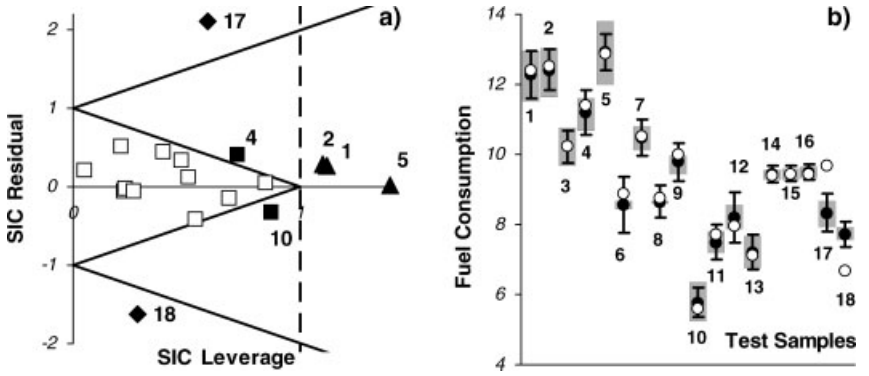

Figure 8. Norwegian cruise ship test data set. (a) SIC object status plot: □, insiders; ■, outsiders; ▲, absolute outsiders; ◆, outliers. (b) SIC prediction: ○, reference values, ▮, SIC prediction intervals; ●, PLS prediction with uncertainty bars.

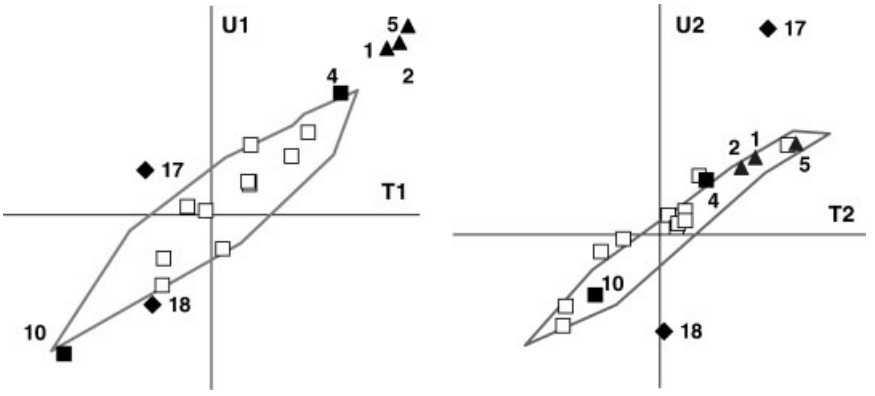

Figure 9. Norwegian cruise ship. t–u plots for PLS model with two PCs. Test set samples: □, insiders; ■, outsiders; ▲, absolute outsiders; ◆, outliers; gray line connects boundary samples.

Figure 9 presents the pertinent model t–u plots in which the test set has been passively projected only. On these plots, all test set samples are annotated in accordance with their OSClas designations. This helps to assess and understand the role of each sample. The gray boundary contours are carried over from Figure 7 (all individual training data set objects are left out for clarity). On the t_1–u_1 plot, one can e.g. graphically appreciate the extreme behavior of samples 1, 2 and 5, while on the t_2–u_2 plot, samples 17 and 18 are evident outliers.

Concluding for this example, we can state that SIC object status classifications not only reveal different kinds of samples among the calibration set but also help to evaluate the test set quality. The results mostly agree with conventional

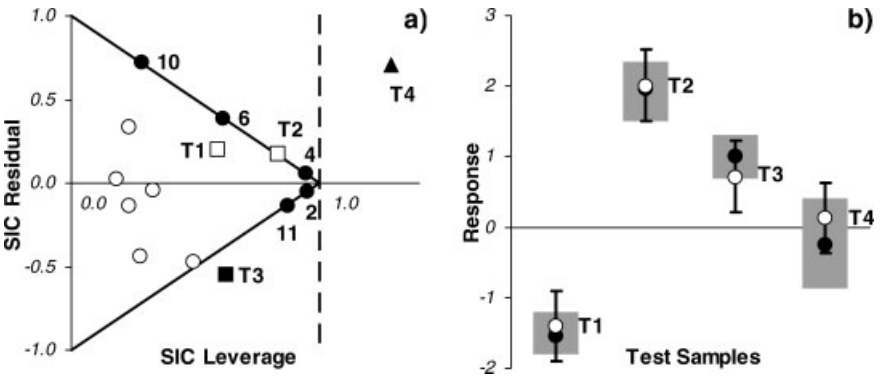

Figure 10. DSC data set. (a) SIC object status plot for training and test sets: ○, insiders; ●, boundary samples; ■, ▲, □, test samples. (b) SIC prediction: ○, reference values with error $\pm b_{SIC}$; ▒, SIC prediction intervals; ●, PCR prediction.

PLS analysis, but the SIC approach yields more detailed information. The *complementary* PLS *t–u* and object status plots reveal the entire object status layout of all objects in both the training set as well as the test set.

7. OUTLIER DETECTION

Once a multivariate calibration model has been established, it is usually used to predict the characteristics of new samples. If a predicted sample is inconsistent with the calibration model, the prediction will be bad (sample with high value for prediction uncertainty); or worse, it will be irrelevant (e.g. sample 4 in Figure 2(a) and samples 17 and 18 in Figure 8(b), for which prediction values and uncertainty intervals are far from reference values). Traditionally, such samples are simply called *outliers*, but there is no consensual theoretical treatment of this class of object in traditional chemometrics, except for a few recent attempts (see Reference [12] and references cited therein). From SIC's point of view, such samples are treated as *absolute outsiders*. For these objects the SIC prediction intervals will be greater than the maximum error deviation β. Statement 4 gives the rule for their determination. This rule is explicit, simple and does not depend on the reference value.

7.1. DSC example

In order to continue the comparison of SIC results and conventional projection methods, we consider as a last example the prediction of antioxidant activity in polypropylene [17]. The **X** matrix includes values of OIT (oxidation initial temperature) measured by differential scanning calorimetry (DSC) at five different heating rates; the **y** vector consists of the corresponding values of long-term heating aging (LTHA), which is a reference characteristic of antioxidant activity. There are again two data sets. The calibration set consists of 11 samples (1–11) with antioxidants of different types. There is a small test set (samples T1–T4) which is used for both validation and prediction. It is not important that both data sets are very small, since their data structure is the decisive factor; also, their small size makes for potentially clear illustration.

We used PCR regression and validated two PCs for this model, which explained 93% of X-variance and 96% of Y-variance. On the basis of this PCR result we construct the

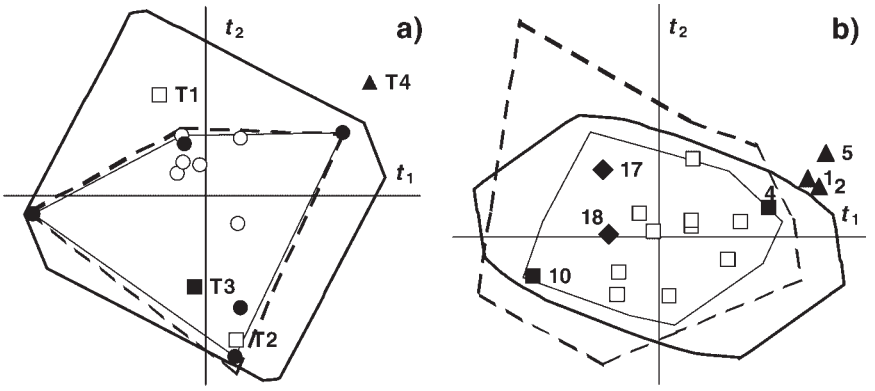

Figure 11. Score plot object status designations (t_1–t_2). (a) DSC example. Training and test sets. (b) Norwegian cruise ship. Test set. Training set: ○, insiders; ●, boundary samples. Test set: □, insiders; ■, outsiders; ▲, absolute outsiders; ◆, outliers. Border of absolute outsiders (—), convex hull (—) and second boundary(— —).

corresponding SIC model and find five boundary samples (Figure 10(a)). Prediction of the test set reveals one absolute outsider, T4 in Figure 10, owing to its abnormally high SIC leverage.

For this example we can explicitly draw the border of absolute outsiders (Statement 5) in the PCA score plot (t–t) and compare this border with the *convex hull* and its *second boundary* as calculated from the comprehensive procedures in Reference [13], which were developed especially for outlier detection in prediction. Figure 11(a) shows this comparison. Sample T4 is detected as abnormal by both methods. However, sample T1, which is designated by SIC as a rather reliable (Figure 10(a)) and well-predicted (Figure 10(b)) object, is *wrongly interpreted* by the convex hull method as an outlier (Figure 11(a)).

7.2. Norwegian cruise ship X-space relationships

We can perform a further X-space outlier detection comparison and return to the Norwegian cruise ship data set. Using this example, we want to show the various outlier areas as they appear in the t–t score plot (Figure 11(b)) as well. Following the rules described in Reference [13], we have constructed the convex hull and its second boundary on the basis of the pertinent training set. However, as we have a rather large data set in this example (23 objects), we only project the test samples in the plot. Abnormal objects are picked up and annotated in the same way as in Figures 8(a) and 9.

Again, samples that are designated as absolute outsiders by SIC (nos 1, 2 and 5) are marked up by the convex hull method as well. Samples 17 and 18 cannot be distinguished as abnormal samples in these score plots by any method, as they differ from the calibration set data with respect to their response values, while they are similar to the calibration samples in their X-structure. Evidently, when analyzing new samples with unknown response values, such a situation cannot be distinguished.

The following reflections pertain to a more detailed comparison of these two methods of outlier (outsider) detection. The border of absolute outsiders is constructed using *all* samples from the training set. We enlarge their score values (Statement 5) in such a way that corresponding SIC leverages become equal to one. Therefore training samples are located *inside* this border (Figure 11(a)). As to the convex hull, it is constructed on the basis of the peripheral objects in the *score space* in such a way that these objects serve as the vertices of the convex hull. As further mentioned in Reference [13], a 'second boundary' is built on the basis of the convex hull, taking into account the uncertainty of the model; in fact, the uncertainty in **X**. In Figure 11(a) this second boundary is very close to the convex hull. This is because we have used the PCR model in which two PCs explain 93% of X-variance. The data set shown in Figure 11(b) (Norwegian cruise ship) represents quite a different PLS model in which the first two components explain only 45% of X-variance. This is why there is a marked visible difference between the convex hull and its second boundary.

The last two examples show that the SIC concept of *absolute outsiders* may be successfully applied for outlier detection in new data sets. In these examples it was possible to make an explicit construction of the border of absolute outsiders, because these models require only two components. Generally, however, we do not need to construct this border explicitly, since we have the simple diagnostic rule (Statement 4) which can be used regardless of the score space dimensionality. This is a significant advantage of SIC.

8. CONCLUSIONS

We have presented a combination of the new SIC approach with the well-known chemometric bilinear projection methods (PCR, PLS), which has been shown to be a powerful and visually simple instrument for detailed analysis of the status of individual objects in both calibration and test data sets. The main thrust of the SIC approach is the object status classification (OSClas), which follows directly from the basic SIC concepts and which is not user-dependent, i.e. it is objective.

The SIC approach offers strict, unequivocal rules for object classification in different cases.

1. All *calibration samples* can be divided into two main classes: *boundary samples*, which are the most influential objects in the modeling, and *insiders*, which are the redundant objects in the training set (Statements 1 and 2).
2. The *test samples* can be SIC typified with detailed and explicit classification (Statements 1, 3 and 4). Thus these objects can be divided into two main classes: *insiders*, which are similar to the calibration samples, and *outsiders*, which are dissimilar to the model. A further distinction can be made between outsiders: *absolute outsiders*, which differ greatly from calibration samples in their X-structure, and *outliers*, which completely contradict the model structure (Statements 1, 3 and 4).
3. For *new samples* we have the strict Statement 4, which also marks up *absolute outsiders* which are inconsistent with the calibration model. It is finally a powerful advantage that the *absolute outsider category* may be found *regardless* of the presence or not of reference y values. This feature is of the highest importance for reliable prediction.

The SIC approach uses no new or extra parameters which cannot be evaluated by the data set and have to be set *a priori*.

9. SOFTWARE

All present SIC calculations were made with software programmed and implemented as an add-in for Excel, including various NIPALS algorithms [1] for bilinear matrix decomposition, a standard Simplex algorithm [28] for optimization, as well as a necessary suite of special procedures, e.g. for preprocessing, transformations, etc. This software is currently in a beta-test version, but all algorithms are of well-known types and may easily be implemented in various standard packages.

REFERENCES

1. Martens H, Naes T. *Multivariate Calibration*. Wiley: New York, 1998.
2. Næs T, Isaksson T, Fearn T, Davies T. *Multivariate Calibration and Classification*. NIR Publications: Chichester, UK, 2002.

3. Esbensen KH. *Multivariate Data Analysis—in Practice* (4th edn). CAMO: Trondheim, 2000.
4. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. *Multi- and Megavariate Data Analysis*. Umetrics: Umeå, 2001.
5. Cook RD. Detection of influential observations in linear regression. *Technometrics* 1977; **19**: 15–18.
6. Cook RD. Influential observations in linear regression. *J. Am. Statist. Assoc.* 1979; **74**: 169–174.
7. Andrews DF, Pregibon D. Finding the outliers that matter. *J. R. Statist. Soc. B* 1978; **40**: 84–93.
8. Draper NR, John JA. Influential observations and outliers in regression. *Technometrics* 1981; **23**: 21–26.
9. Naes T. The design of calibration in near infra-red reflectance analysis by clustering. *J. Chemometrics* 1987; **1**: 121–134.
10. Huber PJ. *Robust Statistics*. Wiley: New York, 1981.
11. Hubert M, Verboven S. A robust PCR method for high-dimensional regressors. *J. Chemometrics* 2003; **17**: 438–452.
12. Hoskuldsson A. *Prediction Methods in Science and Technology*, Vol. 1. Thor Publishing: Copenhagen, 1996.
13. Fernandez Pierna JA, Wahl F, de Noord OE, Massart DL. Methods of outlier detection in prediction. *Chemometrics Intell. Lab. Syst.* 2002; **63**: 27–39.
14. Faber K, Kowalski B. Propagation of measurement errors for the validation of prediction obtained by principal component regression and partial least squares. *J. Chemometrics* 1997; **11**: 181–238.
15. Hoy M, Steen K, Martens H. Review of partial least squares regression prediction error in Unscrambler. *Chemometrics Intell. Lab. Syst.* 1998; **44**: 123–133.
16. Faber K. Comparison of two recently proposed expressions for partial least squares regression prediction error. *Chemometrics Intell. Lab. Syst.* 2000; **52**: 123–134.
17. Pomerantsev AL, Rodionova OYe. Prediction of antioxidants activity using DSC measurements: a feasibility study. In *Aging of Polymers, Polymer Blends and Polymer Composites*, Vol. 2, Zaikov GE, Buchachenko AL, Ivanov VB (eds). Nova Science Publishers: New York, 2002; 19–29.
18. Westad F, Martens H. Variable selection in NIR based on significance testing in partial least squares regression (PLSR). *J. Near Infrared Spectrosc.* 2000; **8**: 117–124.
19. Hoskuldsson A. Variable and subset selection in PLS. *Chemometrics Intell. Lab. Syst.* 2001; **55**: 23–38.
20. Gusnanto A, Pawitan Y, Huang J, Lane B. Variable selection in random calibration of near-infrared instruments: ridge regression and partial least squares. *J. Chemometrics* 2003; **17**: 174–185.
21. Dantas Filho HA, Harrop Galvao RK, Ugulino Araujo MC, da Silva C, Bezerra Saldaha TC, Jose GE, Pasquini C, Raimundo Jr. IM, Rodrigues Rohwedder JJ. A strategy for selecting calibration samples for multivariate modeling. *Chemometrics Intell. Lab. Syst.* 2004; **72**: 83–91.
22. Clancey VJ. Statistical methods in chemical analyses. *Nature* 1947; **159**: 339–340.
23. Rajkó R. Treatment of model error in calibration by robust and fuzzy procedures. *Anal. Lett.* 1994; **27**: 215–228.
24. Gass S. *Linear Programming* (4th edn). McGraw-Hill: New York, 1975.
25. Kuhn HW, Tucker AW. *Linear Inequalities and Related Systems, Vol. 38 of Annals of Mathematics Studies*. Princeton University Press: Princeton, NJ, 1956.
26. Lehmann EL. *Testing Statistical Hypotheses*. Wiley: New York, 1960.
27. Eicker F. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Ann. Math. Statist.* 1963; **34**: 447–456.
28. Dantzig G. *Linear Programming and Extensions*. Princeton University Press: Princeton, NJ, 1963.
29. Taha H. *Operations Research: An Introduction* (3rd edn), Vol. 1. MacMillan: New York, 1982.
30. Jiang J-H, Liang Y-Z, Ozaki Y. On simplex-based method for self-modeling curve resolution of two-way data. *Chemometrics Intell. Lab. Syst.* 2003; **65**: 51–65.
31. Rodionova OYe, Pomerantsev AL. Principles of simple interval calculations. In *Progress in Chemometrics Research* (ISBN: 1-59454-257-0), Pomerantsev AL (ed.). Nova Science Publishers: New York, 2005; 39-48.
32. Gumbel E. *Statistics of Extremes*. Columbia University Press: New York, 1962.
33. Bouveresse E, Massart DL. Standardization of near-infrared spectrometric instruments: a review. *Vibr. Spectrosc.* 1996; **11**: 3–15.
34. Wang Y, Veltkamp DJ, Kowalski BR. Multivariate instrument standardization. *Anal. Chem.* 1991; **63**: 2750–2756.
35. Jouan-Rimbaud D, Massart DL, Saby CA, Puel C. Characterization of the representativity of selected sets in multivariate calibration and pattern recognition. *Anal. Chim. Acta* 1997; **350**: 149–161.
36. Sulima EL, Zubkov VA, Rusinov LA. 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 *Progress in Chemometrics Research* (ISBN: 1-59454-257-0), Pomerantsev AL (ed.). Nova Science Publishers: New York, 2005, 196–203.
37. Camo. *The Unscrambler's User's Guide, Version 8.0*. Camo: Trondheim, 2004.