Anal Bioanal Chem (2004) 380: 397–418
DOI 10.1007/s00216-004-2768-x

REVIEW

**M. Forina · S. Lanteri · M. C. Cerrato Oliveros
C. Pizarro Millan**

Selection of useful predictors in multivariate calibration

Received: 23 February 2004 / Revised: 9 July 2004 / Accepted: 12 July 2004 / Published online: 3 September 2004

Abstract Ten techniques used for selection of useful predictors in multivariate calibration and in other cases of multivariate regression are described and discussed in terms of their performance (ability to detect useless predictors, predictive power, number of retained predictors) with real and artificial data. The techniques studied include classical stepwise ordinary least-squares (SOLS), techniques based on the genetic algorithms, and a family of methods based on partial least-squares (PLS) regression and on the optimization of the predictive ability. A short introduction presents the evaluation strategies, a description of the quantities used to evaluate the regression model, and the criteria used to define the complexity of PLS models. The selection techniques can be divided into conservative techniques that try to retain all the informative, useful predictors, and parsimonious techniques, whose objective is to select a minimum but sufficient number of useful predictors. Some combined techniques, in which a conservative technique is used to perform a preliminary selection before the use of parsimonious techniques, are also presented. Among the conservative techniques, the Westad–Martens uncertainty test (MUT) used in Unscrambler, and uninformative variables elimination (UVE), developed by Massart et al., seem the most efficient techniques. The old SOLS can be improved to become the most efficient parsimonious technique, by means of the use of plots of the *F*-statistics value of the entered predictors and comparison with parallel results obtained with a data matrix with random data. This procedure indicates correctly how many predictors can be accepted and substantially reduces the possibility of overfitting. A possible alternative to SOLS is iterative predictors weighting (IPW) that automatically selects a minimum set of informative predictors. The use of an external evaluation set, with objects never used in the elimination of predictors, or of "complete validation" is suggested to avoid overestimate of the prediction ability.

Keywords Multivariate calibration · Predictor selection

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00216-004-2768-x

M. Forina (✉) · S. Lanteri · M. C. C. Oliveros
Department of Pharmaceutical and Food Chemistry
and Technology, University of Genova,
Via Brigata Salerno (s/n), 16147 Genova, Italy
E-mail: forina@dictfa.unige.it
Tel.: +39-10-3532630
Fax: +39-10-3532684

C. P. Millan
Department of Chemistry,
University of La Rioja, C/Madre de Dios 51,
26006 Logroño (La Rioja), Spain

Introduction

Multivariate regression techniques are widely used for multivariate chemical calibration (determination of chemical quantity from physical measured quantities), multivariate physical calibration (physical quantity as octane index, viscosity from other physical quantities), multivariate sensory calibration (panel scores from physical or chemical quantities), and for study of the relationship between property (retention time, partition coefficient, biological activity) and molecular structure.

In recent years multivariate chemical calibration was first applied in NIR spectroscopy, then in other wavelength intervals, and in "electronic" noses and tongues, whenever the information from the physical instrument is not specific. In many cases, the process to reach conditions of specificity, where the physical quantity is related unequivocally to the chemical quantity, requires time and chemicals, and cost or time are too large compared with the requirements.

In these cases, multivariate calibration solves a complex system of equations and performs a "mathematical separation". From the multivariate regression model the chemical quantity can frequently be estimated with reasonable, sufficient, accuracy, generally with a minimum sample treatment, almost always very quickly and at very low cost.

Calibration requires standards. In real problems where multivariate calibration is applied standards are usually samples (below OBJECTS) for which the chemical quantity (below the RESPONSE) is known, obtained by a reference method. Because of the cost/time of the reference technique the number of such "standards" is not usually very large, often between 50 and 100. For each sample many physical quantities (below the PREDICTORS) are measured, frequently many hundred. So, the classical multivariate regression technique, the ordinary least-squares (OLS) regression, cannot be applied, and a "biased" technique such as principal-components regression (PCR) or partial-least squares (PLS) regression must be applied.

Also ridge regression (RR) can work when the number of predictors is larger than the number of objects, but it is very rarely applied in multivariate chemical calibration, where PLS and, less often, PCR are applied. The use of artificial neural networks (ANN) is less frequent and very often the nets work on the principal components to avoid too heavy overfitting.

A second possibility is that of the selection of a reduced number of predictors. The selection can be based on previous knowledge of the informative predictors (really chemical experience should be always used to select an optimum set of predictors) or by means of another biased technique, e.g. stepwise ordinary least-squares (SOLS) regression.

A second important reason to eliminate predictors is that frequently many predictors are useless, and cause worsening of the predictive ability of the regression model. Sometimes, the elimination of these noisy predictors can be important for economy (e.g. in process control with reduction of the number of sensors) or can help in the interpretation—very important in sensory analysis and in quantitative structure–activity relationships (QSAR).

For this reason many techniques have been developed to eliminate useless predictors. The techniques for the elimination of useless predictors can be classified into three categories:

(a) Subset selection, e.g. by means of SOLS or of genetic algorithms (GA) [1–3] coupled with OLS or with PLS.

(b) Dimension-wise selection. Dimension-wise techniques work on the single dimension (principal component of PCR, latent variable of PLS) of the regression technique. Martens and Naes [4] suggested replacement with zero of the small PLS weights in each latent variable, so that the corresponding predictors are cancelled from the latent variable, but they can be used in one or more of the subsequent steps. Frank [5] improved this procedure in the technique called intermediate least-squares (ILS). Other strategies were used by Kettaneh-Wold et al. [6], Lindgren et al. [7], and Forina et al. [8].

(c) Model-wise elimination. The regression model is developed many times with all the predictors but with only a fraction of the available objects. The useless predictors are eliminated on the basis of the value and/or the dispersion of their regression coefficient b in the regression model:

$$y = b_0 + b_1 x_1 + \cdots + b_v x_v + \cdots + b_V x_V \tag{1}$$

Alternatively, many regression models are developed with all the objects but only a fraction of the predictors, and the useless predictors are eliminated on the basis of their participation (or not) in models with better prediction ability.

(d) Interval or group selection. In spectroscopy "contiguous" predictors correlate very well, because they are the absorbance at close wavelengths. Grid predictors of QSAR measure the interaction between a probe at a point in the space around a molecule and the molecule. Here, the contiguity is a neighbor in space, and contiguous predictors correlate very well. So, in these cases the interest is in a group or interval of predictors, not in the single predictor, and techniques to select clusters of contiguous predictors have been developed for QSAR [9, 10] and for multivariate calibration [11, 12]. Group selection will not be treated here.

Here, we will present and discuss some results obtained with some selection techniques, selected on the basis of availability of related software, of the simplicity of the theory (so that they can be easily understood by chemists without a high chemometric background), and of our experience.

For these reasons, we will treat some methods of subset selection (SOLS, GA–OLS) and of model-wise elimination (ISE, UVE, IPW, GOLPE, MUT, MAXCOR). In some cases, we will compare the results with those obtained on the same data by other authors with least absolute shrinkage and selection operator (LASSO) and variable selection (VS), used for GA coupled with PLS.

With the exception of SOLS, the selection techniques used here are based on evaluation and optimization of the predictive ability of the regression model. Almost all the techniques presented (with the exception of SOLS and GA–OLS) are based on PLS regression.

For all these reasons this paper presents first:

- A short introduction to the evaluation strategies
- A description of the quantities used to evaluate the regression model
- The criteria used to define the complexity of PLS models.

Evaluation

Many strategies are used to evaluate the predictive ability of the regression model. Predictive ability measures the error on objects that are not used to build the regression model.

The usual two-sets strategy divides, one or more times, the available objects (samples with known value of the response) into two sets, the training set, used to compute the model parameters, and the evaluation set, used to measure the error of prediction.

The PLS and PCR use explicitly the evaluation set(s) to evaluate the optimum complexity of the model (how many latent variables, how many principal components). Techniques such as GA, iterative stepwise elimination (ISE), and iterative predictors weighting (IPW) use more or less explicitly the evaluation set(s) to optimize the selection of useful predictors. So, these evaluation sets are really sets for "predictive optimization". This is not very important for use in PLS and PCR. In other cases, the final regression model (after elimination of useless predictors) should be evaluated on a number of samples never used in the development/refinement of the model, the "external" or "third" set, a true evaluation set.

The procedures most frequently used to obtain a suitable evaluation set are:

- Ordering of the samples according to the value of the response. Division of ordered samples into C cancellation groups. Selection of one or more C cancellation groups for the unique evaluation set. The other objects are used also for predictive optimization, frequently by means of cross-validation (CV).
- Casual subdivision by means of the generation of random numbers in two or three sets; sometimes only the true evaluation set is obtained by random generation and the other objects are used for predictive optimization also.
- CV. Many times the total number of objects is divided into sets for calibration and evaluation with the systematic procedure of CV. In turn, each calibration set is divided into training set and optimization set.

Two-sets N-fold CV assigns the objects to N cancellation groups, by ordered numbering 1 2 3 ... k ... N. The objects numbered with k are assigned to the k-th evaluation group. They will be used to constitute the k-th evaluation set, and the other will constitute the k-th training set. The final model is built with all the objects, so that $N+1$ models are computed.

When N is equal to the total number of objects, the evaluation sets contain only one object. Obviously, the leave-one-out or jackknife procedure requires more computing time. Two-sets CV is probably the most used validation technique. Three-sets NM CV build N true evaluation sets. For each subdivision (+1 without evaluation set) uses M-fold CV for the predictive refinement of the model, so that $(M+1)(N+1)$ models are computed. Because of the long time required this procedure is used very rarely.

- Monte Carlo validation or multiple evaluation set. Often the total number of objects is divided into sets for calibration and evaluation by random assignment with a pre-selected probability of assignment to one of the sets. In turn, each calibration set is divided into training set and optimization set. The number of models computed can be very large, with very different combinations of the objects in the sets.
- Methods of experimental design, especially one of the techniques for uniform design, e.g. the Kennard–Stone [13] and twin Kennard–Stone algorithms [14].

Some people indicate with the name "full validation" the leave-one-out procedure. The word "full" has a positive significance, but the leave-one-out procedure has been criticized [15, 16] because it is too optimistic in the evaluation of the predictive ability. We prefer to use the words "full validation" to indicate the repeated use of validation, e.g. with threefold, fivefold, tenfold CV all repeated after random sorting of the objects, or with a large use of Monte Carlo validation.

The results presented here have been obtained with two or three sets. In the case of two sets, the training and evaluation sets have been obtained with CV with five cancellation groups or with the leave-one-out procedure. The three-sets procedure has been used for simulated data only, when it is possible to have an evaluation set with many objects, so that the evaluation of the prediction error is accurate, without the possibility of casual overestimation as can happen for an evaluation set with too few objects. The other objects were in turn divided between training and optimization sets by fivefold or leave-one-out CV.

Quantities used for evaluation of regression models

Many quantities are used as a measure of the prediction error, frequently with different acronyms for the same quantity.

Root-mean-square error of prediction (RMSEP) [17]

(The use of the word "error" to indicate the estimate of a standard deviation is frequent in multivariate calibration.)

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{I_\mathrm{E}} (y_i - \hat{y}_i)^2}{I_\mathrm{E}}} \tag{2}$$

where I_E is the number of samples in the "external" evaluation set (third set).

Root-mean-square error of cross validation (RMSECV) [17]

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{I}} \tag{3}$$

where I is the number of calibration samples; $\hat{y}_i$ is the predicted response when the model is built without sample i. RMSECV, as defined in [17], is really a RMSELOU, leave-one-out.

Frequently, RMSECV is indicated with standard error of prediction (SEP), independently of the number of CV groups, or with RMSEP, as for an external evaluation set, or with standard deviation of the error of prediction (SDEP).

The SEP has been used also for standard error of performance, a strange mixture of prediction and fitting:

$$\mathrm{SEP}^{\mathrm{performance}} = \sqrt{\frac{\sum_{i=1}^{I}(y_i - \hat{y}_i - \mathrm{bias})^2}{I-1}} \quad \text{with bias} = \frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)}{I} \tag{4}$$

Kowalski and Seasholtz [17] suggests also the approximation (valid in the case of OLS with leave-one-out prediction):

$$y_i - \hat{y}_i = \frac{y_i - \tilde{y}_i}{1 - h_i} \tag{5}$$

For the sample i, $\tilde{y}_i$ is the fitting estimate of the response and h_i is the leverage.

Equation (5) is valid for OLS with leave-one-out prediction, and it is largely used in the leave-one-out diagnosis of OLS.

Percentage CV-explained variance (Q^2)

This frequently used quantity is strictly connected with the RMSECV:

$$Q^2 = 100\,\frac{\sum_{i=1}^{I}(y_i - \bar{y}_{\mathrm{g}})^2 - \sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{I}(y_i - \bar{y}_{\mathrm{g}})^2} \tag{6}$$

where $\sum_{i=1}^{I}(y_i - \bar{y}_{\mathrm{g}})^2$ is the before-regression variance. The mean of the response must be evaluated in each CV segment g.

Predictive residual-error sum of squares (PRESS, from RMSECV) [17]

$$\mathrm{PRESS} = \sum_{i=1}^{I}(y_i - \hat{y}_i)^2 \tag{7}$$

$$\mathrm{RMSECV} = \sqrt{\frac{\mathrm{PRESS}}{I}} \tag{8}$$

Mean-square error of prediction [18]

$$\mathrm{MSEP} = E(y - \hat{y})^2 \tag{9}$$

(The operator E indicates the mean of the infinite population.) $\widehat{\mathrm{MSEP}}$ is an estimate of MSEP with I predictions:

$$\widehat{\mathrm{MSEP}} = \frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{I} = \frac{\mathrm{PRESS}}{I} \tag{10}$$

Sometimes MSEP is used to indicate $\widehat{\mathrm{MSEP}}$.

Criteria

Many techniques have been suggested for evaluation of the pseudo-rank of the matrix of the predictors, i.e. the optimum complexity of the regression model. Some criteria can be applied also to compare regression models from different techniques.

1. *Minimum PRESS (CV or Leave-one-out)* The common practice of calculating $\widehat{\mathrm{MSEP}}$ for a score of models and just selecting the model with the lowest value has been called, derogatorily, the "European Song Contest method" (A.G. Steerneman, reported in [18]).
2. *First minimum PRESS* Selects the number of components corresponding to the local minimum in PRESS associated with the smallest possible number of components.
3. *Haaland–Thomas F-statistics [19, 20]*

$$F_a(I, I) = \frac{\mathrm{PRESS}_a}{\mathrm{PRESS}_{\mathrm{m}}} \tag{11}$$

$\mathrm{PRESS}_{\mathrm{m}}$ is the minimum of PRESS, when the number of components is A, which is estimated by means of criterion 1.

The optimum complexity is the smallest value of $a < A$ such that F is not significant.

However, "*While we believe that the procedures outlined above for model selection and comparison of methods will work reasonably well in practice, we admit that they could be improved by taking into consideration the correlation of prediction errors.*" [19].

4. *Osten F-statistics [21]*

$$F_a(1, I - a - 1) = \frac{\mathrm{PRESS}_a - \mathrm{PRESS}_{a+1}}{\mathrm{PRESS}_{a+1}}(I - a - 1) \tag{12}$$

The optimum complexity is the smallest value of a such that F is not significant.

5. *PRESS threshold [21]* Selects the smallest number of components for which the PRESS value is below a set threshold. The threshold is calculated as a percentage between the maximum and minimum observed PRESS values. Osten [21] indicates for this percentage 3 or 5%.
6. *t-Test on the difference between two means of absolute prediction errors* The means of the absolute prediction errors

$$m_a = \frac{\sum_{i=1}^{I} |y_i - \hat{y}_{ai}|}{I} \quad m_b = \frac{\sum_{i=1}^{I} |y_i - \hat{y}_{bi}|}{I} \tag{13}$$

are computed.

As usual the two variances are compared by means of a variance-ratio test. If the variances are not significantly different, the pooled standard deviation s is computed, and the t value for the test is:

$$t = \frac{|m_a - m_b|}{s\sqrt{\frac{2}{I}}} \tag{14}$$

7. Matching pairs t-statistics (residuals) Matching pairs tests take into account the correlation between the prediction errors.

For $a < A$, the differences:

$$d_i = |e_{ai}| - |e_{mi}| = |y_i - \hat{y}_{ai}| - |y_i - \hat{y}_{mi}| \tag{15}$$

are computed.

s_d is the standard deviation of the difference.

The value of the t-statistic is

$$t = \frac{\bar{d}}{s_d/\sqrt{I}} \tag{16}$$

The null hypothesis is that $\bar{d}$ is not significantly different from 0; the alternative hypothesis is that $\bar{d}$ is significantly < 0 (unilateral left).

8. Matching-pairs t-statistics (square residuals) [18] For $a < A$, the differences:

$$d_i = e_{ai}^2 - e_{mi}^2 \tag{17}$$

are computed.

The mean of these differences is:

$$\bar{d} = \frac{\sum_{i=1}^{I} [(y_i - \hat{y}_{ai})^2 - (y_i - \hat{y}_{mi})^2]}{I} = \hat{\mathrm{M}}\mathrm{SEP}_a - \hat{\mathrm{M}}\mathrm{SEP}_m \tag{18}$$

s_d is the standard deviation of these differences.

The value of the t-statistic is:

$$t = \frac{\bar{d}}{s_d/\sqrt{I}} \tag{19}$$

The null hypothesis is that $\bar{d}$ is not significantly different from 0; the alternative hypothesis is that $\bar{d}$ is significantly < 0 (unilateral left).

9. Van der Voet randomization test [18] The t-randomization test of Van der Voet [18] is based on the null hypothesis of equal distribution of the squared residuals.

The algorithm for the one-side (unilateral left) hypothesis $\hat{\mathrm{M}}\mathrm{SEP}_a < \hat{\mathrm{M}}\mathrm{SEP}_m$ is:

1. Calculate the differences $d_i = {e_{ai}}^2 - {e_{mi}}^2$
2. Compute $\bar{d}$
3. Iterate M times:
   (a) Attach random signs to each d_i
   (b) Compute $\bar{d}_m$
4. Rank $\bar{d}$ in position K between the $M\bar{d}_m$ values. Ranking must be from low to high.
5. Calculate the significance level: $p = K/(M+1)$
6. Accept the alternative hypothesis for $p > 10\%$.

The results presented here were obtained:

– with the criterion of the minimum PRESS, according with the statement of Faber [22] in his study comparing many criteria: "... *the conventional methods of optimizing the prediction error estimate obtained from a test set or cross-validation are to be preferred*," or
– with the Osten [21] 5% threshold criterion.

Moreover we will indicate the SDEP with SEP, because this was the acronym first used in multivariate calibration and still used in important packages of chemometric methods.

Methods for selection of predictors

Stepwise ordinary least-squares

Ordinary least-squares regression [23, 24], known also as multi linear regression (MLR) or classical least squares regression (CLS) was the result of the studies of Francis Galton and Karl Pearson more than a century ago.

Stepwise ordinary least-square [23] or stepwise multiple regression is the oldest among the selection techniques used here.

In each step, the "before" sum of the squared residuals with the $V-1$ predictors entered in the previous steps (the mean of the response for the N objects in the first step) is computed:

$$S^2_{\mathrm{before}} = \sum_{i=1}^{N} (y_i - \hat{y}_i^{\mathrm{before}})^2 \tag{20}$$

All the non-entered predictors are tested, and for each predictor p a regression equation is computed with the $V-1$ previously entered predictors and with predictor p. The "after" sum of the squared residuals is:

$$S^2_{\mathrm{after}} = \sum_{i=1}^{N} (y_i - \hat{y}_i^{\mathrm{after}})^2 \tag{21}$$

The ratio:

$$F = S^2_{\mathrm{before}} - S^2_{\mathrm{after}} S^2_{\mathrm{after}} \tag{22}$$

is a Fisher variable with 1 df for the numerator and $N-V-1$ df for the denominator. The predictor with the largest value of F is accepted, provided that F is larger than a critical value (F-to-enter).

Usually the critical value for the F test is 4 (this is a good approximation of $F_{1,v}^{95\%}$ for $N > 20$). A similar test with a critical F-to-remove is used to check the reliability of the entered predictors.

Predictors that produce collinearity are not considered for entering. Their F-to-enter value is considered to be 0. Predictors correlating very well with one or more of the entered predictors do not reduce the variance significantly, so that their F-to-enter value is very small.

Here, SOLS has been used associated with fivefold CV, which is not usual. SOLS was repeated $5+1$ times, with possibly different selection of predictors in each run. The final run was used to select the predictors, but the other runs indicate the "stability" of the selections. The leave-one-out prediction error was obtained from the final run, as a result of the usual leave-one-out diagnostics of OLS regression [24]. Also CV was performed with the predictors selected in the final run. However, different prediction parameters can be obtained by the evaluation sets in the five runs used to evaluate the stability of the selection of useful predictors, and in the five runs the selected predictors are not necessarily the same, or the same as that selected in the final run, with all the objects. Here, these prediction parameters will be reported as "Complete-C.V." and as a remark, because all other techniques perform only once selection of useful predictors using all the objects, so that the modified SOLS seems to be too much penalized in comparison with the other techniques.

Stepwise ordinary least-squares is characterized by the values of the critical F-to-enter and of F-to-remove for the Fisher test. Moreover the user can decide the maximum number of predictors to be selected. Finally, the "goodness" of the information matrix $\mathbf{X}^T\mathbf{X}$ ($\mathbf{X}$ is the matrix of the predictors), measured from its determinant, or from the inflation factor of the predictors, or from the confidence interval (CI) of the regression coefficients, can suggest further reduction of the number of selected predictors.

The maximum number of predictors selected in the use of SOLS is indicated below in parenthesis, e.g. SOLS (20) when the algorithm can select a maximum of 20 predictors.

Interactive stepwise elimination (ISE)

The elimination of predictors with small regression coefficients, (provided that the regression model has been computed with autoscaled predictors) was suggested by Massart and co-workers [25]. ISE [26] is the evolution of the above technique, in the sense that it can be applied also when predictors are non-standardized. ISE is based on the importance of the predictors, defined as:

$$z_v = \frac{|b_v|s_v}{\sum_{v=1}^{V} |b_v|s_v} \qquad (23)$$

It takes into account both the size of the regression coefficient and the dispersion of the predictor, because s_v is the standard deviation of the predictor v. A small value of $|b_v|s_v$ indicates that the range of the contribution of the predictor to the response is small, frequently much less than the error of the reference technique used for determination of the response. After each ISE cycle the predictor with the minimum importance is eliminated, and the model is computed again with the remaining predictors. The final model is that with the maximum predictive ability.

The ISE can also eliminate, after each cycle (the first cycle is that with all the predictors), some predictors, if their importance is less than two to ten times that of the worst predictor, with no more than 3–5% predictors eliminated in each cycle.

Figure 1 shows how predictive ability changes in ISE for the data set *Moisture*, with 19 predictors, described later. In the examples with *Moisture* shown in this section both predictors and response were always simply centered.

The minimum SEP is obtained after elimination of 17 predictors, so that only two predictors are retained in the refined PLS model.

Iterative predictors weighting

Iterative predictors weighting (IPW) [27] is one of the techniques developed to eliminate useless predictors associated with the PLS algorithm. IPW starts with the usual PLS regression with centered or autoscaled variables. The importance, the product of the absolute value of the regression coefficient b_v and the standard deviation s_v (s_v is 1 for autoscaled data) of each predictor, weights each predictor in the next step. So, a predictor with a small b_v will, in the next IPW cycle, have a smaller covariance with the response and consequently a smaller b_v, and finally, after two-ten cycles, it will be eliminated ($b_v = 0$).

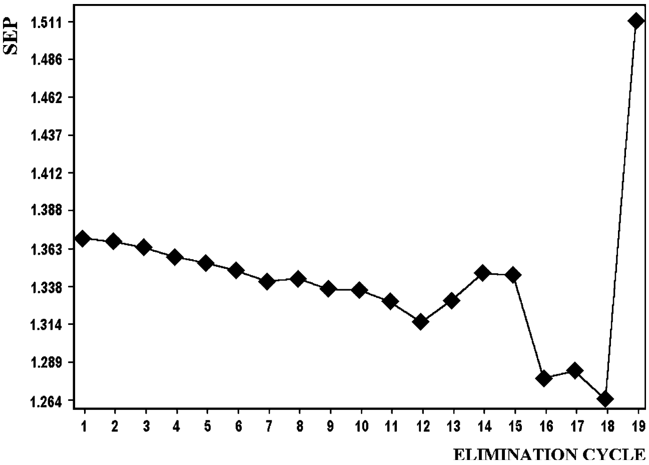


Fig. 1 Changes of predictive ability in ISE cycles (data set *Moisture*)

Figure 2 shows how the importance of the predictors varies in the IPW cycles for the *Moisture* data set.

Uninformative variables elimination

Elimination of uninformative variables (UVE) [28] adds to the original predictors an equal number of random predictors, with very small value (range of about 10^{-10}), so that their influence on the regression coefficients of the original predictors is negligible. The standard deviation of the regression coefficients, s_{b_v} is obtained from the variation of the coefficients b by leave-one-out jack-knifing. The reliability of each predictor v, c_v, is obtained by:

$$c_v = \frac{b_v}{s_{b_v}} \tag{24}$$

The maximum of the absolute value of the coefficient c_v for the added artificial predictors is the cut-off value for elimination of non-informative original predictors. There are some variants of UVE, for example α-UVE in which the cut-off value is the value of the α% quantile of the coefficient c_v of the added artificial predictors. Here we used 90%, 95%, and Normal UVE. Figure 3 shows the results obtained by 90%-UVE on data set *Moisture*.

Generating optimal linear PLS estimation

Generating optimal linear PLS estimations (GOLPE) [9, 10] is a procedure used frequently in QSAR, very rarely in multivariate calibration.

In a first step GOLPE selects predictors according to their position in the PLS loading space, following a D-optimal design criterion (predictors with small loadings on the first PLS components are usually noisy predictors). In QSAR studies, this first selection is performed by removing constant predictors and predictors with few levels, a common procedure for molecular descriptors.

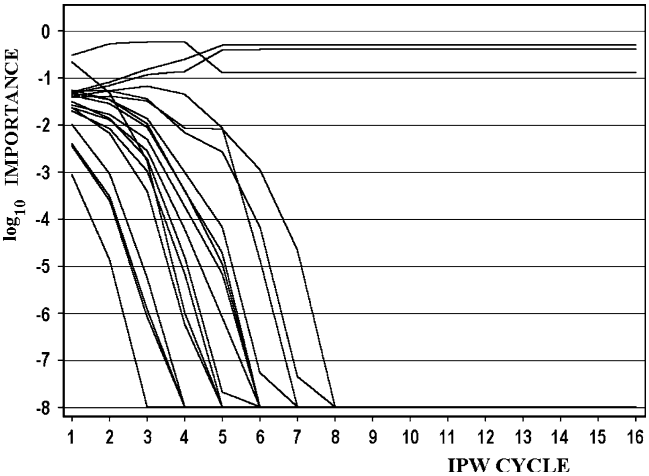


Fig. 2 Evolution of the importance of the predictors in IPW cycles (data set *Moisture*)

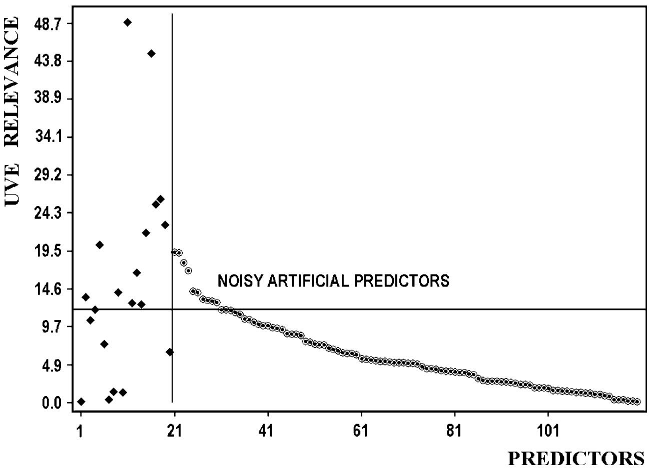


Fig. 3 UVE evaluation of the 19 predictors of data set *Moisture* (100 noisy variables). Noisy added variables are ordered according to their relevance

In a second step, GOLPE builds a large number of "reduced models" similar to the complete model but removing some variables. The predictive ability of each model is evaluated using CV and, from these values, GOLPE relates the predictive ability of the model to the presence or absence of each X-variable.

The strategy used by GOLPE is to make a "design matrix" with M "experiments", following a fractional factorial design (FFD) scheme. According to this matrix, in each experiment some predictors are left-out, so that a "reduced" PLS model is built, and the CV-predicted response measures the predictive ability.

In our implementation the design matrix as built by random generation at a pre-selected probability level p of the predictor condition (p: probability of non-used, $1-p$ probability of used). To equilibrate the design, the range of the predictor condition non-used (expectation pM) was limited between 0.5 and 1.5 pM for each predictor. The M is as large as desired (not a power of two as in FFD) and it is possible to explore more combinations of used predictors.

When the M models have been obtained and SEP for each one has been calculated, GOLPE computes how the presence or absence of a given variable affects the SDEP value. The effect of a predictor is obtained from the equation:

$$E_v = \mathrm{SEP}_{v+} - \mathrm{SEP}_{v-} \tag{25}$$

where E_v is the effect of predictor v; SEP_{v+} is the average SEP for all the models that include the variable, and SEP_{v-} is the average SEP for all the models that do not include the variable.

E is negative when the predictor reduces the standard deviation of the error, i.e. when it increases the predictive ability of the model.

The GOLPE introduces in the design matrix some "dummy" predictors that appear in the design matrix but have no equivalent in the **X**-matrix. The effect of the true predictors is compared with the effect of dummy predictors, by means of the 95% value of the Student distribution.

The confidence interval of the effect of the dummies (CID) is computed:

$$\mathrm{CID} = \sqrt{\frac{\sum_{d=1}^{D} E_d^2}{D}} t_{\mathrm{crit}} \tag{26}$$

where E_d is the effect of the dummy predictor d and D is the number of dummy predictors. This number is limited in the original GOLPE, because the dummies are included in the design matrix. In our implementation the dummy predictors are casually set to present–no present, as the true predictors. Their number has no influence on the number M of experiments, and a very small effect on the computer time; t_{crit} is the 95% critical value of Student distribution.

Instead of the above equation we use:

$$\mathrm{CID} = \sqrt{\frac{\sum_{d=1}^{D} \left(E_d - \bar{E}\right)^2}{D}} t_{\mathrm{crit}} \tag{27}$$

where the standard deviation of the effect of the dummy predictors is computed in the usual way (the mean value of E for dummy predictors is approximately 0 when a large number of dummy predictors is used).

The effect of each variable (E) is compared with the effect of the dummies (CID). From this comparison the variables are labeled as fixed, excluded, or uncertain:

$E > \mathrm{CID}$	the predictor increases SEP substantially (decreases predictivity). *Excluded.*
$E < \mathrm{CID}$ and $E \geq 0$	the predictor increases SEP, but within the uncertainty measured by means of the dummy variables. It is *uncertain.*
$E < \mathrm{CID}$ and $E < 0$	the predictor reduces SEP (increases predictivity). *Retained.*
$E < -\mathrm{CID}$	the predictor substantially reduces SEP. *Excellent.*

We compute PLS models after elimination of excluded predictors (elimination level I), then after elimination of uncertain predictors (elimination level II), and finally only with the excellent predictors (elimination level III).

This last elimination level and the score "Excellent" was introduced in our implementation. Figure 4 shows the results obtained by GOLPE on data set A.

Martens uncertainty test

The Martens uncertainty test (MUT) [29, 30] is based on the standard deviation of the regression coefficients b_v, computed from their values in the cycles of leave-one-out CV. The predictors for which the hypothesis $\beta_v = 0$ is accepted at significance level 5% are eliminated (Student t-test).

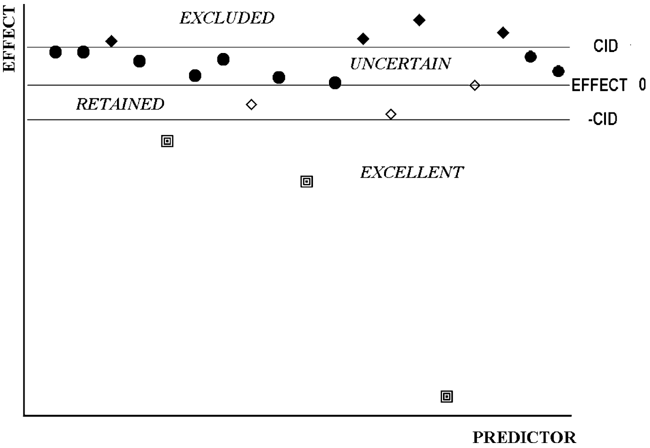


Fig. 4 GOLPE evaluation of the 19 predictors of data set *Moisture* (100 dummy predictors, 500 reduced models. $p = 10\%$, $pM = 50$. Predictors were ignored 30–70 times in the reduced models)

The standard deviation of the regression coefficients in the CV cycles is obtained by the equation:

$$s_{bv}^2 = \frac{\sum_i^N \left(b_{v(i)} - \bar{b}_v\right)^2}{N-1} \tag{28}$$

where N is the number of objects, $b_{v(i)}$ is the value of b_v when object i is left out, and $\bar{b}_v$ is the mean of the N $b_{v(i)}$.

MUT estimates the standard deviation of the regression coefficients by means of the equation [30, 31]:

$$s_{bv,\mathrm{jack}}^2 = \frac{\sum_i^N \left(b_{v(i)} - \bar{b}_v\right)^2}{N}(N-1) \tag{29}$$

the so-called jackknife variance estimator, that corresponds to a value of the estimated standard deviation approximately √N times larger than that obtained by use of Eq. 28. So, taking into account that the perturbation of the leave-one-out procedure is rather small, and that the variability in the coefficients b_v is smaller when the size N of the sample is increased.

MUT was used here with the 95% probability value for the CI.

Figure 5 shows as in the example of *Moisture*, the regression coefficients have in the 60 jackknife repetitions a relatively large jackknife dispersion around the value of the model computed with all the 60 objects. The 95% CI of the regression coefficient of 11 predictors contains the value 0, so that the 11 predictors are eliminated with MUT.

The MUT is the procedure used by Unscrambler [29], a data-analysis package more often used in the world of multivariate calibration, so that MUT is probably the most frequently applied method for

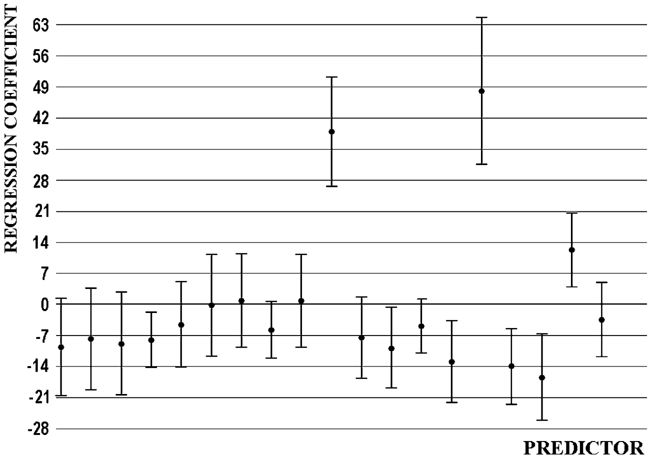


Fig. 5 The MUT procedures illustrated with data set *Moisture*

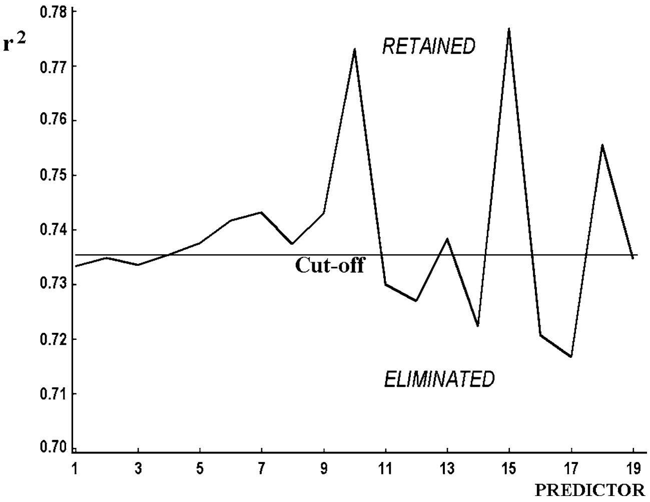


Fig. 6 Optimum cut-off value of the squared correlation coefficient in MAXCOR with the data set *Moisture* (see Fig. 7) and selected predictors

elimination of useless predictors in multivariate calibration.

MAXCOR

With the acronym MAXCOR we indicate here a technique suggested by Höskuldsson [12], "Most Correlated Predictors". The squared correlation coefficients r^2 of the predictors with the response are computed. The maximum and the minimum of r^2 constitute the range of exploration. When the minimum is less than the CI of r^2 the limit of the CI is used instead of the minimum, which means that the predictors with r^2 less than the confidence limit (at 0.01% significance level) are not considered significantly different from 0, so that the corresponding predictors are cancelled.

The confidence limit for r^2 is obtained from the Student distribution. The variable

$$r\sqrt{\frac{N-2}{1-r^2}} \tag{30}$$

is distributed as a Student variable with $v = N{-}2$ degrees of freedom, so that the critical value for r^2 is obtained by means of equation:

$$r^2_{\text{crit}} = \frac{t^2_{\text{crit}}}{t^2_{\text{crit}} + v} \tag{31}$$

The range is divided into, e.g., 20 intervals, and for each corresponding cut-off value the predictors with r^2 less than the cut-off are not used in PLS regression. The optimum cut-off, and consequently the significant predictors, is that with the maximum value of the CV-explained variance. Figure 6 shows the variation of the squared correlation coefficients of the 19 predictors of *Moisture* with response. The range of the squared correlation coefficient was divided into 20 intervals. Figure 7 shows how the fitting and fivefold CV-explained variances change with cut-off level. The maximum prediction ability was obtained with ten predictors selected at the cut-off level shown in Figure 6.

Genetic algorithm-ordinary least-squares

As far as we are aware GA were introduced in analytical chemistry by Lucasius and Kateman [1]. In this paper, the authors indicate possible applications of these algorithms in optimization problems characterized by a number of local maxima. Between these possible applications they indicate the use of GA to find the subset of predictors that produces the best predictive ability with the use of OLS regression.

The genetic algorithm starts with many random guesses of the predictors (the starting population). The

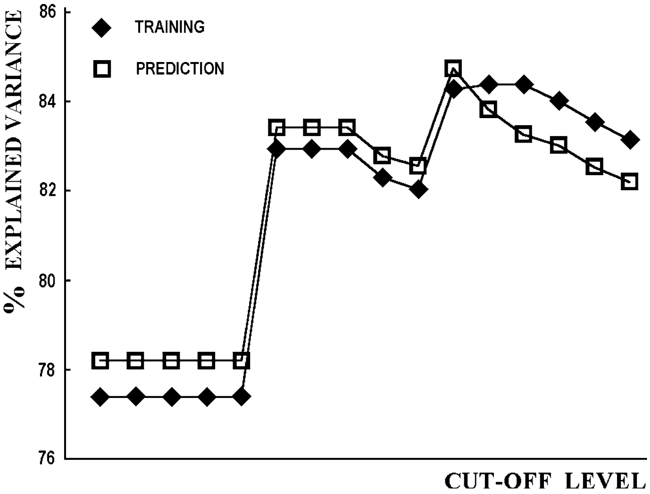


Fig. 7 Explained variance as a function of the cut-off value of the squared correlation coefficient in MAXCOR with the data set *Moisture*

best selections are then "merged" according to the rules of the biological evolution (reproduction, mutation, elitism, migration). The "evolution" continues until a steady state is reached or a stop rule has been verified. GA are essentially a technique of optimization that can be applied advantageously when the all-subset search is too expensive. In practice, the all-subset exploration is more parsimonious than GA when the number of predictors is less than ten to twelve. GA cannot usually find the true optimum but can easily find a near-optimum solution, and generally a set of different but good other solutions.

The GA–OLS uses OLS as regression technique and searches for a subset of variables that can be managed by OLS (which means that their number must be less than the number of objects in the training set and that the information matrix $\mathbf{X}^T\mathbf{X}$ must be invertible, i.e., that the selected predictors must not be correlated too much). For spectroscopic data with very large correlation between predictors this requirement can substantially reduce the number of predictors selected and models evaluated.

VS or GA–PLS

Variable selection (VS) [32] and GA–PLS [2] are the acronyms used for the search of the subset of predictors that produces the maximum prediction ability when the regression technique is PLS. Compared with GA–OLS, VS does not have the constraint of a reduced number of predictors and of limited correlation among them, so it can explore a much larger number of combinations.

Least absolute shrinkage and selection operator

Least absolute shrinkage and selection operator (LASSO) [32, 33] searches for the vector **b** that minimizes the function:

$$\frac{\sum_{i=1}^{N}\left(y_i - \mathbf{b}^T\mathbf{x}_i\right)^2}{N} + \lambda\sum_{v=1}^{V}|b_v| \tag{32}$$

similar to the equation that defines RR:

$$\frac{\sum_{i=1}^{N}\left(y_i - \mathbf{b}^T\mathbf{x}_i\right)^2}{N} + \lambda\mathbf{b}^T\mathbf{b} \tag{33}$$

where **b** is the vector of the estimates of the $V+1$ regression coefficients, **x** is the vector of the predictors, and λ is a parameter.

The difference between LASSO and RR is that for LASSO the constraint due to parameter λ allows solutions for which one or more coefficient is zero or very small; LASSO thus acts as a technique for selection of predictors.

For example, for this small example:

$X1$	$X2$	Y
2	4	6
3	6	9
4	8	12
5	10	15
6	12	18

The RR solution is: $\mathbf{b}=0.6$; 1.2 (the sum of the squares of the coefficient, 1.8, is the minimum among all the equivalent solutions). The Lasso solution is, instead: $\mathbf{b}=0$; 1.5, i.e. that with the minimum value of the sum of the absolute values of the coefficients. RR is frequently performed with autoscaled variables (see pretreatments, below). The same pretreatment was used in LASSO [32].

Combination of methods

Some methods, such as MUT and GOLPE, can be used (according the original literature) as preliminary elimination technique, to be followed by more parsimonious techniques. Nowadays the parsimony principle attracts much attention. A second principle, that of maximum trueness and precision, is considered very important by analytical chemists. This is one reason for the success of PLS, which can use the synergy of very correlated predictors to reduce the error, just as repetition of a measurement reduces the variance of the error. However, PLS exploits the synergy of correlated predictors only when the error of predictors is independent (which does not always happen in spectroscopy) and when the error in the response is negligible compared to the error in the predictors. In contrast, in multivariate calibration the error in the response is often rather large, because of the precision of the reference technique and possible alteration of the sample between classical analysis and acquisition of spectra. In such a situation it is not possible to use synergy; this is probably why SOLS sometimes behaves better than PLS.

A second strategy suggested in the use of GOLPE and MUT is iterative use of the selection of predictors. Here, only some examples of the use of combined methods are shown, all with *Artificial*, the combinations are GOLPE–IPW, MUT–IPW, GOLPE–SOLS and MUT–SOLS. GOLPE was used at the three levels of selection, give an overview of the effect of a more or less heavy pre-selection procedure.

Parameters with effect on the result of selection techniques

The results of all the techniques depend more or less on some selectable criteria and on the values of some parameters:

- the validation procedure (number of cancellation groups, ...) for almost all the techniques
- the criterion used to select the optimum complexity of the model in PLS and related methods

- the number of CV groups in PLS and related methods
- the maximum number of selectable predictors in SOLS and GA-based techniques
- the critical value of *F*-to-enter and of *F*-to-remove in SOLS
- the number of predictors cancelled in each cycle in ISEPLS
- the scaling procedure in LASSO
- the value of α, the number of noisy artificial predictors and the randomization seed in UVE
- the confidence level in MUT
- the number of reduced models and of dummy variables and the cutting level in GOLPE
- the percentage of elitism, the probability of mutation, the stop procedure, and the number of starting populations in GA based techniques.

So, the results are not very indicative of the "value" of a selection technique. Moreover, a regression model must be evaluated not only with reference to the predictive ability, but also on the basis of the number of selected predictors and of its complexity (the number of latent variables in PLS).

However, some general trends can be observed and, finally, the objective of the selection is reliable evaluation of the most important predictors and of the validity of the associated regression models.

Software

All the above techniques were used as implemented in V-PARVUS [34], or in the free version available in Internet, or in the development version (GOLPE, MAXCOR, and MUT). MUT was also used as implemented in Unscrambler [29]. The results of LASSO and of VS for data set *Kalivas* were obtained from the figures in the paper of Ojelund et al. [32]. In this paper, both the response and the predictors were autoscaled and both are referred to the original variables (centering does not modify the scale), so that the results drawn here from Ref. [32] are apparently different from those in the original paper.

Data sets

The results reported here are for only three data sets, that we consider fairly representative of the general behavior of the selection techniques, as evaluated on many other data sets, both for problems of multivariate calibration and for QSAR studies.

Moisture—60 samples of soy flour, spectra measured with a filter instruments, with 19 filters. The response variable is moisture. Details are reported in the original paper [35].

Kalivas—100 wheat samples, spectra measured with a NIR spectrometer, 701 predictors corresponding to wavelengths from 1102 to 2502 nm in 2-nm intervals. Response was moisture. Details are reported in the original paper [36]. These data have been used also in [31].

Artificial—a response, 300 predictors, 400 objects. A total of 300 objects constitute the third set, a true evaluation set, here called generally "external" set. The predictors (all with range 0–1000) were organized into six groups. The first five groups (A–E) contain ten very correlated predictors (correlation coefficient >0.995). The first predictor in groups A–E was obtained by random extraction from a uniform distribution U(0–1000). The other predictors in groups A–E were obtained by addition of random rectangular noise, centered, with range 50 to the first. The 250 predictors in group F were obtained by random extraction from a uniform distribution U(0–1000). The response was obtained by means of the equation

$$\text{Response} = 1.5x_1 + 1.2x_{11} + 0.9x_{21} + 0.6x_{31} + 0.3x_{41}$$

i.e., multiplying the first predictor of each group A–E by the coefficients 1.5, 1.2, 0.9, 0.6, and 0.3, respectively, so that the first E predictor contributes to the value of the response with a value between 0 and 300. Because in each group of ten predictors the predictors from the second to the tenth are a copy of the first, the "theoretical" contribution θ of the first ten predictors is 0.15, of the second ten 0.012, and so on.

Rectangular noise was added to the response, centered around the original value and with range 300. So, the first five groups of predictors make a decreasing contribution to the response, and the last E has a contribution of the same order of the added noise.

After addition of this noise the range of the response was from 550 to 3660 in the training set, from 350 to 3930 in the third set, with a regular distribution.

The 250 predictors in *F* are useless predictors. The "errors" of each group of ten predictors are non-correlated so that these predictors can be regarded as synergetic; this does not always happen for the predictors of multivariate calibration. In spectroscopy, the error in a wavelength is partially transmitted to the next predictors.

Chemists rarely have a quantitative idea of the correlation in their predictors. For the filter instrument the mean correlation coefficient between a predictor and the contiguous next predictor is 0.9980. The minimum correlation coefficient (between predictors 12 and 19) is 0.9921.

For the instrument of data set *Kalivas*, the mean correlation coefficient between contiguous predictors is 0.999984. The minimum correlation coefficient between the first and the last predictor is 0.471. So, the *Artificial* data set, in which the noisy predictors are almost non-correlated (mean of the correlation coefficient between contiguous predictors 0.00006; mean of the absolute correlation coefficient between contiguous predictors 0.039) is not representative of the noise in NIR data, but it represents other analytical calibration problems, as artificial nose and tongue and data for property–structure relationship.

Moreover, *Artificial* was built without interference, with useful signal, proportional to the "interesting analyte" and random noise. In real cases of multivariate calibration interfering species, frequently unknown, contribute with a variable signal to the total signal, and their contribution overlaps more or less with the contribution of the interesting analyte.

Pretreatments

In multivariate calibration many kinds of pretreatment have been applied, depending on the analytical technique.

Column centering, i.e., the subtraction of the mean of each predictor:

$$x_{iv}^{\text{centered}} = x_{iv} - \bar{x}_v = x_{iv} - \frac{\sum_{i=1}^{N} x_{iv}}{N} \qquad (34)$$

and column autoscaling, i.e., the standardization of centered data:

$$x_{iv}^{\text{autoscaled}} = \frac{x_{iv} - \bar{x}_v}{s_v} = \frac{x_{iv} - \bar{x}_v}{\sqrt{\frac{\sum_{i=1}^{N} (x_{iv} - \bar{x}_v)^2}{N-1}}} \qquad (35)$$

are the more general pretreatments. For spectral data other kinds of pretreatment are frequent, for example SNV (row autoscaling, i.e. subtraction of the mean of sample and division for the standard deviation of the sample), derivatives, Fourier transform coefficients, wavelets, ...

Autoscaling must be applied when predictors have different scales. Because for spectra all predictors are on the same scale, autoscaling is not usually applied. Here, we will use generally centered data, sometimes autoscaled data.

The results of the selection procedures depend on the pretreatment, even when it has no theoretical effect, because of numerical problems because of the limited number of digits used in computation. This is the case for OLS, in which autoscaling of predictors can sometimes enable inversion of the information matrix that otherwise cannot be inverted, or can be inverted but at the price of a very small determinant.

Results and discussion

Data set *Moisture*

The results obtained with data set *Moisture* are presented in Table 1 and Fig. 8. We can easily note:

- The large difference between the number of selected predictors, from 2 to 16. Some techniques are rather conservative, as GOLPE I; other techniques are very parsimonious, for example the OLS-based techniques and IPW. Also ISE seems here very parsimonious, but this result is not always confirmed.

Table 1 Data set *Moisture*—results with the selection techniques and five CV groups

Method	Predictors	CV-explained variance (%)	SEP	Components[a]
PLS	19	82.20	1.369	2
MUT	8	82.59	1.354	2
UVE normal	7	82.68	1.350	2
UVE 95%	10	82.81	1.346	2
UVE 90%	12	83.00	1.338	2
GOLPE I	15	83.48	1.319	2
IPW	3	83.77	1.307	3
GOLPE III	3	83.77	1.307	2
MAXCOR	10	84.74	1.268	2
ISE	2	84.82	1.264	2
GOLPE II	6	85.02	1.256	2
SOLS(5)[b]	2	85.68	1.202	2
GA–OLS	2	85.68	1.202	2
PLS LOO	19	83.19	1.302	5
UVE normal LOO	4	85.37	1.215	3

[a] First minimum for PLS-based techniques

[b] Complete-CV: 78.94% CV % explained variance, 1.489 SEP

- The goodness of prediction is not directly correlated with the number of retained predictors. SOLS and GA–OLS, that select the same two predictors, are the techniques with the largest CV-explained variance.
- The bottom two lines in Table 1 indicate that the number of CV groups has an important effect in the definition of the performance of PLS (leave-one-out is slightly optimistic), but also on the optimum complexity of the model, and on the performance of some elimination techniques. Indeed, the very different performance of normal UVE (both in the number of retained predictors and in the explained variance) is because the complexity of the model is different, so the UVE reliability is computed with a very different model.
- The "Complete-CV" parameters of SOLS indicate that the selection of the useful predictors in the validation cycles is different from the final selection. In this case the larger prediction error of Complete-CV is also because of the presence of two outliers (objects 39 and 40, y-outliers, probably because of alteration of the samples). Figure 9 shows the residuals (five CV groups) obtained with the final selection of variables, and those obtained with five or seven CV groups in complete CV. The points fall outside the line when they are predicted with predictors different from those in the final selection.

Data set *Kalivas*

Kalivas predictors were obtained by NIR spectrometry, so the data set is characterized by many highly correlated predictors (more than for the filter instrument for *Moisture*).

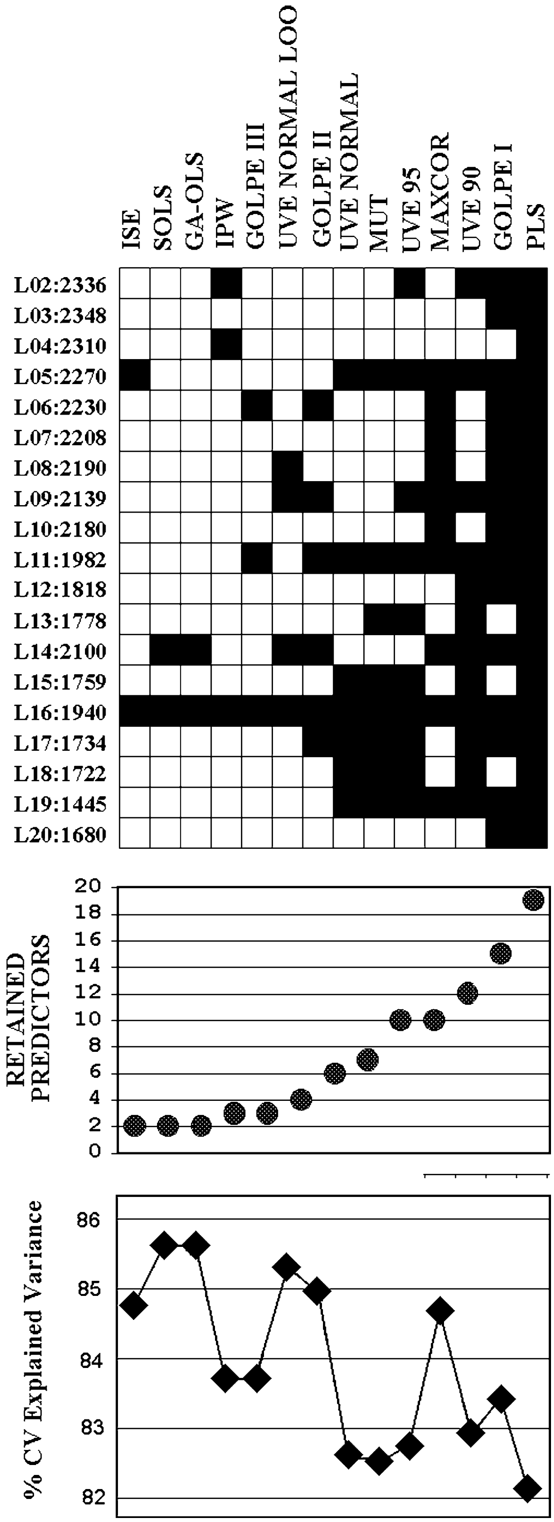


Fig. 8 Data set *Moisture*—selected predictors, number of selected predictors, CV% (five groups) explained variance

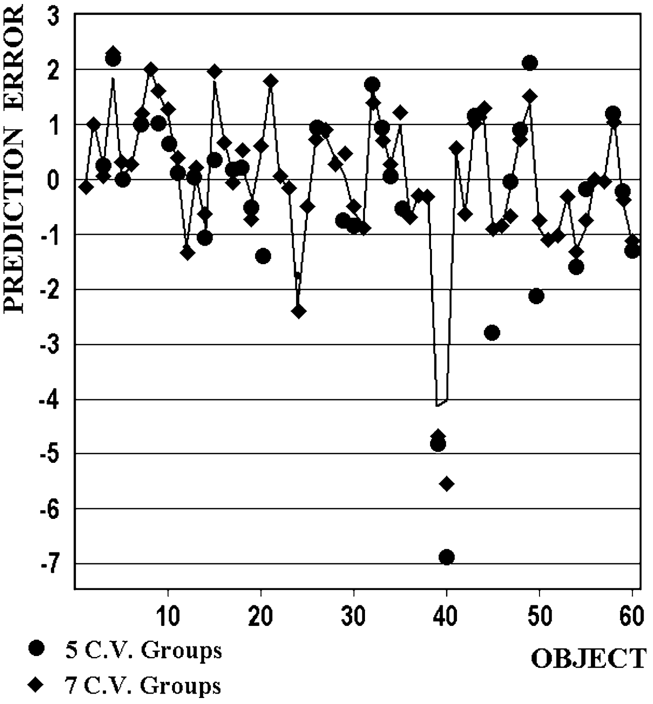


Fig. 9 Data set *Moisture*. Residuals with five CV groups and the final selection of predictors (*line*), and residuals with complete-CV with five or seven validation groups

The results are reported in Table 2 and in Fig. 10a–c:

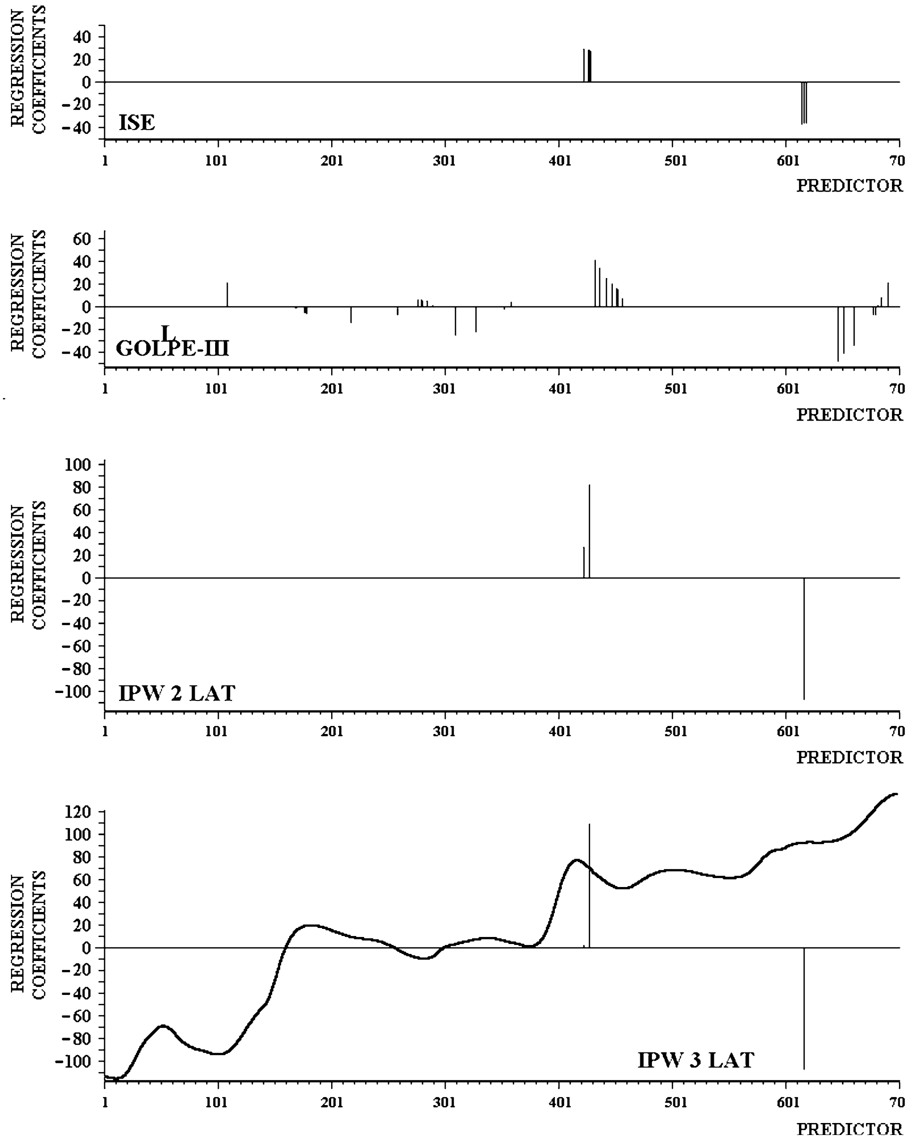
- The techniques seem separated (better than in *Moisture*) between conservative (GOLPE, MAXCOR, MUT, UVE) and parsimonious (GA–OLS, GA–PLS, LASSO, ISE, IPW, SOLS, VS). The prediction performance of the conservative techniques is almost the same as for PLS with all the predictors.
- The predictive ability of the parsimonious techniques seems slightly better than that of the conservative techniques.
- Almost all the selection techniques select predictor 430 (wavelength 1960 nm), or a near predictor, or a

Table 2 Data set *Kalivas*, results with the selection techniques

Method	Predictors	CV-explained variance (%)	SEP	Components[a]
SOLS(2)[b]	2	96.96	0.2408	2
GOLPE III	32	97.39	0.2231	6
UVE 95%	657	97.40	0.2227	5
MUT	575	97.40	0.2227	6
GOLPE I	648	97.43	0.2216	6
PLS	701	97.45	0.2218	5
SOLS(4)[c]	4	97.45	0.2207	4
MAXCOR	684	97.52	0.2217	5
IPW 3 COMP	11	97.52	0.2174	3
GOLPE II	352	97.54	0.2167	6
IPW 2 COMP	11	97.57	0.2155	2
ISE	7	97.57	0.2151	2
LASSO	14	97.58[d]	0.2153[d]	14
GA–OLS a	4	97.61	0.2154	4
VS	14	97.66	0.2111	6
GA–PLS	11	97.74	0.2078	6
GA–OLS b	4	97.75	0.2090	4

[a] First minimum in the case of PLS-based techniques
[b] 96.88% complete-CV: explained variance, 0.2439 SEP
[c] With autoscaled variables; 97.44% complete-CV: explained variance, 0.2208 SEP
[d] Deduced from Ref. [31]

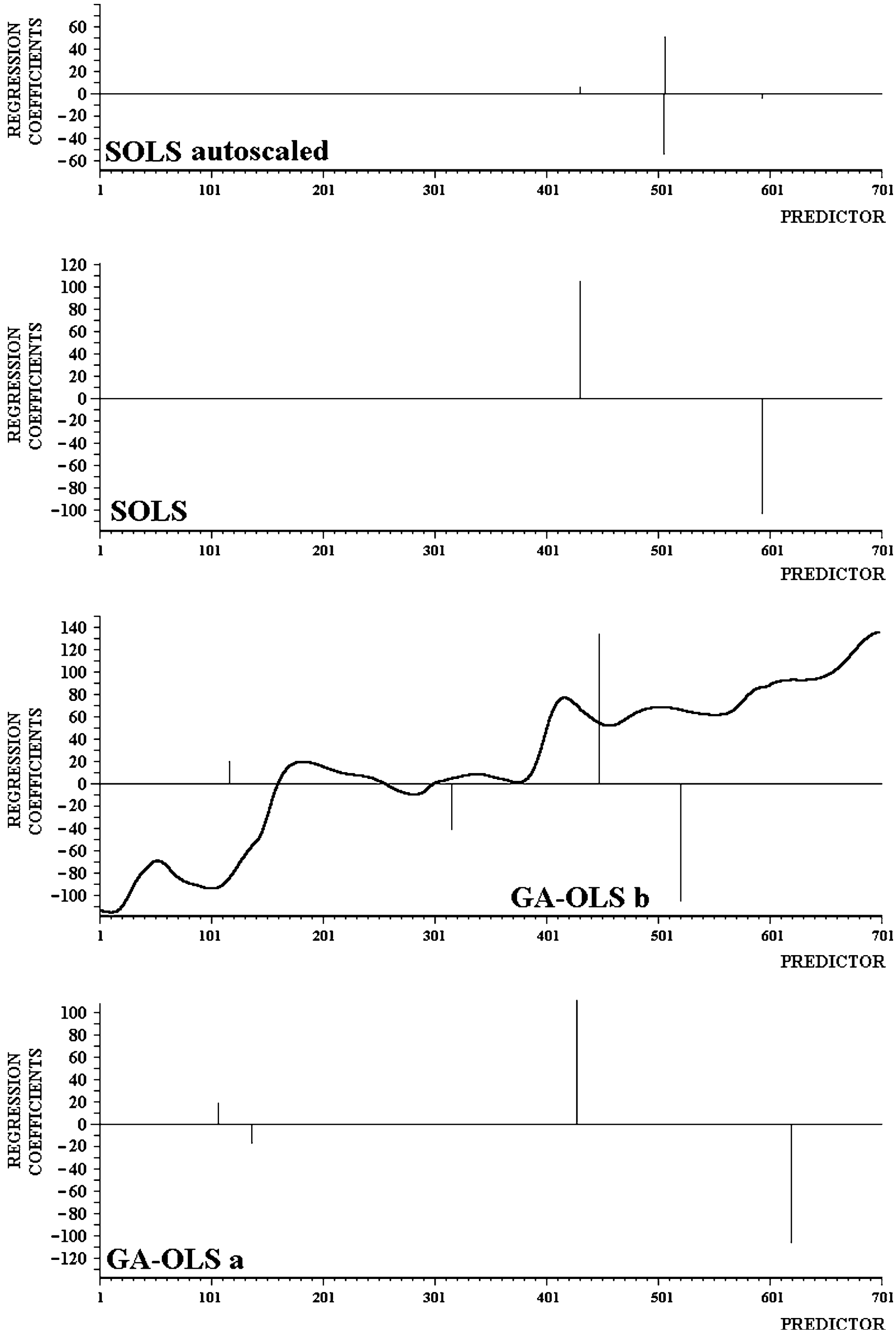
Fig. 10 Data set *Kalivas*. Selected predictors and regression coefficients of **a** ISE, GOLPE III, and IPW; **b** SOLS autoscaled, SOLS, GA–OLS b and GA–OLS a; **c** LASSO, VS and GA–PLS



group of near predictors, with positive regression coefficient (region A), and predictor 615 (2330 nm), or a near predictor, or a group of near predictors, with negative regression coefficient (region B). Only GOLPE III suggests three predictors with negative coefficients at higher wavelengths.

- The IPW oscillates between two solutions, with the same three predictors but with different weights (Tables 3 and 4). Consequently the two near predictors in region A share differently between them the positive contribution. The minimum SEP is at two or three PLS components, but the differences between the predictive performance with two or three components is in both cases very small; also small is the difference from the performance of the model computed when the three predictors are not multiplied by the weight (Table 5).
- The SOLS anticipates region B at predictor 594 (2290 nm approx.). SOLS with the centered data was not able to invert the information matrix with more than two predictors. SOLS autoscaled, with autoscaled data, selected four predictors, but the quality of the information matrix was not excellent. Moreover SOLS autoscaled selected two contiguous (and consequently highly correlated) predictors (506 and 507, approx. 2110 nm, region C) with opposite regression coefficient; which means that the two contributions approximately annul each other. Two or four predictors seem too few to produce good results, so that SOLS has the worst performance.
- The prediction ability with complete-CV was rather good for both SOLS and SOLS autoscaled; this means that the selection of predictors in the five groups of validation is identical with predictors contiguous to those selected in the final run with all the objects in the training set.
- The GA–OLS does not always succeed in individuating regions A and B, and substitutes region B with region C. The predictors selected by LASSO and VS represent both region B and region C (with prevalence

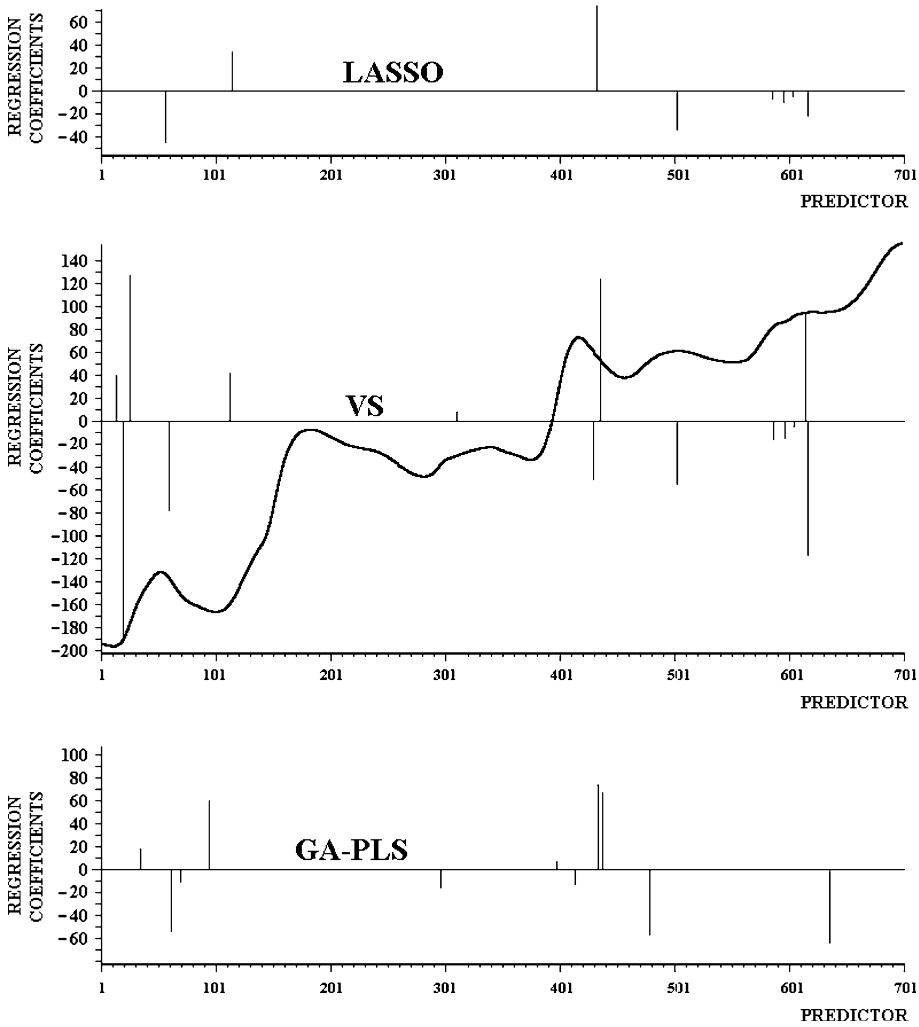
Fig. 10b (Contd.)



of negative regression coefficients) and region A (with prevalence of positive coefficients)

- The GA–SOLS and VS (really the same technique used by us and by Ojelund et al. [32]) select predictors also in the first part of the spectrum. Really, when used with PLS, GA are not limited in choice of predictor other than that imposed by the operator, because the PLS algorithm does not compute and invert the information matrix. This freedom is used by GA–PLS to use more predictors in the effort to obtain the best prediction; this, unfortunately, very often produces an overestimate of the prediction ability ("overprediction" similar to the overfitting of other techniques such as OLS) and selection of useless predictors.
- The MAXCOR selects almost all the predictors, and stops just because 17 predictors have correlation coefficients less than the critical value of the test with hypothesis $r=0$. The reason that the technique is not able to individuate a subset of predictors with better prediction ability is explained with the data in Fig. 11. Here, regions B and C correspond in order to the two peaks of largest correlation. In contrast, region A corresponds to a relative minimum of the correlation.

Fig. 10c (Contd.)



In this case we select only some predictors in region B (cutting at the level indicated with the arrow in Fig. 11) we obtain a "partial" PLS regression model. The residuals with this partial model have in turn the correlation coefficients with the predictors shown in Fig. 11, curve (b), whose maximum corresponds to region A. To select the predictors in this region MAXCOR must reduce the cut-off substantially, with selection of many useless predictors. In contrast SOLS, after selection of only one predictor in region C, searches for the predictor with the largest correlation with the residuals, and finds it in region A.

Table 3 Data set *Kalivas*, weight (importance) applied to the predictors by IPW when it reaches a steady condition oscillating between the two states

Predictor	Wavelength	Weight	
		State low	State high
423	1946	0.049908	0.120490
428	1956	0.427477	0.356845
617	2334	0.522615	0.522665

– The correlation coefficients of the residuals shown in Fig. 11 never are larger than the critical value of r^2 (approx. 0.038). However, in region A there are more than 100 contiguous predictors (from predictor 360 to predictor 480) whose squared correlation coefficient describes a regular peak: this event cannot be a casual result obtained from the repetition, 100 times, of the experiment that consists in the random extraction of the correlation coefficient from an infinite population with correlation coefficient $\rho = 0$, i.e. of the experiment to which the statistics of Eqs. (30) and (31) apply. When the population has $\rho = 0$ a peak of correlation coefficients such as that of the residuals must be considered a very rare event, so the peak is significant.

Data set *Artificial*

First *Artificial* was studied by normal PLS, both with all the 300 predictors and with different subset of predictors, as shown in Table 5.

With progressive addition to the set of predictors used, initially only the first ten, of the first groups of

Table 4 Data set *Kalivas*, results of PLS with the three predictors selected by IPW, in the two states of the oscillation, and without weights applied to the three predictors

Complete	Weight state low		Weight state high		No weight	
	CV residual SD	CV-explained variance (%)	CV residual SD	CV-explained variance (%)	CV residual SD	CV-explained variance (%)
01	1.2137	22.80	1.2078	23.55	1.2428	19.06
02	0.2188	97.49	0.2171	97.53	0.2176	97.522
03	0.2174	97.52	0.2174	97.52	0.2174	97.523

Table 5 Data set *Artificial*—results of PLS (CV% explained variance with five cancellation groups) with the selection of groups of predictors based on the knowledge of their contribution to the response

PLS components	A	A, B	A, B, C	A, B, C, D	A, B, C, D, E	A, B, C, D, E, F	*F*
1	40.023	70.066	85.443	91.946	91.475	82.951	−2.501
2	37.481	71.012	86.159	93.183	93.812	86.987	−8.321
3	36.463	61.356	86.144	93.338	94.120	87.999	−13.121
4	35.677	59.969	82.111	93.435	94.101	87.123	−17.121
5	35.003	59.3139	80.518	91.678	93.838	87.043	−18.901
6	34.984	59.150	79.427	91.699	92.216	86.826	−20.218
7	35.107	58.903	79.420	91.531	90.378	86.476	−20.145
8	35.108	58.941	79.171	91.331	89.452	86.407	−20.204
9	35.110	58.388	78.935	91.099	88.206	86.382	−20.191
10	35.110	58.374	78.473	91.016	87.171	86.387	−20.202
Optimum complexity (threshold 5%)	1	2	2	3	2	3	1
External	47.281	73.710	87.206	92.270	93.391	88.715	−19.719

A predictors 1–10, *B* predictors 11–20, *C* predictors 21–30, *D* predictors 31–40, *E* predictors 41–50, *F* predictors 51–300

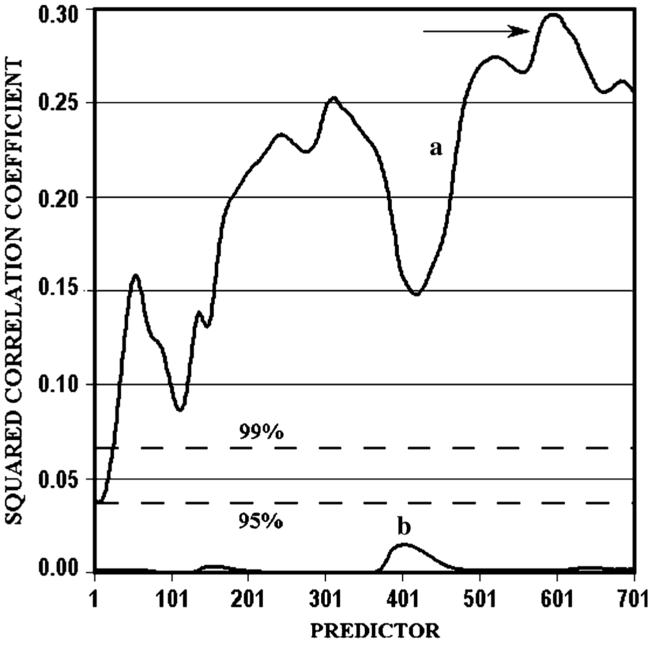


Fig. 11 Data set *Kalivas*—squared correlation coefficients **a** of the original predictors with the response; **b** of the residuals (from PLS regression performed using only the predictors with correlation coefficient with the response larger than that shown by the *arrow*) with the original predictor. The critical value of the squared correlation coefficient is reported for two probability levels

predictors, those with contribution to the response, the CV-explained variance increases, at first very much, then slowly, finally only 0.7%. The results with the external evaluation set of 300 objects confirm it. Moreover, when the number of predictors increases from 40 to 50, with addition of the ten predictors E, the increase for the external evaluation set is 1.1%, more than that evaluated by CV.

The result obtained by PLS with the 50 first predictors, 94.12% CV-explained variance with complexity three and, more important, 93.39% explained variance on the external validation set, is the "ideal" result, the target of the techniques of selection.

When all the potentially useful predictors are used, the regression coefficients of the PLS model are approximately proportional to the "theoretical contribution" $\theta = 0.15$, 0.12, 0.09, 0.06, and 0.03 used to compute the response. The regression coefficients *b* represent, as estimated by PLS, the real contribution of the predictors to the response. Table 6 shows that the real contribution of the predictors in group E is approximately one-half of the theoretical contribution. In turn this means that the noise added to the response partially hides the contribution of the less important useful predictors. Figure 12 shows that the contribution of the regression coefficients in the five groups of significant predictors is almost the same.

Table 6 Regression coefficients of the PLS model with the first 50 predictors

Predictors	Theoretical contribution, θ	Mean PLS regression coefficient b in the group
A	0.15	0.1513
B	0.12	0.1198
C	0.09	0.0956
D	0.06	0.0583
E	0.03	0.0170

Table 7 Regression coefficients of the first 50 predictors in the PLS model with all the 300 predictors

Predictors	Theoretical contribution, θ	Mean PLS regression coefficient b in the group	b/θ
A	0.15	0.1271	0.847
B	0.12	0.1035	0.862
C	0.09	0.0733	0.814
D	0.06	0.0519	0.866
E	0.03	0.0181	0.604

Addition of the 250 noisy predictors substantially reduces the performance of the PLS model, by more than 6% (4.5 for the external evaluation set). The regression coefficients of the useful predictors change, as shown by the data in Table 7 compared with those in Table 6. The mean contribution of noisy predictors is +0.003, but multiplied by 250 the total contribution +0.76 is obtained; this explains the decrease of the correlation coefficient of the most important predictors. Figure 13 shows that some (4) noisy predictors, because of the casual high covariance with the response or its residuals, have a regression coefficient as large as the predictors in group D. Many (53) noisy predictors have a regression coefficient larger than those of group E.

The use of the 250 noisy predictors alone (group F) gives a useless model, with negative explained variance; this means that the mean of the response in the training set is a better estimate of the value in the internal (CV) or external evaluation sets than that computed by the regression model. So, the noisy predictors have no useful information (on the whole, because the single noisy predictor can have an apparently significant fortuitous correlation with the response).

With this knowledge of the true relevance of the predictors it is possible to evaluate the performances of the selection techniques with some elements more than those obtained from real data. However, it must be remembered that the noise of predictors and response in *Artificial* is true random non-correlated noise (which rarely happens with real data) and the number of noisy non-correlated predictors is very large (in the case of real data sometimes there are many noisy predictors, but they are more or less correlated, so that the probability of casual correlation with the response is not as large as in *Artificial*).

Table 8 and Figs. 14, 15, and 16 show some results obtained with the selection techniques. In the case of PLS-based techniques a 5% Osten [21] cut-off is used:

- The results confirm the subdivision between conservative and parsimonious techniques.
- SOLS(20) and the GA-based techniques show the maximum difference between the CV-explained variance and the explained variance measured on the external evaluation set. So, these techniques used without an external evaluation set overestimate the prediction ability (overprediction). The difference (6–10%) is very large, despite all the methods used, with a limit (20) in the maximum number of selectable predictors.
- The IPW is the best parsimonious technique. It works automatically, without a limit in the selectable predictors. The overestimate of the prediction ability is only about 1.5%. It selects one predictor as representative of the correlated predictors in each of the more important groups and no predictors in the group F of noisy. However, IPW is not able to detect the utility of predictors in the fifth group, E.

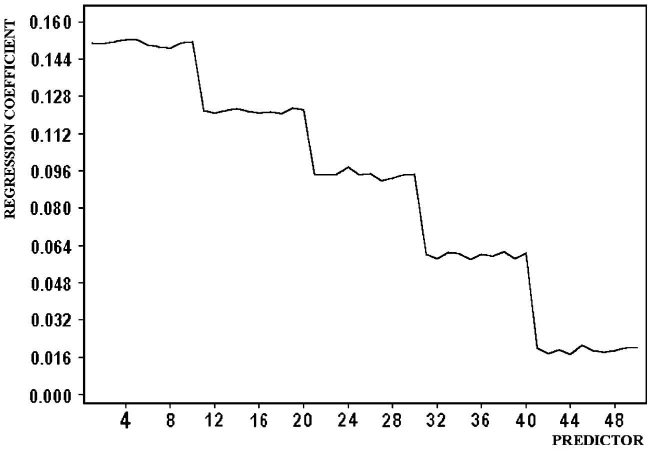


Fig. 12 Data set *Artificial*—regression coefficients of the PLS model with predictors of groups A, B, C, D, and E

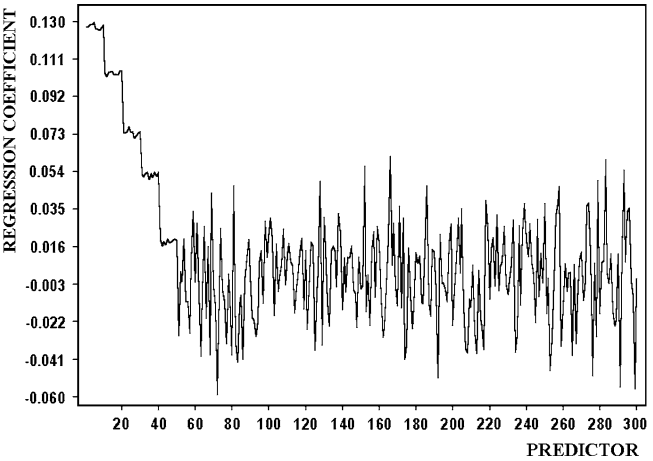


Fig. 13 Data set *Artificial*—regression coefficients of the PLS model with all 300 predictors

Table 8 Data set *Artificial*—results with the selection techniques

Method	Predictors retained	Good predictors retained	Noisy predictors retained	Complexity[a]	CV-explained variance	CV residual SD	Evaluation set-explained variance	Evaluation set residual SD
SOLS(20)[b]	20	7	13	20	97.41	106.1	87.98	240.5
GOLPE I	279	50	229	3	90.11	207.6	89.10	229.1
GA–OLS	10	7	3	10	96.34	126.0	90.10	218.3
GA–PLS	17	6	11	4	96.57	122.2	90.48	214.0
GOLPE II	142	40	102	3	95.43	141.1	91.02	207.9
ISE[c]	109	46	63	3	96.35	122.2	91.13	206.7
GOLPE III	34	22	12	4	93.67	166.1	91.71	199.8
SOLS(4)[d]	4	4	0	4	93.58	166.9	92.08	195.3
MAXCOR	42	40	2	3	93.23	171.7	92.29	192.6
IPW	4	4	0	4	93.88	163.3	92.29	192.6
UVE	67	50	17	3	95.53	139.5	92.54	189.6
ISE[e]	61	44	17	3	96.62	121.3	92.57	189.1
MUT	59	50	9	3	94.98	147.8	92.93	184.5
SOLS(5)[f]	5	5	0	5	94.48	154.8	93.23	180.5

[a] 5% Osten [21] cut-off for PLS-based techniques
[b] 88.68% Complete CV: explained variance, 222.1 SEP
[c] ISE with elimination of more (≤ 10) predictors in each step
[d] 93.55% complete CV: explained variance, 167.6 SEP
[e] ISE with elimination of only one predictor in each step
[f] 93.45% complete CV: explained variance, 168.9 SEP

- The *F*-test used in SOLS causes selection of all 20 permitted predictors. The first five selected predictors are predictors 3, 20, 24, 40, 47, i.e. one for each group of significant predictors. Unfortunately selection based on the *F*-test continues with many noisy entered predictors. Figure 14 shows that after the fourth entered predictor the percentage of explained variance continues to increase almost regularly. The change in the trend, from abrupt to regular, can be used to evaluate the number of significant selected predictors, four or five. A better evaluation can be performed with the plot of the *F*-values of the entered predictors, shown in Fig. 15. Here, the value of *F*, very large with the first four entered predictors, decreases to an almost steady level. The magnified plot in Fig. 15 (below) suggests that also the fifth entered predictor is significant.
- We find it very useful to compare the *F*-plot of the studied data set with the same plot obtained with a similar set (approximately the same number of rows and columns) of purely noisy predictors. In this case, we used the 250 predictors of group F. SOLS(20) accepts 20 predictors in this case also, with a noticeable value of the fitting explained variance (74.7%), but also with apparent good value of the leave-one-out-explained variance (68.4%) and of the CV-explained variance (68.9%). The complete CV-explained variance is negative (−79.5%), as is the percentage of explained variance for the external evaluation set (−105.8%) (which means that the complete CV validation can substitute the use of an external evaluation set). The *F*-plot in Fig. 15 shows that the *F*-value of the entered noisy predictors ranges form nine to five, approximately, for the SOLS cycles, much more than the commonly used *F*-to-enter value. The difference between the *F*-values of the fifth good predictor 47 and the noisy predictors, more than five, seems very indicative of the usefulness of predictor 47. SOLS, with the help of the suggested procedure, SOLS(5), behaves very well. The CV-explained variance is a bit less than that of MUT, but the result with the external evaluation set is excellent, just 0.1% less than that of PLS with the 50 useful predictors. The difference is just the price of the use of the synergism.
- The complete validation performs efficiently. It overestimates, but only by approximately 1%, the percentage of explained variance as evaluated by the external evaluation set. For real data it is rarely possible to have a numerous external set, so that complete validation seems to be a cheap and reliable alternative to the external evaluation set.

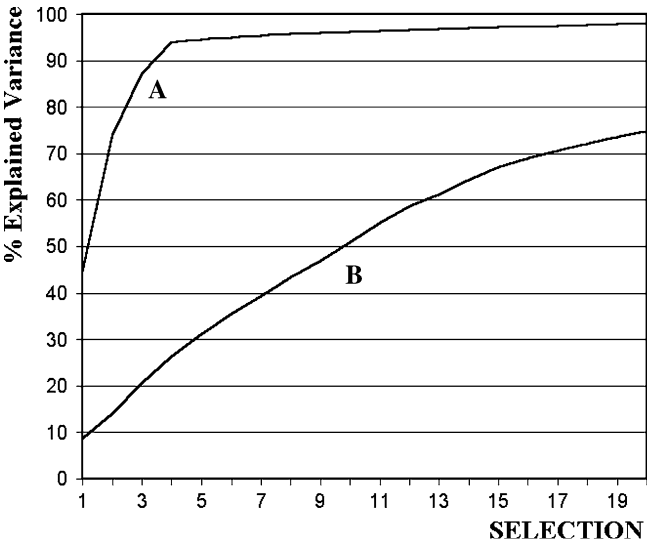


Fig. 14 Data set *Artificial*—fraction of the CV-explained variance as a function of the number of entered predictors: **A** all the predictors; **B** only noisy predictors, group F

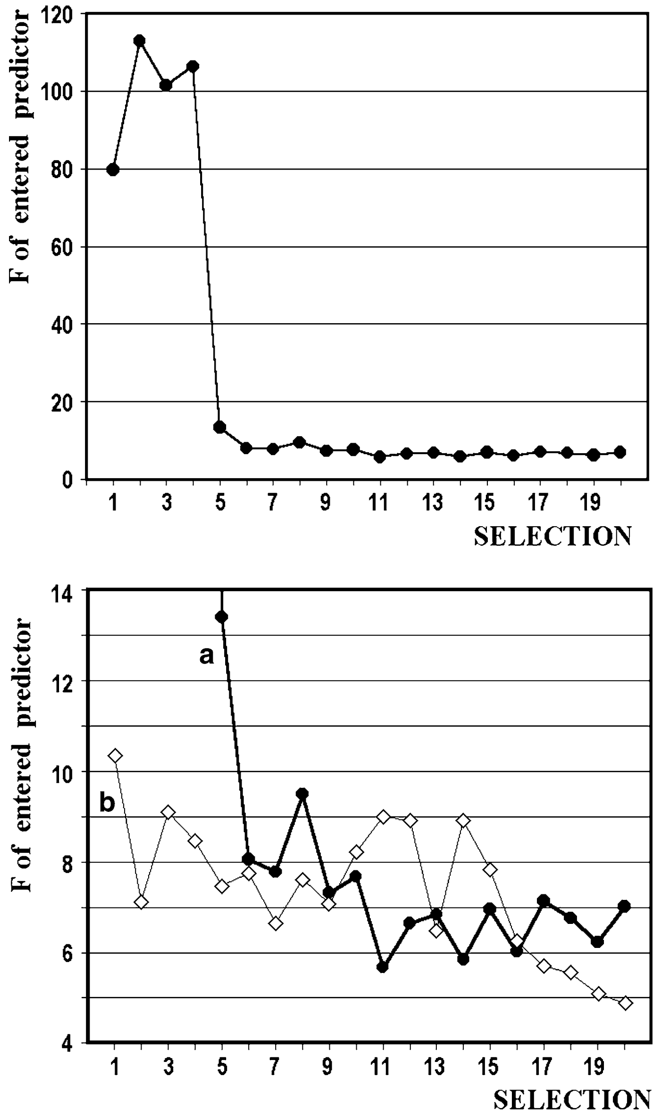


Fig. 15 Data set *Artificial*—*F* ratio of the entered predictors. *Below* magnified view: **a** all the predictors; **b** SOLS performed with only the noisy predictors (group F)

- Among the conservative techniques MUT, UVE, and ISE behave rather well. MUT retains all the good predictors and only nine noisy predictors. UVE retains more noisy predictors. ISE retains fewer useful predictors than UVE. Moreover ISE is more optimistic in the CV-explained variance. Taking into account also the larger computing time it is in this case less efficient.
- The results with GOLPE are disappointing, for all the three levels of selection.
- The MAXCOR eliminates many useless predictors, but is not able to detect the utility of predictors in the fifth group, E. The reason is the very small correlation coefficients of the predictors in group E with the response, shown in Fig. 16. MAXCOR can be regarded as the conservative equivalent of IPW: the two noisy predictors selected compensate for the use of the synergism in the 40 good predictors.

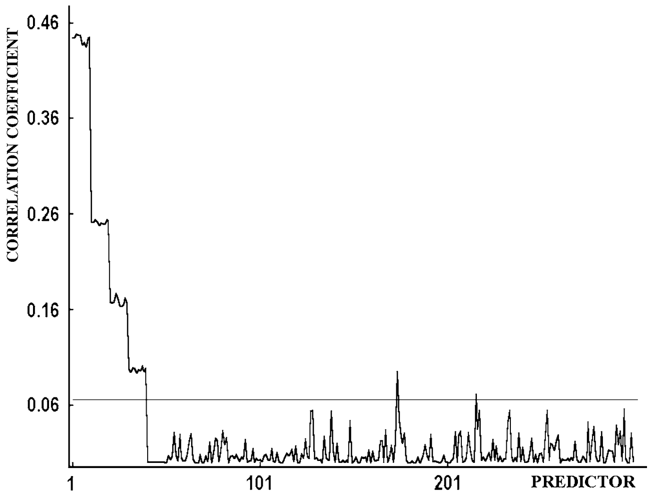


Fig. 16 Correlation coefficient of the predictors with the response and MAXCOR cut-off from the 99% critical value of r^2

- During the IPW cycles, especially in the second and third cycles where all the predictors are used but with very different weights, the CV standard deviation reaches a minimum, 138, less than with all the other techniques (Fig. 17) (The minimum corresponds to five components, but in this case the Osten [21] threshold 5% concludes for four significant components and CV SEP 148.6.)

This intermediate situation of minimum SEP after two to three IPW cycles has been observed frequently (with data sets other than those used here), so demonstrating that weighting the predictors can improve the performance of PLS. In contrast, with the predictors selected after the final IPW cycle the weights are not so important: in this case the percentage explained variance was the same.

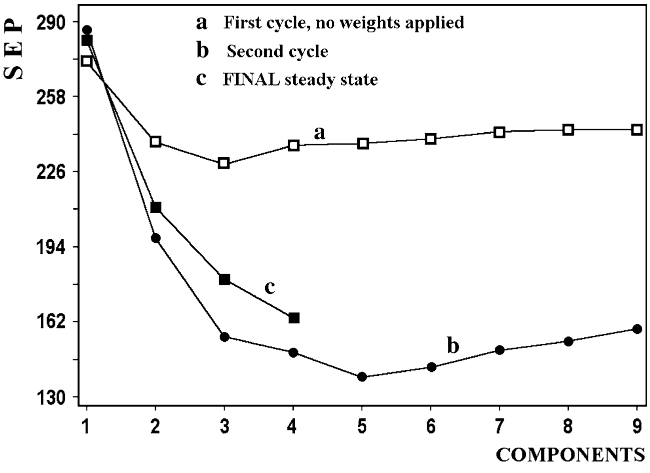


Fig. 17 Data set *Artificial*—plots of CV SEP as a function of the number of latent variables. **a** First IPW cycle (usual PLS); **b** second IPW cycle (all predictors retained, but weighted by the importance obtained in the first cycle); **c** final cycle, with only four predictors

Data set *Artificial*—combined techniques

Tables 1S and 2S and Figs. 1S and 2S in the Electronic Supplementary Material show the results obtained with the combined techniques (MUT and GOLPE combined with SOLS or IPW).

The results of MUT–IPW are exactly the same as obtained with the original 300 predictors.

All the levels of pre-selection of GOLPE have only a negligible (but negative) effect on performances of the final model obtained by IPW.

The evolution (Fig. 1S) from the original model (the usual PLS model with all the predictors) to the final model is different. The intermediate situation where all the predictors are retained but with different weights is more (GOLPE I) or less (GOLPE II and III) evident with GOLPE–IPW. In the case of MUT–IPW the intermediate situation was not observed and SEP increases almost regularly from the excellent value of MUT to the value observed with IPW using all the predictors. Evidently some predictors with positive effect in the intermediate state are not selected by the preselection techniques.

When SOLS is used with the limit 20 for the entered predictors MUT, GOLPE II, and GOLPE III have in this order an increasing positive effect (3–5% on the external evaluation set). When the *F*-plot (Fig. 2S) is used to stop the selection the results are the same as than those obtained by SOLS(5).

GOLPE III, with its drastic reduction of the number of predictors, creates a situation where the value 4 of *F*-to-enter behaves very well. The performance is slightly worse than that of SOLS(5) with all the predictors.

The general conclusion about the combined techniques is that they seem almost useless.

Conclusions

- The selection techniques can be divided into conservative techniques (MUT, UVE, GOLPE I, GOLPE II, MAXCOR) that try to retain all the informative, useful predictors, and parsimonious techniques (SOLS, IPW, GA–SOLS). GOLPE III, ISE, and LASSO are intermediate.
- Among the conservative techniques, MUT, used in Unscrambler [29], and UVE, developed by Massart et al. [23], seem the most efficient techniques. However both behave better with simulated data than with real data.
- ISE is irregular, both in the number of selected predictors and in the quality of the predictive performance.
- SOLS, with the classic strategy of the critical *F*-values, accepts too many predictors with consequent overfitting.
- SOLS can be improved to become the most efficient parsimonious technique, by means of the use of plots of the *F*-statistics value of the entered predictors and comparison with the parallel results obtained with a data matrix with random data. This procedure indicates correctly how many predictors can be accepted and decreases very much the possibility of overfitting. The use of complete validation supplies a reliable measure of the predictive ability.
- A possible alternative to the modified SOLS is IPW that automatically selects a minimum set of informative predictors.
- GA-based techniques generally overestimate the prediction ability and select non-informative predictors in an effort to maximize the prediction. The use of an external evaluation set, with objects never used in the elimination of predictors, or of the "complete validation" is suggested for obtaining a reliable estimate of the efficiency. In the case of GA–PLS a rigid constraint on the maximum number of retained predictors can help to obtain reasonable solutions.
- MAXCOR has two weak points: the use of the test of the significance of the correlation coefficient that does not take into account the correlation between predictors; and frequently, after a first selection of predictors, what is important is the correlation with the residuals.
- Combined techniques do not seem particularly useful, at least in connection with SOLS or IPW.
- LASSO in the unique case studied behaves rather well. It seems useful to collect more information about the performance of this technique.
- Parsimonious techniques lose the benefit of the synergism of good correlate predictors. The results with *Artificial* show that this loss is not important. This is the case (frequent in multivariate calibration) where the main source of error is the determination of the response with a reference technique. In other cases, where the error on the predictors is large compared with that of the response, the use of synergism can be very useful.
- Flexibility and computer time are other elements that here have not been studied in detail. The probability level can noticeably modify the number of predictors selected by UVE, MUT, and GOLPE. This quality has a positive and a negative side. The positive side is the obvious possibility of an optimum probability level. The negative is that usually the chemist has no time to spend in the optimization of a chemometric procedure, with the risk of overestimating performances.
- Computer time is very large for GA-based techniques, for ISE (when only one predictor is eliminated in each cycle) and for GOLPE. With MUT, UVE, and MAXCOR it is almost equivalent. SOLS, with a limit (e.g., ten) in the selectable predictors and IPW (with 10–12 cycles, generally sufficient) can be regarded as the methods that require less computer time.
- Finally, we think that for real problems it is important to have at least two different techniques available, to use a software "transparent" both in the description of the method applied and in the computing details. Care in the refinement of the calibration model can be

a very important element to define the quality of the overall analytical procedure.

Acknowledgements Study developed with funds from the University of Genova.

References

1. Lucasius CB, Kateman G (1991) Trends Anal Chem 10:254–281
2. Leardi R, Boggia R, Terrile M (1992) J Chemometrics 6(5):267–281
3. Brown PJ, Vannucci M, Fearn T (1998) J Chemometrics 12:173–182
4. Martens H, Naes T (eds) (1989) Multivariate calibration. Wiley, Chichester
5. Frank IE (1987) Chemometrics Intell Lab Syst 1:233–242
6. Kettaneh-Wold N, MacGregor JF, Wold S (1994) Chemometrics Intell Lab Syst 23:39–50
7. Lindgren F, Geladi P, Rannar S, Wold S (1994) J Chemometrics 8:349–363
8. Forina M, Drava G, De La Pezuela C (1986) Sixth chemometrics in analytical chemistry conference (CAC), Tarragona, June 25–29, Abstract Book, PII-29
9. Cruciani G, Clementi S, Pastor M (1998) GOLPE-guided region selection. In: Kubinyi H, Folkers G, Martin YC (eds) 3D-QSAR in drug design. Recent advances. Kluwer, Dordrecht
10. GOLPE background, at http://www.miasrl.com/software/golpe/manual/background.html
11. Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB (2000) Applied Spectrosc 54:413–419
12. Höskuldsson A (2001) Chemometrics Intell Lab Syst 55:23–38
13. Kennard RW, Stone LA (1969) Technometrics 11:137–148
14. Snee RD (1977) Technometrics 19:415–428
15. Shao J (1993) J Comput Graph Stat 88:486–494
16. Breiman L, Spector P (1992) Int Stat Rev 60:291–319
17. Kowalski BR, Seasholtz MB (1991) J Chemometrics 5:129–145
18. Van der Voet H (1994) Chemometrics Intell Lab Syst 25:313–323
19. Haaland D, Thomas E (1988) Anal Chem 60:1193–1202
20. Thomas E, Haaland D (1990) Anal Chem 62:1091–1099
21. Osten D (1988) J Chemometrics 2:39–48
22. Faber NM (2001) Anal Chim Acta 432:235–240
23. Massart DL, Vandeginste BGM, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J (eds) (1998) Handbook of chemometrics and qualimetrics, part A. Elsevier, Amsterdam
24. Belsley DA, Kuh E, Welsch RE (eds) (1981) Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New York
25. Garrido Frenich A, Jouan-Rimbaud D, Massart DL, Kuttatharmmakul S, Martinez Galera M, Martinez Vidal JL (1995) Analyst 120:2787–2792
26. Boggia R, Forina M, Fossa P, Mosti L (1997) Quant Struct Activity Relationships (QSAR) 16:201–213
27. Forina M, Casolino C, Pizarro Millán (1999) J Chemometrics 13:165–184
28. Centner V, Massart DL, de Noord OE, de Jong S, Vandeginste BM, Sterna C (1996) Anal Chem 68:3851–3858
29. The Unscrambler, Camo ASA, Oslo
30. Westad F, Martens H (2000) J Near Infrared Spectrosc 8:117–124
31. Efron (eds) (1982) The Jackknife, the bootstrap and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia
32. Ojelund H, Madsen H, Thyregod P (2001) J Chemometrics 15:497–509
33. Tibshirani R (1996) J R Stat Soc Ser B 58:267–288
34. Forina M, Lanteri S, Armanino C, Casolino C, Cerrato Oliveros C (2003) V-PARVUS Release. An extendable package of programs for explorative data analysis, classification and regression analysis. Dip Chimica e Tecnologie Farmaceutiche, University of Genova. Free available at http://www.parvus.unige.it
35. Forina M, Drava G et al (1995) Chemometrics Intell Lab Syst 27:189–203
36. Kalivas JH (1997) Chemometrics Intell Lab Syst 37:255–259