Anal. Chem. **1996,** *68,* 3851−3858

Elimination of Uninformative Variables for Multivariate Calibration

Vítězslav Centner and Désiré-Luc Massart*

ChemoAC, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussel, Belgium

Onno E. de Noord

Koninklijke/Shell International Chemicals B. V., Amsterdam, P.O. Box 38 000, 1030 BN Amsterdam, The Netherlands

Sijmen de Jong and Bernard M. Vandeginste

Unilever Research Vlaardingen, P.O. Box 114, 3130 AC Vlaardingen, The Netherlands

Cécile Sterna

Rhone-Poulenc Industrialisation, C. R. I. T. Décines, 24 Avenue Jean-Jaurès, 69153 Décines Charpieu Cedex, France

A new method for the elimination of uninformative variables in multivariate data sets is proposed. To achieve this, artificial (noise) variables are added and a closed form of the PLS or PCR model is obtained for the data set containing the experimental and the artificial variables. The experimental variables that do not have more importance than the artificial variables, as judged from a criterion based on the *b* coefficients, are eliminated. The performance of the method is evaluated on simulated data. Practical aspects are discussed on experimentally obtained near-IR data sets. It is concluded that the elimination of uninformative variables can improve predictive ability.

The quality of a multivariate calibration model depends, among others, on the quality of the objects and the quality of variables. Potential sources of problems can be the presence of inhomogeneities (outliers or clusters) and the presence of noisy or random variables. The detection of inhomogeneities was the subject of the previous paper;[1] in the present article the attention will be focused on the quality of variables.

A method is proposed to eliminate those variables that are clearly uninformative since they do not contain more information than random variables. Such variables must lead to less precision (higher variance due to an imbedded error[2,3]) and, as shown by Faber et al.,[3,4] to a higher bias in the eigenvalues that reproduce the data matrix **X** by the correct number (A) of eigenvectors. The variance and bias both limit the predictive ability of the model (RMSEP). Therefore, the model built after the elimination of random variables, should be better.

THEORY

(1) Partial Least Squares (PLS). A PLS[5] model expresses the relation between a set of predictors **X** (n, p) and a variable **y** (n, 1) as

$$\mathbf{y} = \mathbf{X} \times \mathbf{b} + \mathbf{e} \tag{1}$$

where ***b*** (1, p) is the vector of PLS regression coefficients and **e** (n, 1) is the vector of errors that cannot be explained by the model.

In the original PLS method all variables are used; PLS is a so-called full-spectrum method. However, one can wonder whether it is useful to include all variables, because some of them may be noisy and/or contain nonrelevant information. We will, in this article, call such variables uninformative. Intuitively it would seem that better results should be obtained if such variables were eliminated. It should be noted that the goal in this article is not a variable selection in the sense that one tries to find the best (small) subset of variables for fitting or prediction of a model, but the elimination of those variables that are useless.

Faber et al.[3] published an error propagation study in principal component analysis. He described how uncertainties are carried over from the data to the estimated parameters and what the influence is of the measurement error and of the number of variables on the bias of the eigenvalues,[3,4] on the model complexity,[6] and on the variance[7] of the eigenvalues. The bias in the ath eigenvalue ($\mathbf{b}_{\lambda_a}$) has been defined[3] as

$$\mathbf{b}_{\lambda_a} = \hat{\lambda}_a - \lambda_a = (n + p - A)\sigma_M^2 \tag{2}$$

where $\hat{\lambda}_a$ is the biased estimate of the ath eigenvalue λ_a, n and p are the numbers of objects and variables, respectively, A is the correct dimensionality (i.e., pseudorank), and σ_M^2 is the measure-

(1) Centner, V.; Massart, D. L.; de Noord, O. E. *Anal. Chim. Acta* **1996**, *330*, 1−17.
(2) Malinowski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*; Wiley: New York, 1980.
(3) Faber, N. M.; Meinders, M. J.; Geladi, P.; Sjostrom, M.; Buydens, L. M. C.; Kateman, G. *Anal. Chim. Acta* **1995**, *304*, 257−271.
(4) Faber, N. M.; Meiders, M. J.; Geladi, P.; Sjostrom, M; Buydens, L. M. C.; Kateman, G. *Anal. Chim. Acta* **1995**, *304*, 273−283.
(5) Martens, H.; Naes, T. *Multivariate Calibration*; Wiley: Chichester, 1989.
(6) Faber, N. M.; Buydens, L. M. C.; Kateman, G. *Anal. Chim. Acta* **1994**, *296*, 1−20.
(7) Faber, N. M.; Buydens, L. M. C.; Kateman, G. *J. Chemom.* **1993**, *7*, 495−529.

S0003-2700(96)00321-6 CCC: $12.00

ment error. It is evident that a large measurement error (σ_M) and a large number (p) of (uninformative) variables will increase the bias in the eigenvalues and therefore also the model bias.

A number of methods to delete uninformative variables have been described in the literature. Martens[5] suggested replacing small loadings by zeros. A similar method called intermediate least squares (ILS) was described by Frank[8] and modified by Lindgren.[9,10] In interactive variable selection (IVS) the uninformative variables are detected for each PLS dimension by applying a threshold procedure to the vector of PLS weights. It was also shown by us[11] that deleting variables with small b coefficients in a model obtained with autoscaled data can be useful. Other methods are more directed toward variable selection, but those that could probably be used in variable elimination are due to Baroni[12,13] and Frank.[14] A weakness of all these methods is the estimation of a suitable number of variables (cutoff level). No explicit rule exists up to now. As a result, all approaches work with a user-defined number of variables or with a user-defined critical value for the considered selection criterion.

(2) Uninformative Variable Elimination by PLS (UVE-PLS). The method proposed here is based on an analysis of the b regression coefficients in eq 1. On the one hand, one can use the absolute values of the b coefficients in a model obtained for autoscaled data (b_w), as we did in ref 11. On the other hand, one can look at the reliability $c = b/s(b)$ of the b coefficient for only centered data. Both methods are studied and compared here, but the theory is concentrated on the latter. The reliability criterion c is based on an analogy with stepwise MLR. The fitness to enter the jth variable in MLR is determined by the ratio of the regression coefficient b_j and its standard deviation $s(b_j)$:

$$c_j = b_j/s(b_j) \quad \text{for} \quad j = 1, ..., p \tag{3}$$

The $s(b_j)$ for PLS coefficients cannot be computed directly. Therefore we propose to estimate b_j as a mean and $s(b_j)$ as a standard deviation from the vector of n b_{ij} coefficients obtained by (leave-one-out) jackknifing ($i = 1, .., n$). A robust variant is discussed further.

Another problem is how to estimate the cutoff level, below which the c_j (or any other criterion) are too small, without having to predefine it. It is proposed here to use artificial random variables, added to the data set and to compute their c values. As such variables should not be included in the model, because they represent (artificially added) noise, their c_j values will be indicative of the values that can be reached by uninformative variables. In this way, one should be able to obtain an appropriate estimate of the cutoff level. The experimental variables j that give c_j smaller than the maximum c value obtained for the artificial variables (c_{artif}) can then be considered uninformative: if $\text{abs}(c_j) < \text{abs}(\max(c_{artif}))$ the jth experimental variable is eliminated from the data (see Figure 1). A variant to this procedure is discussed further.

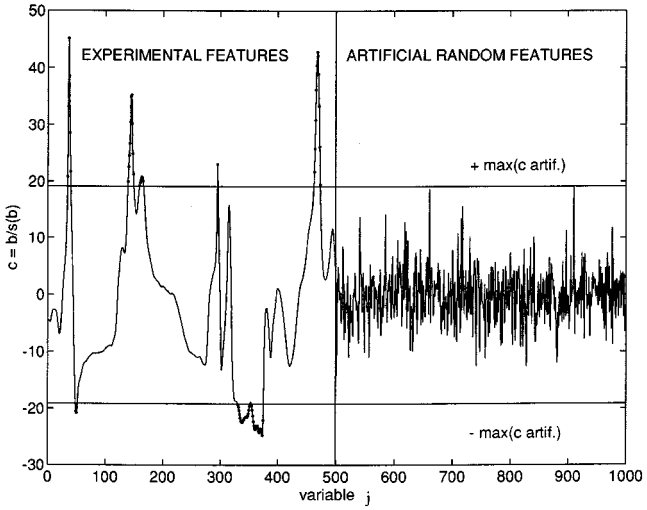

Figure 1. Plot of c for experimental (1−499) and artificial random (500−998) variables. The cutoff level at max(abs(c_{artif})) is indicated by the solid line.

The noise in the added artificial random variables should not be too large so that these variables do not influence the model. Indeed, as shown by Faber et al.[3] the imbedded error and the bias in eigenvalues depend on the measurement error. If it is artificially made too large by adding uninformative variables, then the error in the first eigenvalue would become so large that the second and following eigenvalues would also be affected, thereby having an effect on the b values for the variables in the original **X** matrix. This then could lead to rejection of wrong variables. It is therefore necessary to minimize this phenomenon by the multiplication of the artificial variables by a constant close to zero (for the absorbance data where the magnitude of the real variables is in the range of 0.0−1.0, the constant should be an order of magnitude smaller than the imprecision of the instrument, i.e., 1×10^{-4}. The proposed value here is 1×10^{-10}). The random variation in the artificial variables, needed for the computation of $s(b_j)$ is retained, but its influence on modeling is negligible.

The UVE-PLS algorithm can be summarized as follows:

1. Determination of the optimal model complexity (A) on **X**, with the lowest RMSEP as the criterion[15]

$$\text{RMSEP} = \left(\sum_{i=1}^{n}(\hat{y}_i - y_i)^2/n\right)^{1/2} \tag{4}$$

2. Generation of the artificial variable matrix **R**[16] and its multiplication by a small constant (10^{-10}). This yields the matrix **R** (n, p) with the number of variables p equal to the number of variables in **X**. The a priori probability to make an error in selection, i.e., to eliminate an informative or to retain an uninformative variable is then the same in both **X** and **R**. Inclusion of **R** with **X** (n, p). The resulting matrix is called **XR** (n, $2p$), the p first columns being those of **X** and the p last ones being those of **R**.

3. Calculation of PLS models for **XR** according to a leave-one-out procedure. The number of factors retained (A) is the same as for **X**. This yields n PLS models each with $2p$ regression coefficients b. They are collected in a matrix **B** (n, $2p$).

(8) Frank, I. *Chemom. Intell. Lab. Syst.* **1987**, *1*, 233−242.
(9) Lingren, F.; Geladi, P.; Rannar, S.; Wold, S. *J. Chemom.* **1994**, *8*, 349−363.
(10) Lingren, F.; Geladi, P.; Rannar, S.; Wold, S. *J. Chemom.* **1995**, *9*, 331−342.
(11) Garido Frenich, A.; Jouan-Rimbaud, D.; Massart, D. L.; Martinez Galera, M.; Martinez Vidal, J. L. *Analyst* **1995**, *120*, 2787−2792.
(12) Baroni, M.; Clementi, S.; Cruciani, G.; Constantino, G.; Riganelli, D.; Oberrauch, E. *J. Chemom.* **1992**, *6*, 347−356.
(13) Baroni, M.; Constantino, G.; Cruciani, G.; Riganelli, D.; Validi, R.; Clementi, S. *Quantit. Struct.-Act. Relat.* **1993**, *12*, 9−20.
(14) Frank, I. *Chemometrics95*, Pardubice, plenary lecture.
(15) Thomas, V. *Anal. Chem.* **1994**, *66*, 795−804.
(16) *Matlab, Reference Guide*, The MathWorks, Inc., South Natick, MA, 1992.

4. Determination for each variable j (i.e., both the experimental and random variables) of b_j ($b_j = \sum_{i=1}^{n} b_{ij}/n$), i.e., the mean of the column vector j from **B** and the standard deviation of that column vector

$$s(b_j) = \left(\sum_{i=1}^{n}(b_{ij} - b_j)^2/(n-1)\right)^{1/2} \quad (5)$$

5. Determination for each variable j of the criterion $c_j = b_j/s(b_j)$.

6. Determination of max(abs(c_{artif})), i.e., the highest absolute value of c among all c for artificial variables.

7. Elimination from **X** of the experimental variables for which abs(c_j) < abs(max(c_{artif})), for $j = 1, .., p$. The remaining variables constitute the new **X** matrix, $\mathbf{X_{new}}$.

8. Building of the final PLS leave-one-out cross-validated models on $\mathbf{X_{new}}$ and prediction $\hat{y}$ with A factors.

9. Quantification of the predictive ability of the new model as the cross-validated $RMSEP_{new}$ according to eq 4.

10. If (a) $RMSEP_{new}$ > RMSEP one concludes that the elimination of uninformative variables did not improve modeling and the algorithm is terminated. Otherwise if (b) $RMSEP_{new}$ < RMSEP, one will first wonder whether A was not too large (overfitting), due to the uninformative variables which could have influenced the selection (it is extremely improbable that A was too small due to uninformative variables). In order to check this possibility, the algorithm starting with a new selection on **XR** (point 2) is repeated again for $A = A - 1$ and the original RMSEP is replaced by the $RMSEP_{new}$. When the reduction of A to $A - 1$ does not improve modeling ($RMSEP_{new}$ > RMSEP), the algorithm terminates in 10 (a).

During the development of the algorithm the following variants were tested. (a) UVE-M: a version robust to outliers. The c criterion is replaced by its robust version, c_j = (median(b_j)/interquartile range(b_j)). (b) UVE-α: a variant that eliminates a strong dependence of the cutoff level on the largest c_{artif}. Instead of the max(abs(c_{artif})) value one finds the cutoff level among the ranked abs(c_{artif}) as the value that corresponds to the 99% (95, 90 = α) quantile. With this modification one eliminates somewhat less variables but perhaps also avoids eliminating some informative variables. As a result, broader (spectral) bands are used for modeling. (c) b_w-α: Instead of using c as a criterion to eliminate variables by comparison with artificial noise variables one could try to use the PLS b coefficients for autoscaled data (b_w) in the same way, i.e., compare b_w values for experimental variables and artificial noise variables. Depending on the quantile of $b_{w\,artif}$ used, we will call this method b_w-α-100 when max(abs($b_{w\,artif}$)) is applied and b_w-α-99, -95, or -90, respectively, when the cutoff level is shifted as described above.

(3) Genetic Algorithm (GA). A genetic algorithm is applied as a variable selection method in this study. The GA used here was originally developed by Leardi[17,18] and modified by Jouan-Rimbaud.[19]

(4) Preprocessing. Several preprocessing methods were carried out, namely, centering, autoscaling, off-set correction, standard normal variate (SNV), and multiplicative scatter correction (MSC). The three last methods are concerned with a baseline shift. Centering is the subtraction of the corresponding column mean from each element of the data matrix and is nearly always applied in PLS modeling. Autoscaling is centering combined with normalization (dividing of each matrix element by its corresponding column standard deviation). This preprocessing is applied in the b_w method.[11]

In off-set correction one subtracts, row by row for the whole matrix considered, the row average of a few (1−5) first variables (columns) from each element of the corresponding row of data matrix **X.** SNV transformation[20] corrects each ith spectrum (row) separately by subtraction of the row mean and normalizing in the row direction. In MSC[21] each individual spectrum i (row) is regressed against the mean spectrum in a window or windows of wavelengths not affected by the characteristic or concentration one is determining. The obtained parameters (slope and intercept) are used to correct the spectrum.

EXPERIMENTAL SECTION

(1) Simulated Data. Noise-free data, **SIM:** generation of a matrix of random numbers from 0 to 1 (**S1**) with dimensionality (25, 100); PCA on the centered matrix **S1**, yielding scores and loadings; definition of the complexity, $A = 5$; multiplication of the first five score vectors (25, 5) by the first five loading vectors (5, 100) giving a simulated pure data matrix **SIM** (25, 100) that does not contain any noise; PCA on **SIM** yields relative eigenvalues (%) 23.02, 21.28, 19.50, 18.74, 17.46, 0, 0, ... The complexity of **SIM** is therefore indeed exactly $A = 5$.

A noise-free variable y is defined as $y = 5 \times$ scores(1) + 4 × scores(2) + 3 × scores(3) + 2 × scores(4) + 1 × scores(5), where scores(1) is the vector of scores on PC1.

SIMUI incorporates the noise-free data matrix SIM and additionally an uninformative variable matrix UI: generation of a matrix of random numbers from 0 to 1 (UI) with dimensionally (25,100); attachment of the matrix UI(25,100) to SIM(25,100) results in SIMUI(25,200), SIMUI = [SIM,UI].

SIMUIN is the matrix sum of the SIMUI data and a noise matrix (N): creation of a noise matrix N(25,200), with elements from 0 to 0.005, i.e., small compared to the signal 0−1 in SIMUI-(25,200); summation of SIMUI and N gives SIMUIN (= SIMUI + N).

(2) Experimental Near-IR Data Sets. *CIS−TRANS*: calibration of trans double bond content in a fatty acid mixture; transmittance FT-IR spectra (1600−900 cm^{-1}) of eight standards at concentrations 0, 2, 3, 4, 5, 7.5, and 10% trans. All measurements were duplicated, one sample, at concentration 4, was triplicated. The "window" used for the MSC correction was 1600−1000 cm^{-1}.

This data set contains one outlier (the object at concentration 4). It is also known that the range of wavenumbers 1000−900 cm^{-1} is chemically relevant since the cis−trans content was determined in the past, from the wavenumber at 967 cm^{-1} using an internal standardization procedure. An internal standard was the peak at wavenumber 1435 or at 1465 cm^{-1}.

POLY-DAT: determination of hydroxyl number of polyether polyols by (N)near-IR.

(17) Leardi, R.; Boggia, R.; Terrile, M. *J. Chemom.* **1992**, *6*, 267−281.
(18) Leardi, R. *J. Chemom.* **1994**, *8*, 65−79.
(19) Jouan-Rimbaud, D.; Massart, D. L.; Leardi, R.; de Noord, O. E. *Anal. Chem.* **1995**, *67*, 4295−4301.
(20) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. *Appl. Spectrosc.* **1989**, *43*, 772−777.
(21) Isaksson, T.; Naes, T. *Appl. Spectrosc.* **1988**, *42*, 1273−1284.

Table 1. Data Sets SIM, SIMUI, and SIMUIN: Predictive Ability (RMSEP) of the Models Obtained from Different Calibration Methods[a]

complexity	1	2	3	4	5
SIM					
PLS	1.12	0.12	0.0096	0.0001	1.74×10^{-15} [b]
UVE-PLS	0.66 (39)	0.13 (91)	0.0112 (98)	0.0001 (100)	1.74×10^{-15} (100)
UVE-M	0.79 (29)	0.11 (86)	0.0112 (98)	0.0001 (100)	1.74×10^{-15} (100)
SIMUI					
PLS	2.85	2.32	2.25	2.21	2.19[b]
UVE-PLS	2.04 (11+0)	0.17 (62+0)	0.0170 (45+0)	0.00022 (50+0)	1.68×10^{-15} (68+0)
b_w(100)	1.63 (64+36)	1.12 (66+34)	0.8260 (66+34)	0.7474 (66+34)	0.6627 (66+34)
b_w(50)	1.06 (42+8)	0.82 (42+8)	0.4628 (44+6)	0.3618 (44+6)	0.1873 (45+5)
b_w(30)	0.61 (30+0)	0.25 (30+0)	0.0821 (30+0)	0.0084 (30+0)	1.60×10^{-15} (30+0)
r(100)	1.63 (64+36)	1.31 (64+36)	1.1370 (64+36)	1.0331 (64+36)	0.9550 (64+36)
r(50)	1.06 (42+8)	0.89 (42+8)	0.5730 (42+8)	0.4430 (42+8)	0.2662 (42+8)
r(30)	0.61 (30+0)	0.32 (30+0)	0.0821 (30+0)	0.0084 (30+0)	1.60×10^{-15} (30+0)
SIMUIN					
PLS(1−100)	1.12	0.12	0.0167	0.0129[b]	0.0129
PLS	2.85	2.33	2.25	2.21	2.19
UVE-PLS	1.23 (32+0)	0.17 (48+0)	0.0220 (55+0)	0.0132 (63+0)	0.0133 (43+0)
b_w(100)	1.63 (64+36)	1.12 (66+34)	0.8230 (66+34)	0.7473 (66+34)	0.6623 (66+34)
b_w(50)	1.06 (42+8)	0.81 (42+8)	0.4654 (44+6)	0.3650 (44+6)	0.1933 (45+5)
b_w(30)	0.61 (30+0)	0.25 (30+0)	0.0880 (30+0)	0.0156 (30+0)	0.0129 (30+0)
r(100)	1.63 (64+36)	1.31 (64+36)	1.1327 (64+36)	1.0292 (64+36)	0.9511 (64+36)
r(50)	1.06 (42+8)	0.89 (42+8)	0.5720 (42+8)	0.4419 (42+8)	0.2681 (42+8)
r(30)	0.61 (30+0)	0.32 (30+0)	0.0832 (30+0)	0.0158 (30+0)	0.0122 (30+0)

[a] The number of retained informative (1−100) plus the number of retained uninformative (101−200) variables is shown in parentheses. [b] The optimal complexity of the model.

The data set consists of 74 near-IR spectra (**X**) of polyether polyols and their hydroxyl numbers expressed in mg of KOH/g (**y**). The data were recorded on a NIRSystem Inc. Silver Spring, MD, instrument in the range of wavelengths from 1100 to 2158 nm. The duplicates or triplicates corresponding to one object were averaged. The measurements were off-set corrected and boundary wavelengths eliminated. The final dimension of **X** was 26 × 499.

SOLVENT: calibration of a solvent in a powder by near-IR spectrometry.

Near-IR spectra of 57 samples were measured in the range of wavelengths from 1000 to 2200 nm with a step 2 nm on a Bruker IFS/28 N instrument. Informative wavelengths were expected in the regions 1600−1800 and 2100−2200 nm (strong absorption), 1100−1200 nm (weak absorption).

COMPUTER PROGRAMS

Matlab for Windows, version 4.0 (The MathWorks, Inc.) was used to program all necessary procedures and to generate the normally distributed random matrices (**R**). The stepwise MLR models have been selected with SPSS.[22]

RESULTS

The predictive ability of the model with minimized influence (multiplication by the constant 10^{-10}) of the artificial random variables is found to be exactly the same as the predictive ability of the model without those variables. When the multiplication is not carried out, the predictive ability of the model is worse and the elimination of the original variables can be misleading.

The performance of UVE-PLS is evaluated from the predictive ability of the models after the elimination of the found uninformative variables. Leave-one(object)-out cross validated RMSEP was used to measure this performance and to compare UVE to the other elimination methods, namely, to the b_w method (the number of retained variables N is shown between the brackets), to the elimination based on an absolute univariate correlation coefficient $r(N)$ in decreasing order and to full-spectrum methods.

(1) Simulated Data. In a first step we wanted to keep as many data characteristics as possible under control, such as the data complexity, the experimental noise, the position of the random and nonrandom variables, possible outliers, and clusters or nonlinearity. The simulated data sets **SIM**, **SIMUI,** and **SIMUIN** (see Experimental Section) were prepared for this purpose.

Table 1 compares the predictive ability (RMSEP) of the different PLS variants obtained with successively 1, 2, ..., 5 PLS components, for different selection criteria (c, b_w, and r). The number of informative variables selected (range, 1−100) and the number of random variables selected (range, 101−200) is shown in parentheses. Because the selection based on r does not depend on the complexity (A), the selection of variables is, for a defined number of variables, constant for any A.

SIM. The ability of UVE-PLS to select as many informative variables as possible is evaluated on data set **SIM**. This data set does not include any uninformative variables.

For the true (simulated) complexity $A = 5$ the optimal solution was reached; all 100 variables were selected. Also for a wrongly chosen complexity $A = 4$ or 3 all 100 or 98% of all nonrandom variables was considered correctly. This shows that the algorithm truly retaines variables and does not eliminate any of them from the data when they are meaningful.

SIMUI. Data set **SIMUI** is equal to **SIM** + 100 random variables. Therefore the first 100 variables should not and the second 100 variables should be eliminated. The quality of the models obtained on the **SIMUI** data with different modeling

(22) Norusis, M. J. *SPSS for Windows, Base System User's Guide Release 5.0*, 1992.

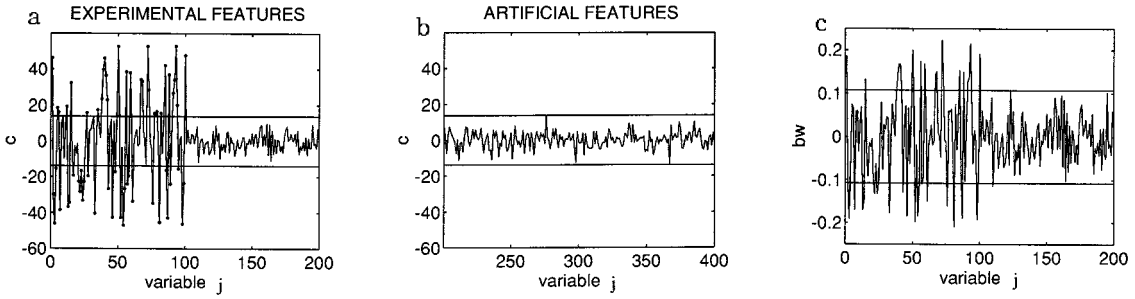

Figure 2. Comparison of the c (plots a and b) and the b_w criterion (plot c) on simulated data (SIMUI). The variables 1−100 are informative ones, and the variables 101−200 are uninformative ones. Plot b shows the cutoff level estimation on artificial random variables.

methods can be compared to the solution for the noise free **SIM** data (see the results for **SIM** in Table 1). The purpose of this exercise is to evaluate the ability of the method to distinguish the uninformative variables (101−200) from the informative ones (1−100).

The influence of the uninformative variables on PLS without variable elimination is, for any complexity, negative. For instance, for five PLS components, one finds RMSEP = 2.19 compared to RMSEP = 1.74×10^{-15} for the subset of all informative variables 1−100. This illustrates the need for eliminating uninformative variables to obtain optimal prediction.

UVE-PLS with five components gives a better solution than the other methods. All random variables and some of the informative ones are ignored. The obtained RMSEP is equal to 1.68×10^{-15}, i.e., similar to that reached for all informative variables 1−100. The values 1.68×10^{-15} and 1.74×10^{-15} are probably within the roundoff error of the computer and the difference between them is therefore not significant.

The b_w or r results are, in some cases (b_w(30), r(30)) comparable, but among the first 100 retained variables (the true number of the informative variables) only 66 are informative. The rest are random variables. As a result, the RMSEP on the $\mathbf{X}_{\mathbf{new}}$ with 100 retained variables (in Table 1 r(100) and b_w(100)) is much higher (0.9550 and 0.6627) than the RMSEP on the subset of all informative variables 1−100. Only when the user-defined number of retained variables is small enough (30 in this case), all random variables are ignored and a good model is found (RMSEP = 1.60×10^{-15}). To reach such a solution one has, however, to ignore 70% of the informative variables. The graphical comparison in Figure 2 shows that the criterion c is more selective than b_w or r.

When one ranks the abs(c) values, the 68 largest ones are due to the informative variables 1−100, the 69th value is the first due to an uninformative variable. When the same is applied to abs(b_w) or r, only the 30 largest values correspond with informative variables and the 31st is due to a random variable.

An estimation of the optimal number of variables to retain can be found by using the strategy with artificial variables. The b_w-α-100 indeed leads to the elimination of all uninformative variables (101−200) and RMSEP equal to 1.7×10^{-15} with 29 relevant variables. The number of retained variables found with b_w-α-100 (29) agrees with the optimal number estimated above (30), but there may be a problem. When one applies the b_w-α method, the evaluation of the useful and the uninformative variables can deteriorate since the random matrix **R** added to the original **X** is autoscaled, therefore having an influence on the selection: the variance of the artificial random variables (**R**) is no longer small compared to that of the real variables.

SIMUIN. This data set was prepared to investigate the influence of noise, added to the experimental informative variables, on the selection. To evaluate results, one should compare the predictive ability of the models after the variable elimination to the RMSEP that is obtained on the subset of all true informative variables 1−100 with the added noise (in Table 1 indicated as PLS(1−100)).

Due to the noise the optimal cross-validated complexity for **SIMUIN** is found to be equal to 4. UVE-PLS eliminates successfully all 100 uninformative variables from the data set together with some of the informative ones. The predictive ability of the final model is comparable to the ideal RMSEP, increases from 0.0129 only to 0.0132, i.e., less than 3%. The b_w(30) or r(30) selection leads to a slightly worse prediction as a result of the elimination of too many (70) informative variables. When, however, one takes into account more b_w or r selected variables, some of them are random ones, so that neither the b_w(50,100) nor the r(50,100) models do lead to a better modeling.

As discussed earlier c is more selective than b_w. To illustrate this fact, further suppose the following simple example. The variable $\mathbf{y} = [0\ 20\ 40\ 60\ 80\ 100]^T$.

The first 10 variables in **X** are created to be collinear with **y** with proportionality coefficients: −0.002 −0.004 −0.001 0.001 0.0015 0.003 0.008 0.005 0.003 0.001.

The last column in **X** is a vector: $[0.009\ 0.001\ 0.007\ 0.008\ 0.006\ 0.004]^T$ that simulates an uninformative random variable.

By applying PCA on **X** as well as on autoscaled **X** one concentrates the meaningful part of the total variance (due to the variables 1−10) into the first latent variable. PC2, on the other hand, explains only the error in data, i.e., the variable number 11. The inspection of relative eigenvalues, however, shows that in the latter case (autoscaled data) there is a strongly increased importance of PC2 (from 0.005 to 8.8%). This indicates that by autoscaling the signal (PC1) to noise (PC2) ratio is lowered from the original 99.995/0.005 to 91.2/8.8 and the b_w selection can therefore easily be influenced by a noise.

It is also noteworthy, that the ratio of the c criteria for the informative (1−10) and the artificial random (12−22) variables $c_{\text{informative}}/c_{\text{artif}}$ is much larger than the $b_{w\ \text{informative}}/b_{w\ \text{artif}}$ ratio. Using c therefore yields better ability to discriminate the useful and the uninformative variables than using b_w. It must be concluded that max(abs($b_{w\ \text{artif}}$)) could be so large that some of the informative experimental variables could be removed.

(2) Experimental Data Sets. *CIS−TRANS.* This data set is an example of a situation where it is known that most of the information is concentrated in a relatively small spectral range and that large areas of the spectrum are uninformative. The

Table 2. Data set CIS−TRANS (700 Variables) RMSEP Obtained for Different Calibration Methods and Different Preprocessing[a]

	preprocessing			
	raw	off-set	SNV	MSC
	CIS−TRANS			
A967	1.92	1.75	0.95	1.03
967/1435	4.09	1.17	0.71	2.38
967/1465	3.97	1.03	1.25	1.81
MLR (successively wavenumbers cm^{-1})	1.90 (966)	1.70 (965)	0.95 (967)	1.03 (967)
	0.85 (975)	0.80 (912)	0.52(985)	0.48 (951)
		0.54 (951)	0.42 (973)	0.40 (985)
PLS (compl)	1.25 (3)	1.10 (4)	1.25 (3)	1.26 (3)
UVE-PLS (compl - retained var)	0.96 (3−43)	0.98 (3−16)	0.73 (3−19)	0.67 (3−23)
	CIS−TRANS-4			
A967	2.21	2.23	0.47	0.60
967/1435	4.74	0.43	0.40	1.50
967/1465	4.64	0.53	0.97	1.94
MLR (successively wavenumbers cm^{-1})	2.18 (966)	2.11 (965)	0.44 (968)	0.57 (968)
	0.87 (911)	0.63 (911)	0.21 (951)	0.21 (951)
	0.84 (1011)	0.46 (951)	0.15 (985)	0.15 (985)
	0.63 (945)			
PLS (complexity)	0.83 (4)	0.83 (3)	0.69 (2)	0.68 (2)
UVE-PLS (compl - retained var) (wavenumber cm^{-1})	1.18 (3−28)	1.25 (3−26)	0.22 (2−20) 978−963 959−956	0.25 (2−27)
b_w (compl - retained var) (wavenumber cm^{-1})	1.05 (3−30)	1.20 (3−20)	0.27 (2−25) 1405−1403 979−958	0.25 (2−35)

[a] For MLR the selected variables are shown in parentheses. For PLS, UVE and b_w are indicated: the model complexity, the number of retained variables, and, for the best models, the list of retained wavenumbers.

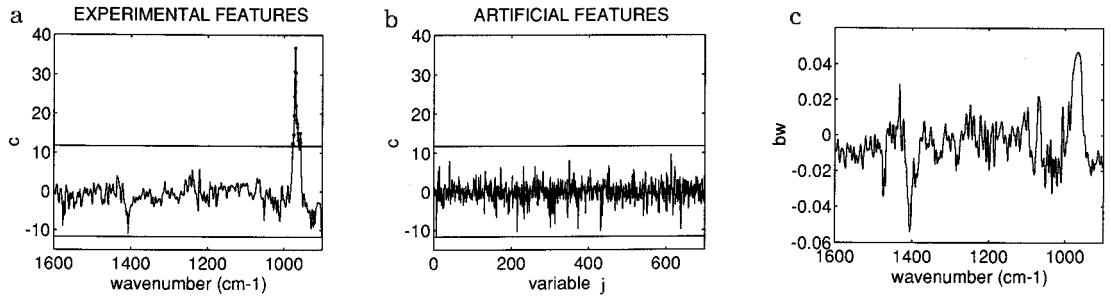

Figure 3. Elimination of uninformative experimental variables in CIS−TRANS data set (plots a and b) on the *c* plot and b_w plot (the outlier was eliminated from data).

original method applied was univariate (wavenumber 967 cm^{-1}) or bivariate (a ratio of two wavenumbers: $\lambda 1/\lambda 2$). As is shown in Table 2, these methods are considerably improved by correcting the data with SNV.

When PLS is applied to raw, SNV, or MSC pretreated data, it yields results that are of similar quality. The best result is, however, one of the bivariate procedures after SNV correction (although it is difficult to state whether the difference is significant or not).

By elimination of uninformative variables with UVE-PLS, one obtains models that are always better than the corresponding full-spectrum PLS models and require only 16 to 43 variables. The region selected is that around 967 cm^{-1}, which is indeed known to be informative, thereby showing that the retained wavenumbers correspond with what one would reasonably expect. We also applied MLR, and interestingly, it should be noted that these models are clearly better.

One object (at the concentration 4) is an outlier; therefore, its elimination improves prediction. The results (Table 2) confirm to a large extent the conclusions obtained with all points. One also concludes that the optimal model complexity has been, by the outlier elimination, reduced from three to two.

Surprisingly, it should be also noted, that the b_w analysis introduces to the subset of the retained variables some of the wavenumbers around 1405 cm^{-1} [compare the b_w plot (Figure 3c) with the *c* plot (Figure 3a)].

It is not clear how to interpret these additional variables, because they do not bring any difference to the model predictive ability compared to UVE.

One interesting feature of the elimination method is the following. The results show that MSC (in the same way as SNV) improves the results. In MSC one usually selects a zone in the spectrum that is not affected by the characteristic being measured to be able to do the correction. Choosing this zone is not always evident. The present method allows one to find zones that contain variables uninformative for the determination of *y* and such zones are useful for the MSC procedure.

POLY-DAT. The investigation of this data set was described earlier.[1,19] The predictive abilities of PLS, MLR, and GA models applied to the full spectrum are shown in Table 3a. In Table 3b

Table 3. Data Set POLY-DAT (499 Variables): RMSEP Obtained for Different Calibration Methods[a]

(a) Modeling Applied to the Original (Full) Data

method	PLS	MLR	GA
complexity	6		
variables	499	6	6
RMSEP	1.86	1.56	1.09
(λ nm)		1430	2062
		1878	1776
		1974	1760
		2064	1838
		1210	1566

(b) Elimination of Uninformative Variables and Final Variable Selection

method	b_w(250)	b_w-α-100	b_w-α-99	b_w-α-95
complexity	6	6	6	6
variables	250	117	162	195
RMSEP	1.70	2.65	1.72	1.73
(λ nm)			1406−1584	
			1832−1840	
			1996−2128	

method	UVE-PLS	UVE-α-99	UVE-α-95	UVE-α-90	MLR	GA
complexity	5	5	5	5		
variables	80	130	247	301	4	4
RMSEP	1.61	1.55	1.54	1.61	1.16	1.15
(λ nm)	1198−1208				2064	2064
	1228−1232				1208	1810
	1408−1426				1810	1208
	1448−1460				1456	1458
	1720					
	1792−1832					
	1840−1880					
	2056−2076					

[a]The optimal complexity and the selected variables or spectral regions (λ, nm) are shown also.

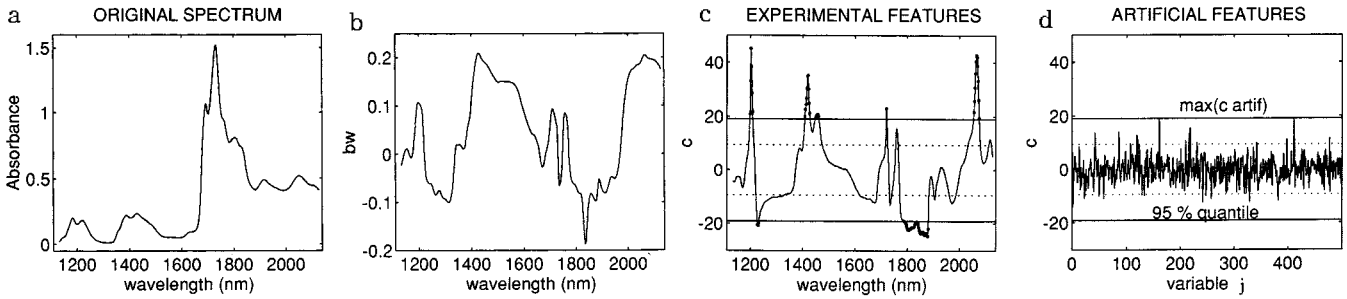

Figure 4. (a) Original spectrum of one object from data set POLY-DAT and (b) the b_w and the c plots for (c) the experimental and (d) the artificial random variables.

are the optimal results obtained with the b_w method, UVE-PLS, its modifications as well as the final MLR and GA models on the UVE-PLS preselected variables.

The RMSEP obtained with six PLS components on the original 499 variables is reduced from 1.86 to 1.61 (1.54) by UVE-PLS (α-95) with five PLS components and 80 (247) variables and to 1.70 by the selection on b_w (250 variables, 6 PLS components). Figure 4 shows (a) the original spectrum of one object and (b) the corresponding b_w plot as well as the c plot for (c) the experimental and (d) the artificial variables.

The b_w-α-100 method leads, probably due to the autoscaling of the extended data matrix **XR**, to a deteriorated distinction between the useful and the uninformative variables. Consequently, only two spectral bands (b_w maximum at 1440 nm and 2070 nm, 117 variables, Figure 4b) are retained and the predictive ability of the model is decreased (RMSEP increases to 2.65). When the cutoff level estimated on $b_{w\,\mathrm{artif}}$ is lowered to 99% a model of similar quality as the optimal b_w(250) model is reached.

The complexity of the PLS models after the elimination of uninformative variables with UVE is lower which seems to indicate overfitting in the original full-spectrum solution. This is indirectly verified also by UVE-α modification. If the cutoff level is lowered and more and more experimental variables considered, the RMSEP at first decreases slightly and then again increases due to the presence of too many variables.

When the genetic algorithm is applied to the full spectrum to select variables for a multiple linear regression model, a solution is found that is similar in dimensionality to the PLS solution but with a lower RMSEP (1.09).

Interestingly the GA applied on the 80 variables remaining after variable elimination on the PLS solution, requires only four variables. Because cross-validation for GAs is difficult, there is some concern that GA variable selection could lead to selection based on chance correlations. Preselection of the variables

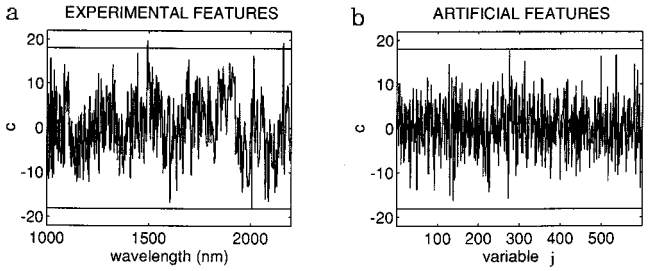

Figure 5. Visualization of the fact that the data set SOLVENT does not contain significanly more information (plot a) than the artificial random variables (plot b).

diminishes this possibility. It is noteworthy that the obtained solution requires only four variables. This lends support to the conviction that full-spectrum PLS can lead to overfitting. It also should be mentioned here that convergence of the GA on the preselected variables was reached after a much smaller number of generations.

Still more interestingly, classical stepwise multiple linear regression (MLR) on these 80 variables also requires 4 variables and the model (with an RMSEP of 1.16) obtained is nearly identical with that obtained by GA since the same variables are selected: only one of them is slightly shifted by about 2 nm. When MLR is applied to the full spectrum it requires six variables and leads to a worse RMSEP than with four. One might conclude from this observation, that there is no real need for the final PLS solution nor in fact for the genetic algorithm. The PLS approach is, however, required in the initial stage to be able to select the interesting parts of the spectrum. Of course, it is not possible to decide that this is always true, but the outcome is intriguing enough to investigate further whether this is generally the case.

SOLVENT. We have tried many preprocessing methods, the elimination of possible outliers, nonlinear modeling, variable selection, etc., to obtain a useful calibration model. The obtained RMSEP was, however, never better than 15% relative. Figure 5 shows that these efforts were doomed to be vain.

Indeed, the *c* plot for UVE-PLS shows that there is no more information in the experimental than in the artificial random variables, so that a good calibration model could not have been expected. By applying this methodology from the start, one could have concluded this immediately and avoided the costly efforts trying to make a model when this is not possible.

CONCLUSION

The application of UVE-PLS allows one to eliminate uninformative variables in multivariate data sets before a final modeling is carried out. The method has two advantages compared to the other selection methods: (1) the used criterion is rather selective; (2) the level to cut the uninformative experimental variables is user-independent.

The simulations show that UVE-PLS is a way to eliminate those variables that clearly have no interest. Compared to other methods, it keeps a larger number of the relevant variables for the final modeling. The experimental data sets lead to the conclusion that the method significantly improves prediction (compared to PLS) in the cases when the data set contains many uninformative variables (for example CIS−TRANS data). In an extreme instance it indicates that there is no more information in the real than in the random variables (SOLVENT). In some cases one eliminates much less experimental variables (POLY-DAT), which shows that many variables carry information. In such a case the improvement of RMSEP is not so evident. Nevertheless, the modeling is simplified and the final dimensionality can be lower than the original.

Of the two criteria tested, the *c* criterion appears to be fundamentally better, but it is more difficult to apply. The b_w criterion can be applied when this is considered to be a problem.

The elimination of the random variables (UVE) can be considered as a general preselection procedure. It avoids problems in the subsequent application of MLR (with or without GA). The results seem to indicate that after UVE the complexity of the GA or MLR model can be smaller than the original one. This means that some overfitting might occur in the models using all data.

The method was applied to PLS, but it can be considered to be equally useful for PCR or other related methods.

ACKNOWLEDGMENT

We thank Mr. Theo Meier from Shell for careful collection of the POLY-DAT data and Dr. B. Walczak for useful discussions and FGWO for financial help.

Received for review April 2, 1996. Accepted July 11, 1996.[⊗]

AC960321M

[⊗] Abstract published in *Advance ACS Abstracts,* August 15, 1996.