JOURNAL OF CHEMOMETRICS
J. Chemometrics **13,** 165–184 (1999)

ITERATIVE PREDICTOR WEIGHTING (IPW) PLS: A TECHNIQUE FOR THE ELIMINATION OF USELESS PREDICTORS IN REGRESSION PROBLEMS

M. FORINA,[1]* C. CASOLINO[1] AND C. PIZARRO MILLAN[2]
[1]*Dipartimento di Chimica e Tecnologie Farmaceutiche ed Alimentari, Università di Genova, Via Brigata Salerno (Ponte), I-16147 Genova, Italy*
[2]*Departamento de Química, Universidad de la Rioja, Obispo Bustamante 3, E-26001 Logroño, Spain*

SUMMARY

A new method for the elimination of useless predictors in multivariate regression problems is proposed. The method is based on the cyclic repetition of PLS regression. In each cycle the predictor importance (product of the absolute value of the regression coefficient and the standard deviation of the predictor) is computed, and in the next cycle the predictors are multiplied by their importance. The algorithm converges after 10–20 cycles. A reduced number of relevant predictors is retained in the final model, whose predictive ability is acceptable, frequently better than that of the model built with all the predictors. Results obtained on many real and simulated data are presented, and compared with those obtained from other techniques. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS: multivariate calibration; feature selection; validation

INTRODUCTION

Regression techniques are widely utilized in chemistry, mainly in analytical chemistry for multivariate calibration and in medicinal chemistry for QSAR problems. In both cases a regression technique searches for a mathematical model which gives the value of a response variable as a function of a number of predictors. Frequently many predictors are useless (only noise), and many useless predictors cause a worsening of the predictive ability of the regression model. For this reason many techniques have been developed to build the regression model only with relevant predictors.

These techniques can be classified in three categories.

(a) *Subset selection.* A number of regression models are built by different subsets of the predictors; the performance of the models is evaluated, and it is used to search for other subsets. The most important example of this class is selection by means of genetic algorithms (GAs).[1] GAs have

* Correspondence to: M. Forina, Dipartimento di Chimica e Tecnologie Farmaceutiche ed Alimentari, Via Brigata Salerno (Ponte), I-16147 Genova, Italy. E-mail: forina@anchem.unige.it

Contract/grant sponsor: European Union; Contract/grant number: SMT4-CT95-2031.
Contract/grant sponsor: University of Genoa.
Contract/grant sponsor: MURST.
Contract/grant sponsor: CNR.

CCC 0886–9383/99/020165–20 $17.50
Received 2 November 1998
Accepted 15 January 1999

Copyright © 1999 John Wiley & Sons, Ltd.

the advantage of exploring fairly well the space of all possible subsets in a large but reasonable time, much less than that required by the study of all possible subsets. Moreover, GAs offer the choice between a number of possible optimal or near-optimal subsets. The model performance is evaluated in predictive optimization, so that it is necessary to use an external set to evaluate the true predictive ability of the selected subsets, with a heavy increase in computing time.[2]

(b) *Dimension-wise selection.* A biased regression technique is built progressively, or by addition of predictors, as in stepwise ordinary least squares (SOLS) regression, or by addition of 'latent variables', the principal components of principal component regression (PCR) or partial least squares (PLS) regression. Dimension-wise techniques work on a single dimension. Martens and Naes[3] suggested replacing with zero the small PLS weights in each latent variable, so that the corresponding predictors are cancelled from the latent variable, but can be used in one or more of the following. Frank[4] improved this procedure in the technique called intermediate least squares (ILS). Different strategies, all working on the PLS weights, were used by Kettaneh-Wold *et al.*,[5] Lindgren *et al.*[6] (interactive variable selection—IVS) and Forina *et al.*[7] (automatic variable selection—AVS)

(c) *Model-wise elimination.* The regression model is developed with all the predictors. Then useless predictors are eliminated on the basis of the value of their regression coefficient b in the regression model

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

The elimination of predictors with small regression coefficients (provided that the regression model has been computed with autoscaled predictors) was suggested.[8] This procedure will be indicated by BAUT.

ISE (iterative stepwise elimination)[9] is based on the importance of the predictors, defined as

$$z_v = \frac{|b_v| s_v}{\sum_{v=1}^{V} |b_v| s_v} \tag{2}$$

where s_v is the standard deviation of the predictor v. In each elimination cycle the predictor with the minimum importance is eliminated, and the model computed again with the remaining predictors. The final model is that with the maximum predictive ability.

UVE-PLS (uninformative variable elimination PLS)[10] 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

$$c_v = \frac{b_v}{s_{b_v}} \tag{3}$$

The maximum of the absolute value of the coefficient c_v for the added artificial predictors is the cut-off value for the elimination of non-informative original predictors. In UVE-α the cut-off value is the α% value of the distribution of the absolute values of c_v, so that more original variables and some uninformative predictors are retained.

Stepwise OLS (below simply SOLS) is based on the use of an F-test. Here SOLS was modified to made the selection predictively. Objects are divided into the cancellation groups of cross-validation.

For each cancellation group, SOLS selects a number of predictors, and for each SOLS step the prediction error is evaluated on the left-out objects. Finally the prediction variance is obtained as a function of the number of entered predictors, and its minimum indicates the suitable number. The corresponding predictors are selected in a final run with all the objects in the training set.

In this paper a new procedure for model-wise elimination of useless predictors is presented. This procedure is based on the cyclic repetition of the PLS algorithm, each time multiplying the predictors by their importance computed in the previous cycle (unity in the first cycle). The performances of this procedure (IPW-PLS—iterative predictor weighting PLS) are shown on several artificial and real data sets, and compared with results obtained by UVE-PLS, ISE-PLS and SOLS.

THEORY

IPW algorithm

The PLS-1 (one response variable) algorithm is based on the marginal regression (straight line through the origin, generally after centring) of each predictor on the response. The vector of slopes $\mathbf{w}$ ($\mathbf{w}^T = \mathbf{y}^T \mathbf{X}/\mathbf{y}^T \mathbf{y}$) is normalized ($\mathbf{w}_{new} = \mathbf{w}_{old}/\|\mathbf{w}_{old}\|$), so that its elements (the PLS weights) become direction cosines, which identify the first PLS latent variable. The weight of a given predictor depends on its correlation coefficient with the response and on its magnitude, so that it is a function of the pre-treatment. A useless predictor always has a correlation coefficient slightly different from zero, so that it influences the latent variables and has a small regression coefficient in the regression equation (1). By reducing its magnitude, by multiplication by its small importance, its contribution to the model decreases. This procedure is repeated many times in IPW, which modifies the PLS algorithm as follows ($\mathbf{Z}$ is the diagonal matrix of importances from equation (2)).

[a] Set $\mathbf{Z} = \mathbf{I}$ (identity matrix).

[b] Set $\mathbf{X}$: original matrix of predictors, $\mathbf{y}$: original vector of response.

[c] Scale predictors and response.

[d] Multiply the matrix of predictors by the diagonal matrix of importances: $\mathbf{X} \Rightarrow \mathbf{XZ}$.

[e] $\mathbf{w}^T = \mathbf{y}^T\mathbf{X}/\mathbf{y}^T\mathbf{y}$.

[f] $\mathbf{w}_{new} = \mathbf{w}_{old}/\|\mathbf{w}_{old}\|$.

[g] $\mathbf{t} = \mathbf{X}\,\mathbf{w}$.

[h] $c = \mathbf{t}^T\mathbf{y}/\mathbf{t}^T\mathbf{t}$.

[i] $\mathbf{p}^T = \mathbf{t}^T\mathbf{X}/\mathbf{t}^T\mathbf{t}$.

[j] $\mathbf{X} \Rightarrow \mathbf{X} - \mathrm{t}\ \mathrm{p}^T$.

[k] $\mathbf{y} \Rightarrow \mathbf{y} - c\ \mathbf{t}$. Go to step [e] to compute the next PLS latent variable. The complexity of the model is obtained by predictive optimization.

[l] Compute $\mathbf{Z}$ with the significant number of PLS components (the regression coefficients $\mathbf{b}$ are referred to the original predictors).

[m] If required, delete predictors with importance less then a cut-off value. Recompute $\mathbf{Z}$. Go to step [b] for the next IPW cycle.

Full validation

Presently it seems well recognized that the predictive ability of a regression technique must be evaluated on objects that have never been used in the development of the regression model, from the pre-treatment (e.g. centring) to the evaluation of the model complexity. This requirement was not very clear when PLS was introduced in chemistry, so that in the first commercial package, cross-validation was used to evaluate the complexity of the model, but the centroid for centring was

Copyright © 1999 John Wiley & Sons, Ltd.

computed with all the objects.

Full validation (FV)[11] is based on the subdivision of objects into three sets: training set, predictive optimization set and evaluation or external set. The training set is used to compute the model parameters. The optimization set, built according to the cancellation groups of cross-validation or with the leave-one-out procedure, is used to evaluate the optimum complexity of the model, and, in the case of the elimination of predictors, also to select the predictors. The evaluation set is used to evaluate the predictive performance of the regression model.

The evaluation set can be a unique 'external' set. In this case it must be constituted by a large number of objects to be representative; an external evaluation set is frequently used in the case of simulated data, as here for data set CRESOLS described below. We will indicate this case simply as that of the external set.

Alternatively, the evaluation set can be built according to cancellation groups. With N objects we can have $E \leq N$ cancellation groups for evaluation. For each evaluation group, M objects are used to develop the regression model ($M = N -$ N/E; in the case where M is not an integer, the number of objects in the cancellation groups is int(M) or int(M)+1). P internal cancellation groups are used for the optimisation, so that each training set contains $T = M$ - M/P objects.

The predictive parameters are indicated as FV–SEP (standard error of prediction or residual standard deviation of prediction), FV explained variance, etc. In the case of the use of the external set the word 'external' is added to the FV parameter.

The estimate of the predictive ability obtained without the evaluation set, only on the basis of the explained variance on the optimization set, is indicated by means of CV–SEP, CV explained variance, etc.

EXPERIMENTAL

Many simulated and real data sets, with number of objects from ten to 100 and number of variables (predictors) from 19 to 601, have been used. No detailed study of the effect of pre-treatments has been made. Generally, data were column centred. In the case of data set CRESOLS, some results with autoscaled data are also reported, to show the performance of IPW in the case of a wrong pre-treatment. In the case of data set ANALOGUES, where the predictors are of a different nature and magnitude, data have been autoscaled.

Computations were made in FORTRAN (Microsoft FORTRAN Powerstation) and Matlab, Version 5·1 (The Math Works, Inc.).

SIMUIN

These simulated data are similar to those used in the paper introducing UVEPLS,[10] with the difference that many data matrices were used, with different seeds from random number generation. The naming of matrices below is that used in Reference 10.

A matrix **S1** ($N \times 100$), with $N = 10$, 25 or 100 (25 in Reference 10), was generated with random numbers from a uniform distribution (0,1). Principal component analysis (PCA) of **S1** (after column centring) produces the matrices of scores **S** and loadings **L**. The first five scores were multiplied by the first five loadings, giving a simulated pure data matrix **SIM** with complexity $A = 5$:

$$\mathbf{SIM}_{N,100} = \mathbf{S}_{N,5}\ \mathbf{L}_{5,100}$$

The vector of response variable y was obtained from the first five scores as

$$\mathbf{y}_{N,1} = \mathbf{S}_{100,5} \begin{bmatrix} 5 \\ 4 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$

A matrix of uninformative predictors **UI** ($N \times 100$) was obtained with random numbers from a uniform distribution (0,1). The matrix **SIMUI** ($N \times 200$) incorporates the noise-free data matrices **SIM** and **UI**:

$$\mathbf{SIMUI} = [\mathbf{SIM} \quad \mathbf{UI}]$$

A noise matrix **NO** ($N \times 200$) was obtained with random numbers from a uniform distribution (0,0·005), with small (compared with those (0–1) in **SIMUI**) elements. The final matrix of predictors **SIMUIN** ($N \times 200$) is the sum of **SIMUI** and **NO**:

$$\mathbf{SIMUIN} = \mathbf{SIMUI} + \mathbf{NO}$$

It is indicated below with the number of objects and with the seed of generating randomization, e.g. as SIMUIN-25-1

In Reference 10, PCA of **SIM** yielded relative (%) eigenvalues 23·02, 21·28, 19·50, 18·74 and 17·46. We obtained for the three different generations of **SIM** (25×100):

SIMUIN-25-1: 22·61, 21·52, 19·38, 18·61, 17·88
SIMUIN-25-2: 23·88, 20·82, 19·79, 18·87, 16·64
SIMUIN-25-3: 22·20, 21·00, 19·82, 19·13, 17·85.

The difference between the first and the fifth eigenvalue diminishes with increasing N. Because of the generation procedure, the objects in these matrices can be considered as being distributed in a five-dimensional hypersphere in the space of the 100 predictors.

CRESOLS

This semi-simulated data set was studied to compare our results with those obtained by Carney and Sanford[12] in 1953 with the use of classical multicomponent analysis on mixtures of the three cresols. The spectra of the three pure cresols at a concentration of about 0·04 g 1^{-1} were recorded in the range 240–300 nm, at 0·1 nm intervals, and were corrected for differences in concentration. These spectra **L** (3×601) are shown in Figure 1. A matrix of concentrations (responses) **Y** (350×3) was obtained by random extraction from a uniform distribution (0,1). Each row was normalized to have sum unity. A noise-free matrix (350×601) of predictors was obtained as **X = YL**. Gaussian noise with standard deviation 0·0002 (about 0·03% of the respective maximum absorbance) was added to this matrix to obtain the final matrix of predictors. The first 50 objects were used to develop the regression model, the other 300 as an external evaluation set.

SOY

This data set consists of 60 samples of soy flour, with spectra measured with a spectrophotometer from 1072 to 2472 nm (176 wavelengths retained from the original 701) (SOY-1) or with a filter instrument with 19 filters (SOY-2). The response variables are moisture, protein and oil. The details are reported in the original paper.[13]

Copyright © 1999 John Wiley & Sons, Ltd.

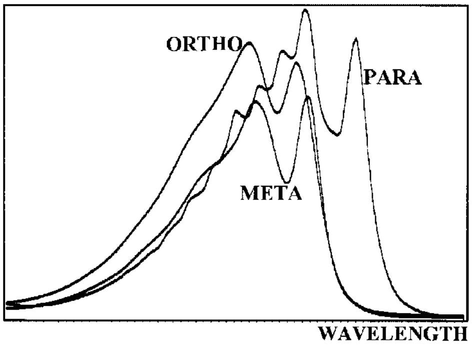

Figure 1. UV spectra of three cresols used for data set CRESOLS

ANALOGUES

Ninety-five conformers of 42 molecules (38 milrinone analogues; two lead compounds, amrinone and milrinone; and two commercial products) were studied[9] with 30 theoretical descriptors (MW, Dmx, DMy, DMz, DM, Hf, IP, EL, qC1, qC2, qC3, qC4, qN5, qH6, qC7, qO8, qH9, Mia, Mib, Mic, VdWA, SA, V, Vand10_1, Vor10_1, Vnot10_1, Vand11_1, Vor11_1, Vnot11_1, cLOGP). The response, ICdt10E-3, is the positive inotropic activity at 10^{-3} M, evaluated using a homogeneous experimental model (the spontaneously beating and electrically driven left atrium from reserpine-treated guinea pigs). In the original paper, four strategies are described for the elimination of conformers and useless predictors. Here this data set was used as in strategy A, step 1, described in the original paper,[9] where all conformers are retained and the descriptors are progressively eliminated by means of ISE-PLS. The final model obtained[9] was with seven descriptors (DMy, Hf, Mia, VdWA, Vor10_1, Vor11_1, cLOGP), two latent variables, 77·9 CV explained variance (20 cancellation groups), 20·80 CV-SEP and 17·2 CV, mean prediction error.

RESULTS AND DISCUSSION

Taking into account the number of predictor matrices and of the response variables, 19 response–predictor combinations **y–X** have been studied. Results are presented as follows:

Set(s)	Responses	Objects	Predictors	Figures	Tables
SIMUI	1	10, 25, 100	200	2–7	1–4
CRESOLS	3	50 + 300	601	8, 9	5, 6
SOY	3	60	176	–	7, 8
ANALOGUES	1	95	30	10	9

Only a limited number of combinations of the number of FV cancellation groups, of the CV internal groups, of the number of IPW cycles and of pre-treatments were studied. The effect of the cut-off value (point [m] of IPW algorithm) was not studied, and always here IPW was applied without this cut-off. Only a part of the results are reported, with the objective to present the performances of IPW in a wide number of conditions, and to compare the results with those of other elimination techniques (UVE, ISE, SOLS). Also, these techniques were applied without exploration of all the possible settings.

Data sets SIMUI were used to show the fundamental behaviour of the IPW algorithm, because of their special nature (possibly useful and surely useless predictors) and their use in the development of

Copyright © 1999 John Wiley & Sons, Ltd.

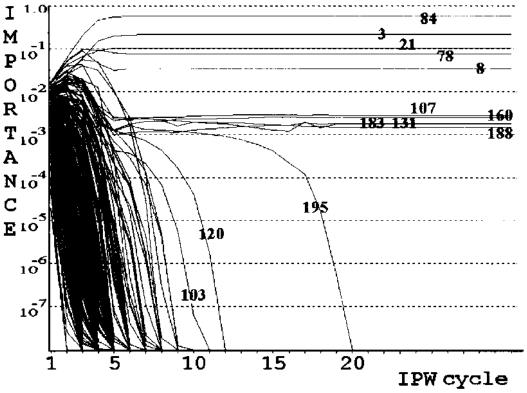

Figure 2. Logarithmic plot of importance of predictors (SIMUIN-25-1, ten internal cancellation groups). The index of the retained and of the last cancelled predictor is indicated

the UVE algorithm.[10] In the case of data set CRESOLS the knowledge of the spectra of pure compounds permits the easy interpretation of the retained predictors. A further characteristic is that the predictors with small range are associated with noise, so that the correct scaling procedure is centring.

The other data sets are examples of real data, and they were used because of their different nature or because of the very different number of predictors and objects.

Data sets SIMUI

Convergence

IPW always converges to a steady state. The convergence is obtained after 10–20 cycles, and only a fraction of the predictors are retained. Sometimes the system oscillates between two almost equivalent states, with the same retained predictors but with a very small difference in their importance. Usually this difference regards only two predictors. Figure 2 shows a typical trend. In the first cycle the predictors (here the 200 predictors of data set SIMUIN-25-1) have importances between about 0·01 and 0·0001. With the iterations the predictors with very low importance rapidly disappeared.

All the predictors are retained in the first cycles. Then the number of retained predictors diminishes rapidly, and finally it becomes constant. Frequently the complexity of the PLS model decreases.

The CV residual standard deviation generally (not always) decreases from the first cycle with the unweighted predictors. It becomes almost constant in the steady state. Sometimes CV-SEP shows a minimum at an intermediate number of cycles, but the following increase is very small. The effect on CV-SEP is usually large in the first cycle, as shown in Figure 3.

At the end of each IPW cycle the PLS algorithm retains A significant latent variables (complexity of the regression model) and (with V predictors and N objects) produces a matrix of weights $\mathbf{W}$ ($A \times V$) and a matrix of scores $\mathbf{T}$ ($N \times A$). The estimate of the N responses is obtained by

$$\mathbf{y} = \mathbf{Tc} = \mathbf{XWc} \tag{4}$$

where $\mathbf{c}$ is the vector of the A regression coefficients computed in step [h] of the algorithm. For a

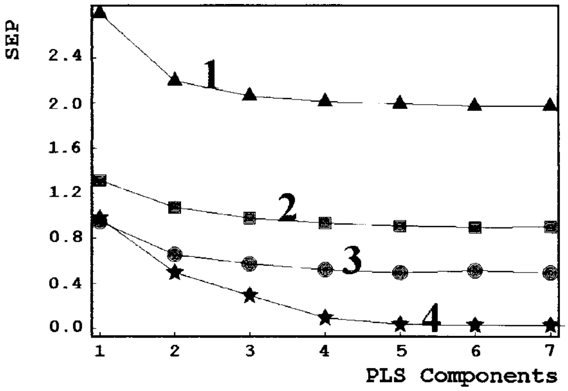

Figure 3. CV standard error of prediction as a function of number of PLS latent variables. Data set SIMUIN-25-1: 1, first IPW cycle (**Z** = **I**); 2, second IPW cycle; 3, third IPW cycle; 4, final (after 15 IPW cycles)

generic object **x** (used or not to build the model),

$$\mathbf{y} = \mathbf{t}^{\mathrm{T}}\mathbf{c} = \mathbf{x}^{\mathrm{T}}\mathbf{w}\mathbf{c} \tag{5}$$

Equation (5) was used to compute, as suggested by Marengo and Todeschini,[14] the regression coefficients **b**. A pseudo-object **x** = **0** introduced in (5) gives the value of the intercept b_0. Others pseudo-objects, such as $\mathbf{x}^{\mathrm{T}} = [0 \quad 0 \quad 0 \quad \ldots 1 \ldots \quad 0 \quad 0 \quad 0]$, give the sum $s_v = b_0 + b_v$, where v is the index of the predictor corresponding to the single ‘1’ in the pseudo-object. Finally the regression coefficient of the predictor v is obtained from

$$b_v = s_v - b_0 \tag{6}$$

The example in Table 1 shows that for a predictor with small importance, gradually s_v and b_0 become very similar, until b_v becomes zero because of the limited number of digits in double-precision variables.

Predictors retained

In the case of data sets SIMUIN the first 100 predictors (those in matrices **SIM**) are the possibly useful predictors. The second 100 predictors, from 101 to 200, are the useless predictors from matrices **UI**. Table 2, shows that IPW retains some useful predictors and also some useless predictors. The importance of the retained useless predictors is generally very small. When the number of objects is small, as in SIMUIN-10-1, many useless predictors can be retained in the final model. In the case of few objects there is the possibility of chance correlation of a useless predictor or of its residuals with the response. For example, for SIMUIN-10-1 the correlation coefficient of predictor 83 (the only useful predictor retained by the final model) was 0·848; the correlation coefficient of predictor 153 (useless predictor retained with large importance from the final model) was 0·841, about the same. As the number of objects increases, the number of useless predictors retained diminishes and their importance becomes very small, so that their total contribution to the response becomes smaller than the residual standard deviation.

When the final model retains some useless predictors with great importance, as for SIMUIN-10-1, the CV residual variance is relatively large, but always with an improvement in comparison with the

Copyright © 1999 John Wiley & Sons, Ltd.

Table 1. Numerical results during IPW iterations for most important predictor (84) and for one of less important predictors (196). Data set SIMUI–25–1

IPW cycle		Predictor 84	Predictor 196
	Initial standard deviation	2·048875133131352E − 001	2·564897139548127E − 001
1	Weight (component 1)	−1·960799873196674E − 001	−8·451312763778680E − 003
	s_v	−1·265042918223083	−1·966486533995359E − 002
	b_0	−1·414618803343700E − 002	−1·414618803343700E − 002
	b_v	−1·250896730189646	−5·518677306516595E − 003
	Importance	1·682111045239174E − 002	9·290148738760088E − 005
2	Weight (component 1)	−2·619498974962820E − 001	−6·235584109844592E − 005
	s_v	−3·357588856779203	2·827773927035849E − 003
	b_0	2·815084113723183E − 003	2·815084113723183E − 003
	b_v	−3·360403940892926	1·268981331266567E − 005
	Importance	5·832545986514270E − 002	2·757251495745004E − 007
3	Weight (component 1)	−4·792827561989897E − 001	−9·765644655992529E − 008
	s_v	−8·677369029608308	2·647770913795973E − 002
	b_0	2·647770925971893E − 002	2·647770925971893E − 002
	b_v	−8·703846738868027	−1·217591964752796E − 010
	Importance	1·956028269473243E − 001	3·425469674289356E − 012
4	Weight (component 1)	−7·754983736763804E − 001	−5·853515805955138E − 013
	s_v	−16·638775154196670	4·182441408902891E − 002
	b_0	4·182441408902891E − 002	4·182441408902891E − 002
	b_v	−16·680599568285690	0·000000000000000E + 000
	Importance	4·576462954430134E − 001	0·000000000000000E + 000

original model with all the predictors. Table 3 shows that the largest residual variances were obtained in the case of data sets SIMUI-10-1 and SIMUI-10-3, where three useless predictors in the first case and one in the second case have importance larger than 0·1. In the case of SIMUI-10-2 only one useless predictor is retained by the final model. Its importance is less than 0·001; the residual variance of the IPW model is very small.

Comparison with other techniques (ISE, UVE, SOLS)

Results with SIMUI allow doing a first comparison with the other elimination techniques.

The final IPW model retains a very small number of predictors, only comparable with that of SOLS (Table 3). ISE retains a medium number of predictors. UVE is the most conservative technique.

When the number of objects is not too small, UVE and ISE have the best CV performance. SOLS has in general the worst performance.

With a small number of objects UVE can give anomalous results, as in the case of SIMUI-10-3 (residual variance of 43·51%). In effect, UVE reliability of the additional predictor 324 is very large (Figure 4), so that only one good predictor (93) is retained. In these cases UVE-α can be used instead, but that obliges one to search for a suitable value of α. In the case of SIMUI-10-3 a satisfactory result was obtained with UVE-α with $\alpha = 0{\cdot}95$.

UVE cannot retain useless predictors. UVE-α in the only case studied here retains four useless predictors and 21 useful predictors, with a CV performance worse that that of IPW with one useless predictor and five informative predictors.

Among the predictors retained by ISE there are very few useless predictors. Generally they have very low importance.

Both IPW and ISE frequently allow the elimination of other predictors at the expense of a

Copyright © 1999 John Wiley & Sons, Ltd.

Table 2. Data sets SIMUIN. Importance of predictors retained by IPW (30 IPW cycles, ten CV cancellation groups). Effect of number of objects and of randomization. Useless predictors underlined

SIMUIN-10-1		SIMUIN-10-2		SIMUIN-10-3	
v	Importance	*v*	Importance	*v*	Importance
83	0·4395	24	0·5303	30	0·2657
115	0·1080	57	0·0459	39	0·1225
153	0·2715	60	0·3906	93	0·1032
162	0·1809	62	0·0156	94	0·2709
		71	0·0166	144	0·0996
		142	0·0010	185	0·1380

SIMUIN-25-1		SIMUIN-25-2		SIMUIN-25-3	
v	Importance	*v*	Importance	*v*	Importance
3	0·2118	35	0·4422	15	0·1552
8	0·0335	45	0·0707	24	0·1830
21	0·1037	50	0·0530	48	0·1174
78	0·0748	51	0·0598	55	0·3270
84	0·5660	77	0·3712	92	0·2088
107	0·0028	150	0·0030	110	0·0023
131	0·0018			168	0·0014
160	0·0025			187	0·0018
183	0·0018			194	0·0014
188	0·0015			195	0·0018

SIMUIN-100-1		SIMUIN-100-2		SIMUIN-100-3	
v	Importance	*v*	Importance	*v*	Importance
5	0·0657	22	0·0569	11	0·1623
17	0·0967	24	0·1053	18	0·1171
30	0·1238	27	0·1152	34	0·0340
34	0·1075	66	0·1849	54	0·0877
36	0·0717	68	0·0935	60	0·1047
37	0·2002	69	0·1128	69	0·1207
38	0·1002	71	0·0611	77	0·1606
50	0·0772	83	0·0639	78	0·1684
53	0·0775	90	0·0409	87	0·0336
63	0·0785	95	0·1259	96	0·0108
143	0·0006	98	0·0383		
183	0·0004	116	0·0005		
		159	0·0010		

negligible increase in CV-SEP.

ISE gives the possibility to observe CV-SEP or CV mean error as a function of the number of eliminated predictors, as shown in Figure 5. In the case of SIMUIN-25-1 after the minimum residual variance, reached with 23 predictors (three useless), the increase in variance is very small. Many other predictors can be eliminated (the first are the three useless predictors) without a significant worsening of performance. With only six predictors the CV residual variance is only 0·0009, one-half that obtained from IPW with ten predictors (five useless). The stepwise elimination of the predictors with

Table 3. Data sets SIMUI. For each elimination technique are reported the number of retained predictors and the CV residual variance. Ten internal cancellation groups

Objects	10			25			100		
Randomization	1	2	3	1	2	3	1	2	3
None	200	200	200	200	200	200	200	200	200
% Res. var.	47·10	51·02	68·13	11·46	13·86	21·63	1·87	1·70	1·81
IPW	4	6	6	10	6	10	12	13	10
% Res. var.	5·1162	0·0003	1·5579	0·0018	0·0020	0·0026	0·0018	0·0019	0·0024
UVE	17	17	25a	63	67	30	89	90	94
% Res. var.	0·9596	0·0006	0·1613	0·0003	0·0005	0·0005	0·0007	0·0008	0·0009
ISE	6	9	6	23	33	26	89	87	79
% Res. var.	0·0018	0·0014	0·0079	0·0003	0·0004	0·0002	0·0006	0·0007	0·0008
SOLS	3	3	4	4	7	5	9	6	30
% Res. var.	0·0068	0·0052	0·0164	0·0065	0·0137	0·0068	0·0070	0·0148	0·0023

[a] This result refers to UVE-α with $\alpha = 0·95$.

smallest importance from the final IPW selection gives a model with five useful predictors and CV residual variance of 0·0033.

Full validation

Figure 6 shows that the decrease with the IPW cycles in the number of predictors depends on the composition of the training–optimization set, in spite of the small perturbation produced by the cancellation of one or two objects. For example, in the case of SIMUI-25-1 the cancellation group 12 (two objects in the prediction set) retains only predictors 3, 8, 21, 78 and 84, i.e. only the good predictors of the final model (see Table 2). The five useless predictors are not retained in this group. Also the quality of the retained predictors can change, more frequently in the case of useless predictors.

In Table 4 the IPW result of full validation is compared with the CV residual standard deviation. FV-SEP is shown also in Figure 7. The first cycle, with all the unweighted predictors, corresponds to the usual PLS. Here the difference between FV-SEP and CV-SEP is due to the fact that the optimum complexity of the partial models computed during the external cancellation cycles can be different from that computed with all the objects in the training–optimization set. This difference, negligible in the example in Table 4 (about 1%), is usually small, not more than 20% in our experience. In the case of data sets SIMUI both CV-SEP and FV-SEP decrease with the iterations and reach an almost stable value, but the FV standard deviation is much larger (100%–300%) than that obtained during the optimization. Thus the evaluation of predictive performances from CV-SEP is for IPW too optimistic. Moreover, the trend shown in Figure 7 is not general, as seen below in the case of data set CRESOLS.

Data set CRESOLS

Because of the small but realistic noise added to the predictors, almost all predictors give useful

Copyright © 1999 John Wiley & Sons, Ltd.

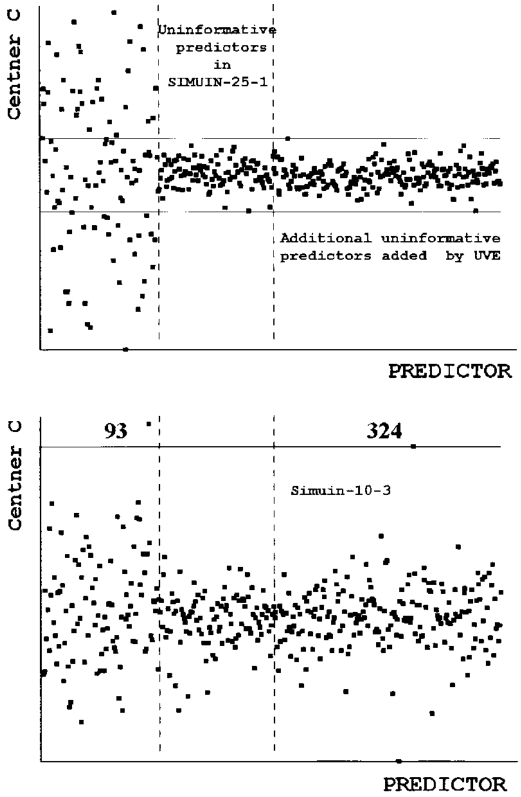

Figure 4. UVE reliability (top, SIMUIN-25-1: bottom, SIMUI-10-3) with indication of uninformative predictors and of predictors added by UVE

information. Thus UVE retains almost all predictors (Table 5, results with centred data), without improvement in the prediction. ISE eliminates 60% of the predictors in the case of *o*-cresol and *m*-cresol with very overlapped spectra, and 93% in the case of *p*-cresol, whose spectrum is more differentiated. The noticeable reduction of predictors is obtained with a very small increase in the prediction error.

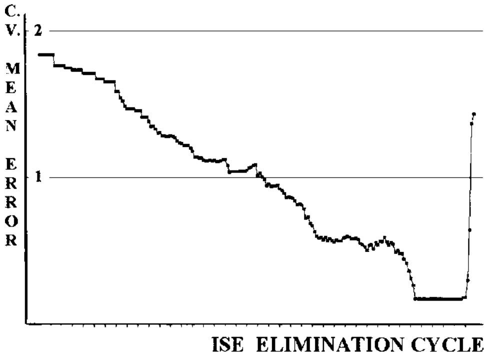

Figure 5. CV mean error in ISE elimination cycles (SIMUIN-25-1, ten internal cancellation groups)

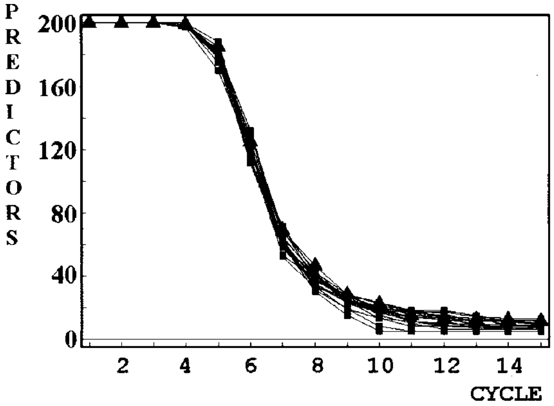

Figure 6. Number of retained predictors as a function of IPW cycle (SIMUIN-25-1, ten internal cancellation groups). Squares refer to the 20 external cancellation groups. Triangles refer to the final cycle without objects in the evaluation set

IPW retains only ten predictors, at the expense of an increase in both the CV prediction error and the FV error, as evaluated on the external set of 300 objects (Figure 8). The increase in CV-SEP during the IPW cycles can be considered anomalous, characteristic of a rare (with real data) situation, with almost all the predictors useful and with very little noise. SOLS retains generally more predictors than IPW, and the prediction error is larger.

Autoscaling increases the effect of useless predictors with small range. Thus the performance of standard PLS with autoscaled data was very bad, with residual standard deviation about ten times that obtained with centred data, as shown in Table 5. Also UVE is very sensitive to the pre-treatment, with a considerable worsening of performance. Both ISE and IPW are relatively insensitive to the pre-treatment, with a moderate worsening of performance. The complexity of the regression model is

Table 4. Data set SIMUI-25-1. Standard errors of prediction in IPW cycles. Twenty external and ten internal cancellation groups

IPW cycle	CV-SEP	FV-SEP
1	2·0158	2·0441
2	0·8898	2·3003
3	0·4713	1·8541
4	0·0934	1·1368
5	0·0354	0·6074
6	0·0238	0·1021
7	0·0220	0·0528
8	0·0221	0·0497
9	0·0220	0·0501
10	0·0218	0·0500
11	0·0220	0·0507
12	0·0221	0·0507
13	0·0221	0·0506
14	0·0222	0·0508
15	0·0222	0·0508

Copyright © 1999 John Wiley & Sons, Ltd.

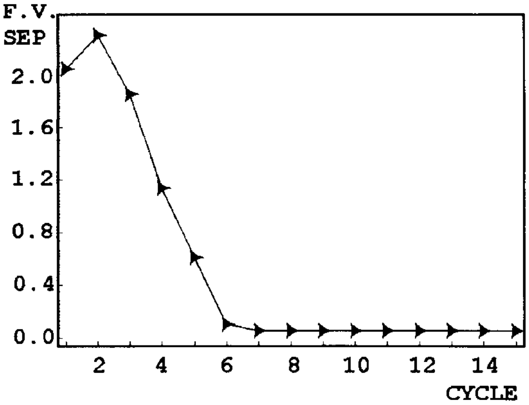

Figure 7. FV standard error of prediction as a function of IPW cycle (SIMUIN-25-1, 20 external and ten internal cancellation groups)

high for all the techniques, about ten latent variables compared with two or three latent variables for the models obtained with centred data.

Comparison with techniques based on original PLS regression coefficients

Data set CRESOLS with response *o*-cresol is used also to show how the selected predictors are related to the original regression vector with all variables. In this case IPW retains only ten predictors.

Table 6, shows the ten predictors with the larger value of the absolute regression coefficient in the original PLS model, the ten predictors with the larger UVE relevance, the ten predictors with the larger importance and the ten predictors with the larger value of the absolute regression coefficient obtained with autoscaled data, procedure BAUT.[8]

Table 5. Data set CRESOLS. Effect of elimination technique on number of retained predictors and on corresponding CV-SEP

		Centred			Autoscaled		
	Elim.	Predictors	10^6 CV-SEP	10^6 FV-SEP (ext.)	Predictors	10^6 CV-SEP	10^6 FV-SEP (ext.)
o-Cresol	None	601	87	95	601	726	587
	IPW	10	330	447	9	311	511
	UVE	594	87	95	540	680	595
	ISE	241	85	100	81	137	171
	SOLS	16	721	533			
m-Cresol	None	601	90	98	601	1285	832
	IPW	10	245	320	17	133	274
	UVE	601	90	98	592	1191	836
	ISE	251	86	105	349	110	485
	SOLS	16	430	514			
p-Cresol	None	601	39	49	601	667	514
	IPW	10	93	107	9	96	128
	UVE	587	39	49	475	478	377
	ISE	44	20	78	38	23	101
	SOLS	6	165	146			

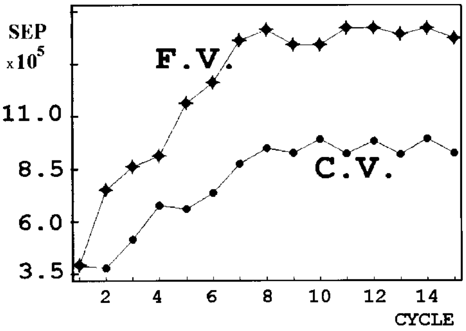

Figure 8. CV-SEP and FV-SEP (external evaluation set) as a function of IPW cycle (CRESOLS, response *p*-cresol, ten CV cancellation groups)

Table 6. Details about selection with different techniques. Data set CRESOLS, response o-cresol. Data are reported only for the first ten predictors ordered according to the order of selection

(a) Regression coefficients, importance and relevance as obtained after standard PLS. Column 4 refers to regression coefficients obtained with autoscaled predictors

Wavelength	Abs(*b*)	Wavelength	Abs(*b*) (autosc.)	Wavelength	UVE relevance	Wavelength	Importance
277·5	0·04581	298·6	0·8746	270·4	14179	277·4	0·00880
277·6	0·04569	299·3	0·8169	269·5	12929	277·3	0·00877
277·4	0·04569	296·7	0·7352	270·2	12039	277·5	0·00874
277·7	0·04533	295·9	0·6937	268·8	11792	277·2	0·00869
277·3	0·04529	298·0	0·6545	277·2	11658	277·6	0·00862
277·2	0·04470	298·9	0·6431	271·0	11621	277·1	0·00853
277·8	0·04469	299·7	0·5574	270·9	11556	277·7	0·00843
277·1	0·04389	298·1	0·4602	277·4	11511	277·0	0·00828
277·9	0·04366	299·9	0·4590	277·7	11303	277·8	0·00816
277·0	0·04277	295·8	0·4311	276·4	10853	276·9	0·00795

(b) Predictors selected by IPW, last ten predictors eliminated by ISE and first ten predictors selected by SOLS. For each technique the third column reports the order of the predictor according to the value of abs(*b*)

IPW-PLS			ISE			SOLS		
Wavelength	IPW final importance	Order Abs(*b*)	Wavelength	ISE CV mean error	Order Abs(*b*)	Wavelength	SOLS *F*-to-enter	Order Abs(*b*)
286·2	0·1451	252	277·4	0·0191843	3	267·1	997259	71
285·8	0·1279	245	286·0	0·0001084	241	267·5	70·1	66
286·5	0·1246	270	277·5	0·0000830	1	267·0	35·5	74
277·4	0·1201	3	285·8	0·0000670	245	267·6	19·0	63
277·2	0·1147	6	277·3	0·0000613	5	267·4	8·1	67
276·8	0·1071	14	285·9	0·0000614	242	271·5	8·7	44
277·8	0·0889	7	277·6	0·0000613	2	267·8	13·7	60
278·0	0·0881	11	286·1	0·0000563	246	299·9	9·9	524
285·9	0·0835	242	277·2	0·0000497	6	299·6	5·4	543
285·0	0·0000027	50	285·7	0·0000490	250	248·8	6·08	374

Copyright © 1999 John Wiley & Sons, Ltd.

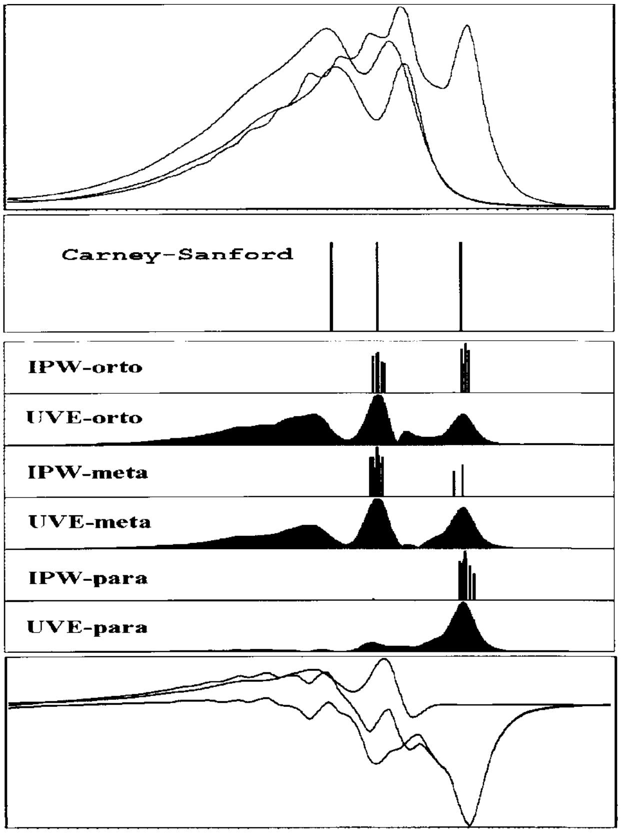

Figure 9. Importance of predictors retained by IPW and UVE for three cresols. The three predictors selected by Carney and Sanford are reported. At the bottom the differences between the spectra (*ortho–para, ortho–meta, meta–para*) are shown

The first ten predictors in the BAUT order are uninformative predictors, selected because of the small difference in the spectrum baseline in the interval 296–300 nm where the three cresols practically do not absorb. PLS with these ten predictors gives the CV-SEP value 23186 × 10^{-6}.

Abs(*b*) and importance values in Table 6 show that the first wavelengths are in the same part of the spectrum, between 277 and 278 nm, where the difference between the absorbances of the three cresols is rather large. PLS with the first ten predictors (abs(*b*) scale) gives the CV-SEP value 5330 × 10^{-6}. PLS with the first ten predictors (abs(*b*) scale) gives the CV-SEP value 5330 × 10^{-6}. PLS with the first ten predictors (importance scale) gives the CV-SEP value 4503 × 10^{-6}.

UVE relevance is large for some wavelengths in the same interval between 277 and 278 nm and for some wavelengths around 270 nm, where the differences between *o*-cresol and *p*-cresol and between *o*-cresol and *m*-cresol are rather large. PLS with the first ten predictors (relevance scale) gives the CV-SEP value 638 × 10^{-6}. The identification of the two most important wavelength intervals explains this rather good result.

In all cases the choice of a reduced number of predictors based on the position according to abs(*b*), importance or relevance has a univariate character, because the correlations and synergism between predictors are not considered.

The importance of the predictors selected by IPW and UVE is shown in Figure 9, where one can

Table 7. Data set SOY-1. Effect of elimination technique on number of retained predictors and on corresponding CV-SEP

Response	Elimination	Predictors	CV-SEP
Moisture	None	176	0·875
	IPW	6	0·762
	UVE	27	0·803
	ISE	11	0·723
	SOLS	3	1·073
Protein	None	176	1·332
	IPW	7	1·289
	UVE	72	1·312
	ISE	13	1·024
	SOLS	5	1·638
Oil	None	176	1·152
	IPW	7	1·040
	UVE	41	1·043
	ISE	8	1·013
	SOLS	3	1·257

compare the three spectra and the differences *ortho–meta*, *ortho–para* and *meta–para*. Both UVE and IPW give great importance to the predictors for which the differences in spectra are large. IPW sacrifices to economy the information in the first part of the spectra, where *o*-cresol is fairly well separated from the other two components.

The choice of Carney and Sanford (in 1953),[12] compared with the choice of IPW and UVE shows that chemometrics with multivariate calibration changed the analysis of multicomponent systems. The choice made in old multicomponent analysis requires the knowledge of the absorptivities of all the components, and it was made on the basis of the spectra of five mixtures of the three cresols. Only the wavelengths corresponding to the most important maxima of individual absorbances were the candidate predictors (six wavelengths). A simple system of three equations was used, with the comparison of the result (fitting ability) for the possible combinations of three among the six candidate predictors. Because in the case of this data set CRESOLS the sum of concentrations is unity, only two equations, i.e. only two wavelengths, are really necessary. The three equations selected two excellent predictors, close to the wavelengths where UVE shows the maxima of importance, and a bad predictor, at a wavelength where the absorbance of *m*-cresol and *p*-cresol is about the same. Obviously this predictor uses the noise to improve the fitting ability.

Data sets SOY

Data set SOY-1 has 176 predictors extracted from the original 701 produced from an NIR spectrophotometer. Some results are reported in Table 7. Full validation was performed only in the case of IPW. With all the predictors FV-SEP is 1·000 for moisture, 1·449 for protein and 1·170 for oil. The final IPW FV-SEP was 0·932, 1·469 and 1·285 respectively. Thus the three cases of FV-SEP diminution, constancy and increase with the IPW cycles are represented here. As usual, UVE retains many predictors but with excellent prediction. ISE retains some predictors more than IPW, with better prediction. SOLS retains the minimum number of predictors but with the worse predictive performance.

Results with data set SOY-2, with only 19 predictors, are reported in Table 8. With two responses, moisture and oil, SOLS behaves fairly well. In the same cases the performances of the elimination

Table 8. Data set SOY-2. Effect of elimination technique on number of retained predictors and on corresponding CV-SEP

Response	Elimination	Predictors	CV-SEP
Moisture	None	19	1·302
	IPW	2	1·220
	UVE	5	1·239
	ISE	4	1·209
	SOLS	2	1·220
Protein	None	19	1·801
	IPW	4	1·750
	UVE	15	1·809
	ISE	7	1·771
	SOLS	3	2·008
Oil	None	19	1·258
	IPW	4	1·128
	UVE	4	1·228
	ISE	6	1·094
	SOLS	5	1·121

techniques are almost equivalent. In the case of protein, UVE retained too many predictors, and the predictive ability of SOLS is not as good as that of the other techniques.

Data set ANALOGUES

This data set is characterized by the large experimental error on the response (the biological activity). Moreover, in this case the elimination of the predictors has the main objective of helping in the interpretation. The agreement between the choice of the elimination techniques is acceptable (Figure 10). UVE retains too many predictors (Table 9; two predictors, qH6 and Mic, have such small importance that the corresponding bar is not visible in Figure 10). IPW show the best performance with the minimum number of predictors. The detailed study,[9] where ISE-PLS was used to eliminate the useless predictors, demonstrated that the activity of this family of molecules depends practically only on volume parameters, so that the presence of charge descriptors (Mullikan charges qC2, qC4, qN5 and qH6 are retained by UVE) hinders the interpretation. In this case ISE and IPW seem almost equivalent, because also the two predictors more retained by ISE (the residual volume referred to a reference molecule, and a component of the moment of inertia) are volume parameters.

CONCLUSIONS

The results presented here demonstrate that IPW can be added to the known tools for elimination of

Table 9. Data set ANALOGUES. Effect of elimination technique on number of retained predictors and on corresponding CV-SEP. Twenty cancellation groups

Elimination	Predictors	CV-SEP
None	30	22·06
IPW	5	20·19
UVE	16	23·25
ISE	7	20·80
SOLS	4	21·49

Copyright © 1999 John Wiley & Sons, Ltd.

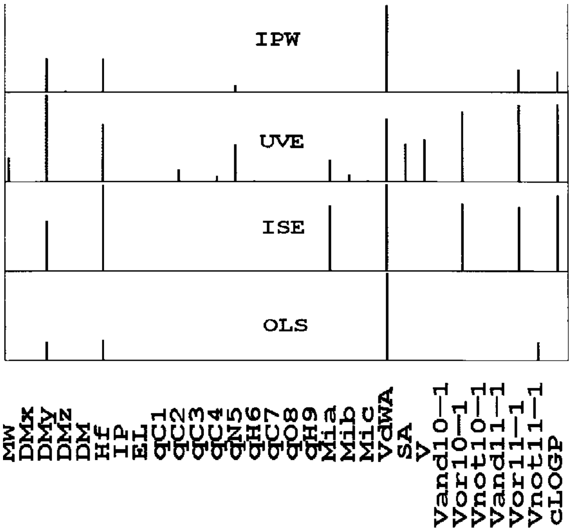

Figure 10. Importance of predictors retained by four elimination techniques (ANALOGUES, 20 internal cancellation groups, 15 cycles for IPW)

useless or redundant predictors in PLS regression, because of its special characteristics to produce acceptable regression models with a very small number of predictors and to be robust against pretreatment. When the predictors in the original data are very noisy, the IPW model has better predictive performance than the original model. The computation time is not very large, so that it is possible to perform full validation and thus to obtain a correct measure of the predictive ability. In some cases the reduced number of relevant predictors can help in the interpretation of the regression model.

The performances of the technique depend on the data set, its dimension, the noise in predictors and in the response. The joint use of UVE, ISE and IPW can probably satisfy the requirements of many real problems.

Probably the IPW algorithm can be further improved and also its speed can be increased, e.g. by the use of a cut-off value of the importance in the elimination of predictors. Work is in progress towards these objectives.

ACKNOWLEDGEMENTS

This work was performed with financial support from the European Union (contract SMT4-CT95-2031) and with partial support from the University of Genoa (60% funds), from MURST and from CNR.

REFERENCES

1. R. Leardi, R. Boggia and M. Terrile, *J. Chemometrics*, **6**, 267 (1992).
2. R. Leardi, *J. Chemometrics*, **8**, 65 (1994).
3. H. Martens and T. Naes, *Multivariate Calibration*, Wiley, Chichester (1989).
4. I. Frank, *Chemometrics Intell. Lab. Syst.* **1**, 233 (1987).
5. N. Kettaneh-Wold, J. F. MacGregor and S. Wold, *Chemometrics Intell. Lab. Syst.* **23**, 39 (1994).
6. F. Lindgren, P. Geladi, S. Rannar and S. Wold, *J. Chemometrics*, **8**, 349 (1994).

Copyright © 1999 John Wiley & Sons, Ltd.

7. M. Forina, C. Drava and C. De La Pezuela, *VI CAC (Chemometrics in Analytical Chemistry Conf.)*, Tarragona, June 1986, Abstract PII–29.
8. A. Garido Frenich, D. Jouan-Rimbaud, D. L. Massart, M. Martinez Galera and J. L. Martinez Vidal, *Analyst*, **120**, 2787 (1995).
9. R. Boggia, M. Forina, P. Fossa and L. Mosti, *QSAR (Quant. Struct.–Act. Relat.* **16**, 201 (1997).
10. V. Centner, D. L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste and C. Sterna, *Anal. Chem.* **68**, 3851 (1996).
11. S. Lanteri, *Chemometrics Intell. Lab. Syst.* **15**, 159 (1992).
12. G. E. Carney and J. K. Sanford, *Anal. Chem.* **25**, 1417 (1953).
13. M. Forina, G. Drava, et al, 'Transfer of calibration function in near-infrared spectroscopy'. *Chemometrics Intell. Lab. Syst.* **27**, 189 (1995).
14. E. Marengo and R. Todeschini, *Chemometrics Intell. Lab. Syst.* **12**, 117 (1992).

Copyright © 1999 John Wiley & Sons, Ltd.