Analytica Chimica Acta, 150 (1983) 61—70
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

A MULTIVARIATE CALIBRATION PROBLEM IN ANALYTICAL CHEMISTRY SOLVED BY PARTIAL LEAST-SQUARES MODELS IN LATENT VARIABLES

MICHAEL SJÖSTRÖM and SVANTE WOLD*

Research Group for Chemometrics, Institute of Chemistry, Umeå University, S-901 87 Umeå (Sweden)

WALTER LINDBERG and JAN-ÅKE PERSSON

Department of Analytical Chemistry, Institute of Chemistry, Umeå University, S-901 87 Umeå (Sweden)

HARALD MARTENS

Norwegian Institute for Food Research, As (Norway)

(Received 26th November 1982)

SUMMARY

The use of partial least squares in latent variables (PLS) for multivariate calibration problems is described. The application is the simultanous determination of ligninsulfonate, humic acid and an optical whitener, from their severely overlapping fluorescence spectra. The predictive performance of the resulting calibration model is tested with a separate set of samples. The PLS method also identifies samples which do not fit the calibration model. The PLS method is compared with principal components analysis combined with multiple regression.

For quantitative analysis of complex samples, fast and cheap spectroscopic methods are preferable to the slow and expensive "wet chemical" approach. However, a disadvantage of the spectroscopic methods is the difficulty of finding frequency regions where the constituents of interest selectively absorb or emit light. This problem can be dealt with by measuring P separate frequencies, of a set of N spectra with known composition. Then some method of data evaluation is applied to find a model which combines the P measurements in $\mathbf{X}$ to give as good a prediction of $\mathbf{Y}$ as possible. Here, the $N \times P$ matrix describing the spectra is denoted by $\mathbf{X}$, and the vector (or matrix) describing the known compositions by y (or $\mathbf{Y}$). In the latter case, Q different properties or constituents of the samples are measured. The calibration model is then used to predict the compositions of new samples n' from their spectra. The data for a calibration problem can be organized as in Fig. 1.

Traditional methods

Usually the calibration problem has been solved by applying multiple regression by a linear model of the vector y in $\mathbf{X}$. With Q y-variables, Q

0003-2670/83/$03.00

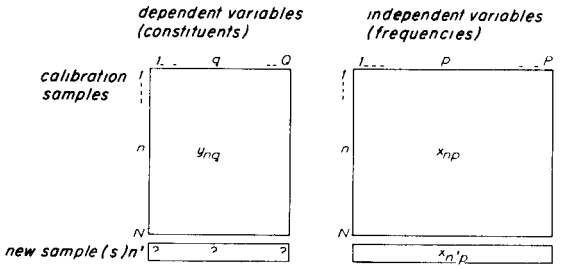

Fig. 1. Organization of data for a multivariate calibration problem.

separate multiple regressions are made, one for each separate y. However, spectroscopic data are not very suitable for ordinary multiple regression where the moment matrix $\mathbf{X}'\mathbf{X}$ is inverted. Because the variables are usually highly correlated, the inversion of $\mathbf{X}'\mathbf{X}$ will result in an almost singular matrix and no statistically stable solution is obtained. The problem can be dealt with by adding a small constant to the diagonal elements in the moment matrix before its inversion, the so-called ridge regression. An alternative is stepwise multiple regression, giving a reduced set of less highly correlated variables [1].

Another way of reducing the dimensionality of the data is to apply principal components analysis to X. Principal components analysis [2] will give a set of uncorrelated new variables (object scores). The object scores can then be used instead of the original variables in multiple regression as shown, for example, by Ho et al. [3]. In the following paragraphs, PC/MR is used to denote principal components analysis combined with multiple regression.

Present approach

A new approach to the multivariate calibration problem is described here. The method is called the partial least-squares model in latent variables (PLS) and was developed by H. Wold and coworkers [4—6]. This method has some features in common with PC/MR because X is described by a principal components type of model, combined with a regression relation between the object scores and Y. However, in PLS the information in Y is also used in the estimation of the object scores for X, which is not the case for PC/MR. Furthermore, the method works in one step and more than one constituent variable (y) at a time can be treated. In this paper, the PLS algorithm described has two blocks, suitable for multivariate calibration problems.

The PLS method will be illustrated by the simultaneous determination of the concentrations of ligninsulfonate, humic acid and a detergent containing a whitener, by molecular fluorescence [7]. Ligninsulfonate is one of the major pollutants emitted from pulp mills into sea water. An attractive method to determine ligninsulfonate is molecular fluorescence, which offers high sensitivity. However, with this method, spectral overlap is a serious limitation, because humic acids and an optical whitener emit energy in the frequency region of interest. Furthermore, the spectra of the pure constituents do

not add up to the expected measured spectrum of a mixture of the three constituents. There are deviations in both intensity and shape. This non-additive behavior places severe demands on the statistical calibration method.

EXPERIMENTAL

The emission spectra of 16 mixtures of the three constituents were recorded between 320 and 540 nm. The emission intensities at 27 equally distributed wavelengths were used. In this way, a 16 × 27 ($N \times P$) matrix (**X**) was formed, which described the emission at P frequencies of the N spectra. The concentrations of the three constituents for the N spectra form a 16 × 3 ($N \times Q$) matrix (**Y**).

Nine additional mixtures were prepared and their spectra were recorded and digitized as before. These samples were not used to calibrate the model, but for testing the predictive properties of the PLS model from **Y** and **X**.

All calculations were done with the SIMCA-3B Basic program for 8-bit microcomputers. The emission spectra were measured on a Perkin-Elmer Model 512 double-beam fluorescence spectrometer.

THE PLS METHOD AND ALGORITHM

In the PLS model, the variation in **X** is explained in terms of the following model (d, e and f are residuals):

$$y_{nq} = \bar{y}_q + \sum_{a=1}^{A} b_{aq} u_{na} + f_{nq} \tag{1}$$

$$x_{np} = \bar{x}_p + \sum_{a=1}^{A} b_{ap} t_{na} + e_{np} \tag{2}$$

$$u_{na} = c_a t_{na} + d_{na} \text{ (for each } a = 1 \ldots A) \tag{3}$$

A geometrical illustration of the PLS method is given in Fig. 2.

Algorithm

(i) When no information about the relative importance of the different y and x variables is available, an initial scaling to variance one is recommended. This is accomplished by dividing each variable in the **X** and **Y** blocks by a scaling factor which is equal to one divided by the standard deviation of the variable. Henceforth, **X** and **Y** refer to the scaled data.

(ii) The **X** and **Y** matrices are centered by subtracting the averages $\bar{y}_q$ and $\bar{x}_p$ for each of the Q columns in case of **Y** and for the P columns in **X**:

$$\bar{y}_q = \sum_n y_{nq}/N \tag{4}$$

$$\bar{x}_p = \sum_n x_{np}/N \tag{5}$$

The zero dimension residuals e_{np} and f_{nq} are then given by

$$f_{nq} = y_{nq} - \bar{y}_q \tag{6}$$

$$e_{np} = x_{np} - \bar{x}_p \tag{7}$$

The subsequent steps (iii)—(viii) provide iterative calculation of the latent variable u_{na} for $a = 1$.

(iii) Starting values for the first iteration of u_{na} are set by the first column in **F**:

$$u_{na} = v f_{nl} \tag{8}$$

Here v is a normalization factor giving u_{na} unit length.

(iv) Calculation of weights w_{ap} for the **X** block are calculated

$$w_{ap} = \sum_n e_{np} u_{na} \tag{9}$$

(v) Latent variable t_{na} is calculated for the **X** block:

$$t_{na} = v \sum_p w_{ap} e_{np} \tag{10}$$

Here v normalizes t to unit length.

(vi) Weights b_{aq} for the **Y** block are calculated from

$$b_{aq} = \sum_n f_{nq} t_{na} \tag{11}$$

(vii) The new latent variable u_{na} is then formed from

$$u_{na} = v \sum_q b_{aq} f_{nq} \tag{12}$$

Here v normalizes u to unit length.

(viii) If the new u_{na} in step (vii) all differ less than one part per million from the u_{na} in the previous round, convergence is reached. Then continue with step (ix), otherwise go back to (iv) for a new round.

(ix) Loadings b_{ap} are done for the **X** block:

$$b_{ap} = \sum_n e_{np} t_{na} \tag{13}$$

(x) The inner relation is

$$c_a = \sum_n u_{na} t_{na} \tag{14}$$

(xi) The new residuals in the **X** and **Y** blocks are formed from

$$e_{np} = e_{np} - b_{ap} t_{na} \tag{15}$$

$$f_{nq} = f_{aq} - c_a b_{aq} t_{na} \tag{16}$$

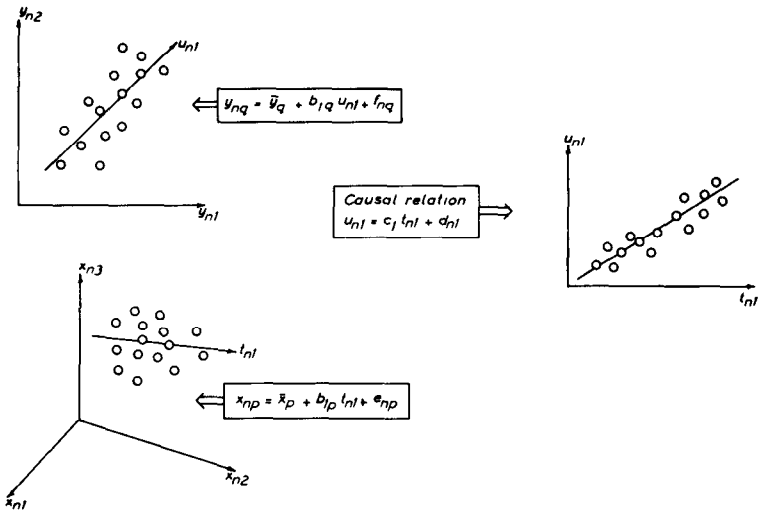

Fig. 2. A geometrical illustration of a PLS model with one cross-term ($A = 1$) for a problem with two constituents (y_{n1} and y_{n2}) and three variables ($x_{n1} - x_{n3}$). In practice, the number of variables in the X block should be much larger and usually $A \gg 1$.

If the last dimension a is deemed insignificant according to, for example, cross-validation (see below), the program is terminated. Otherwise it is continued with an additional dimension ($a + 1$), by going back to step (iii).

The number of dimensions, A. Cross-validation [8] is a method of determining if a cross-term $b_{ap}t_{na}$ is significant or not. With this method, the predictive ability of the ath cross-term is investigated by first deleting, say, one quarter of the calibration samples. The y values of the deleted samples are then predicted and the sum of squared difference (*SS*) between the observed and calculated y values is calculated. Another quarter of the samples is then kept out, a second *SS* is calculated, etc., until all calibration samples have been kept out just once.

If the sum of the four partials *SS* is smaller than the *SS* of Y after $a - 1$ dimensions, then the ath cross-term contains predictive information and is deemed to be statistically significant.

Prediction of new samples

Once a calibration model has been established, the comparison of a new sample n' can be determined as follows. The spectrum is digitized, centered and scaled in the same way as for the calibration set, giving a vector with the elements x_p ($p = 1 \ldots 27$).

The t_a parameters and the residuals are calculated by fitting x_p to b_{ap}, ($a = 1 \ldots A$) by multiple regression:

$$x_p = \sum_{a=1}^{A} t_a b_{ap} + e_p \qquad (17)$$

Equations (1—3) indicate that y_q can be estimated from

$$y_q = \bar{y}_q + \sum_{a=1}^{A} b_{aq} c_a t_a \tag{18}$$

where b_{aq} and c_a are known from the calibration set and t_a from Eqn. (17). Another way, giving slightly different values of y_q, is to use the weights w and calculate t from Eqn. (10).

The predicted values of y are now in scaled and centered form and can be transformed back to the original coordinates by applying the reverse centering and scaling as in the calibration set.

Identification of test samples giving poor prediction

The residual variance for sample n' is given by

$$s_{n'}^2 = \sum_{p} e_p^2/(P-A) \tag{19a}$$

where the residuals e_p are calculated in Eqn. (17). Similarly, the residual variances for a calibration sample n are given by

$$s_n^2 = \sum_{p} e_p^2 \, (N/(P-A)(N-A-1)) \tag{19b}$$

The difference between these equations is due to the different number of degrees of freedom for the calibration samples and the test samples. With an approximate F-test

$$F = s_{n'}^2/S_x^2 \text{ with } (P-A) \text{ and } (P-A)(N-A-1) \text{ degrees of freedom} \tag{20}$$

the variance from Eqn. (19a) can be compared with the overall variance for the calibration set in the X-block, S_x^2:

$$S_x^2 = \sum_{np} e_{np}^2/(P-A)(N-A-1) \tag{21}$$

If this approximate F-test shows that the residual variance of sample n' is significantly larger than the variance of the calibration set, this is an indication not to rely on the predicted y_q values estimated from Eqn. (18). The measurements on sample n' do not fit the model, thus giving a poor estimate of t_a used in the prediction step. However, the approximate nature of this F-test should be stressed. The calibration samples are favored compared to validation samples, and this F-test will give an excessively narrow confidence interval for the validation set. The prediction errors can be expected to increase with decreasing fit of a spectrum to the PLS model. This means that, for large F values, large prediction errors are expected for y, but for F values close to the critical F value, the increase in the prediction errors of y compared to the prediction errors for the calibration set will be small if any. Figure 3 illustrates the procedure for identifying a sample that will give an inferior prediction of y, compared to the calibration samples.

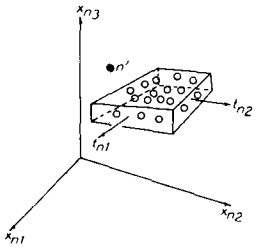

Fig. 3. Illustration of the identification of a test sample not qualified for prediction of y values. With three variables in X, the observations in the calibration set are described as points in a three-dimensional space. With two significant cross-terms, the t_{n1} and t_{n2} vectors describe two new dimensions in this space. The confidence limits from the approximate F-test can be thought of as an envelope containing the calibration samples. If a sample n' lies far outside the confidence limit of the model, this means that the spectrum of n' is unsatisfactorily described by the calibration model ($\bar{x}_p$, b_{1p} and b_{2p}). The resulting $t_{n'1}$ and $t_{n'2}$ will give poor predictions of y.

Principal components analysis with multiple regression

In this approach X is described by a principal component model:

$$x_{np} = \bar{x}_p + \sum_{a=1}^{A} b_{ap} t_{na} + e_{np} \tag{22}$$

In this case the number of significant components is determined by cross-validation. Then for the calibration set the t_{na} vectors are fitted to one y_n vector at a time by multiple regression:

$$y_n = c' + \sum_{a=1}^{A} b'_a t_{na} + e'_n \tag{23}$$

For a validation sample n', its data vector x_p is fitted to the b_{ap} values from Eqn. (22), to give the t_{na} values as in Eqn. (17). By introducing this value in Eqn. (23), where now the b'_a parameters are known, y values can be calculated.

RESULTS AND DISCUSSION

The composition of the data set is given in Table 1 together with the calculated compositions from a PLS model with seven components (A = 7 in Eqns. 1—3). Samples 1—16 are used to calibrate the model and samples 17—25 form a validation set, but their actual compositions are known. Thus the predicted values for samples 17—25 are calculated from Eqns. (17) and (18) with the input of the calibration model. The s_n value is the fit of a spectrum to the X-block part of the model (see Eqns. 19a or b). The F values are found from Eqn. (20).

The validation samples 17, 19, 20 and 21 fit the calibration model well and the sample compositions are predicted just as well as in the calibration

TABLE 1

Sample composition of the calibration and validation set and the prediction errors of a seven-component PLS model

Sample[a]	Composition[b]			Prediction errors[c]			Fit	
	y_1	y_2	y_3	d_1	d_2	d_3	s_n^d	F^e
1	3.011	0	0	0.041	0.028	2.39	0.0042	0.14
2	0	0.401	0	0.067	0.057	0.21	0.0054	0.24
3	0	0	90.63	0.030	0.003	3.67	0.0106	0.91
4	1.482	0.158	40.00	0.098	0.024	0.22	0.0123	1.27
5	1.116	0.410	30.45	0.081	0.020	12.3	0.0125	1.23
6	3.397	0.303	50.82	0.032	0.036	0.41	0.0155	1.95
7	2.428	0.298	70.59	0.041	0.019	0.16	0.0106	0.91
8	4.024	0.115	89.39	0.042	0.053	2.55	0.0094	0.72
9	2.275	0.504	81.75	0.087	0.027	6.60	0.0131	1.40
10	0.959	0.145	101.10	0.018	0.071	2.37	0.0166	1.09
11	3.190	0.253	120.00	0.085	0.012	10.24	0.0149	1.80
12	4.132	0.569	117.70	0.119	0.001	1.70	0.0129	1.35
13	2.160	0.436	27.59	0.019	0.031	3.95	0.0081	0.53
14	3.094	0.247	61.71	0.012	0.057	12.38	0.0126	1.29
15	1.604	0.286	108.80	0.075	0.025	3.40	0.0087	0.62
16	3.162	0.701	60.00	0.037	0.039	2.55	0.0086	0.61
17	2.443	0.289	80.22	0.063	0.018	0.82	0.0141	1.62
18	4.078	0.361	88.52	0.074	0.054	14.37	0.0377	11.6
19	1.065	0.234	69.23	0.084	0.009	1.22	0.0082	0.54
20	3.317	0.123	40.13	0.065	0.067	16.26	0.0136	1.50
21	0.998	0.416	30.74	0.089	0.070	3.0	0.0100	0.81
22	2.983	0.403	120.0	0.083	0.103	2.56	0.0155	1.95
23	5.132	0.229	49.05	0.334	0.041	14.64	0.0243	4.80
24	5.058	0.000	51.06	0.181	0.008	12.24	0.0195	3.10
25	0.0	0.735	99.57	0.443	0.755	31.63	0.231	433.0

[a]Samples 1—16 form the calibration set and samples 17—25 are the validation set. [b]y_1—y_3 are the sample compositions ($\mu g\ ml^{-1}$) for humic acid, ligninsulfonate and the whitener, respectively. [c]Prediction errors are the absolute values of the difference between the actual composition and the calculated value. For samples 1—16 the calculated values are from Eqn. (1) and for samples 17—25 from Eqn. (18). [d]The fit of each sample to the model is given by s_n (Eqn. 19a, b). [e]The F-test is according to Eqn. (20): $F_{0.01,\ crit.} = 2.0$.

set. Slightly enhanced F values are observed for samples 18, 23 and 24; for samples 23 and 24, the prediction for ligninsulfonate is slightly worse. Sample 25 has a high F value ($F = 443$) and the prediction errors are large for all three constituents compared to the prediction errors of the calibration set.

Thus the ability of the PLS method to predict the compositions of the validation samples from their highly overlapped spectra is demonstrated. Further, the PLS method also detects samples which will give poor predictions.

For the present data set, multiple regression methods are inferior because the number of variables is large compared to the number of samples. Step-

wise multiple regression might be used in this problem if about two thirds of the variables were deleted, which would lead to loss of information. It should also be noted that multiple regression methods, in contrast to PLS, will give no indication of the strange behavior of the spectrum of sample 25.

The individual spectra for specific concentrations do not add up to the spectrum of the mixture with the same concentrations of the constituents. This means that multiple regression with the three individual spectra as independent variables also gives poor predictions of y.

Comparison with the PC/MR method

The PLS method is compared with the PC/MR approach, and the prediction errors for the validation set are presented, in Table 2. Principal components analysis combined with cross-validation of X shows that eight components are needed to describe the systematic information in the data; this is one component more than in the PLS method. Thus the results for the two methods are presented with both seven and eight components.

The PLS method gives slightly better prediction for ligninsulfonate and the whitener compared to PC/MR with both $A = 7$ and $A = 8$. It should be noted that these are the constituents that contribute less to the spectra. This comparison with the PC/MR method shows that PLS makes a good prediction of the constituents that make a small contribution to the overall fluorescence spectra, at the cost of a very slight decrease of the precision of the most strongly-emitting compound.

Conclusion

The PLS approach has some obvious advantages over the traditional approach. First, the information in Y is used; hence if X contains a structure which has predictive relevance for Y, this will appear in the PLS solution, which is not necessarily the case for the PC/MR method, multiple regression or ridge regression. Secondly, once the PLS model has been determined, it is possible to classify a new sample as similar to the calibration set or not. This means that the information is obtained whether or not the calibration set is qualified to determine the composition of a particular new sample. Such information is not given by the multiple or ridge regression techniques.

TABLE 2

Comparison of the predictive ability of the PLS and PC/MR methods for the validation set. The predictive ability is expressed as the sum of the squared prediction errors ($\Sigma_n (y_n - y_{n,pred.})^2$)

Constituent	PLS ($A = 7$)	PC/MR ($A = 7$)	PLS ($A = 8$)	PC/MR ($A = 8$)
Humic acid	0.376	0.335	0.419	0.347
Ligninsulfonate	0.595	0.669	0.492	0.621
Whitener	1.85×10^3	3.1×10^3	1.88×10^3	1.92×10^3

Finally, the PLS algorithm is easy to program for microcomputers and calibration problems can be solved for large data sets in a fraction of the time needed for the MR approach.

Grants from the Swedish Natural Science Research Council (NFR) and the Swedish Board of Research Councils are gratefully acknowledged.

REFERENCES

1 N. R. Draper and H. Smith, Applied Regression Analysis, 2nd edn., Wiley, New York, 1981.
2 See, e.g., D. L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier, Amsterdam, 1980, Ch. 19.
3 C.-N. Ho, G. D. Christian and E. R. Davidson, Anal. Chem., 52 (1980) 1071.
4 H. Wold, in J. Gani (Ed.), Perspectives in Probability and Statistics. Papers in honour of M. S. Bartlett, Academic Press, London, 1975.
5 H. Wold, in K. G. Jöreskog and H. Wold (Eds.), Systems under Indirect Observation, North-Holland, Amsterdam, 1982.
6 S. Wold, H. Wold, W. J. Dunn III, A. Ruhe, Report UMINF, 83 (1980).
7 W. Lindberg and J-Å. Persson, Anal. Chem., 55 (1983) in press.
8 S. Wold, Technometrics, 20 (1978) 379.