Aquatic Sciences 57/3, 1995

1015–1621/030217-25 $ 1.50 + 0.20/0

Multivariate analysis of aquatic toxicity data with PLS

Lennart Eriksson[1], Joop L.M. Hermens[2], Erik Johansson[1], Henk J.M. Verhaar[2] and Svante Wold[3]

[1] Umetri AB, P.O. Box 7960, 90719 Umeå, Sweden
[2] Research Institute of Toxicology, Utrecht University, P.O. Box 80176, 3508 TD Utrecht, The Netherlands
[3] Research Group for Chemometrics, Department of Organic Chemistry, Umeå University, 90187 Umeå, Sweden

Key words: PLS, multivariate analysis, experimental design, aquatic toxicity, QSAR

ABSTRACT

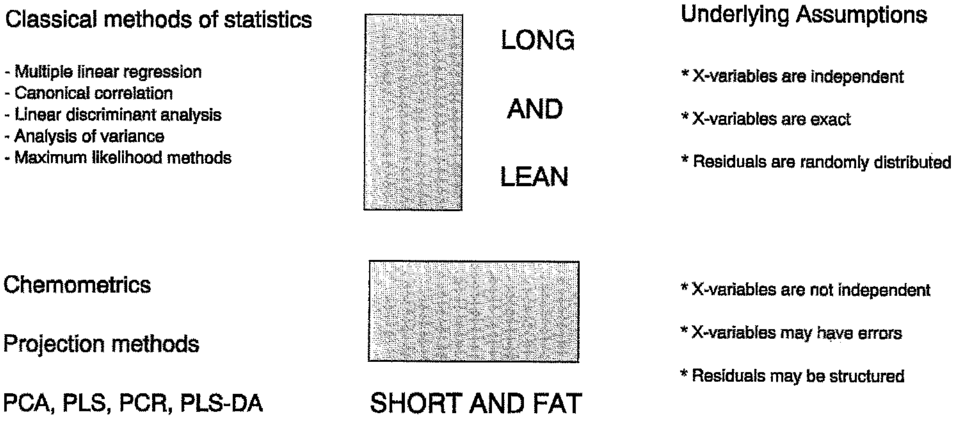
A common task in data analysis is to model the relationships between two sets of variables, the descriptor matrix *X* and the response matrix **Y**. A typical example in aquatic science concerns the relationships between the chemical composition of a number of samples (X) and their toxicity to a number of different aquatic species (Y). This modelling is done in order to understand the variation of **Y** in terms of the variation of **X**, but also to lay the ground for predicting *Y* of unknown observations based on their known **X**-data. Correlations of this type are usually expressed as regression models, and are rather common in aquatic science. Often, however, the multivariate **X** and **Y** matrices invalidate the use of multiple linear regression (MLR) and call for methods which are better suited for collinear data. In this context, multivariate projection methods represent a highly useful alternative, in particular, partial least squares projections to latent structures (PLS). This paper introduces PLS, highlights its strengths and presents applications of PLS to modelling aquatic toxicity data. A general discussion of regression, comparing MLR and PLS, is provided.

1 Introduction

1.1 The evolution of data matrices

In the early days of this century it was difficult to make extensive measurements on a series of investigated samples. Thus, data tables usually had many more observations (rows) than variables (columns), see for instance (Fisher, 1936). This type of data arrangement with more observations than variables is representative for most data tables that arose in scientific applications at that time (Wold et al., 1984). We refer to such matrices as "long and lean" (Fig. 1).

Today, however, reality for experimentalists has changed. It is no longer difficult and time-consuming to measure variables. Due to the introduction of modern electronics, a vast array of technical instruments (spectrophotometers, chromatographs,



Classical methods of statistics

- Multiple linear regression
- Canonical correlation
- Linear discriminant analysis
- Analysis of variance
- Maximum likelihood methods

LONG AND LEAN

Underlying Assumptions

* X-variables are independent

* X-variables are exact

* Residuals are randomly distributed

Chemometrics

Projection methods

PCA, PLS, PCR, PLS-DA

SHORT AND FAT

* X-variables are not independent

* X-variables may have errors

* Residuals may be structured

Figure 1. Two shapes of data matrices, long and lean, and short and fat, and some assumptions of common data analytical methods. Some methods, such as ridge regression, have an intermediate position between the classical and chemometric techniques

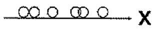
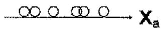
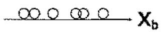
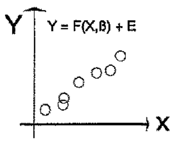
Problem	Uni- and bivariate	Multivariate
Summarize a data set (overview)	 $\bar{x}$, s^2 Histogram Median	$\bar{x}$, s^2 PCA + plots
Compare two or more groups (classification)	 $\bar{x}_a$, s_a^2  $\bar{x}_b$, s_b^2 ANOVA	PCA PLS-DA NN KNN LDA
Compare two sets of variables, X and Y (quantification and prediction)	 LR MLR	PLS PCR NN KNN MLR RR CC

Figure 2. The three problem types of data analysis and some pertinent methods

etc.) have been devised, which are capable of outputting hundreds or thousands of variables within a short period of time, reflecting the characteristics of a sample. The number of observations in a data table, on the other hand, is comparatively difficult to increase because nowadays new regulations apply regarding costs, time and ethics (individuals, animal testing, etc.), constraining the number of samples. The practical consequence of this is that data matrices are no longer typically "long and lean" but rather "short and fat" (Fig. 1). This fact raises new demands on data analytical techniques (Wold, 1995).

1.2 Three types of data analytical problems

Once experimental data have been acquired they must be analysed to separate information form noise. In principle, data analytical problems can be divided into three major types, regardless of application. These are: (i) summarizing a data set, (ii) comparing groups, aiming at classification of unknown samples, and (iii) modelling of relationship between variables or sets of variables for quantification and prediction purposes (Fig. 2). In the univariate, bivariate and few variate (less than, say, five variables) cases, (i) and (ii) can be accomplished by calculating variable averages, standard deviations and covariances and evaluating these. In the multivariate case, however, this approach becomes tedious and inefficient, and other alternatives must be sought for. Moreover, with case (iii), one can plot and compare one pair of variables at a time, or try to find a mathematical expression linking a predictor variable to a response variable using multiple linear regression. Anew, the analyst runs into difficulties in the multivariate case, because not only are there many pairwise variable comparisons to make, but the risk for coincidental correlations increases quadratically with increasing number of variables (Topliss and Edwards, 1979).

In science in general, and certainly in aquatic science, many applications are of type (iii). For instance, numerous examples are found in the literature regarding the determination of toxicity of chemicals to aquatic species, in which quantitative relationships are explored between chemical properties of compounds and toxicological responses (Blum and Speece, 1990). In the present contribution we are concerned with data analytical methods suitable for finding and probing such quantitative relationships, and in particular our aim is to introduce partial least squares (PLS) regression analysis and illuminate its utility in aquatic toxicity research. But before we enter this discussion, we shall briefly review some general points regarding regression analysis – case (iii) – and point out some of the problems that might occur when applying multiple linear regression (MLR) to the "short and fat" data structures that are predominant today.

1.3 Consequences of "short and fat" data structures on regression

The classical approach to regression problems is MLR. For MLR to work properly, however, experimental data must fulfill certain statistical conditions, which reflect the basic assumptions underlying the technique (Draper and Smith, 1981; Wold,

1995) (Fig. 1). Notably, the predictor variables, normally called **X**, are assumed mathematically independent, implying that a change in one X-variable is not strongly coupled to a change in another X-variable. Mathematical independence means that the rank of **X** is **K** (i. e. equals the number of X-variables). This assumption may be reasonable for "long and lean" data matrices, but violated as soon as matrices are multivariate and include collinear variables (Wold, 1995; Topliss and Edwards, 1979). Such multicollinearity occurs whenever some predictor variables are linear functions of other predictor variables, a feature which is typical for instance in spectroscopic data. This appears automatically with short and fat data (few observations, many variables) regardless of origin and regardless of moderate pairwise correlations. If MLR is applied to data sets exhibiting collinearities, the calculated regression coefficients get unstable and their interpretability breaks down (Draper and Smith, 1981; Topliss and Edwards, 1979; Lindgren, 1994). For example, certain coefficients may be much larger than expected, or they may even have the wrong sign (Lindgren, 1994; Mulllet, 1976). In fact, the problem of sign inversion with respect to an anticipated correlation structure is not uncommon, and will here be exemplified with a small data set from the literature (see below). Furthermore, stepwise multiple linear regression (SMLR) with variable deletion is sensitive to collinear data structures. SMLR models give rise to misleading interpretation and poor predictions (Frank and Friedman, 1993; Topliss and Edwards, 1979).

1.4 The need for multivariate projection methods

One way to circumvent the dilemma of multicollinearity is to take benefit from it, by employing multivariate projection methods, such as partial least squares projections to latent structures, PLS. This method is particularly apt at handling the situation when the number of variables exceeds the number of observations. This is because projections to latent variables in multivariate space tend to become more distinct and stable the more variables are involved (Wold, 1995; Lindgren, 1994). PLS is a recently developed generalization of regression and gives identical results to MLR in situations when X has full rank. In most other cases, PLS gives a solution that is reminiscent to that of MLR. However, in addition PLS provides a set of score and loading plots that inform about the correlation structure between predictor and response variables, and regarding similarities among the observations. Model interpretation is also faciliated by these plots. PLS will be described in detail below.

We note that there exist alternatives to PLS for the multivariate analysis of aquatic science data. Some of these methods are principal components regression (PCR), canonical correspondence analysis (CCA), correspondence analysis scaling (CAS, for discrete data), redundancy analysis (RA) and ridge regression (RR) (Jackson, 1991; Jongman et al., 1987). In fact, CCA of Ter Braak (Jongman et al., 1987) is similar to PLS in the same way as correspondence analysis is similar to principal component analysis of data with an appropriate scaling. We do not give a detailed account of these alternatives to PLS but the interested reader is referred to the appropriate references.

2 Examples

In order to introduce PLS to aquatic science and highlight some of its useful features, we shall consider three data sets from the literature. Two of the data sets are directly connected with aquatic toxicology, whereas one is not related to aquatic science, but is included to be simple and yet illustrative of typical problems associated with MLR when applied to data sets containing collinear variables.

2.1 Energy of protein unfolding (Data Set I)

The first data set concerns a series of 19 proteins (tryptophane synthase α unit of bacteriophage T4 lysosome, modified in position 49, Table 1). The altered amino

Table 1. Chemical descriptor data and biological response for data set I

		x_1	x_2	x_3	x_4	y_1
Protein no	Amino acid	PIF	DGR	SAC	MR	DDGTS
1	Ala	0.31	-0.55	254.2	2.126	8.5
2	Asn	-0.6	0.51	303.6	2.994	8.2
3	Asp	-0.77	1.2	287.9	2.994	8.5
4	Cys	1.54	-1.4	282.9	2.933	11
5	Gln	-0.22	0.29	335	3.458	6.3
6	Glu	-0.64	0.76	311.6	3.243	8.8
7	Gly	0	0	224.9	1.662	7.1
8	His	0.13	-0.25	337.2	3.856	10.1
9	Ile	1.8	-2.1	322.6	3.35	16.8
10	Leu	1.7	-2	324	3.518	15
11	Lys	-0.99	0.78	336.6	2.933	7.9
12	Met	1.23	-1.6	336.3	3.86	13.3
13	Phe	1.79	-2.6	366.1	4.638	11.2
14	Pro	0.49	-1.5	288.5	2.876	8.2
15	Ser	-0.04	0.09	266.7	2.279	7.4
16	Thr	0.26	-0.58	283.9	2.743	8.8
17	Trp	2.25	-2.7	401.8	5.755	9.9
18	Tyr	0.96	-1.7	377.8	4.791	8.8
19	Val	1.22	-1.6	295.1	3.054	12
	Correlation	PIF	DGR	SAC	MR	DDGTS
	PIF	1				
	DGR	-0.96832	1			
	SAC	0.416383	-0.46264	1		
	MR	0.555481	-0.58201	0.955283	1	
	DDGTS	0.711445	-0.64764	0.267735	0.290469	1

x_1 (PIF) = lipophilicity constant.
x_2 (DGR) = polarity measure.
x_3 (SAC) = accessible surface area [Å].
x_4 (MR) = molecular refractivity.
y_1 (DDGTS) = the unfolding free energy in water of the 19 modified proteins [kcal/mol].
All variables are taken from references Wold 1995 and El Tayar et al. 1992.

Table 2. The eight chemical descriptors and the eight aquatic toxicity responses for data set II

		x_1	x_2	x_3	x_4	x_5	x_6
No	Compound	Bp	Mp	D	log P	σ	HOMO
1	nitrobenzene	210.8	5.7	1.2037	1.89	0	−10.5615
2	1-chloro-2-nitrobenzene	246	34.5	1.348	2.26	0.27	−10.3348
3	1-chloro-3-nitrobenzene	235	43	1.534	2.49	0.37	−10.3668
4	1-chloro-4-nitrobenzene	242	83.6	1.298	2.35	0.27	−10.474
5	1,2-dichloro-3-nitrobenzene	257	61	1.449	3.01	0.64	−10.2826
6	1,3-dichloro-4-nitrobenzene	258	30		2.9	0.54	−10.4768
7	1,4-dichloro-2-nitrobenzene	267	56	1.669	2.9	0.64	−10.2177
8	1,3-dichloro-5-nitrobenzene		65.4	1.692	3.13	0.74	−10.4143
9	2-nitrotoluene	221.7	−9.5	1.1629	2.3	−0.15	−10.1716
10	3-nitrotoluene	232.6	16	1.1571	2.4	−0.07	−10.1972
11	4-nitrotoluene	238.3	54.5	1.392	2.34	−0.15	−10.3039
12	4-chloro-2-nitrotoluene	240	38		3.05	0.22	−10.0528
13	2-chloro-6-nitrotoluene	238	35		3.09	0.22	−10.1267
14	2,3-dimethylnitrobenzene	240	15	1.1402	2.83	−0.22	−9.94105
15	3,4-dimethylnitrobenzene	254	30	1.112	2.91	−0.22	−10.0749

x_1 (Bp) = boiling point [C]; x_2 (Mp) = melting point [C]; x_3 (D) = density [ ]; x_4 (log P) = log octanol/water partition coefficient; x_5 (σ, sigm) = sigma minus of Hansch & Leo; x_6 (HOMO) = energy of highest molecular orbital [eV]; x_7 (LUMO) = energy of lowest unoccupied molecular orbital [eV]; x_8 (η, Eta) = energy difference between (HOMO-LUMO)/2 [eV].
y_1 (DMI48h) = log conc causing immobilization of 50% of *D Magna* after 48h [umol/l]; y_2 (DMI21d) = log conc causing immobilization of 50% of *D Magna* after 21 days [umol/l].
y_3 (DMRm) = log lowest conc causing significantly lowered population growth of *D Magna* after 21 days [umol/l]; y_4 (DMle) = log lowest conc causing significantly lowered mean length of D magna after 21 days [umol/l]; y_5 (CPEC50) = log conc causing 50% decrease in population density of *C pyrenoidosa* after 96 h [umol/l]; y_6 (PHEC50) = log conc causing 50% decrease in bioluminescense of *Ph phosphoreum* after 15 min [umol/l]. y_7 (PoeLC50) = log conc causing 50% lethality of *P Reticulata* after 14 days (umol/l]; y_8 (BCF) = log bioconcentration factor for *P Reticulata.* For more details on these data reference is made to the original literature [Deneer et al. 1987, Deneer et al. 1989].

acids are described by four predictor variables, namely lipophilicity (PIF, x_1), polarity (DGR, x_2), molecular surface area (SAC, x_3) and molecular refractivity (MR, x_4). The response variable of interest is the energy of unfolding of these modified proteins. It should be noted that two pairs of variables, x_1/x_2 and x_3/x_4, are highly correlated, with $r^2 > 0.9$. For more details, reference is made to the literature (Wold, 1995; El Tayar et al., 1992).

2.2 *Aquatic toxicity of mono-nitrobenzene derivatives (Data Set II)*

In the second example, quantitative structure-activity relationship (QSAR) modelling is attempted for a set of 15 mono-nitrobenzene derivatives (Table 2). The goal in this study is to be able to model and predict the aquatic toxic profiles of the 15 chemicals based on information concerning their chemical properties. The 15 compounds were multivariately characterized using an ensemble of eight experimental

Table 2 (continued)

x_7	x_8	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
LUMO	η	DMI48h	DMI21d	DMRm	DMIe	CPEC50	PHEC50	PoeLC50	BCF
−1.06761	4.74693	2.43	2.29	2.16	2.16	2.16	2.16	2.7	1.47
−1.0817	4.626555	2.18	1.83	1.8	1.8	1.64	1.46	2.28	2.29
−1.2849	4.540965	2.1	1.77	1.05	1.8	1.08	1.92	1.99	2.42
−1.34278	4.565585	1.63	1.46	1.05	1.31	1.49	2.33	1.58	2.46
−1.21595	4.533345	1.34	1.26	0.97	0.72	1.18	0.89	1.34	3.01
−1.51518	4.4808	1.34	1.36	0.72	1.22	1.1	0.95	1.54	3.02
−1.29702	4.460335	1.76	1.3	0.97	1.22	1.04	1.64	1.41	2.92
−1.48489	4.464725	1.59	1.15	0.46	0.72	0.49	1.97	1.47	3.01
−1.01153	4.580055	1.9	1.73	1.86	1.86	2.54	1.13	2.38	2.28
−1.01392	4.591625	1.74	1.78	1.37	1.12	2.01	1.46	2.34	2.31
−1.0442	4.62987	2.14	1.71	1.61	1.61	2.21	1.9	2.43	2.37
−1.2255	4.413635	1.73	1.6	1.02	1.27	1.54	1.45	1.56	3.02
−1.20624	4.46022	1.39	1.3	1.02	1.27	1.6	0.71	1.48	3.09
−0.96153	4.48976	1.44	1.4	1.33	1.33	1.62	0.55	1.61	2.86
−0.99881	4.538065	2.02	1.59	1.33	1.33	1.77	1.15	1.79	2.84

and quantum-chemically derived descriptor variables (Table 2). Variables like boiling point (Bp), melting point (Mp) and density (D) were taken from standard reference compilations (Weast, 1987), whereas log P and σ^- (sigm) were obtained from the original work (Deneer et al., 1987; Deneer et al., 1989), and the three theoretical descriptors (HOMO, LUMO, hardness (η)) from semi-empirical molecular orbital calculations (Stewart, 1990). In total, eight biological responses were available for this set of compounds. These are primarily related to toxicity towards the four aquatic species *Poecilia reticulata*, *Daphnia magna*, *Chlorella pyrenoidosa* and *Phytobacterium phosphoreum* (Table 2). With the exception of the response BCF, lower measured response values imply higher toxicity (Table 2).

2.3 Identification of sources of acute toxicity in produced water (Data Set III)

On offshore oil production platforms, large volumes of so called produced water are discharged into the sea. In addition to dispersed oil, such produced water contains dissolved hydrocarbons, organic acids, phenols, salts of heavy metals and traces of chemicals added along the process line (Johnsen et al., 1994). In order to understand the possible environmental impact of produced water emissions, it is important to uncover the causes of observed toxicological effects. In this example artificial water samples were tested for their acute toxicity using the Microtox test. The produced water samples were made artificially based on detailed knowledge of the composition of "real" produced water. In the produced water samples, the influence of five constituents (chemical factors) on aquatic toxicity was studied using statistical experimental design (Box et al., 1978). In total, 24 mixtures of produced water were blended according to a statistical experimental design in the five chemical factors (Table 3). The five factors were x_1 (Aro) representing the dissolved crude oil

Table 3. Factors and responses of the experimental design underlying data set III

	x_1	x_2	x_3	x_4	x_5	y_1	y_2	y_3	y_4
Run	Aro	Phen	HM	Inhib	Flocc	V18	V14	V12	V11
1	0.1	0.05	0.0064	0.5	15	0.5	1	2	9
2	10	0.05	0.0064	0.5	0.5	44.75	55.5	64.5	77
3	0.1	5	0.0064	0.5	0.5	30	38.5	44.5	48.75
4	10	5	0.0064	0.5	15	51.5	58.75	65.75	70.75
5	0.1	0.05	0.32	0.5	0.5	1.5	2.25	6.5	11.25
6	10	0.05	0.32	0.5	15	34.25	46	56	61.5
7	0.1	5	0.32	0.5	15	31	40.25	48.75	50.5
8	10	5	0.32	0.5	0.5	40	50	61.5	67
9	0.1	0.05	0.0064	15	0.5	0	1	10.25	34.25
10	10	0.05	0.0064	15	15	44.25	58.25	63.5	74.25
11	0.1	5	0.0064	15	15	26.25	36	47.25	59.75
12	10	5	0.0064	15	0.5	50	58.5	66.25	74.5
13	0.1	0.05	0.32	15	15	0	0	4.5	14
14	10	0.05	0.32	15	0.5	39.5	54.5	62.5	70.75
15	0.1	5	0.32	15	0.5	25.5	36.5	50.25	62.25
16	10	5	0.32	15	15	51.25	62	68.5	72.75
17	5	2.5	0.16	7.5	7.5	38.5	48.5	57.25	65.5
18	5	2.5	0.16	7.5	7.5	38.5	47	56	62
19	5	2.5	0.16	7.5	7.5	36.5	46.75	55.75	61.25
20	5	2.5	0.16	7.5	7.5	37	46.75	55.75	62.25
21	0.1	2.5	0.16	7.5	7.5	18.25	27	35.25	43.75
22	10	2.5	0.16	7.5	7.5	45.5	57	66.5	69.75
23	5	0.05	0.16	7.5	7.5	30	41.25	49.5	60
24	5	5	0.16	7.5	7.5	44	53.5	58.5	65.25

x_1 (Aro) = aromatics representing dissolved crude oil fraction [ppm]; x_2 (Phen) = mixture of phenols and C1–C4 alkylated phenols [ppm]; x_3 (HM) = mixture of water soluble salts of the heavy metals Cr, Mn, Cu, Zn, Hg and Pb [ppm]; x_4 (Inhib) = corrosion inhibitor [ppm]; x_5 (Flocc) = flocculant used in oil production [ppm].

Microtox responses are given as absolute reduction in light emissions. y_1 (V11) = undiluted sample; y_2 (V12) = sample diluted 1:2 with sea water; y_3 (V14) = sample diluted 1:4 with sea water; y_4 (V18) = sample diluted 1:8 with sea water. For more details, see Johnsen et al. 1994.

aromatic fraction, x_2 (Phen) corresponding to the phenolic fraction, x_3 (HM) representing a mixture of water soluble salts of various heavy metals, x_4 (Inhib) depicting the most toxic corrosion inhibitor employed at the oil fields, and x_5 (Flocc) corresponding to the most toxic flocculant added along the production line. Among these five factors, $x_1 - x_3$ are unavoidable constituents when producing oil, whereas x_4 and x_5 represents contaminants from artificial additives. For each one of these 24 water samples, four toxicity responses were registered in the Microtox assay, and the endpoints were expressed as absolute reduction in light emission at four different dilutions of the water mixtures (Johnsen et al., 1994). In summary, the objective was to investigate whether the artificial constituents (x_4 and x_5) caused significant aquatic toxicity.

3 The linear PLS model

The development of a PLS model can be described as follows: For a certain set of observations – compounds in data sets I and II and mixed samples in data set III – appropriate response variables are monitored. These form the N×M response data matrix **Y**, where N and M are the number of observations and responses, respectively. Moreover, for the same set of observations, relevant predictor variables are gathered to constitute the N×K predictor matrix **X**, where N is the same as above and K the number of predictor variables. The Y-data are then modelled by the X-data using PLS. A geometric representation of PLS is given in Fig. 3. The observations can be seen as points in two spaces, that of **X** with K dimensions and that of **Y** with M dimensions. PLS finds lines, planes or hyperplanes in **X** and **Y** that map the shapes of the point-swarms as well as possible.

PLS has two primary objectives, namely to well approximate **X** and **Y** and to model the relationship between **X** and **Y**. This is accomplished by making the bilinear projections

$$X = TP' + E \tag{1}$$

$$Y = UC' + G \tag{2}$$

and connecting X and Y through the inner relation

$$U = T + H \tag{3}$$

where **E**, **G** and **H** are residual matrices. Here **T** is N×A, **P** is K×A and **C** is M×A, where A is the number of PLS components. A more detailed account of the PLS algorithm is given in the appendix.

PLS simultaneously projects the X and Y-variables onto the same subspace, **T**, in such a manner that there is a good relation between the position of one observation on the X-plane and its corresponding position on the Y-plane (Fig. 3). Moreover, this relation is asymmetric ($X \Rightarrow Y$), which follows from equation (3). In this respect, PLS differs from, e.g., canonical correlation where this relation is symmetric. In essence, each PLS model dimension consists of the X score vector **t**, the Y score vector **u**, the X loading vector **p**, the X weight vector *w* and the Y weight vector **c** (see appendix). The weight vectors **w** and **c** are used for interpreting which X-variables are influential for modelling the Y-variables.

Another way to see PLS is that it forms "new x-variables", **t**, as linear combinations of the old ones, and therafter uses these new t's as predictors of Y. Only as many new t's are formed as are needed, and this is assessed from their predictive power (see below).

3.1 Interpretation

Once a PLS model has been derived, it is important to construe its meaning. For this, the scores **t** and **u** are considered. They contain information about the observations and their similarities/dissimilarities in X- and Y-space with respect to the given problem and model. The X-weights **w** and the Y-weights **c** provide informa-

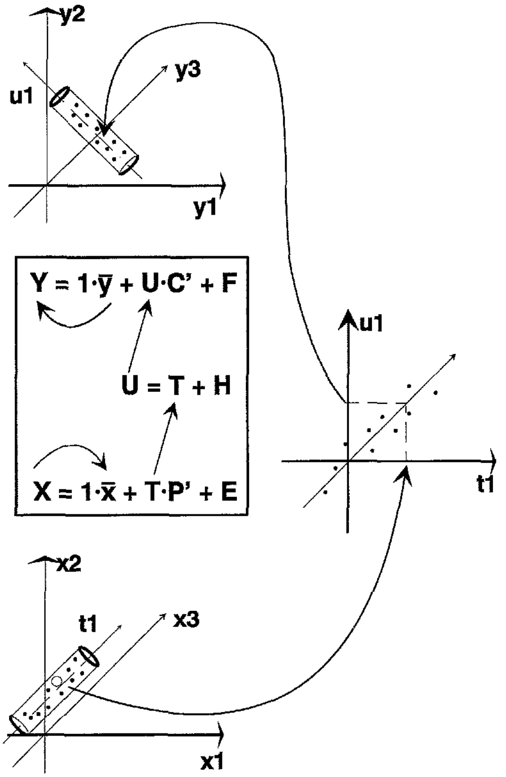


Figure 3. A geometrical representation of PLS

tion about how the variables combine to form **t** and **u**, which in turn express the quantitative relation between *X* and *Y*. Hence, these weights are essential for the understanding of which X-variables are important for modelling **Y** (numerically large **w**-values), which X-variables that provide common information (similar profiles of **w**-values), and for the interpretation of the scores **t**.

Sometimes it may be quite taxing to overview the PLS weights, especially if the number of latent variables to consider is larger than about 3. In such circumstances, PLS provides a powerful alternative, the **VIP** (*v*ariable *i*nfluence on *p*rojection) parameter, which informs about the relevance of each X-variable pooled over all dimensions and Y-variables (Wold, 1995). Thus, in principle, VIP in square is a weighted sum of squares of the PLS weights, **w**, taking into account also the amount of Y-variance explained by each latent variable. We note that for a one-dimensional PLS model, VIP-values are proportional to the values of **w**.

Alternatively, the PLS solution may be transferred into a regression-like model:

$$\mathrm{Y} = \mathrm{X}\,\mathrm{B}_{\mathrm{PLS}} + \mathrm{F} \tag{4}$$

Here $\mathbf{B}_{\mathrm{PLS}}$ corresponds to the regression coefficients. Thus, these coefficients are determined from the underlying PLS model and can be used for interpretation, in the same way as coefficients originating from MLR. However, with collinear variables we must remember that these coefficients are not independent.

The parts of the data that are not explained by the model, the *residuals,* are of diagnostic interest. Large Y-residuals indicate that the model is inadequate, and a normal probability plot of the residuals of a single Y-variable is useful for identifying outliers. In PLS we also get residuals for **X**, the part not used in the modelling of **Y**. Such X-residuals are useful for identifying outliers in the X-space, i.e., observations which do not conform with the model.

3.2 Incomplete X and Y matrices (missing data)

PLS tolerates moderate amounts of missing data both in **X** and **Y**. With missing data in **Y**, it must be multivariate, i.e. **Y** must have at least two columns. The larger the matrices **X** and **Y** are, the higher the proportion of missing data that may be tolerated. Here, the three examples have 19, 15 and 24 observations, and for such ordinary sized matrices around 10 to 20% missing data elements can be handled, provided that they are not missing according to some systematic pattern. The PLS algorithm accounts for the missing values, in principle by iteratively substituting the missing values with predictions from the model. This corresponds to giving the missing data values that have zero residuals and thus have no influence on the model parameters.

3.3 One Y-variable at a time, or all in the same model?

PLS has the ability to model and analyze several Y-variables together. This is favorable when the Y-variables are correlated, because the analyst only obtains one model to interpret and not one model for each single variable. If the Y's really measure different things, however, and are fairly independent, one gains little by analyzing them in the same model. On the contrary, with fairly independent Y-variables the PLS model tends to have many components and hence be difficult to interpret. The separate modelling of the Y's then gives a set of simpler models with fewer dimensions, which are easier to interpret.

To judge whether the Y-variables are correlated or not, it is recommended to precede the PLS analysis with a principal component analysis (PCA) of the Y-matrix. This will inform about the practical rank of **Y**, A, i.e., the number of components of the PC model. If A is small compared to the number of Y-variables (M), and if we can understand the resulting components, we can conclude that the Y's are correlated, and a PLS model of all responses together is warranted. Often, however, one finds from the PCA that the Y's cluster in two or three groups according to the nature of activity they measure. Then this is an indication for one separate PLS model for each such group of Y-variables.

3.4 The number of PLS components, A

It is essential to determine the correct complexity of a PLS model. With many X-variables there is a substantial risk for "overfitting", i.e., getting a well fitting model with little or no predictive power. Hence a strict test of the statistical significance of each consecutive PLS component is necessary. This test is used to determine where to stop, when components start to be non-significant.

Cross-validation (CV) is a practical and reliable way to test this significance (Wold, 1995; Lindgren, 1994; Wold, 1978), and one that has become standard in PLS analysis. A good discussion of the subject was recently given in (Wakeling and Morris, 1993). Basically, CV is performed by dividing the data in a number of groups, say, seven, and then developing a number of parallel models from the reduced data with one of the groups deleted. It should be noted that having the number of CV groups equal to N, i.e., the so called leave-one-out approach, is debatable (Shao, 1993; Wold and Eriksson, 1995). In practice, between five and ten groups works well.

After developing a model, the deleted data are used as a test set, and differences between actual and predicted Y-values are calculated for the test set. The sum of squares of these differences are computed and collected from all the parallel models to form *PRESS* (predictive residual sum of squares), which is a measure of the predictive ability of the model. Usually, PRESS is reexpressed as Q^2 (the "cross-validated R^2") which is (1-PRESS/SS) were SS is the sum of squares of Y, corrected for the mean. This can be compared with $R^2 =$ (1-RSS/SS), where RSS is the residual sum of squares. In models with several Y's, one obtains also R^2_m and Q^2_m for each Y-variable. The explained variance, R^2 or more strictly adjusted for degrees of freedom, R^2_{adj}, varies between 0 and 1, where 1 means a perfect model and 0 a model of no relevance at all. Normally, the predicted variance, Q^2, varies between 0 and 1 as well, but negative values indicating nonsense models, may be obtained occasionally. As a rule of thumb, R^2 is normally 5–20% higher than Q^2, and substantially larger differences is a warning for overfitting, or many irrelevant X-variables.

3.5 Softwares

PLS is incorporated into many commercially available statistical packages including SIRIUS, SCAN, UNSCRAMBLER, PIROUETTE, MODDE and SIMCA. We use MODDE 2.1 for Windows (Modde manual 1994) for examples 1 and 3, and SIMCA P 2.1 for Windows (Simca P manual 1994) for example 2. MODDE contains an MLR option (example 1) and statistical experimental design support (example 3), whereas SIMCA offers multivariate projection methods (example 2).

4 Results

4.1 Data set I

As said above, the aim of the first example is not so much to deal with PLS and aquatic data, but rather to highlight some advantages of PLS compared to MLR when dealing with collinear variables. These advantages include features such as stability of the model and believability of its result. This data set has four X-variables and one response. When applying MLR to the data a model with $R^2 = 0.66$,

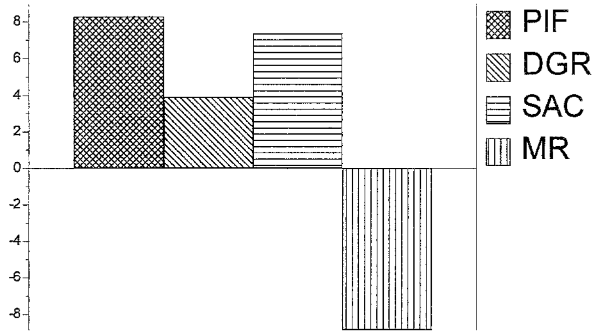


Figure 4. Regression coefficients of scaled and centered variables for data set I (MLR)

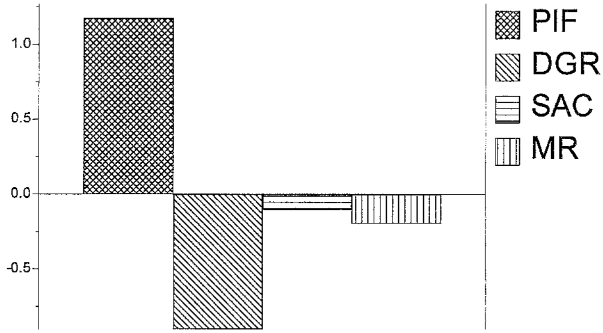


Figure 5. Regression coefficients of scaled and centered variables for data set I (PLS)

$R^2_{adj} = 0.57$ and $Q^2 = 0.29$ resulted, while the corresponding values for PLS were $R^2 = 0.48$, $R^2_{adj} = 0.34$ and $Q^2 = 0.30$. At first sight MLR seems to outperform PLS. However, since the Q^2's are of similar size the tentative conclusion is that MLR shows overfit.

Next, we examine the regression coefficients of both models (Figs. 4 and 5). Interestingly, there are some discrepancies, both with regards to the size of the coefficients as well as their sign. Let us take a closer look at variables x_2 (DGR) and x_4 (MR). According to the MLR model (Fig. 4), DGR has a positive and MR a negative relation to the response variable. As opposed to this result, the PLS model (Fig. 5) suggests that DG is negatively related with the response variable, and that MR has no modelling influence whatsoever. Furthermore, x_1 and x_2 are strongly negatively correlated (Table 1) and thus their coefficients ought to have opposite signs in the MLR model. Astonishingly, however, these two variables are estimated by MLR to relate positively to this response. A similar contradiction can also be traced for the variable pair x_3/x_4. It should be noted thought that – due to their collinearity – the coefficients of x_3 and x_4 are statistically insignificant.

All these MLR-results are puzzling and we must try to find out what is real and what is misleading. The easiest way to elucidate this is to look at scatter plots of raw

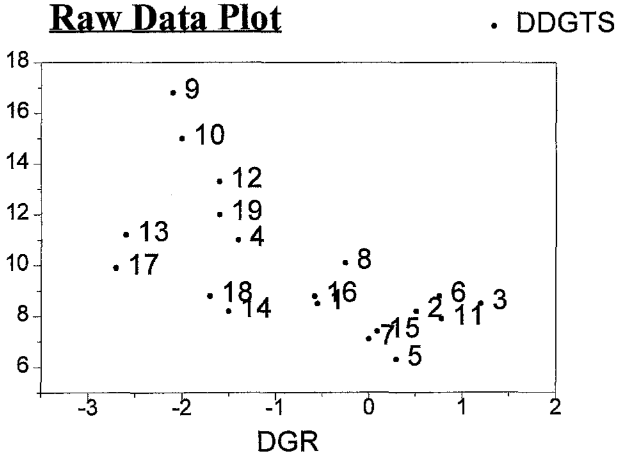


Figure 6. Raw data plot of the response DDGTS versus the predictor DGR. Notation as in Table 1

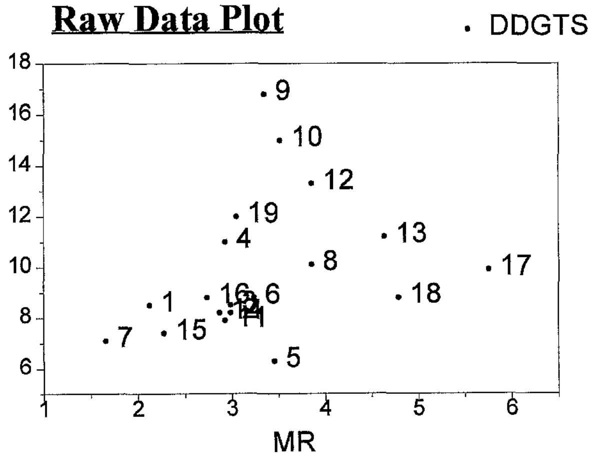


Figure 7. Raw data plot of the response DDGTS versus the predictor MR. Notation as in Table 1

data and scrutinize how, for instance, DGR and MR correlate with the response variable in reality. Figure 6 shows that DGR has a negative correlation (r = −0.65) and Fig. 7 that MR has, if any, a weak positive correlation (r = 0.29) with respect to the endpoint. Thus, this supports the PLS model and contradicts the MLR output. The reason why MLR produces erroneous coefficients is that the X-variables are strongly correlated. Not only are the coefficients of wrong sign, but the model is likely overfitted as well. Hence, PLS is preferred to MLR for this and other data sets with correlated predictor variables.

4.2 Data set II

The PLS analysis of the second data set, with eight X-variables and eight Y-variables, resulted in a two-component model with $R^2 = 0.76$, $R^2_{adj} = 0.72$ and $Q^2 = 0.67$, which is an excellent result taking into account biological variability and the fact that eight responses are handled simultaneously. It is also possible to extract similar information for each Y-variable separately, which is presented in Fig. 8. Evidently, the individual R^2's vary between 0.53–0.95 and Q^2's range from 0.38–0.92, and it is clear that five endpoints (DM21d, DMRm, CPEC50, PoeLC50 and BCF) are well modelled by this battery of eight predictor variables, and that some improvements would be desirable for three responses (DM48h, DMe and PHEC50). However, the

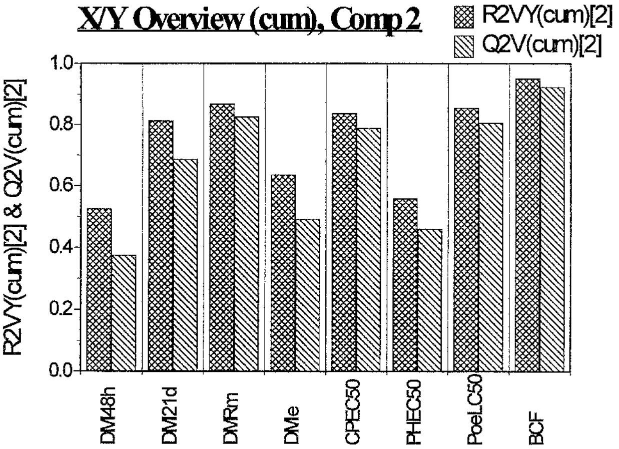


Figure 8. Individual R^2Y (explained sum of squares) and Q^2 (predicted variance) of each response

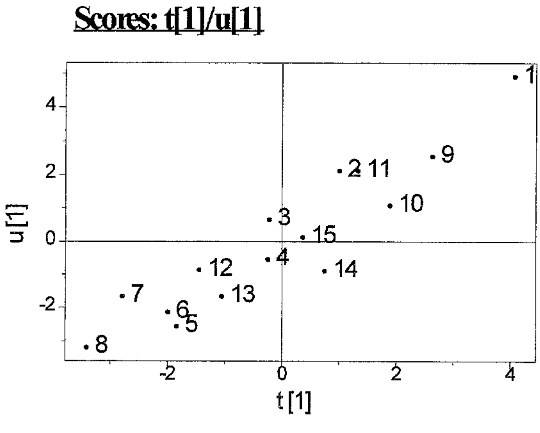


Figure 9. First pair of latent variables for the PLS model of data set II. Notation as in Table 2

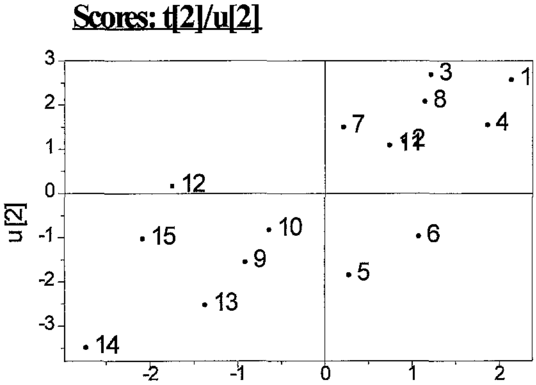


Figure 10. Second pair of latent variables for the PLS model of data set II. Notation as in Table 2

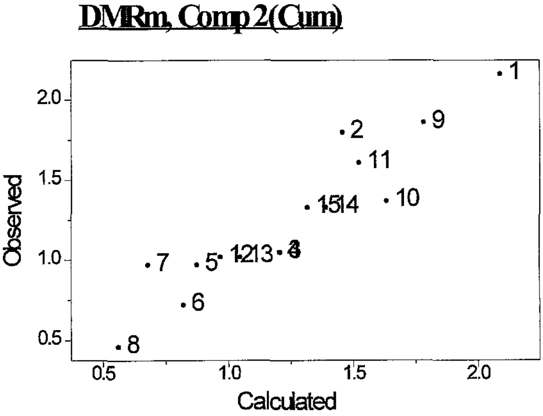


Figure 11. Observed versus calculated values for the response DMRm (data set II). Notation as in Table 2

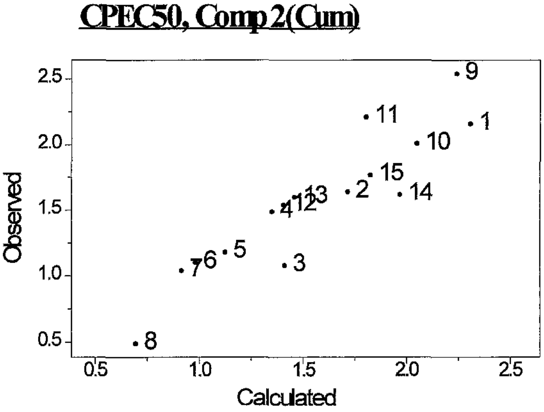


Figure 12. Observed versus calculated values for the response CPEC50 (data set II). Notation as in Table 2

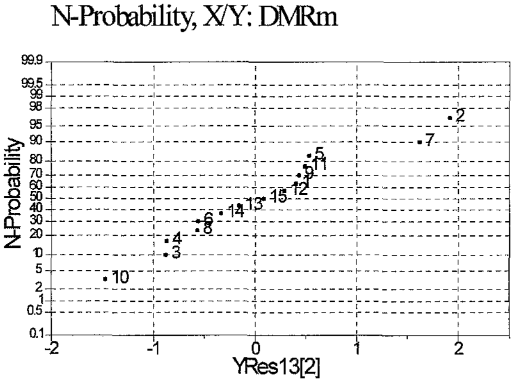


Figure 13. Normal probability plot of the residuals of DMRm after two PLS components (data set II). Notation as in Table 2

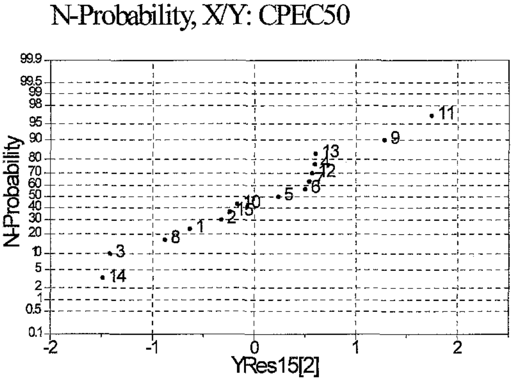


Figure 14. Normal probability plot of the residuals of CPEC50 after two PLS components (data set II). Notation as in Table 2

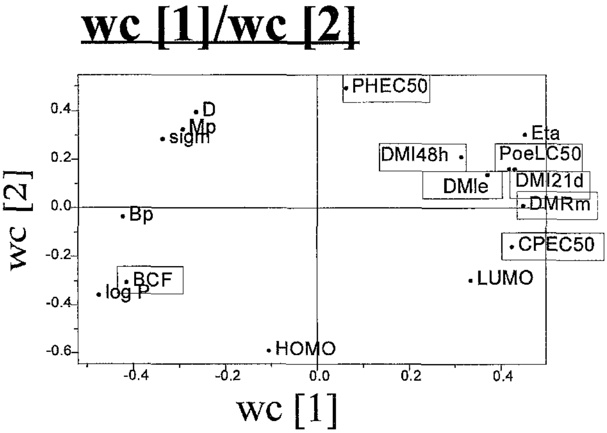


Figure 15. The second PLS weight vector plotted against the first for the PLS model of data set II. Notation as in Table 2. The eight responses are boxed

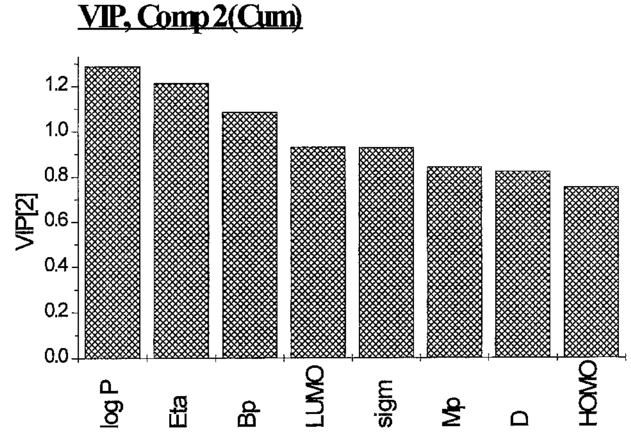


Figure 16. Variable influence on projection (VIP) for the predictor variables of dat set II. The higher the value the more influential the variable. Notation as in Table 2

general relationship between X and Y, as expressed by the latent variables **t** and **u** (Figs. 9 and 10), is stable and the conclusion is that the multivariate QSAR is well founded and warranted, and has a good predictive power. To explore the fit of the QSAR, we consider Figs. 11 and 12 in which good relationships between observed and calculated responses for DMRm and CPEC50 are displayed. The normal probability plots (Figs. 13 and 14) of the Y-residuals corroborate this model, since no strong deviants are found (the residuals lie well within ± 2 SD.s).

For the interpretation of this QSAR model we consider the PLS weights to see how the X- and Y-variables are interrelated (Fig. 15). Figure 15 indicates that all X-variables load strongly in the two model dimensions, and that D, Mp, σ^- (sigm) and LUMO are closely related. A second group is formed by log P, Bp and η (Eta); whereas HOMO provides information different from these two groups. The VIP plot is displayed in Fig. 16, and this column plot reveales that log P is the most important variable, followed by eta, Bp, and so on. This may be interpreted saying that hydrophobic properties of the nitrobenzene derivatives are of crucial importance for the toxic effects they elicit.

Now that we tentatively have interpreted the QSAR, we may attempt to get some feedback from the PLS score plot in Fig. 9. (Fig. 10 accounts for a minor portion of the Y-variance, and is neglected for this sake). This graph well summarizes the distribution of the nitrobenzene derivatives along the various toxicity scales. Altogether, nitrobenzene (no. 1) is the least toxic compound to these aquatic organisms and at the same time exhibits the lowest bioconcentration factor, whereas 1,3-dichloro-5-nitrobenzene (no. 8) is the most potent chemical in the same test systems. Actually, nitrobenzene is the least hydrophobic compound (lowest value of log P) and 1,3-dichloro-5-nitrobenzene the most, which corroborate the previous interpretation concerning the significance of log P. For a deeper toxicological account of these endpoints, we refer to the original literature (Deneer et al., 1987; Deneer et al., 1989). In summary, this example underlines the suitability of PLS for modelling highly correlated X- and Y-variables, and at the same time point out the possibilities of predicting aquatic toxicity of environmental pollutants based on knowledge of their chemical and structural characteristics.

4.3 Data set III

The third example is different from the two preceding ones in that statistical experimental design was used to plan the experiments. One great advantage of designing the X-matrix is the possibility of evaluating model inadequacies in terms of a lack of fit estimate. In addition, with designed data, the X-variables are independent, which justifies the use of MLR, although we prefer PLS because it can accommodate all four responses in the same model.

The current investigation was done in two-step procedure. From the beginning 19 experiments (Table 3) were set up according to a 2^{5-1} fractional factorial design with three center-points. This design supports the estimation of linear and interaction terms. The first PLS analysis of the resulting data suggested that only the interaction effect Aro*Phen was meaningful besides the five linear terms. Thereafter PLS was applied to these six X-variables and the four Y-responses, yielding a two-component model with $R^2 = 0.95$, $R^2_{adj} = 0.92$ and $Q^2 = 0.80$.

The relationship between X and Y for this model is linear with respect to the first pair of latent variables, but is non-linear in the second (Fig. 17). This model deficiency is also seen in the lack of fit of 26.2, which is large value indicating model imperfections. The interpretation of this model revealed that x_1 (Aro) and x_2 (Phen) and their joint interaction are the only significant variables, and that the influence of x_3-x_5 is negligible. This was uncovered by inspecting resulting regression coefficients and their confidence limits (no plots provided).

The investigators wished to better account for the non-linearity observed (cf. Fig. 17). Hence, some experimental trials were added to enable estimation of the squared terms of the two primary variables, Aro and Phen. Thus, the original design was augmented with five trials (runs 20–24 in Table 3) laid out so as to allow Aro^2

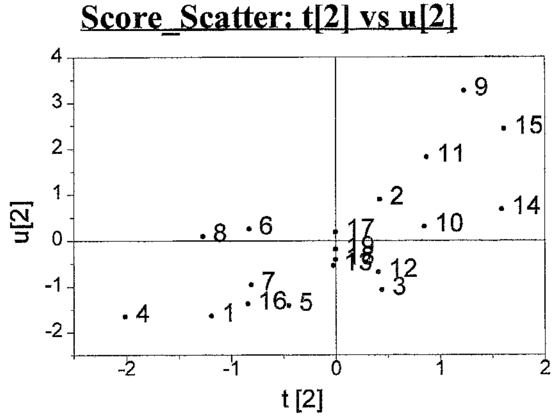


Figure 17. PLS t2/u2 score plot for model one (19 observations) of data set III. Notation as in Table 3. Note the non-linear relationship between t and u (and hence X and Y)

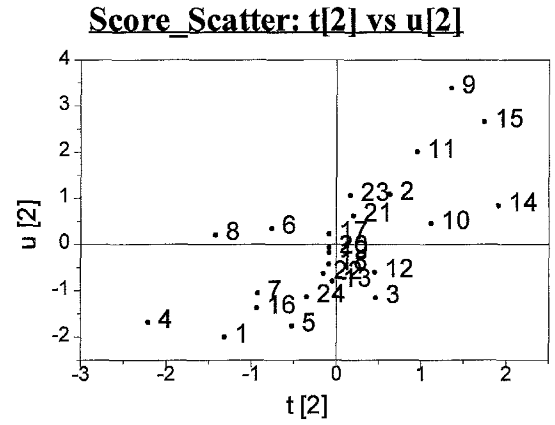


Figure 18. PLS t_2/u_2 score plot for model two (24 observations) of data set III. Notation as in Table 3. Note that the curvature has diminished (cf Figure 17)

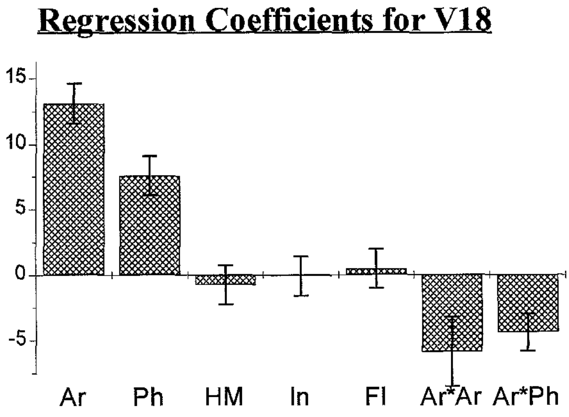


Figure 19. PLS regression coefficients of response V18 (with 95 % confidence bars) for model two of data set III. Notation as in Table 3

and Phen^2 to be estimated. In the PLS analysis of this data set, it was found that of these only Aro^2 was significant. The "final" model thus included seven terms, five linear, one interaction and one square term, with the overall statistics of $R^2 = 0.97$, $R^2_{adj} = 0.95$ and $Q^2 = 0.82$. We see in Fig. 18 that the non-linearity in t_2/u_2 has been eliminated due to the inclusion of Aro^2. The lack of fit is reduced to 13.2. Figure 19 shows the significance of Aro^2 for the response V18, and Fig. 20 displays the response surface for V18 in the two significant factors.

In summary, this investigation revealed that only the two factors x_1 and x_2 critically influence the toxicity responses, and that they do so in a non-additive manner. This is modelled by incorporating one interaction and one square term. It is also evident that this kind of information would have been difficult to acquire without the statistical experimental design of the data.

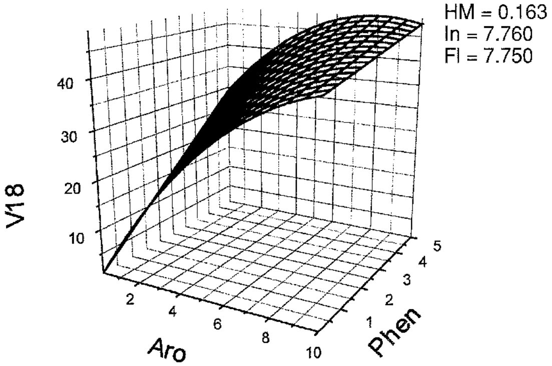


Figure 20. Response surface plot of the second PLS model of data set III, showing how the response V18 changes as a function of x_1 (Aro) and x_2 (Phen), with x_3–x_5 held constant at their center level

5 Discussion

Data analysis carried out with the intention of linking a set of predictor variables, matrix X, to a set of response variables, matrix Y, for quantification and prediction purposes, is a common task in scientific and industrial research and development. In chemistry, aquatic science, and so on, these data tables X and Y often are multicollinear, because they are not generated in adherence to a statistical experimental design protocol, and because the manner in which these tables are produced means that they have many more variables than observations. The classical approach to a regression-type problem formulation relies on methods such as MLR or canonical correlation. As discussed and shown above, however, MLR (and the like) will not work properly when applied to multivariate collinear data. Inevitably, this will only yield models of low relevance and poor reliability because the derived regression coefficients are highly uncertain.

Being a bilinear projection method, PLS provides a rational methodology for modelling the quantitative and often complex relationships between the multivariate matrices X and Y. The assumptions underlying PLS – correlated X-variables, X-variables with errors, residuals may be structured (cf. Fig. 1) – are more realistic than those underlying MLR. Hence models developed with PLS generally will have greater practical applicability and be more realistic. In addition, the diagnostics of PLS and similar methods (PCR, CCA, etc.), the scores, loadings, coefficients, and VIP plots and crossvalidation, supply information about the data structure and model complexity that is not attainable in traditional MLR. This facilitates model interpretation and aids the detection of inhomogeneities and inconsistencies in data (Verhaar et al., 1994). The connection between PLS weights, coefficients and VIPs is overviewed in Fig. 21.

The aim with the three examples has been to introduce the concept of multivariate projections and the method of PLS, to illustrate the utility of PLS in aquatic science (as well as many other fields), and to motivate present and presumptive

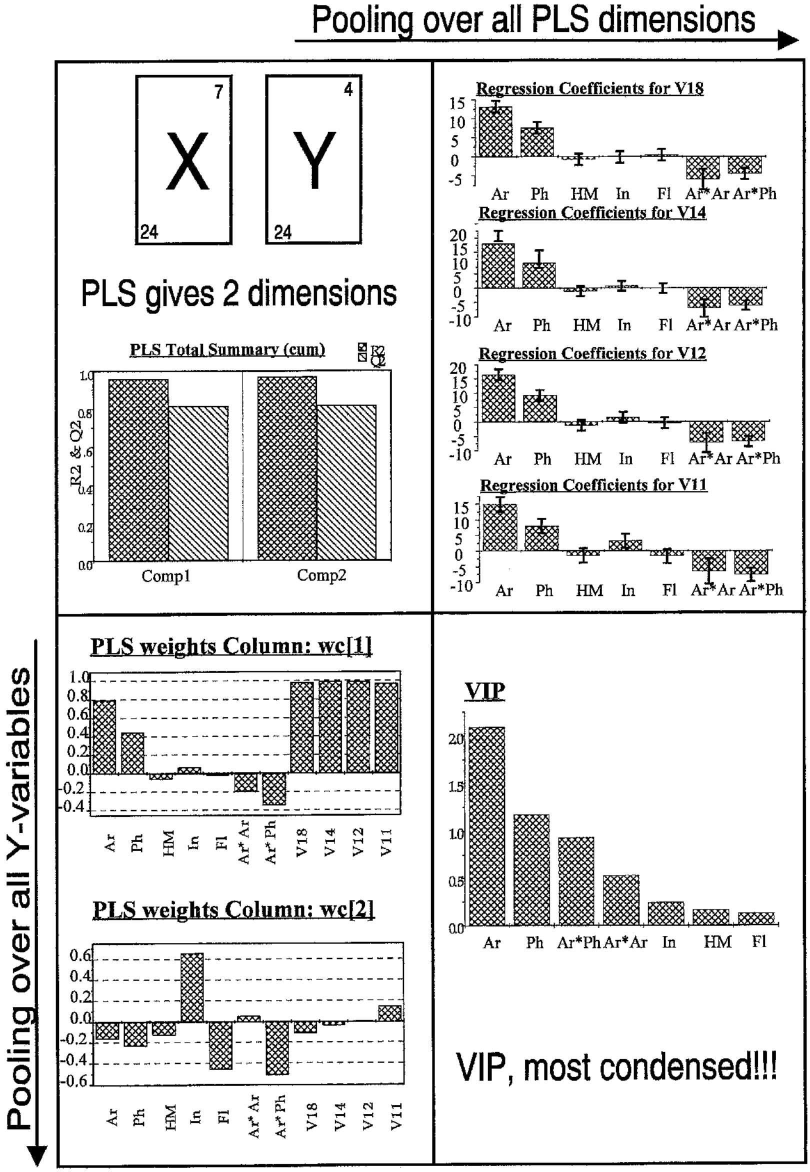


Figure 21. An overvies of PLS weights, coefficients and VIPs. This may be considered as a two-way phenomenon. The calculation of weights is a pooling over Y-variables, and the computation of regression coefficients is a pooling over components. VIPs, however, are obtained pooling both ways. Thus, the VIP parameter is the most condensed way of expressing the variable information

users of PLS of its analytical power. The first case addressed was an example in which the PLS solution cast some doubts on the MLR model, and did so by indicating that the upper (and more reliable) limits of R^2 and R^2_{adj} were lower than what could be expected judging from MLR. Also, the regression coefficients of PLS were also more in line with reality than those of MLR (cf. Figs. 4–7). Secondly, PLS was used for QSAR analysis, with the final goal of being capable of modelling and predicting toxicity profiles of mono-nitrobenzenederivatives from their chemical properties. This data set exemplified two features of PLS that are lacking in MLR, viz. the ability to treat several responses (here eight) in one single model, and the capability to cope with incomplete data matrices (missing data). Besides PLS working well in this application, we note that multivariate QSAR modelling is very useful in aquatic science (Hermens, 1989; Blum and Speece, 1990). Finally, PLS was used to explore which among five chemical factors significantly influenced the aquatic toxicity of produced water discharged from oil production. With the statistical experimental design used as a foundation, this example underlines how a complicated system – here the relationship between mixture composition and aquatic toxicity – can be mapped efficiently and intelligently. It is indisputable that only two of the chemical factors considered were influencing the responses. The question of revealing which mixture constituents adversely affect certain species or environmental compartments, cannot be adequately resolved and quantified unless such experimental planning is utilized.

6 Concluding remarks

PLS is a rational data analytical tool to gain insight into complex systems encountered in aquatic science. The graphical representation of PLS parameters and residuals enables evaluation of developed models and facilitates their interpretation. Since the assumptions of PLS are more realistic than those of MLR, it is our belief that PLS will be increasingly used in all kinds of complicated scientific applications.

Appendix

In the outline of PLS below, the index a (a = 1, 2, . . . , A) represents the number of PLS components, index i (i = 1, 2, . . . , N) the number of observations, index k (k = 1, 2, . . . , K) the number of X-variables and index m (m = 1, 2, . . . , M) the number of Y-variables. The linear PLS model finds A "new" variables, latent variables, denoted by $\mathbf{t}_a$. These scores are linear combinations of the original variables $\mathbf{x}_k$ with the coefficients, "weights", $\mathbf{w}^*_{ka}$.

$$t_{ia} = \Sigma_k w^*_{ka} x_{ik} \tag{1}$$

PLS computes the X-scores ($\mathbf{t}_a$'s) to have certain advantageous properties. First of all, they are good predictors of Y, so that

$$y_{im} = \Sigma_a c_{ma} t_{ia} + f_{im} \tag{2}$$

$$Y = TC' + F \tag{2a}$$

in which $\mathbf{c}_{ma}$ are the PLS Y-weights and $\mathbf{f}_{im}$ the Y-residuals. The latter formulation (2a) expresses the model in matrix form. The residuals, $\mathbf{f}_{im}$, express the deviations between the observed and modelled data, and comprise the elements of the Y residual matrix, **F** in (2a). Because of (1) and (2), the latter can be rewritten as a regression model:

$$y_{im} = \Sigma_a c_{ma} \Sigma_k w^*_{ka} x_{ik} + f_{im} = \Sigma_k b_{mk} x_{ik} + f_{im} \tag{3}$$

The PLS regression coefficients, $\mathbf{b}_{mk}$, can be written as:

$$b_{mk} = \Sigma_a c_{ma} w^*_{ka} \tag{4}$$

Secondly, the X-scores are few (A in number) and orthogonal, and are good summaries of **X**, so that the X-residuals, $\mathbf{e}_{ik}$, in (5) are “small”:

$$x_{ik} = \Sigma_a t_{ia} p_{ka} + e_{ik} \tag{5}$$

$$X = TP' + E \tag{5a}$$

In (5) above $\mathbf{p}_{ka}$ are the X-variable loadings. The latter equation (5a) is the X-model in matrix form. With multivariate Y (when M > 1), the Y-scores are good summaries of Y, so that the residuals, $\mathbf{g}_{im}$, in (6) are “small”:

$$y_{im} = \Sigma_a u_{ia} c_{ma} + g_{im} \tag{6}$$

$$Y = UC' + G \tag{6a}$$

Here, $\mathbf{u}_{ia}$ denotes the Y-scores and $\mathbf{c}_{ma}$ corresponds to the Y-weights. The latter (6a) is the Y-model in matrix form.

After each dimension, **a**, the X-matrix is updated by subtracting $\mathbf{t}^*_{ia}\,\mathbf{p}_{ka}$ from the element $\mathbf{x}_{ik}$. This makes the PLS model alternatively be expressed in weights $\mathbf{w}_a$ referring to the residuals after previous dimension, $\mathbf{E}_{a-1}$, instead of relating to the X-variables themselves (weights $\mathbf{w}^*$ in eqn. 1). Thus, instead of (1), we have (7):

$$t_{ia} = \Sigma_k w_{ka} e_{ik,a-1} \tag{7}$$

$$e_{ia,a-1} = e_{ik,a-2} - t_{i,a-1} p_{k,a-1}$$

$$e_{ik,0} = x_{ik}$$

However, the weights, **w**, can be transformed to $\mathbf{w}^*$, which directly relate to X, giving (1) above. The relation between the two is given by:

$$W^* = W (P' W)^{-1} \tag{8}$$

REFERENCES

Blum, D.J.W. and R.E. Speece, 1990. Determining chemicals toxicity to aquatic species. Environ. Sci. Technol. 24:284–293.

Box, G.E.P., W.G. Hunter and J.S. Hunter, 1978. Statistics for Experimenters, J. Wiley and Sons, N.Y.

Deneer, J.W., T.L. Sinnige, W. Seinen and J.L.M. Hermens, 1987. Quantitative structure-activity relatinships for the toxicity and bioconcentration factor of nitrobenzene derivatives towards the guppy (*Poecilia reticulata*). Aquatic Toxicol. 10:115–129.

Deneer, J.W., C.J. van Leeuwen, W. Seinen, J.L. Maas-Diepeveen and J.L.M. Hermens, 1989. QSAR study of the toxicity of nitrobenzene derivatives towards *Daphnia magna*, *Chlorella pyrenoidosa* and *Photobacterium phosphoreum*. Aquatic Toxicol. 15:83–98.

Draper, N.R. and H. Smith, 1981. Applied regression analysis, J. Wiley and Sons, N.Y.

El Tayar, N., R.S. Tsai, P.A. Carrupt and B. Testa, 1992. Octan-1-ol water partition coefficients of zwitterionic amino acids. Determination by centrifugal partition chromatography and factorization into steric/hydrophobic and polar components. J. Chem. Soc. Perkin Trans. 2:79–84.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugenics 7:179–188.

Frank, I.E. and J.H. Friedman, 1993. A statistical view of some chemometric regression tools. Technometrics 35:109–148

Hermens, J.L.M., 1989. Quantitative structure-activity relationships of environmental pollutants. In: (ed.) O. Hutzinger, The handbook of environmental chemistry, Vol. 2E, Springer Verlag, Berlin, Germany, pp. 111–162.

Jackson, J.E., 1991. A users guide to principal components, Wiley-Interscience.

Jongman, R.G.H., C.J.F. ter Braak and O.F.R. van Tongeren, 1987. Data analysis in community and landscape ecology, Pudoc, Wageningen, The Netherlands.

Johnsen, S, A.T. Smith, J. Brendenhaug, H. Riksheim and A.L. Gjose, 1994. Identification of sources of acute toxicity in produced water. SPE 27138, pp. 1–8.

Lindgren, F., 1994. Third generation PLS – Some elements and applications. Ph. D. Thesis, Umeå Univesity, Umeå, Sweden.

MODDE 2.1 manual 1994, Umetri AB, P.O. Box, 90719 Umeå, Sweden.

Mullet, G.M., 1976. Why regression coefficients have the wrong sign. J. Qual. Technol. 8:121–126.

Shao, J., 1993. Linear model selection by cross-validation. J. Amer. Stat. Assoc. 88:486–494.

Stewart, J.J.P., 1990. MOPAC manual, version 6.0. Frank J. Seiler Research Laboratory, U.S. Air Force Academy, CO.

SIMCA P 2.1 manual 1994, Umetri AB, P.O. Box 7960, 90719 Umeå, Sweden.

Topliss, J.G. and R.P. Edwards, 1979. Chance factors in studies of quantitative structure-activity relationships. J. Med. Chem. 22:1238–1244.

Verhaar, H.J.M., L. Eriksson, M. Sjöström, G. Schüürmann, W. Seinen and J.L.M. Hermens, 1994. Modelling the toxicity of organophosphates: A comparison of the multiple lineare regression and PLS regression methods. Quant. Struct.-Act. Relat. 13:133–143.

Wakeling, I.N. and J.J. Morris, 1993. A test of significance for partial least squares (PLS). J. Chemometrics 7:281–304.

Weast, R.C., 1987. Handbook of chemistry and physics, 67th ed., CRC Press, Bocan Raton, FL.

Wold, S., 1978. Cross-validatory estimation of the number of components in factor and principal component models. Technometrics 20:387–405.

Wold, S., C. Albano, W.J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg and M. Sjöström, 1984. Multivariate data analysis in chemistry. IN: B.R. Kowalski (ed.), Chemometrics – Mathematics and statistics in chemistry, D. Reidel Publishing Company, Dordrecht, Holland, pp. 1–79.

Wold, S., 1995. PLS for multivariate linear modelling. In: H. van de Waterbeemd (ed.), QSAR: Chemometric methods in molecular design, Methods and principles in medicinal chemistry, Vol. 2, Verlag Chemie, Weinheim, Germany, pp. 195–218.

Wold, S. and L. Eriksson, 1995. Validation Tools. In: H. van de Waterbeemd (ed.), QSAR: Chemometric methods in molecular design, Methods and principles in medicinal chemistry, Vol. 2, Verlag Chemie, Weinheim, Germany, pp.309–318.

Received 23 January 1995;
revised manuscript accepted 30 May 1995.