ELSEVIER

Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

PLS-regression: a basic tool of chemometrics

Svante Wold [a,*], Michael Sjöström [a], Lennart Eriksson [b]

[a] *Research Group for Chemometrics, Institute of Chemistry, Umeå University, SE-901 87 Umeå, Sweden*
[b] *Umetrics AB, Box 7960, SE-907 19 Umeå, Sweden*

Abstract

PLS-regression (PLSR) is the PLS approach in its simplest, and in chemistry and technology, most used form (two-block predictive PLS). PLSR is a method for relating two data matrices, **X** and **Y**, by a linear multivariate model, but goes beyond traditional regression in that it models also the structure of **X** and **Y**. PLSR derives its usefulness from its ability to analyze data with many, noisy, collinear, and even incomplete variables in both **X** and **Y**. PLSR has the desirable property that the precision of the model parameters improves with the increasing number of relevant variables and observations.

This article reviews PLSR as it has developed to become a standard tool in chemometrics and used in chemistry and engineering. The underlying model and its assumptions are discussed, and commonly used diagnostics are reviewed together with the interpretation of resulting parameters.

Two examples are used as illustrations: First, a Quantitative Structure–Activity Relationship (QSAR)/Quantitative Structure–Property Relationship (QSPR) data set of peptides is used to outline how to develop, interpret and refine a PLSR model. Second, a data set from the manufacturing of recycled paper is analyzed to illustrate time series modelling of process data by means of PLSR and time-lagged X-variables.

Keywords: PLS; PLSR; Two-block predictive PLS; Latent variables; Multivariate analysis

1. Introduction

In this article we review a particular type of multivariate analysis, namely PLS-regression, which uses the two-block predictive PLS model to model the relationship between two matrices, **X** and **Y**. In addition PLSR models the "structure" of **X** and of **Y**, which gives richer results than the traditional multiple regression approach. PLSR and similar approaches provide *quantitative* multivariate modelling methods, with inferential possibilities similar to multiple regression, *t*-tests and ANOVA.

The present volume contains numerous examples of the use of PLSR in chemistry, and this article is merely an introductory review, showing the development of PLSR in chemistry until, around, the year 1990.

1.1. General considerations

PLS-regression (PLSR) is a recently developed generalization of multiple linear regression (MLR) [1–6]. PLSR is of particular interest because, unlike MLR, it can analyze data with strongly collinear (correlated), noisy, and numerous X-variables, and also simultaneously model several response variables, **Y**, i.e., *profiles of performance*. For the meaning of the PLS acronym, see Section 1.2.

* Corresponding author. Tel.: +46-90-786-5563; fax: +46-90-13-88-85.
E-mail address: svante.wold@umetrics.com (S. Wold).

0169-7439/01/$ - see front matter
PII: S0169-7439(01)00155-1

The regression problem, i.e., how to model one or several dependent variables, responses, **Y**, by means of a set of predictor variables, **X**, is one of the most common data-analytical problems in science and technology. Examples in chemistry include relating **Y** = properties of chemical samples to **X** = their chemical composition, relating **Y** = the quality and quantity of manufactured products to **X** = the conditions of the manufacturing process, and **Y** = chemical properties, reactivity or biological activity of a set of molecules to **X** = their chemical structure (coded by means of many X-variables). The latter models are often called QSPR or QSAR. Abbreviations are explained in Section 1.3.

Traditionally, this modelling of **Y** by means of **X** is done using MLR, which works well as long as the X-variables are fairly few and fairly uncorrelated, i.e., **X** has full rank. With modern measuring instrumentation, including spectrometers, chromatographs and sensor batteries, the X-variables tend to be many and also strongly correlated. We shall therefore not call them "independent", but instead "predictors", or just X-variables, because they usually are correlated, noisy, and incomplete.

In handling numerous and collinear X-variables, and response profiles (**Y**), PLSR allows us to investigate more complex problems than before, and analyze available data in a more realistic way. However, some humility and caution is warranted; we are still far from a good understanding of the complications of chemical, biological, and economical systems. Also, quantitative multivariate analysis is still in its infancy, particularly in applications with many variables and few observations (objects, cases).

1.2. A historical note

The PLS approach was originated around 1975 by Herman Wold for the modelling of complicated data sets in terms of chains of matrices (blocks), so-called path models, reviewed in Ref. [1]. This included a simple but efficient way to estimate the parameters in these models called NIPALS (Non-linear Iterative PArtial Least Squares). This led, in turn, to the acronym PLS for these models (Partial Least Squares). This relates to the central part of the estimation, namely that each model parameter is iteratively estimated as the slope of a simple bivariate regression (least squares) between a matrix column or row as the y-variable, and another parameter vector as the x-variable. So, for instance, the PLS weights, **w**, are iteratively re-estimated as $\mathbf{X}'\mathbf{u}/(\mathbf{u}'\mathbf{u})$ (see Section 3.10). The "partial" in PLS indicates that this is a partial regression, since the **x**-vector (**u** above) is considered as fixed in the estimation. This also shows that we can see any matrix–vector multiplication as equivalent to a set of simple bivariate regressions. This provides an intriguing connection between two central operations in matrix algebra and statistics, as well as giving a simple way to deal with missing data.

Gerlach et al. [7] applied multi-block PLS to the analysis of analytical data from a river system in Colorado with interesting results, but this was clearly ahead of its time.

Around 1980, the simplest PLS model with two blocks (**X** and **Y**) was slightly modified by Svante Wold and Harald Martens to better suit to data from science and technology, and shown to be useful to deal with complicated data sets where ordinary regression was difficult or impossible to apply. To give PLS a more descriptive meaning, H. Wold et al. have also recently started to interpret PLS as Projection to Latent Structures.

1.3. Abbreviations

AA	Amino Acid
ANOVA	ANalysis Of VAriance
AR	AutoRegressive (model)
ARMA	AutoRegressive Moving Average (model)
CV	Cross-Validation
CVA	Canonical Variates Analysis
DModX	Distance to Model in X-space
EM	Expectation Maximization
H-PLS	Hierarchical PLS
LDA	Linear Discriminant Analysis
LV	Latent Variable
MA	Moving Average (model)
MLR	Multiple Linear Regression
MSPC	Multivariate SPC
NIPALS	Non-linear Iterative Partial Least Squares
NN	Neural Networks

PCA	Principal Components Analysis
PCR	Principal Components Regression
PLS	Partial Least Squares projection to latent structures
PLSR	PLS-Regression
PLS-DA	PLS Discriminant Analysis
PRESD	Predictive RSD
PRESS	Predictive Residual Sum of Squares
QSAR	Quantitative Structure–Activity Relationship
QSPR	Quantitative Structure–Property Relationship
RSD	Residual SD
SD	Standard Deviation
SDEP, SEP	Standard error of prediction
SECV	Standard error of cross-validation
SIMCA	Simple Classification Analysis
SPC	Statistical Process Control
SS	Sum of Squares
VIP	Variable Influence on Projection

1.4. Notation

We shall employ the common notation where column vectors are denoted by bold lower case characters, e.g., **v**, and row vectors shown as transposed, e.g., $\mathbf{v}'$. Bold upper case characters denote matrices, e.g., **X**.

$*$	multiplication, e.g., $\mathbf{A}*\mathbf{B}$
$'$	transpose, e.g., $\mathbf{v}',\mathbf{X}'$
a	index of components (model dimensions); $(a = 1,2,\ldots,A)$
A	number of components in a PC or PLS model
i	index of objects (observations, cases); $(i = 1,2,\ldots,N)$
N	number of objects (cases, observations)
k	index of X-variables $(k = 1,2,\ldots,K)$
m	index of Y-variables $(m = 1,2,\ldots,M)$
$\mathbf{X}$	matrix of predictor variables, size $(N * K)$
$\mathbf{Y}$	matrix of response variables, size $(N * M)$
$\mathbf{b}_m$	regression coefficient vector of the mth y. Size $(K * 1)$
$\mathbf{B}$	matrix of regression coefficients of all Y's. Size $(K * M)$
$\mathbf{c}_a$	PLSR Y-weights of component a
$\mathbf{C}$	the $(M * A)$ Y-weight matrix; $\mathbf{c}_a$ are columns in this matrix
$\mathbf{E}$	the $(N * K)$ matrix of X-residuals
$\mathbf{f}_m$	residuals of mth y-variable; $(N * 1)$ vector
$\mathbf{F}$	the $(N * M)$ matrix of Y-residuals
G	the number of CV groups $(g = 1,2,\ldots,G)$
$\mathbf{p}_a$	PLSR X-loading vector of component a
$\mathbf{P}$	Loading matrix; p_a are columns of P
R^2	multiple correlation coefficient; amount Y "explained" in terms of SS
R^2_X	amount X "explained" in terms of SS
Q^2	cross-validated R^2; amount Y "predicted"
$\mathbf{t}_a$	X-scores of component a
$\mathbf{T}$	score matrix $(N * A)$, where the columns are $\mathbf{t}_a$
$\mathbf{u}_a$	Y-scores of component a
$\mathbf{U}$	score matrix $(N * A)$, where the columns are $\mathbf{u}_a$
$\mathbf{w}_a$	PLSR X-weights of component a
$\mathbf{W}$	the $(K * A)$ X-weight matrix; $\mathbf{w}_a$ are columns in this matrix
$\mathbf{w}_a^*$	PLSR weights transformed to be independent between components
$\mathbf{W}^*$	$(K * A)$ matrix of transformed PLSR weights; $\mathbf{w}_a^*$ are columns in $\mathbf{W}^*$.

2. Example 1, a quantitative structure property relationship (QSPR)

We use a simple example from the literature with one Y-variable and seven X-variables. The problem is one of QSPR or QSAR, which differ only in that the response(s) **Y** are chemical properties in the former and biological activities in the latter. In both cases, **X** contains a quantitative description of the variation in chemical structure between the investigated compounds.

The objective is to understand the variation of y = DDGTS = the free energy of unfolding of a protein (tryptophane synthase a unit of bacteriophage T4 lysosome) when position 49 is modified to contain each of the 19 coded amino acids (AA's) except arginine. The AA's are described by seven highly correlated X-variables as shown in Table 1. Computational and other details are given in Ref. [8]. For al-

Table 1
Raw data of example 1

	PIE	PIF	DGR	SAC	MR	Lam	Vol	DDGTS
(1) Ala	0.23	0.31	−0.55	254.2	2.126	−0.02	82.2	8.5
(2) Asn	−0.48	−0.60	0.51	303.6	2.994	−1.24	112.3	8.2
(3) Asp	−0.61	−0.77	1.20	287.9	2.994	−1.08	103.7	8.5
(4) Cys	0.45	1.54	−1.40	282.9	2.933	−0.11	99.1	11.0
(5) Gln	−0.11	−0.22	0.29	335.0	3.458	−1.19	127.5	6.3
(6) Glu	−0.51	−0.64	0.76	311.6	3.243	−1.43	120.5	8.8
(7) Gly	0.00	0.00	0.00	224.9	1.662	0.03	65.0	7.1
(8) His	0.15	0.13	−0.25	337.2	3.856	−1.06	140.6	10.1
(9) Ile	1.20	1.80	−2.10	322.6	3.350	0.04	131.7	16.8
(10) Leu	1.28	1.70	−2.00	324.0	3.518	0.12	131.5	15.0
(11) Lys	−0.77	−0.99	0.78	336.6	2.933	−2.26	144.3	7.9
(12) Met	0.90	1.23	−1.60	336.3	3.860	−0.33	132.3	13.3
(13) Phe	1.56	1.79	−2.60	366.1	4.638	−0.05	155.8	11.2
(14) Pro	0.38	0.49	−1.50	288.5	2.876	−0.31	106.7	8.2
(15) Ser	0.00	−0.04	0.09	266.7	2.279	−0.40	88.5	7.4
(16) Thr	0.17	0.26	−0.58	283.9	2.743	−0.53	105.3	8.8
(17) Trp	1.85	2.25	−2.70	401.8	5.755	−0.31	185.9	9.9
(18) Tyr	0.89	0.96	−1.70	377.8	4.791	−0.84	162.7	8.8
(19) Val	0.71	1.22	−1.60	295.1	3.054	−0.13	115.6	12.0
Correlation matrix								
PIE	1.000	0.967	−0.970	0.518	0.650	0.704	0.533	0.645
PIF	0.967	1.000	−0.968	0.416	0.555	0.750	0.433	0.711
DGR	−0.970	−0.968	1.000	−0.463	−0.582	−0.704	−0.484	−0.648
SAC	0.518	0.416	−0.463	1.000	0.955	−0.230	0.991	0.268
MR	0.650	0.555	−0.582	0.955	1.000	−0.027	0.945	0.290
Lam	0.704	0.750	−0.704	−0.230	−0.027	1.000	−0.221	0.499
Vol	0.533	0.433	−0.484	0.991	0.945	−0.221	1.000	0.300
DDGTS	0.645	0.711	−0.648	0.268	0.290	0.499	0.300	1.000

The lower half of the table shows the pair-wise correlation coefficients of the data. PIE and PIF are the lipophilicity constant of the AA side chain according to El Tayar et al. [8], and Fauchere and Pliska, respectively; DGR is the free energy of transfer of an AA side chain from protein interior to water according to Radzicka and Woldenden; SAC is the water-accessible surface area of AA's calculated by MOLSV; MR molecular refractivity (from Daylight data base); Lam is a polarity parameter according to El Tayar et al. [8]. Vol is the molecular volume of AA's calculated by MOLSV. All the data except MR are from Ref. [8].

ternative ways, see Hellberg et al. [9] and Sandberg et al. [10].

3. PLSR and the underlying scientific model

PLSR is a way to estimate parameters in a scientific model, which basically is linear (see Section 4.3 for non-linear PLS models). This model, like any scientific model, consists of several parts, the philosophical, conceptual, the technical, the numerical, the statistical, and so on. We here illustrate these using the QSPR/QSAR model of example 1 (see above), but the arguments are similar in most other modelling in science and technology.

Our chemical thinking makes us formulate the influence of structural change on activity (and other properties) in terms of "*effects*"—lipophilic, steric, polar, polarizability, hydrogen bonding, and possibly others. Similarly, the modelling of a chemical process is interpreted using "effects" of thermodynamics (equilibria, heat transfer, mass transfer, pressures, temperatures, and flows) and chemical kinetics (reaction rates).

Although this formulation of the scientific model is not of immediate concern for the technicalities of PLSR, it is of interest that PLSR modelling is consis-

tent with "effects" causing the changes in the investigated system. The concept of *latent variables* (Section 4.1) may be seen as directly corresponding to these effects.

3.1. The data—*X* and *Y*

The PLSR model is developed from a training set of N observations (objects, cases, compounds, process time points) with K X-variables denoted by $\mathbf{x}_k$ $(k = 1, \ldots, K)$, and M Y-variables $\mathbf{y}_m$ $(m = 1,2, \ldots, M)$. These training data form the two matrices **X** and **Y** of dimensions $(N * K)$ and $(N * M)$, respectively, as shown in Fig. 1. In example 1, $N = 19$, $K = 7$, and $M = 1$.

Later, predictions for new observations are made based on their X-data. This gives predicted X-scores (t-values), X-residuals, their residual SD's, and y-values with confidence intervals.

3.2. Transformation, scaling and centering

Before the analysis, the X- and Y-variables are often transformed to make their distributions be fairly symmetrical. Thus, variables with range of more than one magnitude of 10 are often logarithmically transformed. If the value zero occurs in a variable, the fourth root transformation is a good alternative to the logarithm. The response variable in example 1 is already logarithmically transformed, i.e., expressed in thermodynamic units. No further transformation was made of the example 1 data.

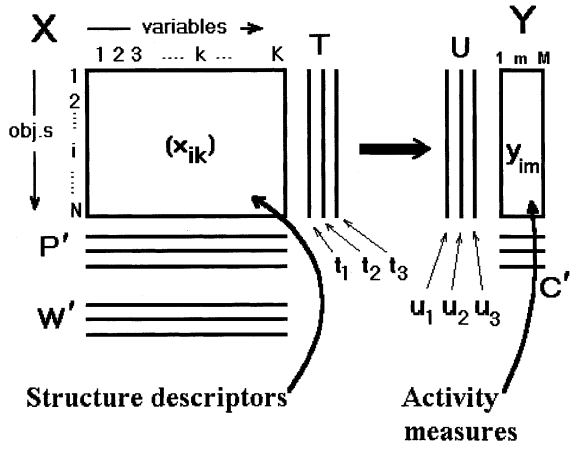

Fig. 1. Data of a PLSR can be arranged as two tables, matrices, **X** and **Y**. Note that the raw data may have been transformed (e.g., logarithmically), and usually have been centered and scaled before the analysis. In QSAR, the X-variables are descriptors of chemical structure, and the Y-variables measurements of biological activity.

Results of projection methods such as PLSR depend on the *scaling* of the data. With an appropriate scaling, one can focus the model on more important Y-variables, and use experience to increase the weights of more informative X-variables. In the absence of knowledge about the relative importance of the variables, the standard multivariate approach is to (i) scale each variable to unit variance by dividing them by their SD's, and (ii) center them by subtracting their averages, so-called *auto-scaling*. This corresponds to giving each variable (column) the same weight, the same prior importance in the analysis.

In example 1, all variables were auto-scaled, except x_6 = Lam, the auto-scaling weight of which was multiplied by 1.5 to make it not be masked by the others (so-called block-scaling).

In CoMFA and GRID-QSAR, as well as in multivariate calibration in analytical chemistry, auto-scaling is often not the best scaling of **X**, but non-scaled X-data or some intermediate between auto-scaled and non-scaled may be appropriate [5]. In process data, the acceptable interval of variation of each variable can form the basis for the scaling.

For ease of interpretation and for numerical stability, it is recommended to *center* the data before the analysis. This is done—either before or after scaling—by subtracting the averages from all variables both in **X** and **Y**. In process data other reference values such as set point values may be subtracted instead of averages. Hence, the analysis concerns the deviations from the reference points and how they are related. The centering leaves the interpretation of the results unchanged.

In some applications it is customary to normalize also the *observations* (objects). In chemistry, this is often done in the analysis of chromatographic or spectral profiles. The normalization is typically done by making the sum of all peaks of one profile be 100 or 1000. This removes the *size* of the observations (objects), which may be desirable if size is irrelevant. This is closely related to correspondence analysis [11].

3.3. The PLSR model

The linear PLSR model finds a few "new" variables, which are estimates of the LV's or their rotations. These new variables are called X-scores and denoted by $\mathbf{t}_a$ ($a = 1,2,\ldots,A$). The X-scores are predictors of **Y** and also model **X** (Eqs. (4) and (2) below), i.e., both **Y** and **X** are assumed to be, at least partly, modelled by the same LV's.

The X-scores are "few" (A in number), and orthogonal. They are estimated as linear combinations of the original variables $\mathbf{x}_k$ with the coefficients, "weights", w^*_{ka} ($a = 1,2,\ldots,A$). These weights are sometimes denoted by r_{ka} [12,13]. Below, formulas are shown both in element and matrix form (the latter in parentheses):

$$t_{ia} = \sum_k W^*_{ka} X_{ik}; \qquad (\mathbf{T} = \mathbf{X}\mathbf{W}^*). \tag{1}$$

The X-scores ($\mathbf{t}_a$'s) have the following properties: (a) They are, multiplied by the loadings p_{ak}, good "summaries" of **X**, so that the X-residuals, e_{ik}, in Eq. (2) are "small":

$$X_{ik} = \sum_a t_{ia} p_{ak} + e_{ik}; \qquad (\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}). \tag{2}$$

With multivariate **Y** (when $M > 1$), the corresponding "Y-scores" ($\mathbf{u}_a$) are, multiplied by the weights c_{am}, good "summaries" of **Y**, so that the residuals, g_{im}, in Eq. (3) are "small":

$$y_{im} = \sum_a u_{ia} c_{am} + g_{im} \qquad (\mathbf{Y} = \mathbf{U}\mathbf{C}' + \mathbf{G}), \tag{3}$$

(b) the X-scores are good predictors of **Y**, i.e.:

$$y_{im} = \sum_a c_{ma} t_{ia} + f_{im} \qquad (\mathbf{Y} = \mathbf{T}\mathbf{C}' + \mathbf{F}). \tag{4}$$

The Y-residuals, f_{im} express the deviations between the observed and modelled responses, and comprise the elements of the Y-residual matrix, **F**.

Because of Eq. (1), Eq. (4) can be rewritten to look as a multiple regression model:

$$y_{im} \sum_a c_{ma} \sum_k w^*_{ka} x_{ik} + f_{im} = \sum_k b_{mk} x_{ik} + f_{im}$$

$$(\mathbf{Y} = \mathbf{X}\mathbf{W}^*\mathbf{C}' + \mathbf{F} = \mathbf{X}\mathbf{B} + \mathbf{F}). \tag{5}$$

The "PLS-regression coefficients", b_{mk} (**B**), can be written as:

$$b_{mk} = \sum_a c_{ma} w^*_{ka} \qquad (\mathbf{B} = \mathbf{W}^*\mathbf{C}'). \tag{6}$$

Note that these b's are *not independent* unless A (the number of PLSR components) equals K (the number of X-variables). Hence, their confidence intervals according to the traditional statistical interpretation are infinite.

An interesting special case is at hand when there is a single y-variable ($M = 1$) and $\mathbf{X}'\mathbf{X}$ is diagonal, i.e., **X** originates from an orthogonal design (fractional factorial, Plackett–Burman, etc.). In this case there is no correlation structure in **X**, and PLSR arrives at the MLR solution in one component [14], and the MLR and PLS-regression coefficients are equal to $\mathbf{w}_1\mathbf{c}'_1$.

After each component, a, the **X**-matrix is "deflated" by subtracting $t^*_{ia} p_{ka}$ from x_{ik} ($\mathbf{t}_a\ \mathbf{p}'_a$ from **X**). This makes the PLSR 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. Thus, instead of Eq. (1), we can write:

$$t_{ia} = \sum_k w_{ka} e_{ik,a-1} \qquad (\mathbf{t}_a = \mathbf{E}_{a-1}\mathbf{W}_a) \tag{7a}$$

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

$$(\mathbf{E}_{a-1} = \mathbf{E}_{a-2} - \mathbf{t}_{a-1}\mathbf{p}'_{a-1}) \tag{7b}$$

$$e_{ik,0} = X_{ik} \qquad (\mathbf{E}_0 = \mathbf{X}). \tag{7c}$$

However, the weights, **w**, can be transformed to $\mathbf{w}^*$, which directly relate to **X**, giving Eq. (1) above. The relation between the two is given as [14]:

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

The **Y**-matrix can also be "deflated" by subtracting $\mathbf{t}_a\ \mathbf{c}'_a$, but this is not necessary; the results are equivalent with or without Y-deflation.

From the PLSR algorithm (see below), one can see that the first weight vector ($\mathbf{w}_1$) is the first eigenvector of the combined variance–covariance matrix, $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$, and the following weight vectors (component a) are eigenvectors to the deflated versions of the same matrix, i.e., $\mathbf{Z}'_a\mathbf{Y}\mathbf{Y}'\mathbf{Z}'_a$, where $\mathbf{Z}_a = \mathbf{Z}_{a-1} - \mathbf{T}_{a-1}\ \mathbf{P}'_{a-1}$. Similarly, the first score vector ($\mathbf{t}_1$) is an eigenvector to $\mathbf{X}\mathbf{X}'\mathbf{Y}\mathbf{Y}'$, and later X-score vectors ($\mathbf{t}_a$) are eigenvectors of $\mathbf{Z}_a\mathbf{Z}'_a\mathbf{Y}\mathbf{Y}'$.

These eigenvector relationships also show that the vectors $\mathbf{w}_a$ form an ortho-normal set, and that the vectors $\mathbf{t}_a$ are orthogonal to each other. The loading vectors ($\mathbf{p}_a$) are not orthogonal to each other, and neither are the Y-scores, $\mathbf{u}_a$. The **u**'s and the **p**'s are orthogonal to the **t**'s and the **w**'s, respectively, one and more components earlier, i.e., $\mathbf{u}'_b\mathbf{t}_a = 0$ and $\mathbf{p}'_b\mathbf{w}_a = 0$, if $b > a$. Also, $\mathbf{w}'_a\mathbf{p}_a = 1.0$.

3.4. Interpretation of the PLSR model

One way to see PLSR is that it forms "new x-variables" (LV estimates), $\mathbf{t}_a$, as linear combinations of the old x's, and thereafter uses these new t's as predictors of **Y**. Hence, PLSR is based on a linear model (see, however, Section 4.3). Only as many new t's are formed as are needed, as are predictively significant (Section 3.8).

All parameters, **t**, **u**, **w** (and $\mathbf{w}^*$), **p**, and **c** (see Fig. 1), are determined by a PLSR algorithm as described below. For the *interpretation* of the PLSR model, the scores, **t** and **u**, contain the information about the objects and their similarities/dissimilarities with respect to the given problem and model.

The weights $\mathbf{w}_a$ or the closely similar $\mathbf{w}_a^*$ (see below), and $\mathbf{c}_a$, give information about how the variables combine to form the quantitative relation between **X** and **Y**, thus providing an interpretation of the scores, $\mathbf{t}_a$ and $\mathbf{u}_a$. Hence, these weights are essential for the understanding of which X-variables are important (numerically large w_a-values), and which X-variables that provide the same information (similar profiles of w_a-values).

The PLS weights $\mathbf{w}_a$ express both the "positive" correlations between **X** and **Y**, and the "compensation correlations" needed to predict **Y** from **X** clear from the secondary variation in **X**. The latter is everything varying in **X** that is *not* primarily related to **Y**. This makes $\mathbf{w}_a$ difficult to interpret directly, especially in later components ($a > 1$). By using an orthogonal expansion of the X-parameters in O-PLS, one can get the part of $\mathbf{w}_a$ that primarily relates to **Y**, thus making the PLS interpretation more clear [15].

The part of the data that are not explained by the model, *the residuals*, are of diagnostic interest. Large Y-residuals indicate that the model is poor, and a normal probability plot of the residuals of a single Y-variable is useful for identifying outliers in the relationship between **T** and **Y**, analogously to MLR. In PLSR we also have residuals for **X**; the part not used in the modelling of **Y**. These X-residuals are useful for identifying outliers in the X-space, i.e., molecules with structures that do not fit the model, and process points deviating from "normal" process operations. This, together with control charts of the X-scores, $\mathbf{t}_a$, is used in multivariate SPC (MSPC) [16].

3.5. Geometric interpretation

PLSR is a projection method and thus has a simple geometric interpretation as a projection of the X-matrix (a swarm of N points in a K-dimensional space) down on an A-dimensional hyper-plane in such a way that the coordinates of the projection ($\mathbf{t}_a$, $a = 1,2,\ldots,A$) are good predictors of **Y**. This is indicated in Fig. 2.

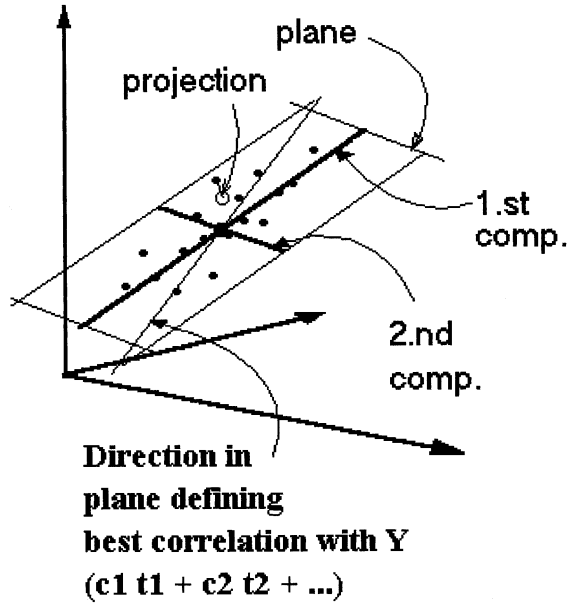

Fig. 2. The geometric representation of PLSR. The **X**-matrix can be represented as N points in the K dimensional space where each column of **X** ($\mathbf{x}_k$) defines one coordinate axis. The PLSR model defines an A-dimensional hyper-plane, which in turn, is defined by one line, one direction, per component. The direction coefficients of these lines are p_{ak}. The coordinates of each object, i, when its data (row i in **X**) are projected down on this plane are t_{ia}. These positions are related to the values of **Y**.

The direction of the plane is expressed as slopes, p_{ak}, of each PLS direction of the plane (each component) with respect to each coordinate axis, $\mathbf{x}_k$. This slope is the cosine of the angle between the PLS direction and the coordinate axis.

Thus, PLSR develops an A-dimensional hyperplane in X-space such that this plane well approximates **X** (the N points, row vectors of **X**), and at the same time, the positions of the projected data points on this plane, described by the scores t_{ia}, are related to the values of the responses, activities, Y_{im} (see Fig. 2).

3.6. Incomplete *X* and *Y* matrices (missing data)

Projection methods such as PLSR tolerate moderate amounts of missing data both in **X** and in **Y**. To have missing data in **Y**, it must be multivariate, i.e., have at least two columns. The larger the matrices **X** and **Y** are, the higher proportion of missing data is tolerated. For small data sets with around 20 observations and 20 variables, around 10% to 20% missing data can be handled, provided that they are not missing according to some systematic pattern.

The NIPALS PLSR algorithm automatically accounts for the missing values, in principle by iteratively substituting the missing values by predictions by the model. This corresponds to, for each component, giving the missing data values that have zero residuals and thus have no influence on the component parameters $\mathbf{t}_a$ and $\mathbf{p}_a$. Other approaches based on the EM algorithm have been developed, and often work better than NIPALS for large percentages of missing data [17,18]. One should remember, however, that with much missing data, any resulting parameters and predictions are highly uncertain.

3.7. One Y at a time, or all in a single model?

PLSR has the ability to model and analyze several Y's together, which has the advantage to give a simpler overall picture than one separate model for each Y-variable. Hence, when the Y's are correlated, they should be analyzed together. If the Y's really measure different things, and thus are fairly independent, a single PLSR model tends to have many components and be difficult to interpret. Then a separate modelling of the Y's gives a set of simpler models with fewer dimensions, which are easier to interpret.

Hence, one should start with a PCA of just the **Y**-matrix. This shows the practical rank of **Y**, i.e., the number of resulting components, A_{PCA}. If this is small compared to the number of Y-variables (M), the Y's are correlated, and a single PLSR model of all Y's is warranted. If, however, the Y's cluster in strong groups, which is seen in the PCA loading plots, separate PLSR models should be developed for these groups.

3.8. The number of PLS components, A

In any empirical modelling, it is essential to determine the correct complexity of the model. With numerous and correlated X-variables there is a substantial risk for "over-fitting", i.e., getting a well fitting model with little or no predictive power. Hence, a strict test of the predictive significance of each PLS component is necessary, and then stopping when components start to be non-significant.

Cross-validation (CV) is a practical and reliable way to test this predictive significance [1–6]. This has become the standard in PLSR analysis, and incorporated in one form or another in all available PLSR software. Good discussions of the subject were given by Wakeling and Morris [19] and Clark and Cramer [20].

Basically, CV is performed by dividing the data in a number of groups, G, say, five to nine, and then developing a number of parallel models from reduced data with one of the groups deleted. We note that having $G = N$, i.e., the leave-one-out approach, is not recommendable [21].

After developing a model, differences between actual and predicted Y-values are calculated for the deleted data. The sum of squares of these differences is computed and collected from all the parallel models to form the predictive residual sum of squares (PRESS), which estimates the predictive ability of the model.

When CV is used in the sequential mode, CV is performed on one component after the other, but the peeling off (Eq. (7b), Section 3.3) is made only once on the full data matrices, where after the resulting residual matrices **E** and **F** are divided into groups for the CV of next component. The ratio $\mathrm{PRESS}_a/\mathrm{SS}_{a-1}$ is calculated after each component, and a component is judged significant if this ratio is smaller than around 0.9 for at least one of the Y-variables. Slightly

sharper bonds can be obtained from the results of Wakeling and Morris [19]. Here SS_{a-1} denotes the (fitted) residual sum of squares *before* the current component (index a). The calculations continue until a component is non-significant.

Alternatively with "total CV", one first divides the data into groups, and then calculates PRESS for each component up to, say 10 or 15 with separate "peeling" (Eq. (7b), Section 3.3) of the data matrices of each CV group. The model with the number of components giving the lowest $\text{PRESS}/(N-A-1)$ is then used. This "total" approach is computationally more taxing, but gives similar results.

Both with the "sequential" and the "total" mode, a PRESS is calculated for the final model with the estimated number of significant components. This is often re-expressed as Q^2 (the cross-validated R^2) which is $(1-\text{PRESS}/\text{SS})$ where SS is the sum of squares of **Y** corrected for the mean. This can be compared with $R^2=(1-\text{RSS}/\text{SS})$, where RSS is the *fitted* residual sum of squares. In models with several Y's, one obtains also R^2_{m} and Q^2_{m} for each Y-variable, y_m.

These measures can, of course, be equivalently expressed as residual SD's (RSD's) and predictive residual SD's (PRESD's). The latter is often called standard error of prediction (SDEP or SEP), or standard error of cross-validation (SECV). If one has some knowledge of the noise in the investigated system, for example ± 0.3 units for $\log(1/C)$ in QSAR's, these predictive SD's should, of course, be similar in size to the noise.

3.9. Model validation

Any model needs to be validated before it is used for "understanding" or for predicting new events such as the biological activity of new compounds or the yield and impurities at other process conditions. The best validation of a model is that it consistently precisely predicts the Y-values of observations with new X-values—a *validation set*. But an independent and representative validation set is rare.

In the absence of a real validation set, two reasonable ways of model validation are given by cross-validation (CV, see Section 3.8) which simulates how well the model predicts new data, and model re-estimation after data randomization which estimates the chance (probability) to get a good fit with random response data.

3.10. PLSR algorithms

The algorithms for calculating the PLSR model are mainly of technical interest, we here just point out that there are several variants developed for different shapes of the data [2,22,23]. Most of these algorithms tolerate moderate amounts of missing data. Either the algorithm, like the original NIPALS algorithm below, works with the original data matrices, **X** and **Y** (scaled and centered). Alternatively, so-called kernel algorithms work with the variance–covariance matrices, $\mathbf{X}'\mathbf{X}$, $\mathbf{Y}'\mathbf{Y}$, and $\mathbf{X}'\mathbf{Y}$, or association matrices, $\mathbf{X}\mathbf{X}'$ and $\mathbf{Y}\mathbf{Y}'$, which is advantageous when the number of observations (N) differs much from the number of variables (K and M).

For extensions of PLSR, the results of Höskuldsson [3] regarding the possibilities to modify the NIPALS PLSR algorithm are of great interest. Höskuldsson shows that as long as the steps (C) to (G) below are unchanged, modifications can be made of **w** in step (B). Central properties remain, such as orthogonality between model components, good summarizing properties of the X-scores, $\mathbf{t}_a$, and interpretability of the model parameters. This can be used to introduce smoothness in the PLSR solution [24], to develop a PLSR model where a majority of the PLSR coefficients are zero [25], align **w** with à priori specified vectors (similar to "target rotation" of Kvalheim et al. [26]), and more.

The simple NIPALS algorithm of Wold et al. [2] is shown below. It starts with optionally transformed, scaled, and centered data, **X** and **Y**, and proceeds as follows (note that with a single y-variable, the algorithm is non-iterative).

(A) Get a starting vector of **u**, usually one of the **Y** columns. With a single y, $\mathbf{u}=\mathbf{y}$.

(B) The X-weights, **w**:

$$\mathbf{w}=\mathbf{X}'\mathbf{u}/\mathbf{u}'\mathbf{u},$$

(here **w** can now be modified) norm **w** to $\|\mathbf{w}\|=1.0$

(C) Calculate X-scores, **t**:

$$\mathbf{t}=\mathbf{X}\mathbf{w}.$$

(D) The Y-weights, **c**:

$$\mathbf{c}=\mathbf{Y}'\mathbf{t}/\mathbf{t}'\mathbf{t}.$$

(E) Finally, an updated set of Y-scores, **u**:

$$\mathbf{u} = \mathbf{Yc}/\mathbf{c}'\mathbf{c}.$$

(F) Convergence is tested on the change in **t**, i.e., $\|\mathbf{t}_{\text{old}} - \mathbf{t}_{\text{new}}\|/\|\mathbf{t}_{\text{new}}\| < \varepsilon$, where ε is "small", e.g., 10^{-6} or 10^{-8}. If convergence has *not* been reached, return to (B), otherwise continue with (G), and then (A). If there is only one y-variable, i.e., $M = 1$, the procedure converges in a single iteration, and one proceeds directly with (G).

(G) Remove (deflate, peel off) the present component from **X** and **Y** use these deflated matrices as **X** and **Y** in the next component. Here the deflation of **Y** is optional; the results are equivalent whether **Y** is deflated or not.

$$\mathbf{p} = \mathbf{X}'\mathbf{t}/(\mathbf{t}'\mathbf{t})$$

$$\mathbf{X} = \mathbf{X} - \mathbf{tp}'$$

$$\mathbf{Y} = \mathbf{Y} - \mathbf{tc}'.$$

(H) Continue with next component (back to step A) until cross-validation (see above) indicates that there is no more significant information in **X** about **Y**.

Chu et al. [27] recently has reviewed the attractive properties of matrix decompositions of the Wedderburn type. The PLSR NIPALS algorithm is such a Wedderburn decomposition, and hence is numerically and statistically stable.

3.11. Standard errors and confidence intervals

Numerous efforts have been made to theoretically derive confidence intervals of the PLSR parameters, see, e.g., Ref. [28]. Most of these are, however, based on regression assumptions, seeing PLSR as a biased regression model. Only recently, in the work of Burnham et al. [12], have these matters been investigated with PLSR as a *latent variable* regression model.

A way to estimate standard errors and confidence intervals directly from the data is to use jack-knifing [29]. This was recommended by Wold [1] in his original PLS work, and has recently been revived by Martens and Martens [30] and others. The idea is simple; the variation in the parameters of the various sub-models obtained during cross-validation is used to derive their standard deviations (called standard errors), followed by using the t-distribution to give confidence intervals. Since all PLSR parameters (scores, loadings, etc.) are linear combinations of the original data (possibly deflated), these parameters are close to normally distributed, and hence jack-knifing works well.

4. Assumptions underlying PLSR

4.1. Latent Variables

In PLS modelling, we assume that the investigated system or process actually is influenced by just a few underlying variables, *latent variables (LV's)*. The number of these LV's is usually not known, and one aim with the PLSR analysis is to estimate this number. Also, the PLS X-scores, $\mathbf{t}_a$, are usually not direct estimates of the LV's, but rather they span the same space as the LV's. Thus, the latter (denoted by **V**) are related to the former (**T**) by a, usually unknown, rotation matrix, **R**, with the property $\mathbf{R}'\mathbf{R} = 1$:

$$\mathbf{V} = \mathbf{TR}' \quad \text{or} \quad \mathbf{T} = \mathbf{RV}.$$

Both the X- and the Y-variables are assumed to be realizations of these underlying LV's, and are hence *not assumed to be independent*. Interestingly, the LV assumptions closely correspond to the use of microscopic concepts such as molecules and reactions in chemistry and molecular biology, thus making PLSR philosophically suitable for the modelling of chemical and biological data. This has been discussed by, among others, Wold et al. [31,32], Kvalheim [33], and recently from a more fundamental perspective, by Burnham et al. [12,13]. In spectroscopy, it is clear that the spectrum of a sample is the sum of the spectra of the constituents multiplied by their concentrations in the sample. Identifying the latter with **t** (Lambert–Beers "law"), and the spectra with **p**, we get the latent variable model $\mathbf{X} = \mathbf{t}1\mathbf{p}1' + \mathbf{t}2\mathbf{p}2' + \ldots = \mathbf{TP}'$ + noise. In many applications this interpretation with the data explained by a number of "factors" (components) makes sense.

As discussed below, we can also see the scores, **T**, as comprised of derivatives of an unknown function underlying the investigated system. The choice of the interpretation depends on the amount of knowledge

about the system. The more knowledge we have, the more likely it is that we can assign a latent variable interpretation of the X-scores or their rotation.

If the number of LV's actually equals the number of X-variables, *K*, then the X-variables are independent, and PLSR and MLR give identical results. Hence, we can see PLSR as a generalization of MLR, containing the latter as a special case in situations when the MLR solution exists, i.e., when the number of X- and Y-variables is fairly small in comparison to the number of observations, N. In most practical cases, except when **X** is generated according to an experimental design, however, the X-variables are not independent. We then call **X** *rank deficient*. Then PLSR gives a "shrunk" solution which is statistically more robust than the MLR solution, and hence gives better predictions than MLR [34].

PLSR gives a model of **X** in terms of a bilinear projection, plus residuals. Hence, PLSR assumes that there may be parts of **X** that are unrelated to **Y**. These parts can include noise and/or regularities non-related to **Y**. Hence, unlike MLR, PLSR tolerates noise in **X**.

4.2. Alternative derivation

The second theoretical foundation of LV-models is one of Taylor expansions [35]. We assume the data **X** and **Y** to be generated by a multi-dimensional function $F(\mathbf{u},\mathbf{v})$, where the vector variable **u** describes the change between observations (rows in **X**) and the vector variable **v** describes the change between variables (columns in **X**). Making a Taylor expansion of the function F in the **u**-direction, and discretizing for i = observation and k = variable, gives the LV-model. Again, the smaller the interval of **u** that is modelled, the fewer terms we need in the Taylor expansion, and the fewer components we need in the LV-model. Hence, we can interpret PCA and PLS as models of similarity. Data (variables) measured on a set of similar observations (samples, items, cases, ...) can always be modelled (approximated) by a PC- or PLS model. And the more similar are the observations, the fewer components we need in the model.

We hence have two different interpretations of the LV-model. Thus, real data well explained by these models can be interpreted as either being a linear combination of "factors" or according to the latter interpretation as being measurements made on a set of similar observations. Any mixture of these two interpretations is, of course, often applicable.

4.3. Homogeneity

Any data analysis is based on an assumption of *homogeneity*. This means that the investigated system or process must be in a similar state throughout all the investigation, and the mechanism of influence of **X** on **Y** must be the same. This, in turn, corresponds to having some limits on the variability and diversity of **X** and **Y**.

Hence, it is essential that the analysis provides *diagnostics* about how well these assumptions indeed are fulfilled. Much of the recent progress in applied statistics has concerned diagnostics [36], and many of these diagnostics can be used also in PLSR modelling as discussed below. PLSR also provides additional diagnostics beyond those of regression-like methods, particularly those based on the modelling of **X** (score and loading plots and X-residuals).

In the first example, the first PLSR analysis indicated that the data set was inhomogeneous—three aromatic amino acids (AA's) are indicated to have a different *type* of effect than the others on the modelled property. This type of information is difficult to obtain in ordinary regression modelling, where only large residuals in Y provide diagnostics about inhomogeneities.

4.4. Non-linear PLSR

For non-linear situations, simple solutions have been published by Höskuldsson [4], and Berglund and Wold [37]. Another approach based on transforming selected X-variables or X-scores to qualitative variables coded as sets of dummy variables, the so-called GIFI approach [38,39], is described elsewhere in this volume [15].

5. Results of example 1

5.1. The initial PLSR analysis of the AA data

The first PLSR analysis (linear model) of the AA data gives one significant component explaining 43% of the Y-variance ($R^2 = 0.435$, $Q^2 = 0.299$). In contrast, the MLR gives an R^2 of 0.788, which is equiv-

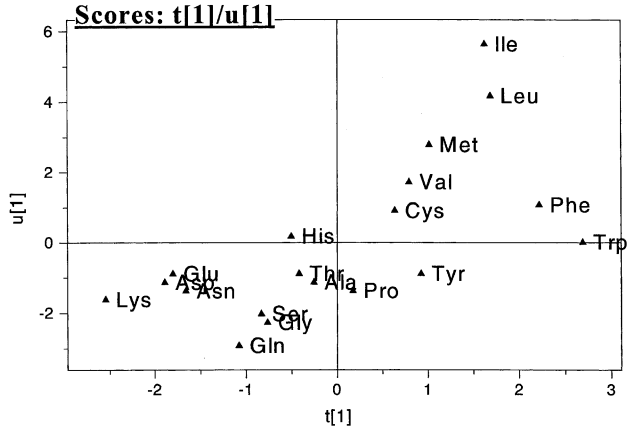

Fig. 3. The PLS scores $\mathbf{u}_1$ and $\mathbf{t}_1$ of the AA example, 1st analysis.

alent to PLSR with $A = 7$ components. The full MLR solution, however, has a Q^2 of -0.215, indicating that the model is poor, and does not predict better than chance.

With just one significant PLS component, the only meaningful score plot is that of **y** against **t** (Fig. 3). The aromatic AA's, Trp, Phe, and, may be, Tyr, show a much worse fit than the others, indicating an inhomogeneity in the data. To investigate this, a second round of analysis is made with a reduced data set, $N = 16$, without the aromatic AA's.

5.2. Second analysis

The modelling of $N = 16$ AA's with the same linear model as before gives a substantially better result with $A = 2$ significant components and $R^2 = 0.783$, $Q^2 = 0.706$. The MLR model for these 16 objects gives a R^2 of 0.872, and a Q^2 of 0.608. This strong improvement indicates that indeed the data set now is more homogeneous and can be properly modelled.

The plot in Fig. 4 of $\mathbf{u}_1$ (**y**) vs. $\mathbf{t}_1$ shows, however, a fairly strong curvature, indicating that squared terms in lipophilicity and, may be, polarity are warranted. In the final analysis, the squares of these four variables were included in the model, which indeed gave better results. Two significant PLS components and one additional with border-line significance were obtained. The resulting R^2 and Q^2 values are for $A =$ 2: 0.90 and 0.80, and for $A = 3$: 0.925 and 0.82, respectively. The $A = 3$ values correspond to RSD = 0.92, and PRESD (SDEP) = 1.23, since the SD of Y

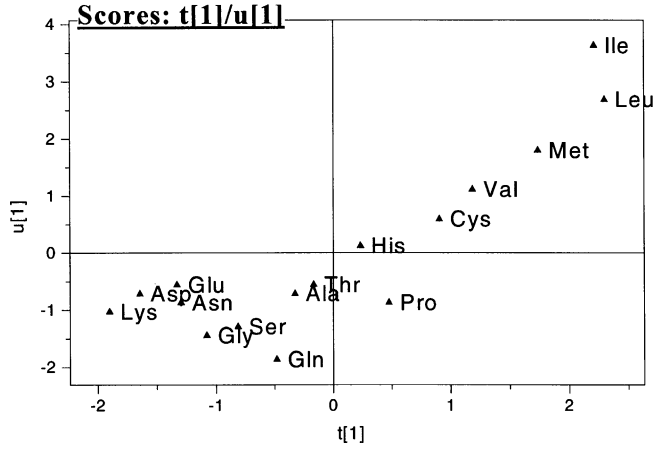

Fig. 4. The PLS scores, $\mathbf{u}_1$ and $\mathbf{t}_1$ of the AA example, 2nd analysis.

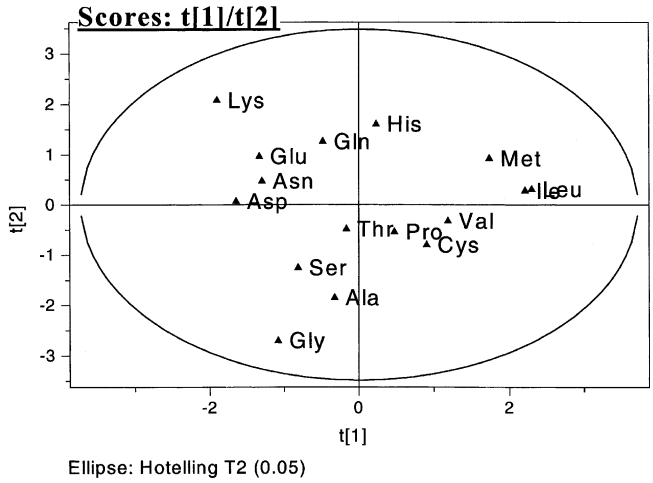

Fig. 5. The PLS scores $\mathbf{t}_1$ and $\mathbf{t}_2$ of the AA example, 2nd analysis. The overlapping points up to the right are Ile and Leu.

is 2.989. The full MLR model gives $R^2 = 0.967$, but with much worse $Q^2 = 0.09$.

5.2.1. X-scores ($\mathbf{t}_a$) show object similarities and dissimilarities

The plot of the X-scores ($\mathbf{t}_1$ vs. $\mathbf{t}_2$, Fig. 5) shows the 16 amino acids grouped according to polarity from upper left to lower right, and inside each group, according to size and lipophilicity.

*5.2.2. PLSR weights **w** and **c***

For the interpretation of PLSR models, the standard is to plot the $\mathbf{w}^*$'s and $\mathbf{c}$'s of one model dimension against another. Alternatively, one can plot the w's and c's; the results and interpretation is similar. This plot shows how the X-variables combine to form the scores $\mathbf{t}_a$; X-variables important for the ath component fall far from the origin along the ath axis in the **wc**-plot. Analogously, Y-variables well modelled by the ath component have large coefficients, c_{am}, and hence fall far from the origin along the ath axis in the same plot.

The example 1 weight plot (Fig. 6, often also called loading plot) shows the first PLS component dominated by lipophilicity and polarity (PIF, PIE, and Lam on the positive side, and DGR on the negative), and the second component being a mixture of size and polarity with MR, Vol, and SAC on the positive (high) side, and Lam on the negative (low) side.

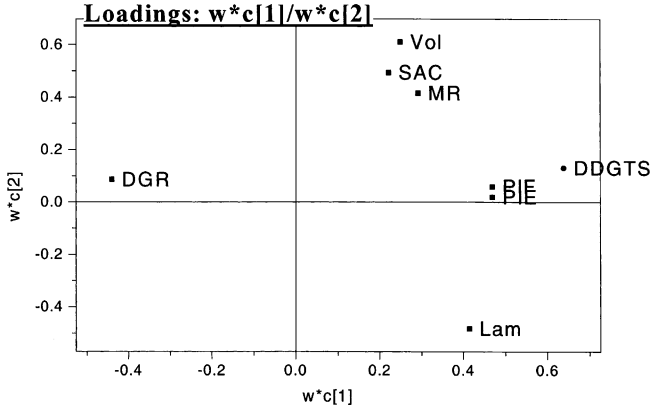

Fig. 6. The PLS weights, $\mathbf{w}^*$ and $\mathbf{c}$ for the first two dimensions of the 2nd AA model.

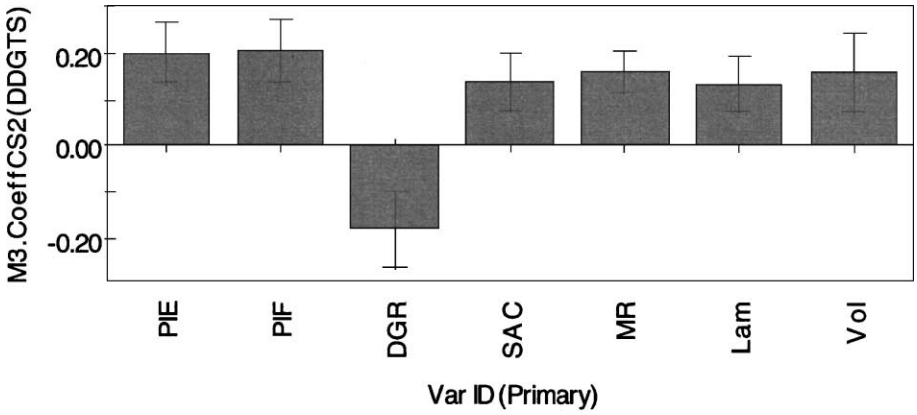

Fig. 7. PLS regression coefficients after $A = 2$ components (second analysis). The bars indicate 95% confidence intervals based on jack-knifing.

The *c*-values of the response, **y**, are proportional to the linear variation of Y explained by the corresponding dimension, i.e., R^2. They define one point per response; in the example with a single response, this point (DDGTS) sits far to the right in the first quadrant of the plot.

The importance of a given X-variable for **Y** is proportional to its distance from the origin in the loading space. These lengths correspond to the PLS-regression coefficients after $A = 2$ dimensions (Fig. 7).

5.3. Comparison with multiple linear regression (MLR)

Comparing the PLS model ($A = 2$) and the MLR model (equivalent to the PLS model with $A = 7$) shows that the coefficients of x_3 (DGR) and x_4 (SAC) change sign between the PLSR model (Fig. 7) and the MLR model (Fig. 8). Moreover, the coefficients of x_4, x_5, and x_7 which are moderate in the PLSR ($A = 2$) model become large and with opposite signs in the MLR model ($A = 7$), although they are strongly positively correlated to each other in the raw data.

The MLR coefficients clearly are misleading and un-interpretable, due to the strong correlations between the X-variables. PLSR, however, stops at $A = 2$, and gives reasonable coefficient values both for $A = 2$ and $A = 3$. With correlated variables one cannot assign "correct" values to the individual coefficients, the only thing we can estimate is their *joint* contribution to **y**.

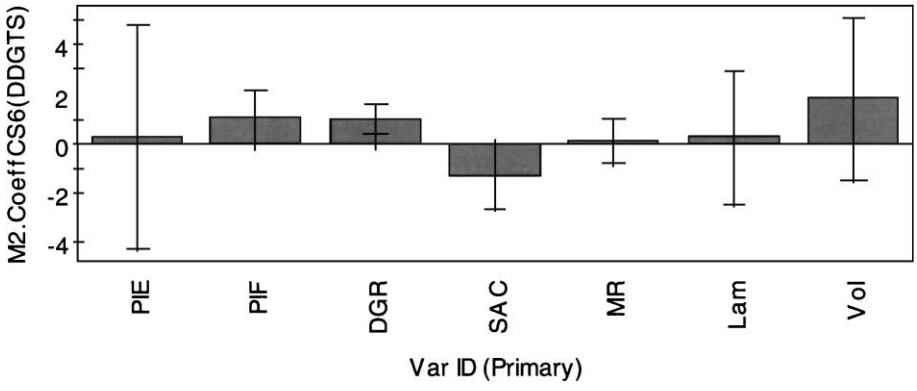

Fig. 8. The MLR coefficients corresponding to analysis 2. The bars indicate 95% confidence intervals based on jack-knifing. Note the differences in the vertical scale to Fig. 7.

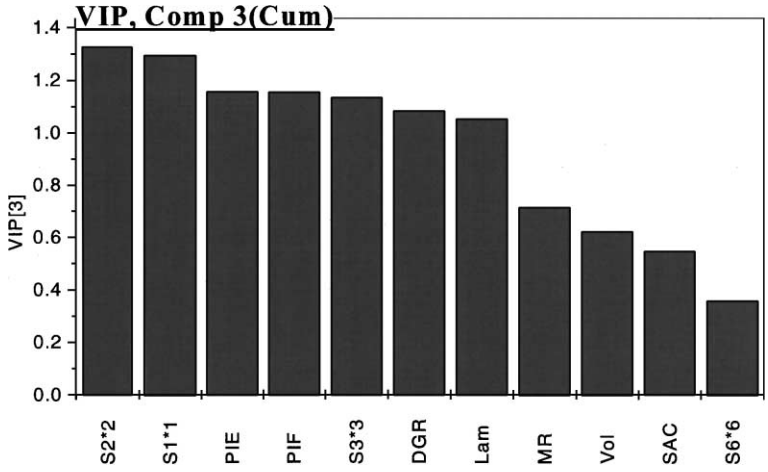

Fig. 9. VIP of the X-variables of the three-component PLSR model, 3rd analysis. The squares of x_1 = PIE, x_2 = PIF, x_3 = DGR, and x_6 = Lam, are denoted by S1∗1, S2∗2, S3∗3, and S6∗6, respectively.

5.4. Measures of variable importance

In PLSR modelling, a variable ($\mathbf{x}_k$) may be important for the modelling of **Y**. Such variables are identified by large PLS-regression coefficients, b_{mk}. However, a variable may also be important for the modelling of **X**, which is identified by large loadings, p_{ak}. A summary of the importance of an X-variable for *both* **Y** and **X** is given by VIP_k (variable importance for the projection, Fig. 9). This is a weighted sum of squares of the PLS-weights, w^*_{ak}, with the weights calculated from the amount of Y-variance of each PLS component, *a*.

5.5. Residuals

The residuals of **Y** and **X** (**E** and **F** in Eqs. (2) and (4) above) are of diagnostic value for the quality of the model. A normal probability plot of the Y-residuals (Fig. 10) of the final AA model shows a fairly straight line with all values within ±3 SD's. To be a serious outlier, a point should clearly deviate from the line, and be outside ±4 SD's.

Since the X-residuals are many ($N * K$), one needs a summary for each observation (compound) not to drown in details. This is provided by the residual SD of the X-residuals of the corresponding row

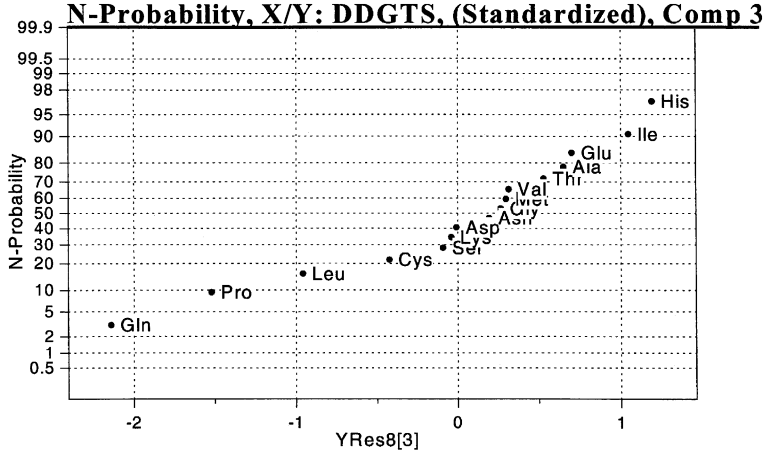

Fig. 10. Y-residuals of the three-component PLSR model, 3rd analysis, normal probability plot.

of the residual matrix **E**. Because this SD is proportional to the distance between the data point and the model plane in X-space, it is also often called DModX (distance to the model in X-space). A DModX larger than around 2.5 times the overall SD of the X-residuals (corresponding to an *F*-value of 6.25) indicates that the observation is an outlier. Fig. 11 shows that none of the 16 compounds in example 1 has a large DModX. Here the overall SD = 0.34.

5.6. Conclusions of example 1

The initial PLSR analysis gave diagnostics (score plots) that indicated in-homogeneities in the data. A much better model was obtained for the $N = 16$ non-aromatic AA's. A remaining curvature in the score plot of $\mathbf{u}_1$ vs. $\mathbf{t}_1$ lead to the inclusion of squared terms, which gave a good final model. Only the squares in the lipophilicity variables are significant in the final model.

If additional aromatic AA's had been present, a second separate model could have been developed for this type of AA's. This would provide insight in how this aromatic group differs from the non-aromatic AA's. This is, in a way, a non-linear model of the changes in the relationship between structure and activity when going from non-aromatic to aromatic AA's. These changes are too large and non-linear to be modelled by a linear or low degree polynomial model. The use of two separate models, which do not directly model the change from one group to another, provides a simple approach to deal with these non-linearities.

6. Example 2 (SIMCODM)

High and consistent product quality combined with "green" plant operation is important in today's competitive industrial climate. The goal of process data modelling is often to reduce the amount of down time and eliminate sources of undesired and deleterious process variability. The second example shows the investigation of possibilities to operate a process industry in an environment-friendly manner [40].

At the Aylesford Newsprint paper-mill in Kent, UK, premium quality newsprint is produced from 100% recycled fiber, i.e., reclaimed newspapers and magazines. The first step of the process is one of de-inking. This has stages of cleaning, screening, and flotation. First the wastepaper is fed into rotating drum pulpers, where it is mixed with warm water and chemicals to initiate the fiber separation and ink removal. In the next stage, the pulp undergoes pre-screening and cleaning to remove light and heavy contaminants (staples, paper clips, sand, plastics, etc.). Thereafter the pulp goes to a pre-flotation stage. After the pre-flotation the pulp is thickened and dispersed and the brightness is adjusted. After post-flotation and washing the de-inked pulp (DIP) is sent via

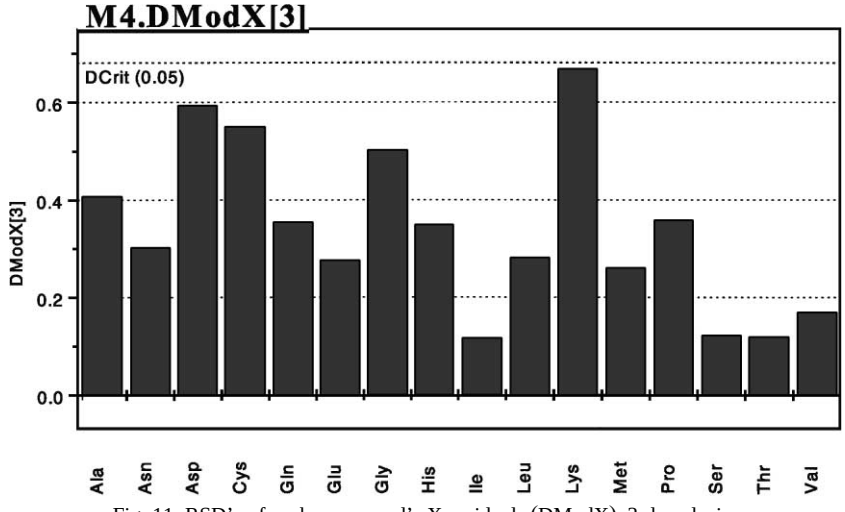

Fig. 11. RSD's of each compound's X-residuals (DModX), 3rd analysis.

storage towers to the paper machines for production of premium quality newsprint.

The wastewater discharged from the de-inking process is the main source of chemical oxygen demand (COD) of the mill effluent. COD estimates the oxygen requirements of organic matter present in the effluent. The COD in the effluent is reduced by 85% prior to discharge to the river Medway. Stricter environmental regulations and costs of COD reduction made the investigation of the COD sources and possible control strategies one of the company's environmental objectives for 1999. A multivariate approach was adopted.

Thus, data from more than a year were used in the multivariate study. The example data set contains 384 daily observations, with one response variable (COD load, y_1) and 54 process variables ($x_2 - x_{55}$). Half of the observations (process time points) are used for model training and half for model validation.

6.1. Overview by PCA

PCA applied to the entire data set (**X** and **Y**) gave an eight-component model explaining $R^2 = 79\%$ and predicting $Q^2 = 68\%$ of the data variation. Scores and loadings of the first two components accounting for 54% of the sum of squares are plotted in Figs. 12 and 13.

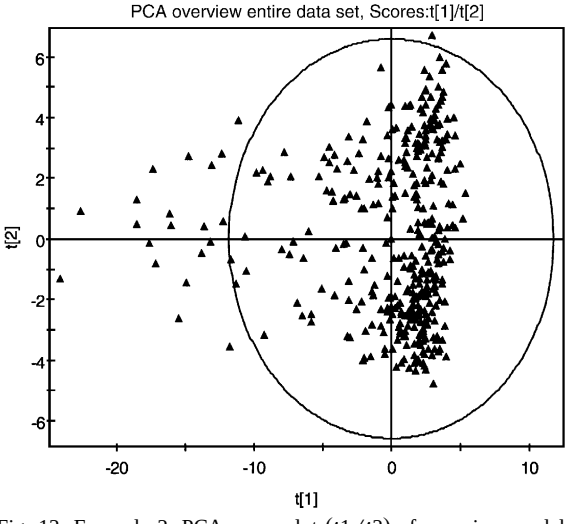

Fig. 12. Example 2. PCA score plot (**t**1/**t**2) of overview model. Each point represents one process time point.

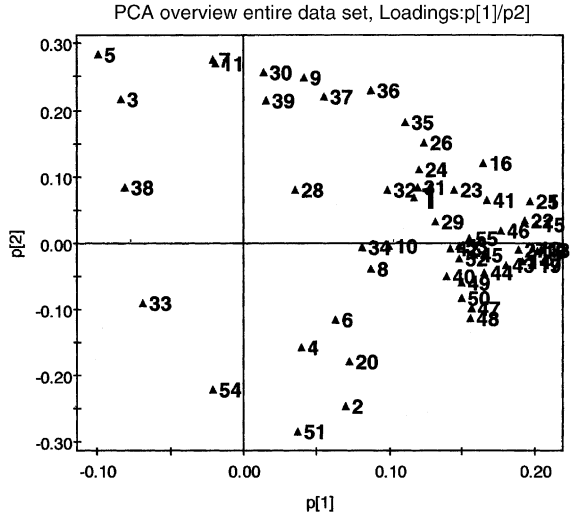

Fig. 13. Example 2. PCA loading plot (**p**1/**p**2) corresponding to Fig. 12. The position of the COD load, variable number 1, is highlighted by an increased font size.

Fig. 12 reveals a main cluster in the right-hand part of the score plot, where the process has spent most of its time. Occasionally, the process has drifted away towards the left-hand area of the score plot, corresponding to a lower production rate, including planned shut downs in one of the process lines. The loading plot shows most of the variables having low numerical values (i.e., low material flow, low addition of process chemicals, etc.) for the low production rates.

This analysis shows that the data are not strongly clustered. The low production process time points deviate from the main cluster, but not seriously. Rather, the inclusion of the low production time points in the subsequent PLSR modelling give a good spanning of key features such as brightness, residual ink, and addition of external chemicals.

6.2. Selection of training set for PLSR and lagging of data

The parent data set, with preserved time order, was split into two halves, one training set and one prediction set. Only forward prediction was applied. Process insight, initial PLSR modelling, and inspection of cross-correlations, indicated that lagging 10 of the variables was warranted to catch process dynamics.

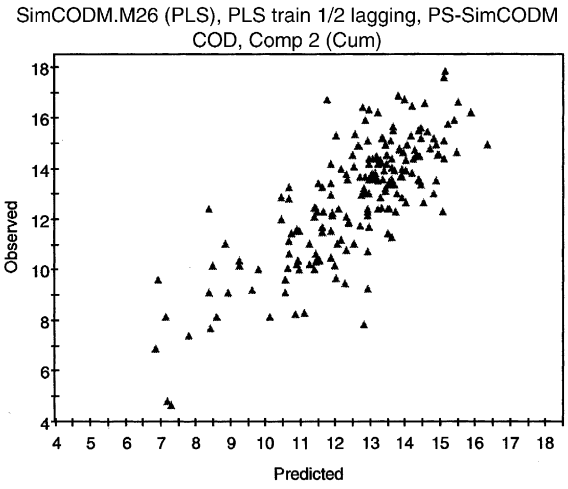

Fig. 14. Example 2. Observed verses predicted COD for the prediction set.

These were: lags 1–4 of COD, and lags 1 and 2 of variables x_{23}, x_{24} (addition of two chemicals), and x_{35} (a flow), totally 10 lagged X's. Lagging a variable L steps means that the same variable shifted L steps in time is included as an additional variable in the data set.

6.3. PLSR modelling

PLSR was utilized to relate the 54 + 10 process variables to the COD load. A two-component PLSR model was obtained, with $R_X^2 = 0.49$, $R_Y^2 = 0.60$, Q_Y^2 (CV) = 0.57, and Q_Y^2 (ext) = 0.59 (external validation set). Fig. 14 shows relationships between observed and predicted levels of COD for the validation set. The picture for the training set is very similar.

6.4. Results

The collected process variables together with their lags predict COD load with $Q^2 > 0.6$. This is a satisfying result considering the complexity of the process and the great variations that occur in the starting material (recycled paper).

The PLS-regression coefficients (Fig. 15), show the process variables with the strongest correlation to the COD load to be x_{23}, x_{24} (addition of chemicals), x_{34}, x_{35} (flows), and x_{49} and x_{50} (temperatures), with coefficients exceeding 0.04. Only some of these can be controlled, namely x_{23}, x_{24}, and x_{49}. More-

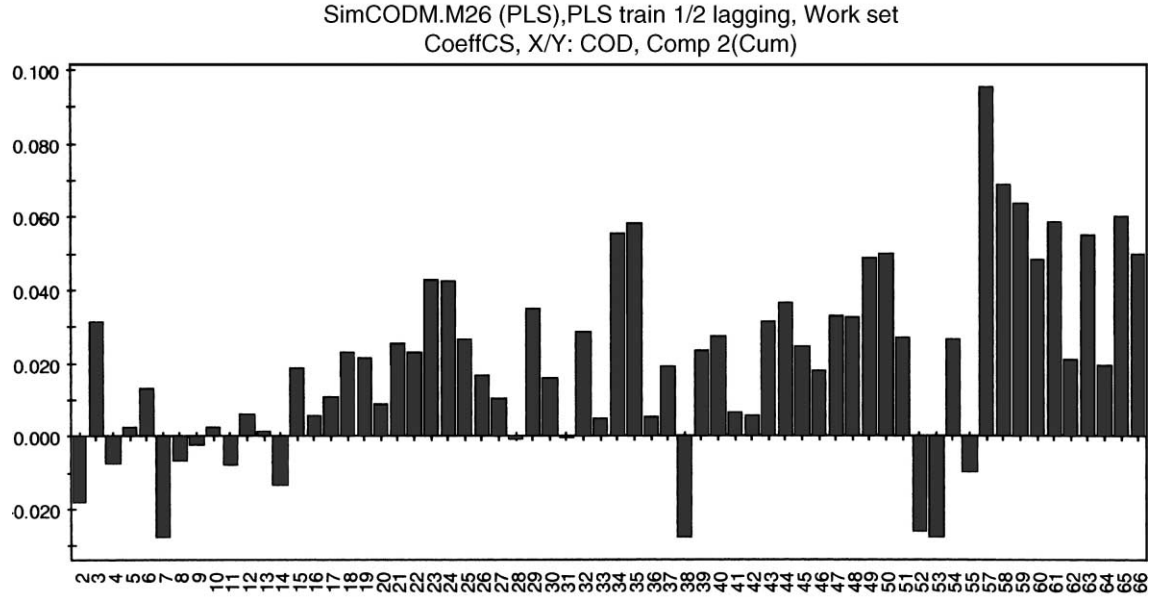

Fig. 15. Example 2. Plot of regression coefficients of scaled and centered variables of PLS model. Coefficients with numbers 2 to 55 represent the 54 original process variables. Coefficients 57–60 depict lags 1–4 of COD, and coefficients 61–66 the two first lags of variables x_{23}, x_{24}, and x_{35}.

over, the lagged variables are important. Variables 57–60 are the four first lags of the COD variable. Obviously, past measurements of COD are useful for modelling and predicting future levels of COD—the change in this variable is slow. The variables 61–66 are lags 1 and 2 of x_{23}, x_{24}, and x_{35}. The dynamics of these process variables are thus related to COD.

Unfortunately, none of the three controlled variables (x_{23}, x_{24}, and x_{49}) offer a means to decrease COD. The former two variables describe the addition of caustic and peroxide to the dispergers, and cannot be significantly reduced without endangering pulp quality. The last variable is the raw effluent temperature which cannot be lowered without negative effects on pulp quality.

This result is typical for the analysis of historical process data. The value of the analysis is mainly to detect deviations from normal operation, not as a means to process optimization. The latter demands data from a statistically designed set of experiments with the purpose of optimizing the process.

7. Summary; how to develop and interpret a PLSR model

(1) Have a good understanding of the stated problem, particularly which responses (properties), **Y**, that are of interest to measure and model, and which predictors, **X**, that should be measured and varied. If possible, i.e., if the X-variables are subject to experimental control, use statistical experimental design [41] for the construction of **X**.

(2) Get good data, both **Y** (responses) and **X** (predictors). Multivariate Y's provide much more information because they can first be separately analyzed by a PCA. This gives a good idea about the amount of systematic variation in **Y**, which Y-variables that should be analyzed together, etc.

(3) The first piece of information about the model is the number of significant components, A—the complexity of the model and hence of the system. This number of components gives a lower bound of the number of *effects* we need to postulate in the system.

(4) The goodness of fit of the model is given by R^2 and Q^2 (cross-validated R^2). With several Y's, one obtains also R^2_m and Q^2_m for each $\mathbf{y}_m$. The R^2's give an upper bound of how well the model explains the data and predicts new observations, and the Q^2's give lower bounds for the same things.

(5) The (**u**,**t**) score plots for the first two or three model dimensions will show the presence of outliers, curvature, or groups in the data.

The (**t**,**t**) score plots—windows in X-space—show in-homogeneities, groups, or other patterns. Together with this, the (**w** * **c**) weight plots gives an interpretation of these patterns.

(6a) If problems are seen, i.e., small R^2 and Q^2 values, or outliers, groups, or curvature in the score plots, one should try to remedy the problem. Plots of residuals (normal probability and DModX and DModY) may give additional hints for the sources of the problems.

Single outliers should be inspected for correctness of data, and if this does not help, be excluded from the analysis only if they are non-interesting (i.e., low activity).

Curvature in (**u**,**t**) plots may be improved by transforming parts of the data (e.g., log), or by including squared and/or cubic terms in the model.

After possibly having transformed the data, modified the model, divided the data into groups, deleted outliers, or whatever warranted, one returns to step 1.

(6b) If no problems are seen, i.e., R^2 and Q^2 are OK, and the model is interpretable, one may try to mildly prune the model by deleting unimportant terms, i.e., with small regression coefficients and low VIP values. Thereafter a final model is developed, interpreted, and validated, predictions are made, etc. For the interpretation the (**w** * **c**) weight plots, coefficient plots, and contour or 3D plots with dominating X's as plot coordinates, are invaluable.

8. Regression-like data-analytical problems

A number of seemingly different data-analytical problems can be expressed as regression problems with a special coding of **Y** or **X**. These include linear discriminant analysis (LDA), analysis of variance (ANOVA), and time series analysis (ARMA and similar models). With many and collinear variables (rank-deficient **X**), a PLSR solution can therefore be formulated for each of these.

In linear discriminant analysis (LDA), and the closely related canonical variates analysis (CVA), one has the **X**-matrix divided in a number of classes, with index $1,2,\ldots,g,\ldots,G$. The objective of the analysis is to find linear combinations of the X-variables (discriminant functions) that discriminate between the classes, i.e., have very different values for the classes [11]. Provided that each class is "tight" and occupies a small and separate volume in X-space, one can find a plane—a discriminant plane—in which the projected observations are well separated according to class. With many and collinear X-variables, a PLS version of LDA (PLS-DA) is useful [42,43].

However, when some of the classes are not tight, often due to a lack of homogeneity and similarity in these non-tight classes, the discriminant analysis does not work. Then other approaches, such as, SIMCA have to be used, where a PC or PLSR model is developed for each tight class, and new observations are classified according to their nearness in X-space to these class models.

Analogously, when **X** contains qualitative variables (the ANOVA situation), these can be coded using dummy variables, and the data then analyzed using MLR or PLSR. The latter is advantageous if the coded **X** is unbalanced and/or rank deficient (or close to). With several Y-variables, this provides a simple approach to MANOVA (multiple responses ANOVA) [11].

Auto-regressive and transfer function models in time-series analysis are easily analyzed by first constructing an expanded X-matrix that contains the appropriate lags of Y- and X-variables, and then calculating a PLSR solution. If the disturbances $\mathbf{a}_t$ are included in the models (ARMA, etc.), a two-step analysis is needed, first calculating estimates of $\mathbf{a}_t$, and then using these in lagged forms in the final PLSR model.

In the modelling of mixtures, for instance of chemical ingredients in paints, pharmaceutical, cosmetic, and polymer formulations, beverages, etc., the sum of the X-variables is 1.0, since ingredients sum to 100%. This makes **X** rank deficient, and a number of special regression approaches have been developed for the analysis of mixture data. With PLSR, however, the rank deficiency presents no difficulties, and the data analysis becomes straight forward, as shown by Kettaneh-Wold [44].

The iterative calculations in non-linear regression often involve a regression-like updating step (e.g., the Gauss–Newton approach). The **X**-matrix of this step is often highly rank-deficient, and ridge-regression is used to remedy the situation (Marquardt–Lefwenberg algorithms). PLSR provides a simple and interesting alternative, which also is computationally very fast.

9. Conclusions and discussion

PLSR provides an approach to the quantitative modelling of the often complicated relationships between predictors, **X**, and responses, **Y**, that with complex problems often is more realistic than MLR including stepwise selection variants. This because the assumptions underlying PLS—correlations among the X's, noise in **X**, model errors—are more realistic than the MLR assumptions of independent and error free X's.

The diagnostics of PLSR, notably cross-validation and score plots (**u**, **t** and **t**, **t**) with corresponding loading plots, provide information about the correlation structure of **X** that is not obtained in ordinary MLR. In particular, PLSR results showing that the data are inhomogeneous (like the AA example looked at here), are hard to obtain by MLR. In complicated systems, non-linearities so strong that a single polynomial model cannot be constructed, seem to be rather common. Hence, a flexible approach to modelling is often warranted with separate models for different mechanistic classes. And there is no loss of information with this approach in comparison with the single model approach. A new observation (object) is first classified with respect to its X-values, and predicted response values are then obtained by the appropriate class model.

The ability of PLSR to analyze *profiles* of responses, makes it easier to device response measurements that are relevant to the stated objective of the investigation; it is easier to capture the behavior of a complicated system by a battery of measurements than by a single variable.

PLSR can be extended in various directions, to non-linear modelling, hierarchical modelling when the variables are very numerous, multi-way data, PLS time series, PLS-DA, etc. Many recently developed

PLS based approaches are discussed in other articles in this volume. The statistical understanding of PLSR has recently improved substantially [12,13,34].

We feel that the flexibility of the PLS-approach, its graphical orientation, and its inherent ability to handle incomplete and noisy data with many variables (and observations) makes PLS a simple but powerful approach for the analysis of data of complicated problems.

Acknowledgements

Support from the Swedish Natural Science Research Council (NFR), and from the Center for Environmental Research in Umeå (CMF) is gratefully acknowledged. Dr. Erik Johansson is thanked for helping with the examples.

References

[1] H. Wold, Soft modelling, The basic design and some extensions, in: K.-G. Jöreskog, H. Wold (Eds.), Systems Under Indirect Observation, vols. I and II, North-Holland, Amsterdam, 1982.
[2] S. Wold, A. Ruhe, H. Wold, W.J. Dunn III, The collinearity problem in linear regression, The partial least squares approach to generalized inverses, SIAM J. Sci. Stat. Comput. 5 (1984) 735–743.
[3] A. Höskuldsson, PLS regression methods, J. Chemom. 2 (1988) 211–228.
[4] A. Höskuldsson, Prediction Methods in Science and Technology, vol. 1, Thor Publishing, Copenhagen, 1996, ISBN 87-985941-0-9.
[5] S. Wold, E. Johansson, M. Cocchi, PLS—partial least squares projections to latent structures, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design, Theory, Methods, and Applications, ESCOM Science Publishers, Leiden, 1993, pp. 523–550.
[6] M. Tenenhaus, La Regression PLS: Theorie et Pratique, Technip, Paris, 1998.
[7] R.W. Gerlach, B.R. Kowalski, H. Wold, Partial least squares modelling with latent variables, Anal. Chim. Acta 112 (1979) 417–421.
[8] N.E. El Tayar, R.-S. Tsai, P.-A. Carrupt, B. Testa, Octan-1-ol-water partition coefficients of zwitterionic a-amino acids, Determination by centrifugal partition chromatography and factorization into steric/hydrophobic and polar components, J. Chem. Soc., Perkin Trans. 2 (1992) 79–84.
[9] S. Hellberg, M. Sjöström, S. Wold, The prediction of bradykinin potentiating potency of pentapeptides, An example of a peptide quantitative structure–activity relationship, Acta Chem. Scand., Ser. B 40 (1986) 135–140.
[10] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, S. Wold, New chemical dimensions relevant for the design of biologically active peptides, A multivariate characterization of 87 amino acids, J. Med. Chem. 41 (1998) 2481–2491.
[11] J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.
[12] A.J. Burnham, J.F. MacGregor, R. Viveris, Latent variable regression tools, Chemom. Intell. Lab. Syst. 48 (1999) 167–180.
[13] A. Burnham, R. Viveros, J. MacGregor, Frameworks for latent variable multivariate regression, J. Chemom. 10 (1996) 31–45.
[14] R. Manne, Analysis of two partial least squares algorithms for multivariate calibration, Chemom. Intell. Lab. Syst. 1 (1987) 187–197.
[15] J. Trygg et al., This issue.
[16] P. Nomikos, J.F. MacGregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (1995) 41–59.
[17] P.R.C. Nelson, P.A. Taylor, J.F. MacGregor, Missing data methods in PCA and PLS: score calculations with incomplete observation, Chemom. Intell. Lab. Syst. 35 (1996) 45–65.
[18] B. Grung, R. Manne, Missing values in principal component analysis, Chemom. Intell. Lab. Syst. 42 (1998) 125–139.
[19] I.N. Wakeling, J.J. Morris, A test of significance for partial least squares regression, J. Chemom. 7 (1993) 291–304.
[20] M. Clark, R.D. Cramer III, The probability of chance correlation using partial least squares (PLS), Quant. Struct.-Act. Relat. 12 (1993) 137–145.
[21] J. Shao, Linear model selection by cross-validation, J. Am. Stat. Assoc. 88 (1993) 486–494.
[22] F. Lindgren, P. Geladi, S. Wold, The kernel algorithm for PLS I, Many observations and few variables, J. Chemom. 7 (1993) 45–59.
[23] S. Rännar, P. Geladi, F. Lindgren, S. Wold, The kernel algorithm for PLS II, Few observations and many variables, J. Chemom. 8 (1994) 111–125.
[24] K.H. Esbensen, S. Wold, SIMCA, MACUP, SELPLS, GDAM, SPACE and UNFOLD: the ways towards regionalized principal components analysis and subconstrained N-way decomposition—with geological illustrations, in: O.J. Christie (Ed.), Proc. Nord. Symp. Appl. Statist. Stavanger, 1983, pp. 11–36, ISBN 82-90496-02-8.
[25] N. Kettaneh-Wold, J.F. MacGregor, B. Dayal, S. Wold, Multivariate design of process experiments (M-DOPE), Chemom. Intell. Lab. Syst. 23 (1994) 39–50.
[26] O.M. Kvalheim, A.A. Christy, N. Telnaes, A. Bjoerseth, Maturity determination of organic matter in coals using the methylphenantrene distribution, Geochim. Cosmochim. Acta 51 (1987) 1883–1888.
[27] M.T. Chu, R.E. Funderlic, G.H. Golub, A rank-one reduction formula and its applications to matrix factorizations, SIAM Rev. 37 (1995) 512–530.
[28] M.C. Denham, Prediction intervals in partial least squares, J. Chemom. 11 (1997) 39–52.
[29] B. Efron, G. Gong, A leisurely look at the bootstrap, the jackknife, and cross-validation, Am. Stat. 37 (1983) 36–48.
[30] H. Martens, M. Martens, Modified jack-knife estimation of

parameter uncertainty in bilinear modeling (PLSR), Food Qual. Preference 11 (2000) 5–16.
[31] S. Wold, C. Albano, W.J. Dunn III, U. Edlund, B. Eliasson, E. Johansson, B. Norden, M. Sjöström, The indirect observation of molecular chemical systems, in: K.-G. Jöreskog, H. Wold (Eds.), Systems Under Indirect Observation, vols. I and II, North-Holland, Amsterdam, 1982, pp. 177–207, Chapter 8.
[32] S. Wold, M. Sjöström, L. Eriksson, PLS in chemistry, in: P.v.R. Schleyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer III, P.R. Schreiner (Eds.), The Encyclopedia of Computational Chemistry, Wiley, Chichester, 1999, pp. 2006–2020.
[33] O. Kvalheim, The latent variable, an editorial, Chemom. Intell. Lab. Syst. 14 (1992) 1–3.
[34] I.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, With discussion, Technometrics 35 (1993) 109–148.
[35] S. Wold, A theoretical foundation of extrathermodynamic relationships (linear free energy relationships), Chem. Scr. 5 (1974) 97–106.
[36] D.A. Belsley, E. Kuh, R.E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, New York, 1980.
[37] A. Berglund, S. Wold, INLR, implicit non-linear latent variable regression, J. Chemom. 11 (1997) 141–156.
[38] S. Wold, A. Berglund, N. Kettaneh, N. Bendwell, D.R. Cameron, The GIFI approach to non-linear PLS modelling, J. Chemom. 15 (2001) 321–336.
[39] L. Eriksson, E. Johansson, F. Lindgren, S. Wold, GIFI-PLS: modeling of non-linearities and discontinuities in QSAR, QSAR 19 (2000) 345–355.
[40] L. Eriksson, P. Hagberg, E. Johansson, S. Rännar, D. Sarney, O. Whelehan, A. Åström, T. Lindgren, Multivariate process monitoring of a newsprint mill, Application to modelling and predicting COD load resulting from de-inking of modeling, J. Chemom. 15 (2001) 337–352.
[41] G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters, Wiley, New York, 1978.
[42] M. Sjöström, S. Wold, B. Söderström, PLS discriminant plots, Proceedings of PARC in Practice, Amsterdam, June 19–21, 1985, Elsevier, North-Holland, 1986, pp. 461–470.
[43] L. Ståhle, S. Wold, Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study, J. Chemom. 1 (1987) 185–196.
[44] N. Kettaneh-Wold, Analysis of mixture data with partial least squares, Chemom. Intell. Lab. Syst. 14 (1992) 57–69.