Chemometrics and Intelligent Laboratory Systems, 8 (1990) 109–119
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

A Nonlinear PLS Model

ILDIKO E. FRANK
Jerll, Inc., 790 Esplanada, Stanford, CA 94305 (U.S.A.)

(Received 15 July 1989; accepted 5 December 1989)

ABSTRACT

Frank, I.E., 1990. A nonlinear PLS model. *Chemometrics and Intelligent Laboratory Systems*, 8: 109–119.

A nonlinear extension of the PLS (partial least squares regression) method is introduced. The algorithm connects the predictor and response latent variables by a smooth but otherwise unrestricted nonlinear function. Similarities and differences between the linear and nonlinear PLS models are discussed. The performance of the new nonlinear PLS method is illustrated on three chemical data sets with different covariance structures.

INTRODUCTION

PLS (partial least squares) regression has been a popular method for years in chemometrics [1–4]. PLS, similar to ridge and principal component regressions, decreases the prediction error by introducing bias to the least squares solution. PLS is a viable regression model for situations where the prediction error is dominated by the variance, e.g. highly collinear predictors or low observation/variable ratio.

PLS has found applications in many fields of chemistry. For example in organic chemistry the structure–activity relationship of compounds are modeled. In food chemistry quality and origin of products are predicted from chemical measurements. Multivariate calibration is a typical PLS application in analytical chemistry. In environmental chemistry polluting sources are revealed by relating source profiles to aerosol composition.

Nonlinear models offer better prediction than linear ones when the prediction error is mainly due to the bias associated with model misspecification, i.e. when the linear model is a poor approximation. Usually these situations are well determined with high observation/variable ratio. One of the most flexible nonlinear methods is ACE (Alternating Conditional Expectations), which models the response as a sum of nonlinear functions of the predictor variables [5].

It might seem at first that PLS and nonlinear regression models are most suitable in opposite contexts, therefore a nonlinear PLS model would not be advantageous. Often when PLS is appropriate, there are not enough degrees of freedom left for fitting a nonlinear model. However, we have found a few data sets where the point cloud of the inner relationship plots show curvature. This indicates nonlinear relationship between the predictor and response blocks.

In the first part of this paper a nonlinear PLS model, called NLPLS is presented; motivation and details of the algorithm are discussed. In the second part PLS and NLPLS models are com-

pared in three data sets from various fields of chemistry.

METHODS

PLS is a linear regression method that relates a set of predictor variables (predictor matrix is denoted **X** with row index $i = 1, I$ and variable index $j = 1, J$) to a set of response variables (response matrix is denoted **Y** with row index $i = 1, I$ and variable index $k = 1, K$). It performs a least squares regression on a set of uncorrelated, so called latent variables (denoted **T** with row index $i = 1, I$ and latent variable index $m = 1, M$), which are linear combinations of the original predictor variables. PLS models the responses as:

$$\mathbf{Y}[I,K] = \mathbf{T}[I,M]\,\mathbf{B}[M,M]\,\mathbf{Q}[M,K] + \mathbf{E}[I,K] \quad (1)$$

where **B** is a diagonal matrix containing the least squares coefficients of the latent variables, matrix **Q** contains the weights of the responses in each component, and **E** is the error matrix. PLS calculates one latent variable, i.e. one component at a time. For the mth component a latent variable $\boldsymbol{t}_m$ and also a linear combination of the responses $\boldsymbol{u}_m$ are calculated with coefficients $\boldsymbol{w}_m^T$ and $\boldsymbol{q}_m^T$, respectively

$$\boldsymbol{t}_m = \mathbf{X}\boldsymbol{w}_m^T \quad (2)$$

$$\boldsymbol{u}_m = \mathbf{Y}\boldsymbol{q}_m^T \quad (3)$$

$\boldsymbol{u}_m$ is regressed on $\boldsymbol{t}_m$

$$\boldsymbol{u}_m = b_m \boldsymbol{t}_m \quad (4)$$

and finally residuals are taken in both **X** and **Y** matrices

$$\mathbf{X} = \mathbf{X} - b_m \boldsymbol{t}_m \boldsymbol{p}_m \quad (5)$$

$$\mathbf{Y} = \mathbf{Y} - b_m \boldsymbol{t}_m \boldsymbol{q}_m \quad (6)$$

where $\boldsymbol{p}_m$ is the projection of **X** on $b_m \boldsymbol{t}_m$.

The calculation of the latent variables is called the outer relationship, while the least squares fit between the two latent variables is called the inner relationship. PLS is similar to the principal component regression (PCR), in that it also performs least squares regression on uncorrelated linear combinations of the predictor variables. In PCR the only criterion for a new orthogonal linear combination is to maximize its variance, while in PLS the criterion for the calculation of the latent variables also includes the correlation between the responses and the latent variable.

$$\text{PCR criterion:} \quad \max\left[\operatorname{var}\left(\mathbf{X}\boldsymbol{w}_m^T\right)\right] \quad (7)$$

$$\text{PLS criterion:} \quad \max\left[\operatorname{corr}^2\left(\mathbf{Y}, \mathbf{X}\boldsymbol{w}_m^T\right) \cdot \operatorname{var}\left(\mathbf{X}\boldsymbol{w}_m^T\right)\right] \quad (8)$$

Due to the linearity of both outer and inner relationships in PLS, the coefficients **W**, **P**, **Q**, and **B** can be combined into one set of regression coefficients **A**. Therefore the final PLS model, similar to a linear least squares model, can be written as

$$y_k = \sum_{j=1,J} a_{jk} x_j + e_k \quad k = 1 \ldots K \quad (9)$$

In both PLS and PCR the optimal number of linear combinations M has to be determined. It is usually done by cross-validation. If the underlying dimensionality equals the number of predictors ($M = J$), both the PLS and the PCR models give the same results as the least squares solution.

The PLS model is linear in both regression coefficients and predictors. There have been few attempts to extend PLS into nonlinearity. If the **X** matrix includes not only the predictors themselves but also their squares and cross terms, then the model becomes quadratic in the variables, but still linear in the regression coefficients [6].

$$y_k = \sum_j a1_{jk} x_j + \sum_j a2_{jk} x_j^2 + \sum_l a3_{lk} x_{j1} x_{j2} + e_k$$
$$j = 1 \ldots J; \quad k = 1 \ldots K;$$
$$l = 1 \ldots (J * J - J)/2 \quad (10)$$

Of course this is not a viable solution when there are a large number of predictor variables.

Wold et al. [7] suggest nonlinearizing the inner relationships of the PLS model in each component by including the quadratic term $\boldsymbol{t}_m^2$

$$\boldsymbol{u}_m = b_{0m} + b_{1m}\boldsymbol{t}_m + b_{2m}\boldsymbol{t}_m^2 \quad (11)$$

This 'quadratic' QPLS model is linear in predictors, but quadratic in the coefficients. This model

cannot be written in a simple form as in (9) or (10), and it is less parsimonious than any linear model since it generally involves several linear combinations of high order powers of X, whereas the linear PLS model can be represented by a single linear combination.

Our nonlinear generalization of PLS (we call it NLPLS) is motivated by a regression model called SMART (Smooth Multiple Additive Regression Technique) [8]. SMART is a nonlinear model in the form

$$y_k = \sum_{m=1,M} q_{mk} f_m(\mathbf{X}\mathbf{w}_m^T) + e_k \qquad k = 1 \ldots K \tag{12}$$

where the coefficients w_m and q_{mk} and the functions f_m are simultaneously estimated by minimizing the least squares criterion. M is the optimal number of functions (components). SMART models each response as a (usually) different linear combination of the predictor functions f_m. Each predictor function is taken as a (smooth but otherwise unrestricted) function of a (usually) different linear combination of the predictor variables.

In SMART, the least squares solution is obtained by simultaneously estimating the linear coefficients and the nonlinear functions [8]. In each component there are three steps that are iterated. Given the linear combination of predictors w_m and the function f_m the weights of the responses q_m are determined regressing each response separately on the functions. Given the linear combination of the predictors w_m and of the responses q_m the function f_m is obtained by a smoothing procedure. Finally, given the response weights q_m and the function f_m the coefficients of the predictor linear combination w_m are calculated by the Gauss–Newton method.

Comparing these three steps of SMART and the PLS algorithm, the following analogies and differences can be found. The first step is similar to the way that PLS calculates the response weights in the outer relationship, except that the latent variable is used in place of the function. The second step corresponds to the PLS inner relationship calculation. The smoothing procedure (discussed later) replaces the linear fit and (as will be seen) offers a nonlinear extention to PLS. The third step is similar to the predictor latent variable calculation in PLS. However, due to the nonlinear inner relationship, the coefficients w_m can no longer be estimated by simple linear covariances. The Gauss–Newton step in SMART takes into account not only the correlation between predictors and the nonlinear function, but also the correlation between predictors and the derivative of the nonlinear function. In PLS, because the inner relationship is linear, its derivative is constant and the Gauss–Newton estimate of w_m becomes simply the linear covariance between a predictor and the response latent variable.

The results of the above comparison inspired our NLPLS model, which also can be defined in closed form by (12). The only difference from the SMART model is the way the parameters are estimated. The block diagram of the NLPLS algorithm including the PLS algorithm is presented in Fig. 1.

Just as in the PLS method, the components are calculated in turn one by one. In contrast to PLS, however, no residuals are taken from the X matrix, i.e. all latent variables are linear combinations of the original predictor variables. Taking residuals in X causes the latent variables to be uncorrelated. This has several advantages in the linear model. It ensures that the PLS solution converges in J steps to the least squares solution. The inner relationship coefficients, calculated one component at a time, are the multivariate least squares regression coefficients of the latent variables. Unfortunately, these nice properties would not hold in NLPLS. Also taking residuals in X would prohibit us from defining the model in a simple form (12), making the interpretation of the components impossible.

In each NLPLS component, first the linear combinations of predictors and of the responses (latent variables) are estimated the same way as in PLS. Only after the outer relationship has converged, is the nonlinear inner relationship calculated by the smoothing procedure as

$$u_m = f_m(t_m) \tag{13}$$

This means that the estimates of w_m are based on a linear approximation, although the inner rela-

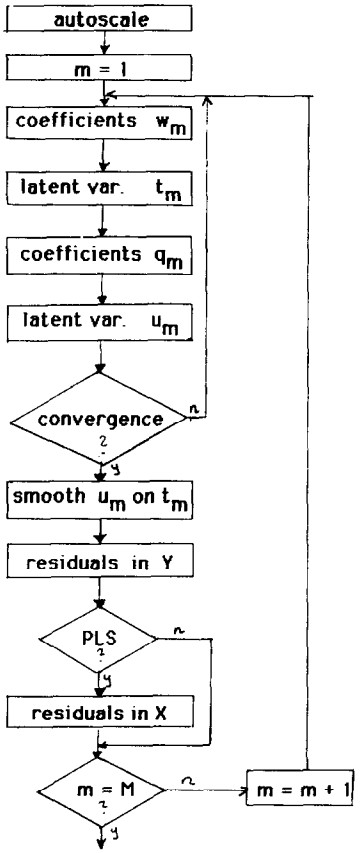

Fig. 1. Block diagram of the nonlinear PLS algorithm.

tionship is nonlinear. What is gained and what is lost with this simplification? In PLS, each predictor (or its residual) enters into the latent variable according to its covariance with the responses. Extending the same concept to nonlinearity, w_m represents the association among predictors and the response latent variable. If the function is monotone, the association values will be closely related to the covariance values. It is quite rare that the direction based on linear function is very different from the direction that was estimated by a Gauss–Newton step (starting from linearity). PLS tries to decrease the prediction error by biasing the solution away from the least squares model. Calculating the direction based on linear function is just another way of trading bias for variance in estimating the latent variables. The advantage of this simplification is that the calculation of the outer and inner relationships can be separated (as in linear PLS) and so that the iteration is much faster. Also, as mentioned above, interpretation is facilitated by the simple form of the resulting model.

The real novelty of the NLPLS algorithm is the estimation of the inner relationships by the smoothing procedure [9]. A smoother is a procedure applied to bivariate data to summarize the association between the two variables, in our case between the predictor and the response latent variables. Instead of describing the association between $\boldsymbol{u}_m$ and $\boldsymbol{t}_m$ in a functional form, which is restricted, like for example (4) in PLS or (11) in QPLS, the smoothing procedure constructs a smooth but otherwise unrestricted nonlinear function in terms of a set of point pairs. To examine and interpret the nonlinearity, the set of point pairs are plotted. The smoother can be regarded as a procedure for estimating the conditional expectation

$$f_m(t_{im}) = E[\boldsymbol{u}_m \mid \boldsymbol{t}_m = t_{im}] \tag{14}$$

A straightforward estimation of this conditional expectation is a conditional average

$$\hat{f}_m(t_{im}) = \text{ave}[\boldsymbol{u}_m \mid t_{im}] = u_{im} \tag{15}$$

Because in our case there is only one u_{im} value for each t_{im}, this estimate has a very high variance. A more reasonable estimate is based on local averaging, i.e. averaging all u_{nm} values that fall in a neighborhood N_i

$$\hat{f}_m(t_{im}) = \text{ave}[\boldsymbol{u}_m \mid t_{nm} \in N_i] \tag{16}$$

This estimate, however, also has serious shortcomings. It does not reproduce straight lines if the abscissa values are not equispaced, and it does not behave well near the boundaries. Therefore in our NLPLS algorithm we use a local linear fit instead of a local average (fitting a constant) in the neighborhood. The smoothed value of the function $f_m(t_{im})$ is taken to be the value of the straight line at t_{im} fitted by least squares to the points in the neighborhood N_i. The most important choice in both local averaging and local linear fit smoother is the choice of the size of the neighborhood over which the averaging or the linear fit takes place. The parameter SPAN determines the

size of the neighborhood as a fraction of the total number of points ($0 < \text{SPAN} \leqslant 1$). SPAN determines also the smoothness of the function. In case of SPAN = 1, i.e. when the local neighborhood containes all the points, the resulting function is linear and NLPLS calculates the PLS model (in this case residuals in **X** are taken). The smaller the SPAN, the more nonlinearity can be captured by the smoother, but also the function estimate has more variance. The higher the observation/predictor ratio, i.e. the larger the number of degrees of freedom, the smaller SPAN can be chosen. For practical purposes SPAN is usually taken between 0.2 and 0.4.

The NLPLS model is defined by the linear coefficients **W**, **Q**, and by the M nonlinear functions, each stored as I point pairs. A new observation i is predicted by the following steps. First the M latent variables are calculated as

$$\boldsymbol{t}_i = \boldsymbol{x}_i \mathbf{W}^T \tag{17}$$

Then the values of the M nonlinear functions have to be looked up on the $\boldsymbol{u}_m - \boldsymbol{t}_m$ curves. If the new t_{im} falls exactly on one of the abscissa points, the new u_{im} is the corresponding ordinate value. If the new t_{im} falls between two points on the abscissa, the new u_{im} is calculated by linear interpolation. Finally the predicted y_{ik} values are calculated as

$$y_{ik} = \sum_{m=1,M} u_{im} q_{mk} \tag{18}$$

The NLPLS algorithm, that includes PLS as a special case, was implemented in a FORTRAN program.

DATA SETS

Three data sets were chosen from various fields of chemistry to demonstrate the performance of the NLPLS method. The first set [10], called 'Gambino' is an example for modeling molecular structure–activity relationship. A sample of 37 compounds was collected to examine this relationship in 6-anilinouracils as inhibitors of *Bacillus subtilis* DNA polymerase III and of the mutant enzyme pol III/azp-12. Four parameters were measured on each compound: hydrophobic constant and group size of the *meta* and *para* substituents. The two responses are the logarithm of the inverse concentrations of 6-anilinouracil required to achieve 50% inhibition of the enzyme and the mutant enzyme activities. This data set is well determined, and has a relatively high observation/predictor ratio. Also in the literature strong nonlinearity was indicated in the structure–activity model. This data set was a good example for the ACE method [5], so it offers an interesting comparison with the NLPLS application.

The second data set is a product quality control example from analytical chemistry [11]. It was collected at 3M Company, hence called '3M', to model the quality of adhesive tapes. The predictor block consists of IR spectra measured at 219 wavelengths between 600 and 4000 cm^{-1}. As responses, two quality control parameters were obtained on each of the 34 samples. In order not to reveal proprietary information they are referred to as quality variable A and quality variable B. This set is a typical case of low observation/variable ratio and high collinearity among predictors for which a biased regression method is appropriate.

Finally the third data set, called 'Wine', comes from food chemistry [12]. 38 Pinot Noir wine samples were subjected to elemental analysis by atomic emission spectrometry and to sensory evaluation. Our goal here was to predict three sensory scores: aroma character, sugar concentration, and flavor character from the contents of 17 elements: Cd, Mo, Mn, Ni, Cu, Al, Ba, Cr, Sr, Pb, B, Mg, Si, Na, Ca, P, and K. We suspect nonlinearity in this relationship, because the wine samples originate from three different geographical regions.

RESULTS

There were two models calculated for each data set: the linear PLS and the nonlinear NLPLS models. The goodness of fit of the models is reflected by the R2 value (1 − sum of squared error/sum of squared response) and the goodness of predictive capability of the models is characterized by the CR2 value, which is the crossvalidated R2. The crossvalidation procedure was

TABLE 1

Goodness-of-fit and goodness-of-prediction of PLS and NL-PLS models in Gambino data set ($I = 37$, $J = 4$, $K = 2$, $M = 4$, SPAN = 0.2)

ICOMP	PLS				NLPLS			
	R2	CR2	R2	CR2	R2	CR2	R2	CR2
1	0.28	0.00	0.22	0.00	0.62	0.29	0.50	0.17
2	0.28	0.00	0.38	0.09	0.65	0.37	0.84	0.42
3	0.36	0.05	0.46	0.18	0.81	0.61	0.86	0.45
4	0.36	**0.05**	0.46	**0.20**	0.86	**0.65**	0.88	**0.57**

TABLE 2

Goodness-of-fit and goodness-of-prediction of PLS and NL-PLS models in 3M data set ($I = 34$, $J = 219$, $K = 2$, $M = 4$, SPAN = 0.4)

ICOMP	PLS				NLPLS			
	R2	CR2	R2	CR2	R2	CR2	R2	CR2
1	0.74	0.63	0.44	0.30	0.86	0.77	0.53	0.36
2	0.94	0.90	0.56	0.45	0.96	0.94	0.64	0.50
3	0.97	0.95	0.63	0.49	0.98	0.95	0.65	0.53
4	0.98	**0.96**	0.70	**0.45**	0.98	**0.98**	0.71	**0.60**
5	0.98	0.97	0.74	0.32	0.98	0.96	0.74	0.60
6	0.98	0.97	0.78	0.23	0.98	0.96	0.75	0.58

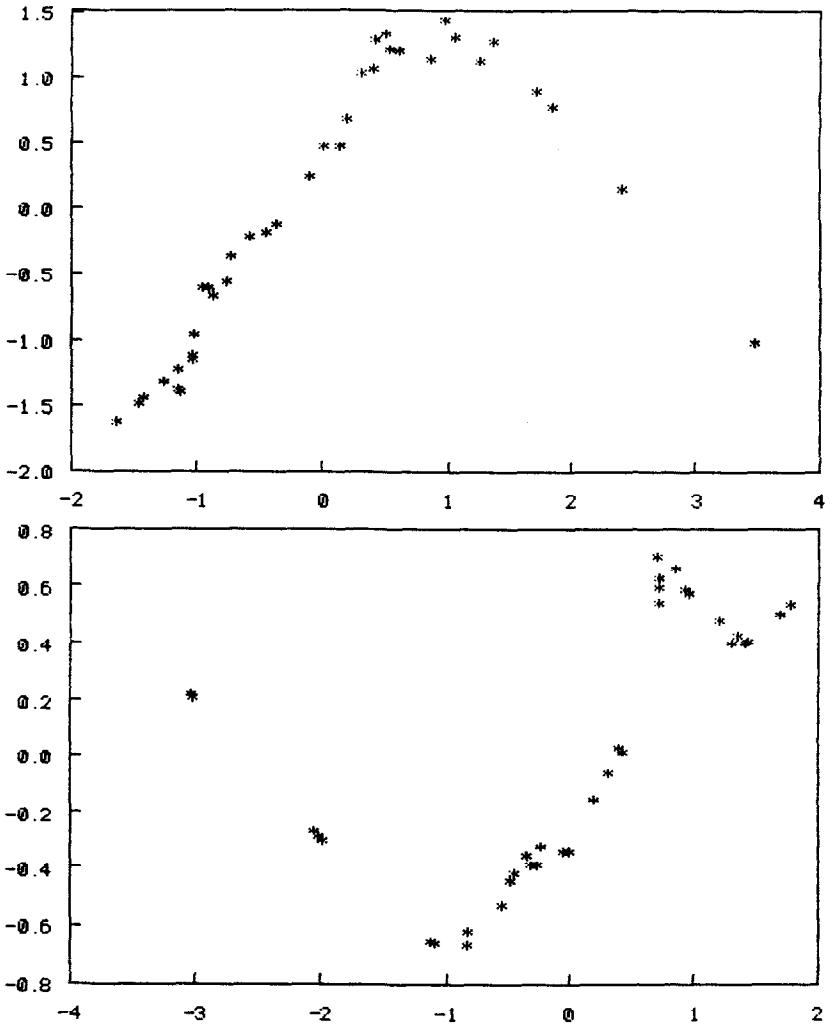

Fig. 2.

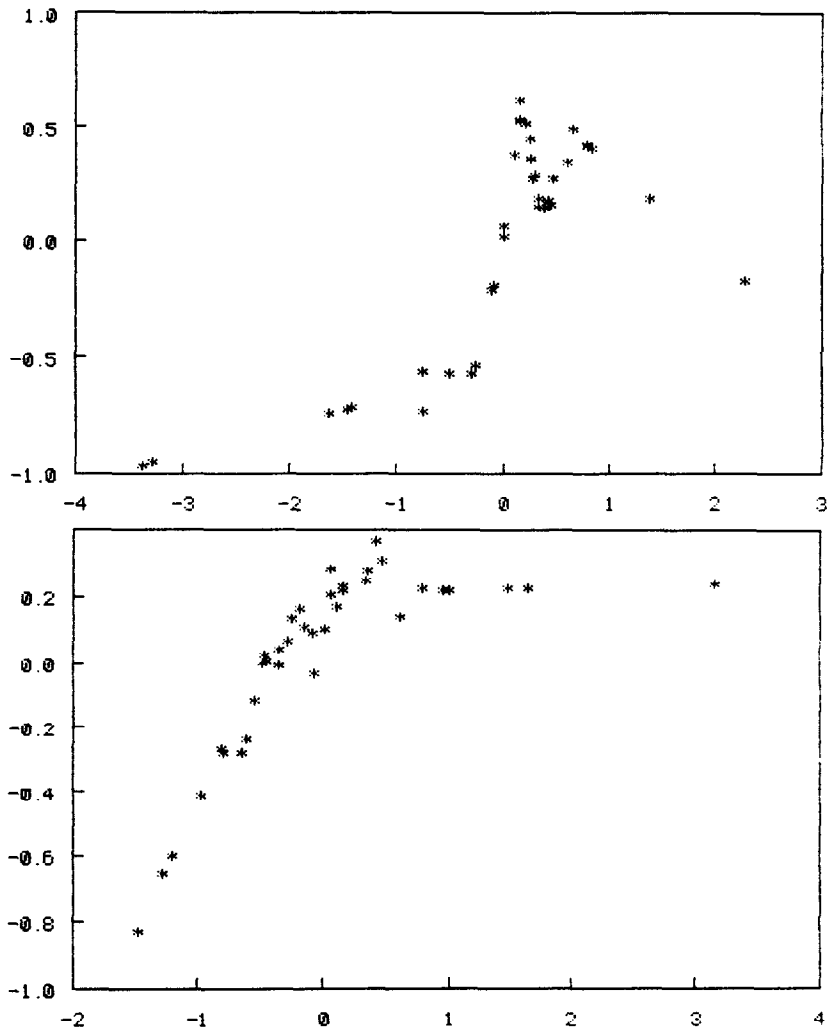

Fig. 2. Nonlinear association between predictor and response latent variables in four components of the Gambino data set. $I = 37$, $J = 4$, $K = 2$, $M = 4$, SPAN = 0.2.

performed leaving out one observation at a time. The R2 and CR2 values for each response and each component are shown in Table 1 for the Gambino data, in Table 2 for the 3M data, and in Table 3 for the Wine data. The nonlinear inner relationship functions of the optimal M compo-

TABLE 3

Goodness-of-fit and goodness-of-prediction of PLS and NLPLS models in Wine data set ($I = 38$, $J = 17$, $K = 3$, $M = 2$, SPAN = 0.4)

ICOMP	PLS						NLPLS					
	R2	CR2	R2	CR2	R2	CR2	R2	CR2	R2	CR2	R2	CR2
1	0.58	0.41	0.01	0.00	0.32	0.14	0.67	0.52	0.02	0.00	0.38	0.20
2	0.62	**0.47**	0.61	**0.38**	0.36	**0.16**	0.72	**0.60**	0.59	**0.37**	0.42	**0.24**
3	0.70	0.40	0.69	0.37	0.48	0.05	0.72	0.54	0.65	0.34	0.43	0.18

nents are presented in Fig. 2 for the Gambino data ($M = 4$), in Fig. 3 for the 3M data ($M = 4$), and in Fig. 4 for the Wine data ($M = 2$).

The Gambino data set is well determined with relatively high observation/predictor ratio. The best PLS model is the least squares model. The small R2 and CR2 values indicate that the linear model is inadequate to describe this structure–activity relationship, i.e. the prediction error is mainly due to the bias component. There are enough degrees of freedom to calculate a more complex nonlinear model.

The NLPLS model offers significant improvement in both responses in each component. The best NLPLS model has four components. The optimal span of the smoother is small (SPAN = 0.2), because there are enough observations to closely fit the curve. The nonlinearity of the inner relationships are mainly due to few compounds, that do not fall on the straight line determined by the majority of the compounds. Comparing only the PLS and NLPLS models leaves us with a success story. Unfortunately, we have applied also the ACE model to this data set [5], that gave

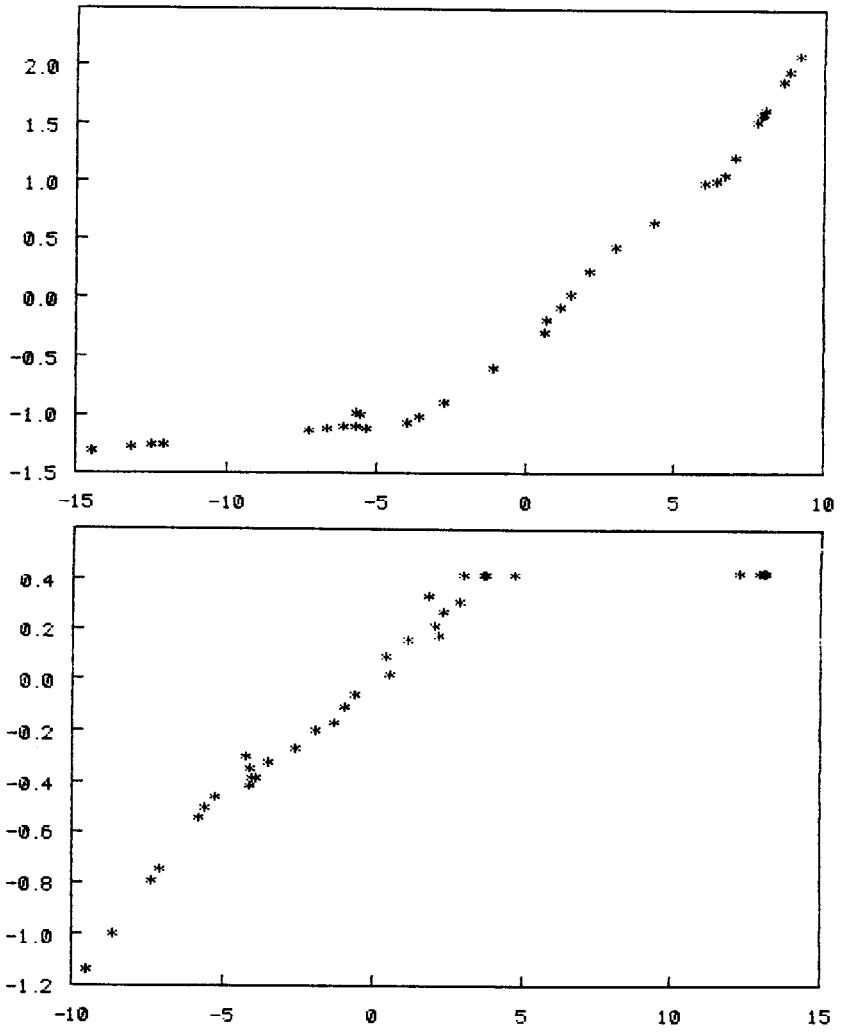

Fig. 3.

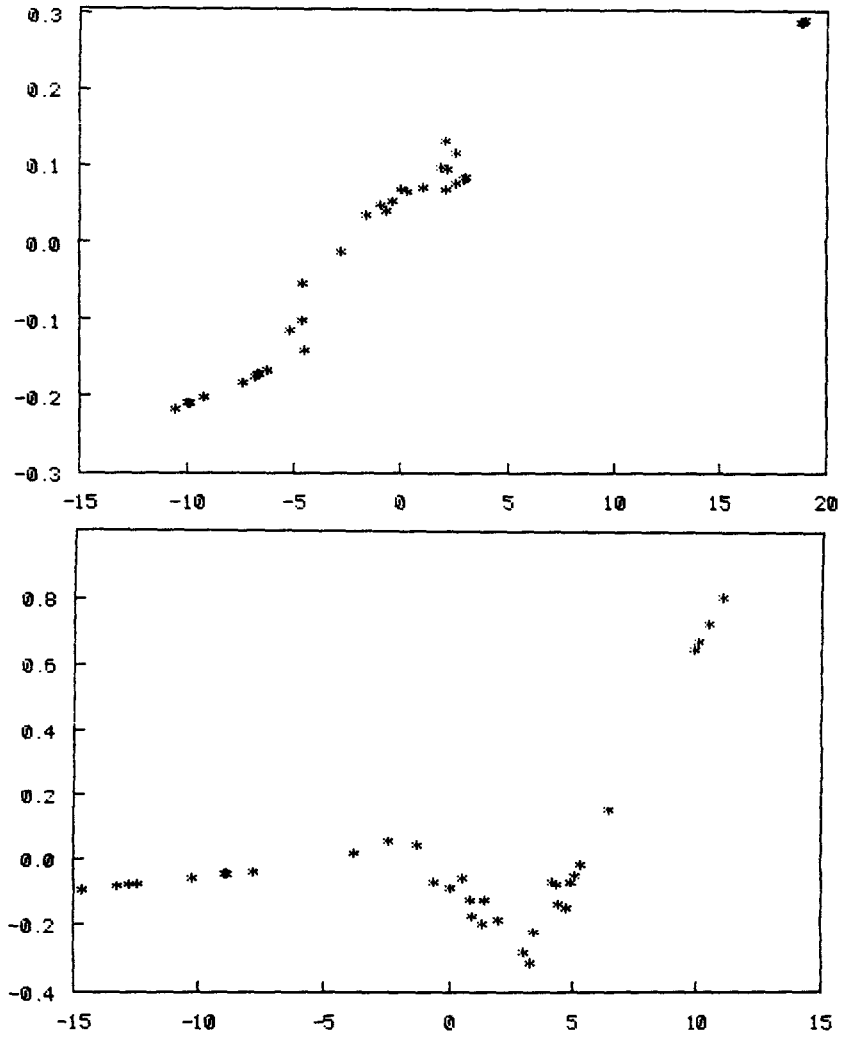

Fig. 3. Nonlinear association between predictor and response latent variables in four components of the 3M data set. $I = 34$, $J = 219$, $K = 2$, $M = 4$, SPAN = 0.4.

R2 = 0.94 and CR2 = 0.78 for the first response and R2 = 0.92 and CR2 = 0.77 for the second response. This supports our previous remark that NLPLS is probably not the best nonlinear method for well determined cases, where not much can be gained by the bias–variance trade off.

The 3M data set is a typical PLS application: high collinearity among the predictors, and low observation/variable ratio. The best PLS model of three components gives excellent prediction for the first response but a mediocre one for the second response. The low underlying dimensionality justifies trying the NLPLS model here.

The four component NLPLS model gives significantly better prediction than the PLS model for both responses. Due to scarsity of degrees of freedom, a larger span (SPAN = 0.4), i.e. smoother function estimate is appropriate in this case. Again, the nonlinearity of the inner relationships is caused mainly by few points at the end of the range. The shape of the first two curves suggests that this data set might be modeled by two separate linear

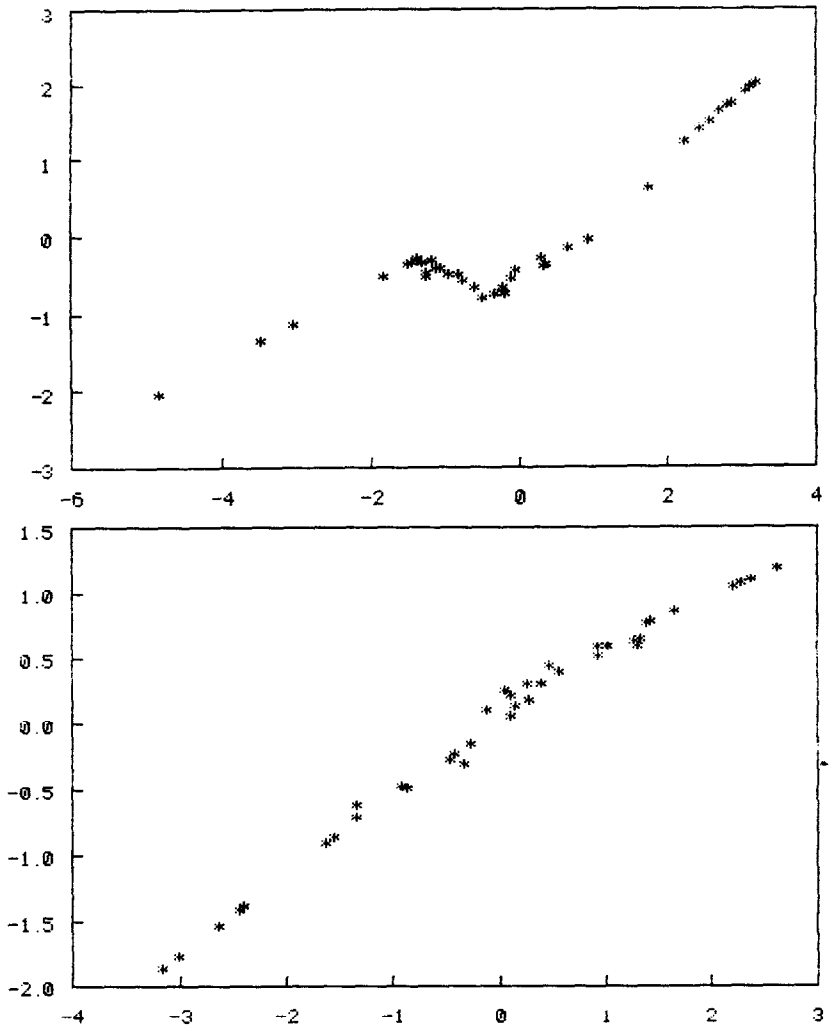

Fig. 4. Nonlinear association between predictor and response latent variables in two components of the Wine data set, $I = 38$, $J = 17$, $K = 3$, $M = 2$, SPAN = 0.4.

models covering low and high quality, respectively.

The PLS model gives quite bad prediction results for the Wine data set. We were hoping that the large errors are due to linear bias, and introducing nonlinearity would improve the model. Again, the underlying dimensionality is low ($M = 2$) compared to the number of predictors.

A significant improvement has happened only in the first response. The other two responses are related linearly to the predictors. This was discovered by running NLPLS only with one response at a time. This means that the dominating component of the model error was variance, not bias. The inherent dimensionality of two components persisted also in NLPLS. The nonlinearity of the first component is interesting, because the curvature is in the middle of the range of the latent variables.

These three applications show that there are some situations where the PLS model can be improved by estimating nonlinear inner relationships. However, if the observation/variable ratio is high, it is probable that other nonlinear models

that use the full least squares coefficient estimates, like ACE or SMART, offer better prediction. On the other hand, a small number of degrees of freedom in underdetermined data sets prohibits the fitting of nonlinear functions with acceptable variance.

Estimating smooth but otherwise unrestricted functions in the inner relationship instead of forcing a specific parametric form on the data, enables NLPLS to capture a wide range of nonlinearity. It also allows the model to remain approximately linear if in fact the relationship between **X** and **Y** is linear.

ACKNOWLEDGEMENT

I would like to gratefully acknowledge the helpful discussions with Jerome H. Friedman.

REFERENCES

1 P. Geladi and B.R. Kowalski, Partial least squares regression (PLS): a tutorial, *Analytica Chimica Acta*, 185 (1986) 1–17.
2 A. Lorber, L.E. Wangen and B.R. Kowalski, A theoretical foundation for the PLS algorithm, *Journal of Chemometrics*, 1 (1987) 19–31.
3 I.E. Frank, Intermediate least squares regression method, *Chemometrics and Intelligent Laboratory Systems*, 1 (1987) 233–242.
4 A. Hoskuldsson, PLS regression methods, *Journal of Chemometrics*, 2 (1988) 211–228.
5 I.E. Frank and S. Lanteri, ACE: A nonlinear regression model, *Chemometrics and Intelligent Laboratory Systems*, 3 (1988) 301–313.
6 O.M. Kvalheim, Latent-structure decompositions (projections) of multivariate data, *Chemometrics and Intelligent Laboratory Systems*, 2 (1987) 283–290.
7 S. Wold, N. Kettaneh-Wold and B. Skagerberg, Nonlinear PLS modeling *Chemometrics and Intelligent Laboratory Systems*, 7 (1989) 53–65.
8 J.H. Friedman, Stanford University, Department of Statistics, SMART, Technical Report No. 1, 1984.
9 J.H. Friedman, Stanford University, Department of Statistics, A variable span smoother, Technical Report No. 5, 1984.
10 G.E. Wright and J.J. Gambino, Quantitative structure–activity relationship of 6-anilinouracils as inhibitors of *Bacillus subtilis* DNA polymerase III, *Journal Medical Chemistry*, 27 (1984) 181–185.
11 I.E. Frank, J. Feikema, N. Constantine and B.R. Kowalski, Prediction of product quality from spectral data using the partial least squares method, *Chemical Information and Computer Science*, 24 (1984) 20–24.
12 I.E. Frank and B.R. Kowalski, Prediction of wine quality and geographic origin from chemical measurements by partial least squares regression modeling, *Analytica Chimica Acta*, 162 (1984) 241–251.