ELSEVIER

Chemometrics and Intelligent Laboratory Systems 55 (2001) 101–108

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Direct Orthogonalization: some case studies

J.A. Fernández Pierna [a], D.L. Massart [a,*], O.E. de Noord [b], Ph. Ricoux [c]

[a] *ChemoAC, Pharmaceutical Institute, Department of Pharmaceutical and Biomedical Analysis, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*

[b] *Shell International Chemicals B.V., Shell Research and Technology Centre, P.O. Box 38000, 1030 BN Amsterdam, Netherlands*

[c] *Elf-Centre de Recherche, B.P. 22, F-69360 Solaize, France*

Received 6 July 2000; accepted 21 November 2000

Abstract

The effects of a Direct Orthogonalization before applying PCR and PLS are studied for several data sets. In all cases the number of PLS factors needed to obtain the optimal model decreases but the number of PLS and DO factors together is the same as when PLS alone is used. However, the quality of the calibration model (measured as RMSECV) is usually not better when using DO, nor does the predictive quality (RMSEP) change significantly in most cases. The method may be used, however, to obtain a better understanding of the variation present in the data.

Keywords: Pre-processing methods; Direct Orthogonalization; Multivariate calibration

1. Introduction

Multivariate calibration establishes a relationship between a variable to be predicted, for instance concentration, and several predicting variables such as absorbances at different wavelengths. This relationship is obtained in two steps. In the first one, the values of the predicted and predicting variables are known and used to build a calibration model. In the second one this model is used to predict new samples. Methods as PCR or PLS are used but it often happens that some of the variables do not (only) contain relevant information. They can therefore increase the imprecision of the latent variable model [1].

Two main methods have been proposed to cope with this problem. The first is to decrease the number of variables. "Intermediate Least Squares" [2] replaces small loadings by zeros by using a prepecified threshold value. This means that certain variables are eliminated or become less important. "Uninformative Variable Elimination in PLS" [3] eliminates variables that are clearly uninformative.

The second method is to apply pre-treatments such as multiplicative signal correction (MSC) [1,4] or second-order derivatives. They eliminate background information but can also eliminate some information important to build the model.

Recently, Orthogonal Signal Correction (OSC) [5–7], and Direct Orthogonalization (DO) were pro-

* Corresponding author. Tel.: +32-2-477-4734; fax: +32-2-477-4735.

E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

0169-7439/01/$ - see front matter
PII: S0169-7439(00)00121-0

Table 1
RMSECV values using from 1 to 11 PLS components for the polyether polyol data

Number of DO components	Number of PLS components										
	1	2	3	4	5	6	7	8	9	10	11
0	9.780	4.676	4.164	3.427	2.631	2.102	**1.719**	1.860	2.196	2.451	2.444
1	4.648	4.107	3.336	2.621	2.058	**1.684**	1.793	2.068	2.266	2.200	2.155
2	4.265	3.308	2.403	1.911	**1.556**	1.705	1.932	2.095	2.013	1.957	1.549
3	3.543	2.411	1.940	**1.504**	1.655	1.813	1.970	1.904	1.812	1.483	1.408
4	2.794	2.324	**1.587**	1.683	1.724	1.847	1.737	1.553	1.210	1.777	1.102

posed [8]. The main idea of DO consists of removing a number of factors from the data by building an orthogonal model with scores independent of the predicted variables being modelled. This orthogonal part is removed from the original data in order to make a regression model on the remaining part of the data. Orthogonality means that these removed factors should account for as much as possible of the variation in the data not related to the concentration. Therefore, DO is a pre-treatment such that the variation in X that is orthogonal to y is subtracted in order to make a signal correction that does not remove relevant information from the data.

This method is studied in this article on several data sets, one or several DO components are obtained at a time and removed from the original data in order to obtain a better understanding of the variation present in the data.

2. Experimental

For this study, four different data sets were used. The first data set consists of NIR data collected on gasoil derivatives in order to predict the cloud point. The second one consists of polymer data to measure the fluidity index. The third data set consists of NIR data for polyether polyols to model the hydroxyl number. This data set was repeatedly studied for other purposes [9–11]. The fourth data set consists of polymer data to model the viscosity. Each data set was split into a calibration set and a test set by using the Duplex algorithm [12]. All calculations were implemented in MATLAB™ for WINDOWS™ version 5.2 (The MathWorks).

2.1. The cloud point data

This data set consists of 92 gasoil spectra. The instrument used to collect this data was an ABB Bomem MB154S Spectrometer with a resolution of 4 cm^{-1}. A one-point baseline correction at 4780 cm^{-1} was applied. Fifty-two sample spectra were used for calibration, while the remaining 40 were used for model validation.

2.2. The polymer 1 data

This data set consists of 609 spectra measured between 400 and 2500 nm, each 2 nm, collected using

Table 2
RMSECV values using from 1 to 11 PLS components for the polymer 1 data

Number of DO components	Number of PLS components										
	1	2	3	4	5	6	7	8	9	10	11
0	0.548	0.522	0.461	0.411	0.373	**0.369**	0.367	0.366	**0.365**	0.365	0.362
1	0.527	0.467	0.413	0.373	**0.369**	0.368	0.366	**0.365**	0.366	0.362	0.361
2	0.493	0.417	0.373	**0.370**	0.368	0.366	**0.365**	0.366	0.362	0.361	0.366
3	0.431	0.375	**0.371**	0.369	0.367	**0.366**	0.366	0.363	0.361	0.366	0.359

Table 3
RMSECV values using from 1 to 11 PLS components for the gasoil data to model the cloud point

Number of DO components	Number of PLS components										
	1	2	3	4	5	6	7	8	9	10	11
0	2.257	1.947	1.937	1.800	1.548	**1.456**	1.511	1.520	1.469	1.460	1.445
1	2.010	1.991	1.838	1.570	**1.468**	1.520	1.498	1.447	1.411	1.451	1.455
2	2.613	2.184	1.700	**1.620**	1.635	1.548	1.494	1.414	1.455	1.366	1.316
3	2.507	1.723	**1.648**	1.650	1.541	1.442	1.386	1.421	1.323	1.256	1.259
4	1.716	**1.622**	1.633	1.541	1.410	1.357	1.395	1.290	1.224	1.219	1.167

a NIRSystem 6500 in granulated samples. For calculation of a calibration model, 400 sample spectra were used. The remaining 209 sample spectra were used for model validation.

2.3. Polyether polyols

This data set consists of 84 samples measured between 1100 and 2158 nm, each 2 nm. These spectra were measured on a Pacific Scientific 6250 Scanning Spectrometer (NIRSystem, Silver Spring, MD). For calibration, 60 sample spectra were used to model the hydroxyl number in mg KOH/g and the remaining 24 sample spectra were used to validate the model.

2.4. The polymer 2 data

This data set consists of 305 NIR samples of polymer measured at 700 wavelengths to model the viscosity. These spectra were collected using a NIRSystem 5000 in powdered samples. During the building of the model, 18 samples were detected as outliers and they were removed. Finally, 210 sample spectra were used for calibration and the remaining 77 sample spectra were used to validate the model.

3. Results and discussion

The data sets were split into two independent subsets, a calibration set and a test set. The number of factors to retain in the model was calculated on the calibration set, and the test set was used to obtain prediction errors.

The calibration model was obtained in the following way. First, one DO component was obtained, and after subtraction from the raw data, a PLS or PCR model was constructed. The number of PLS or PCR factors was determined by leave-one-out cross-validation. This was repeated for two DO factors, etc. The PLS/PCR models were also obtained for the raw data, i.e. without subtraction of any DO component.

The root-mean-square error (RMSE) of residuals from the final PLS or PCR model for calibration and

Table 4
RMSECV values using from 1 to 14 PLS components for the polymer 2 data

Number of DO components	Number of PLS components													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0.1096	0.0490	0.0414	0.0356	0.0309	0.0272	0.0239	0.0227	0.0198	0.0170	0.0163	0.0145	0.0139	**0.0134**
1	0.0490	0.0413	0.0355	0.0307	0.0270	0.0238	0.0226	0.0196	0.0169	0.0162	0.0144	0.0138	**0.0133**	0.0132
2	0.0490	0.0393	0.0317	0.0277	0.0242	0.0228	0.0200	0.0170	0.0162	0.0144	0.0138	**0.0132**	0.0131	0.0129
3	0.0491	0.0345	0.0293	0.0253	0.0233	0.0210	0.0170	0.0163	0.0143	0.0137	**0.0132**	0.0130	0.0129	0.0127
4	0.0420	0.0331	0.0290	0.0243	0.0225	0.0174	0.0165	0.0143	0.0137	**0.0132**	0.0131	0.0129	0.0127	0.0125

test data sets was calculated by using the following equation:

$$\mathrm{RMSE} = \sqrt{\sum_{n} (\hat{y} - y)^2 / n}$$

where $\hat{y}$ is the value of y predicted for each object at each model complexity considered, y is the response and n is the number of objects from the considered data set.

An internal validation can be performed using the calibration data set. Tables 1–4 show the RMSECV (root-mean-square error in cross-validation) values obtained after the use of PLS. The first row shows the values when no DO component is removed and the boldfaced RMSECV value indicates the optimal complexity of the model that has been determined as the complexity at which the first minimum RMSECV is encountered. The next rows in the table show the RMSECV values when one to four DO components are removed. It is evident from the table that each time a DO component is removed, the optimal number of PLS factors decreases.

One can see in Table 1 that for the polyether polyol data, the optimum RMSECV value using three DO factors and four PLS factors decreases a little (~ 12.2%) compared to PLS without any DO pre-treatment. It is noteworthy that for each model, the sum of the number of DO factors and the number of PLS factors yields an optimal RMSECV constant and equal to 7.

In Table 2 one can see the results for the polymer 1 data. It shows that the RMSECV value for raw data by using PLS without DO for the optimal complexity (9) is 0.365 (or 0.369 with an optimal complexity of 6 if a simple randomization test [13,14] is used to determine this optimal complexity). Each time a DO component is removed, almost the same value is obtained and here also the total complexity remains always the same (9 or 6). The same results are obtained in Table 3 (total complexity of 6) for the cloud point data and in Table 4 (complexity of 14) for the polymer 2 data.

Therefore, in all cases, each time a DO component is removed, the RMSECV value changes but the total dimensionality (DO + PLS) remains the same.

As a first conclusion, one can say that when only internal validation is considered, the results do not improve very much with increasing number of DO's (in our examples only the first one improves as compared to the original data). The total number of factors in the full model (retained PLS-factors + subtracted DO's) is always the same. The complexity of the model does not really change.

An external validation was performed using the test data set. Tables 5–8 show the values for the RMSEP using the optimal PLS-model for each number of DO's. In (Tables 5, 6 and 8), one can see that the RMSEP value does not really change with the total number of factors and Table 7 shows that for this case, the prediction becomes worse with an increasing number of DO components.

Table 5
RMSEP for PLS models for the polyether polyol data

Number of DO components	Number of PLS components	RMSEP
0	7	1.748
1	6	1.747
2	5	1.744
3	4	1.731
4	3	1.715

Table 6
RMSEP for PLS models for the polymer 1 data

Number of DO components	Number of PLS components	RMSEP
0	9	0.376
1	8	0.377
2	7	0.379
3	6	0.379

Table 7
RMSEP for PLS models for the gasoil data to model cloud point

Number of DO components	Number of PLS components	RMSEP
0	6	1.469
1	5	1.474
2	4	1.621
3	3	1.847
4	2	1.836

Table 8
RMSEP for PLS models for the polymer 2 data

Number of DO components	Number of PLS components	RMSEP
0	14	0.0155
1	13	0.0155
2	12	0.0157
3	11	0.0159
4	10	0.0160

DO provides some insight in the data, but no new information was obtained as compared to the information already available from studying PCA and PLS scores and loadings. Consider the polyether polyol data. The PC1–PC2 score plot of these data is shown in Fig. 1. Two main clusters can be seen because different product types are present, with different chemical structures, making the data set inhomogeneous.

The heterogeneity of the data is due to differences in the chemical structure of the samples. A previous study performed by Jouan-Rimbaud et al. [9] showed that PC1 was a descriptor of the type of CH groups, the second PC was a descriptor of the OH groups (and therefore related to the hydroxyl number, the characteristic that is being determined) and that the PC3 described the presence of water in samples. These three PCs explain more than 95% of the variance in the data.

In Fig. 2a and b the first and second PCs when no DO component was removed are plotted vs. the hydroxyl number. For the second PC, the obtained correlation is better as could be expected. Clusters are also evident along both PCs. The first PC explains almost 60% of the variance and the second PC (related to the response) almost 33%. When one DO component was subtracted, most of the existing variance before DO (79.34%) has been removed. PC1 now becomes the PC most correlated with the hydroxyl number with 82% of explained variance compared to 8% for the PC2 (Fig. 3a and b).

Jouan-Rimbaud et al. [9] determined the wavelengths most important in relation to the correlation between the absorbance and the hydroxyl number.

They found that the wavelength with the highest correlation coefficient was 1430, and the effect of applying the DO method can also be seen when the absorbance at this wavelength is plotted against the

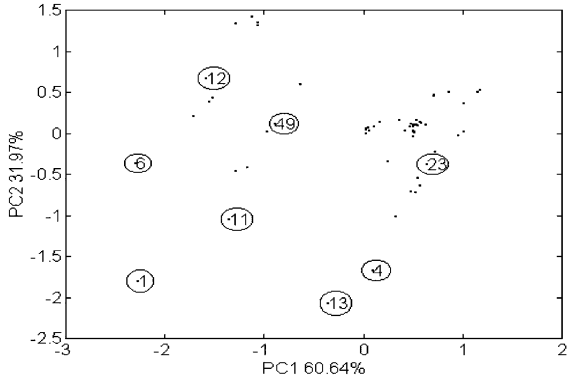

Fig. 1. PC plot (PC1 vs. PC2) for the polyether polyol data. Circled objects: objects discussed in the text.

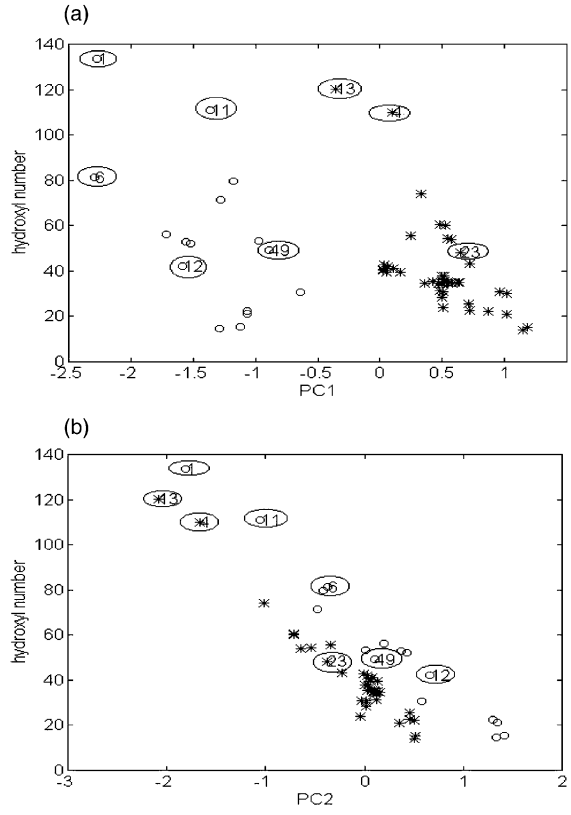

Fig. 2. (○) Objects of cluster1, (∗) objects of cluster2. (a) Hydroxyl number vs. PC1 when no DO component was removed. (b) Hydroxyl number vs. PC2 when no DO component was removed. Circled objects: objects discussed in the text.

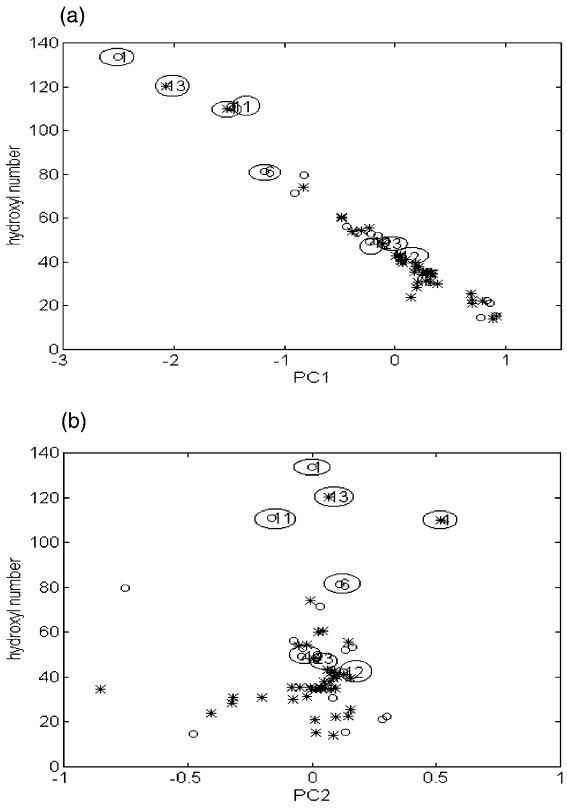

Fig. 3. (○) Objects of cluster1, (∗) objects of cluster2. (a) Hydroxyl number vs. PC1 when one DO component was removed. (b) Hydroxyl number vs. PC2 when one DO component was removed. Circled objects: objects discussed in the text.

hydroxyl number (Fig. 4a and b). The improvement is also evident.

It is interesting to investigate what the effect of the DO method is on the variation in the data. To do this, PCA (Principal Component Analysis) is applied to the original matrix and to the matrix after removing several DO components.

PCA score plots are obtained for the centered raw data and the data after applying DO. In Fig. 1 one can see the score plots for the original data. Two main clusters are evident in the PC1–PC2 score plot. Each time a DO component is removed, a PC plot is obtained and these score plots are also shown in Fig. 5a–c.

After removal of DO1 (first DO component) most of the points corresponding to different clusters merge together in one cluster. To check this grouping, several points belonging to both clusters are outlined to follow their changes. In Fig. 1, objects can be seen in the original data space. Fig. 5a–c shows that after DO, these points merge in the same group. Objects 1, 4, 6, 11 and 13 are extreme points of several elongated clusters in Fig. 1 and after DO, take on extreme positions in the main cluster. The fact that they are extreme in hydroxyl number remains, but the differences, not relevant for hydroxyl number, have disappeared. Objects 12, 23 and 49, for instance, are part of different clusters in the original data space, but their differences disappear after subtraction of one DO. The percentage of variance explained for each PC is shown in the plots and it is easy to see that the first PC becomes more and more important each time a DO component is removed. This was also demonstrated by Andersson [8].

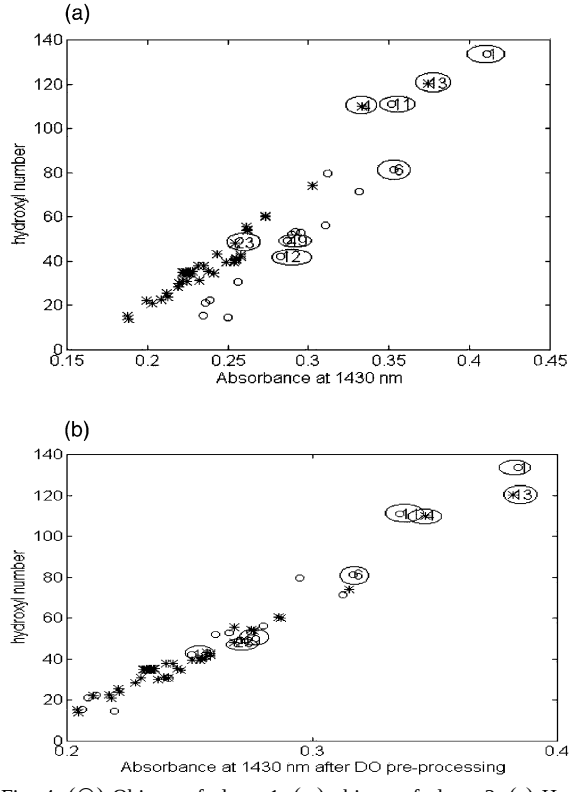

Fig. 4. (○) Objects of cluster1, (∗) objects of cluster2. (a) Hydroxyl number vs. the absorbance at 1430 nm without DO component. (b) The hydroxyl number vs. the absorbance at 1430 nm with one DO component. Circled objects: objects discussed in the text.

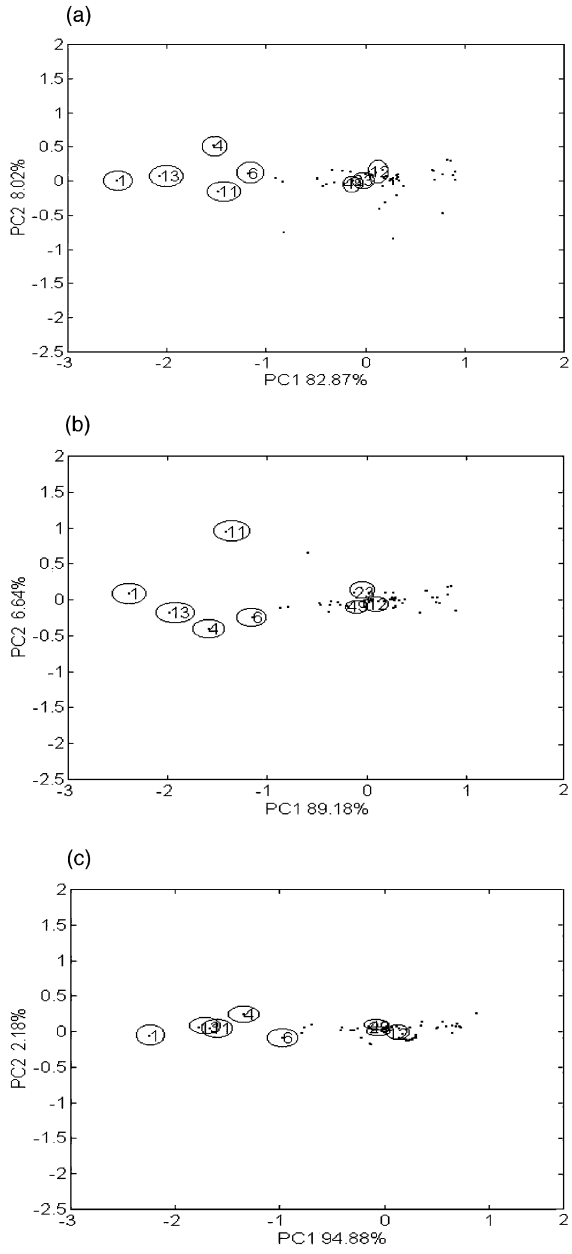

Fig. 5. PC plot (PC1 vs. PC2) for the polyether polyol data. (a) With one DO component, (b) with two DO components, (c) with three DO components. Circled objects: objects discussed in the text.

In Fig. 6 the loading plot for the first PC is shown. This plot allows to identify spectral regions that are important in describing the data. The peak at 1690 nm

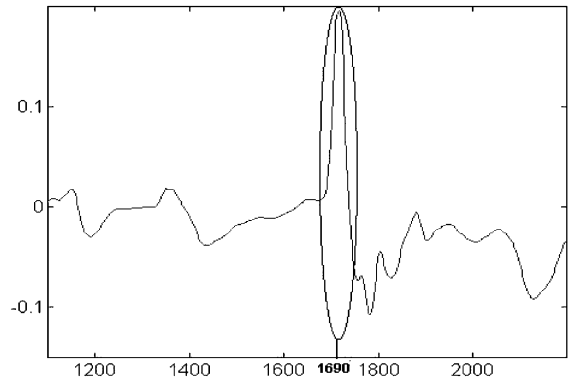

Fig. 6. Loading plot for PC1.

is responsible for the inhomogeneity of the data: the objects belonging to one of the clusters show a peak at this wavelength, the objects belonging to the other cluster do not. The cluster identity of a sample depends on the chemical structure. Absorption at 1690 nm indicates first the CH stretch overtone. The role of this variable is therefore to bring into account the difference between the two clusters.

After the subtraction of one DO component, the variation in the spectra is reduced especially for the wavelength more important for the clustering tendency, the variables that are not relevant for *y* have been removed and differences between samples disappear. The loading plot in Fig. 7 shows high loadings for the peaks related to OH (around 1430 and 2100 nm).

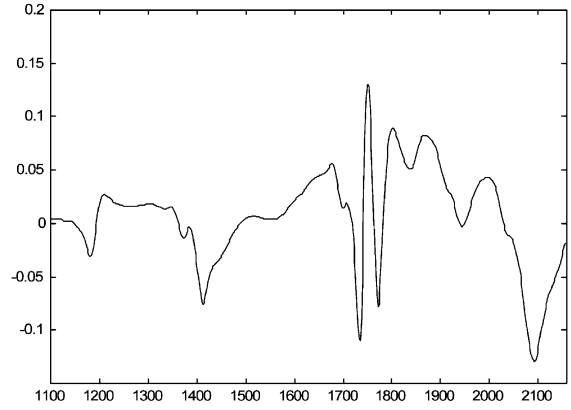

Fig. 7. Loading plot for PC1 after removing one DO component.

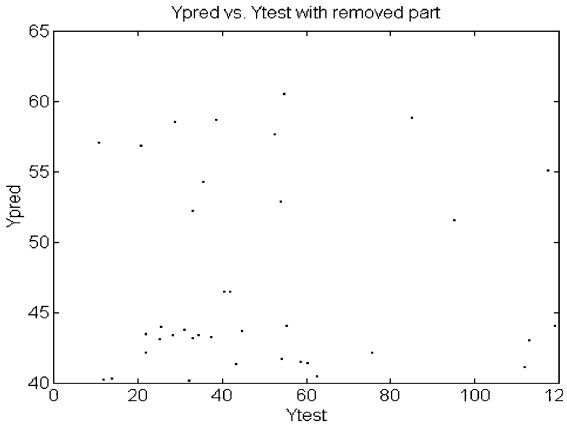

Fig. 8. *Y*pred. vs. *Y*test with the removed part for the polyether polyol data.

In order to verify that the removed part contains only irrelevant information concerning the response, a calibration model was made by using only information in the first DO component. The RMSECV value for this model is 83.9 and Fig. 8 shows the predicted *y* value vs. *y* test. It shows that there is no relationship between them, they are indeed uncorrelated and the DO component represents information not related to *y*.

4. Conclusions

Direct Orthogonalization (DO) shows similar results for the four data sets. The number of PLS factors decreases compared to the number of factors in a PLS without DO pre-treatment, but at the expense of increasing the number of DO factors that are necessary to remove irrelevant information. The total number of factors to be used remains constant for almost the same RMSECV value when only calibration is considered. Using DO with PCR leads to worse RMSECV values compared to PLS. For the same model complexity, more DO components should be removed, and therefore the total number of factors (PCR + DO) increases considerably.

The results for PLS and PCR in prediction do not change or become worse. With DO no better results are obtained than with classical pre-processing methods applied in PLS or PCR regression. Removing *Y*-orthogonal factors allows, however, to obtain a better understanding of the variation present in the data, which may be useful.

References

[1] O.E. De Noord, Chemom. Intell. Lab. Syst. 23 (1994) 65–70.
[2] I.E. Frank, Chemom. Intell. Lab. Syst. 1 (1987) 233–242.
[3] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851–3858.
[4] P. Geladi, D. MacDougall, H. Martens, Appl. Spectrosc. 39 (1985) 491–500.
[5] S. Wold, H. Antti, F. Lindgren, J. Ohman, Chemom. Intell. Lab. Syst. 44 (1998) 175–185.
[6] J. Sjoblom, O. Svensson, M. Josefson, H. Kullberg, S. Wold, Chemom. Intell. Lab. Syst. 44 (1998) 229–244.
[7] T. Fearn, Chemom. Intell. Lab. Syst. 50 (2000) 47–52.
[8] C.A. Andersson, Chemom. Intell. Lab. Syst. 47 (1999) 51–63.
[9] D. Jouan-Rimbaud, D.L. Massart, R. Leardi, O.E. De Noord, Anal. Chem. 67 (1995) 4295–4301.
[10] D. Jouan-Rimbaud, D.L. Massart, O.E. De Noord, Chemom. Intell. Lab. Syst. 35 (1996) 213–220.
[11] V. Centner, D.L. Massart, O.E. De Noord, Anal. Chim. Acta 330 (1996) 1–17.
[12] R.D. Snee, Technometrics 19 (1977) 415–428.
[13] H. van der Voet, Chemom. Intell. Lab. Syst. 25 (1994) 313–323.
[14] H. van der Voet, Chemom. Intell. Lab. Syst. 28 (1995) 315.