Received: 1 October 2007, Revised: 7 December 2007, Accepted: 17 December 2007, Published online in Wiley InterScience: 2008

(www.interscience.wiley.com) DOI: 10.1002/cem.1199

Multi-block methods in multivariate process control[†]

Jarno Kohonen[a]*, Satu-Pia Reinikainen[a], Kari Aaljoki[b], Annikki Perkiö[c], Taito Väänänen[c] and Agnar Höskuldsson[d]

In chemometric studies all predictor variables are usually collected in one data matrix X. This matrix is then analyzed by PLS regression or other methods. When data from several different sub-processes are collected in one matrix, there is a possibility that the effects of some sub-processes may vanish. If there is, for instance, mechanic data from one process and spectral data from another, the influence of the mechanic sub-process may not be detected. An application of multi-block (MB) methods, where the X-data are divided into several data blocks is presented in this study. By using MB methods the effect of a sub-process can be seen and an example with two blocks, near infra-red, NIR, and process data, is shown. The results show improvements in modelling task, when a MB-based approach is used. This way of working with data gives more information on the process than if all data are in one X-matrix. The procedure is demonstrated by an industrial continuous process, where knowledge about the sub-processes is available and X-matrix can be divided into blocks between process variables and NIR spectra. Copyright © 2008 John Wiley & Sons, Ltd.

Keywords: multi-block PLS; priority regression; CovProc; process control; oil refining

1. INTRODUCTION

The amount of data available from processes is nowadays very large and collecting data has become easier due to process computers and automatic instrumentation. Multivariate methods are widely applied to the process control in industry. Even the quality of a product is nowadays more and more found through multivariate methods instead of traditional measurements. Consequently, the information from the multivariate methods can be utilized in several ways. The main fields of application can be divided into two categories; one is to predict the quality of a product in a manner that is reliable and fast. The second category is process control, where deviations from normal conditions, and the reasons for them are wanted to be detected in an early stage, so that corrective actions can be taken on as soon as possible.

Sometimes, however, these tasks are not facile to perform. It can sometimes be difficult to see, which are the variables or conditions that are to be revised in order to correct the process cycle. Data from industry often consist of different types of variables and the numbers of variables in the different types can variate massively. Spectral data, for example near infra-red (NIR), usually have more than thousand variables, whereas the process variables, for example, temperature and pressure, can often be limited to tens or less. Spectral data can be available from different parts of the process, as can be the process variables. If the data are handled in a single matrix **X**, the influence of one block can easily be hidden by the influence of another. The tribulation in this is that the hidden effect could be useful in process control. Also when the number of variables in **X** is large, the monitoring and diagnosing of the process will become tedious. More information of the different variables can be found by dividing the data into meaningful blocks either by the types of variables or by the part of the process they originate from. This can increase the amount of extracted information from the data and lead to more accurate process control. In case of Kourti *et al.* [1], multi-block (MB) based control charts were found as an effective way to assign the cause of deviant behaviour of a batch process.

The predictor variables in **X** can often be partitioned into smaller blocks. Data can be divided into several blocks by the type of variables or according to the part of the process. The matrix **X** often has different types of variables, for example, variables from spectral instruments and process variables. In the case of data from a cement production, Reference [2], the matrix **X** contains 163 variables, which can be divided into four blocks; chemical variables, superficial microstructures, variables describing particle size distribution and process variables. In this way, the contributions of different blocks to the modelling task could be

* *Lappeenranta University of Technology, P.O. Box 20, 53851 Lappeenranta, Finland.*
E-mail: jarno.kohonen@lut.fi

[a] *J. Kohonen, S.-P. Reinikainen*
Lappeenranta University of Technology, P.O. Box 20, 53851 Lappeenranta, Finland

[b] *K. Aaljoki*
Neste Engineering, P.O. Box 310, 06101 Porvoo, Finland

[c] *A. Perkiö, T. Väänänen*
Neste Oil, P.O. Box 310, 06101 Porvoo, Finland

[d] *A. Höskuldsson*
Centre for Advanced Data Analysis, Eremitageparken 301, 2800 Kgs Lyngby, Denmark

[†] *This article was previously published in Volume 22 Issue 3-4 with the DOI: 10.1002/cem.1120.*

studied and the modelling could focus on the important blocks. In the study of a wastewater treatment process, Reference [3], the variables are divided into blocks according to the part of the process; influent part, equilibrium tank and aeration tank. The block method was found efficient in monitoring complex biochemical processes.

Modelling of data in blocks can be handled in numerous ways. One way to handle a situation where there are several blocks of variables is the priority regression where the maximum amount of response variables' variation is explained with one block and the rest with the others. Another way is to utilize MB method. In the MB method, weights are found separately for each block in a manner where the resulting score vectors maximize the covariance with the response variable. The multi-block PLS method is highly applicable to very different types of situations and thus has increased growing interest. The principles behind MB data analysis methods originate from Wold [4] and Wangen and Kowalski [5].

The objective of this paper is to demonstrate the efficiency and difference of multi-block PLS method in relation to normal PLS (partial least squares or projection to latent structure) regression and priority regression. The data in these examples come from an oil refinery plant and consist of NIR measurements of the product, process variables and the quality variable from the product as the response variable.

2. MATHEMATICAL METHODS

2.1. PLS regression

PLS regression has been the major regression technique for multivariate data analysis for a long time. In the industry as in the nature, the interactions between the variables are complex and the available data are usually also noisy. If such data are wanted to explain with a linear model, the PLS method is a supreme choice [6]. PLS is an invaluable tool especially when underlying factors have little or no physical meaning. An important feature of PLS in comparison to simpler methods, for example PCR, is that it takes into account errors both in response matrix and predictor matrix [7]. As in PCA, principal component analysis, the matrices **X** and **Y** are decomposed into matrices of lower dimensions, Equations (1) and (2). However, the score vectors in **T**, Equation (3), are found by rotating **X** by a weight vectors **w** in such a way that the score vectors are as **Y**-relevant as possible, that is, covariance structure is maximized [8].

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{1}$$

where **X** is the predictor matrix of instrumental data; **T** the X-score matrix; **P** the X variable loading matrix and **E** the X residual matrix.

$$\mathbf{Y} = \mathbf{U} \cdot \mathbf{Q}^{\mathbf{T}} + \mathbf{F} \tag{2}$$

where **Y** is the response matrix; **U** the Y-score matrix; **Q** the Y-loading matrix and **F** the Y residual matrix.

$$\mathbf{t}_i = \mathbf{X}_{i-1}\,\mathbf{w}_i \tag{3}$$

where t_i is the score vector; $\mathbf{X}_{i-1}$ the predictor matrix at the ith step and $\mathbf{w}_i$ the weight vector.

PLS regression is an iterative method where at each iteration a weight vector $\mathbf{w}_i$ is determined. When score and loading vectors have been computed, **X** is adjusted by the results, $\mathbf{X}_i = \mathbf{X}_{i-1} - \mathbf{t}_i\mathbf{p}_i^{\mathrm{T}}$.

2.2. CovProc methods

A fairly new approach to modelling is the CovProc (*Covariance Procedure*) method. The method is a combination of PLS method and classical regression methods. In PLS the aim is to balance the fit and the prediction, while the classical methods focus on the fit of the model. For many types of data there are too many variables and it may not be appropriate to use all samples. The aim of CovProc methods is to find the appropriate subset of variables and samples that contribute to the modelling of the response variable. This is done by securing both large score vectors and requiring that the measure being used (measure of fit, of cross-validation etc.) is as large as possible [9].

In the CovProc method used here, the matrix **X** is treated according to the following procedure [9]:

1. The weight vector **w** can for example, be found through PLS regression.
2. The weights are sorted according to their absolute values, $|w_i|$.
3. Variables are included stepwise starting from the highest $|w_i|$. Corresponding $\mathbf{w}_r$,$\mathbf{t}_r$ and the measure used, here fit, $|\mathbf{Y}^{\mathrm{T}}\mathbf{t}_r|^2/(\mathbf{t}^{\mathrm{T}}{}_r\mathbf{t}_r)$, are computed at each step.

   $\mathbf{w}_r = (0,\ldots,0,\ w_i,0,\ldots,\ w_j,\ldots,0,\ w_k,0,\ldots),\ \mathbf{t}_r = \mathbf{X}\cdot\mathbf{w}_r$

   The non-zero values in $\mathbf{w}_r$ are the r numerically largest values. Variables are selected based on the maximum of the fit value found.
4. **X** is adjusted by the corresponding score vector, $\mathbf{t}_r$.
5. More variables are selected by starting again at Step 1.

Here the CovProc method is used only as a pre-processing method for the other methods with the purpose of finding appropriate subset of variables and samples that should be used in the analysis. More information on the method can be found in the Reference [10].

2.3. Priority regression

Industrial processes produce today a lot of information. Data can be available in different sources and in different types. An advanced way to handle this kind of data is to model separately these blocks as an alternative to unfolding them into PLS. In priority analysis each variable is given a priority number, for example, 1, 2,3. The regression analysis is then carried out for the variables with priority 1. When the modelling task is no more improved using the variables in priority 1 or no more significant score vectors are found, the second block is introduced and the modelling task is moved to the second block with priority number 2. This is done as long as all the variables have been included. In priority regression, the weights of the other variables are assumed zero, and only the variables in the current priority group are assumed to be non-zero. Thus, the score vectors found refer only to the current block of variables [11].

Priority regression may assist the scientists in assessing the importance of the variables. Priority regression is useful when there are groups of variables and the modelling task is wanted to be studied in light of these groups. It is often useful to apply priority regression to diagnostics work, medical, engineering and economic cases. There are various questions, which may be of interest for the engineer [11]:

1. From what stage can the quality be adequately predicted?
2. When has the process (in stages) been stabilized?
3. At what stage of the process can the remaining process values be predicted?
4. How well the quality values can be predicted from the process variables?
5. How the choice of the initial process values at stage 1 can be improved? [11]

In latent structure regression, the regression analysis is based on the score vectors, which can be calculated according to Equation (3). In priority regression the weights are restricted to the data block that is being treated. It is possible to show that regression coefficients are linear combinations of the weights. Therefore, when working with priority group 1, only those regression coefficients are calculated. These coefficients will change when later groups are included. In the computations, matrix **R** is computed such that $\mathbf{T} = \mathbf{XR}$. From Equation (4), the regression coefficients can be seen for the process variables. The method is explained more thoroughly in Reference [11].

$$\mathbf{y}_e = \mathbf{TQ}^{\mathrm{T}} = \mathbf{X}(\mathbf{RQ}^{\mathrm{T}}) \tag{4}$$

2.4. Selection of score vectors

In the presented methods the score vectors of a data block are selected as long as they are 'significant'. A criterion that is used is the following one: select a score vector $\mathbf{t}_{A+1}$, if it satisfies the criterion

$$(y^{\mathrm{T}}\mathbf{t}_{A+1})^2/(\mathbf{t}_{A+1}^{\mathrm{T}}\mathbf{t}_{A+1}) > s_{A+1}^2(2N-1)/(N+A) \tag{5}$$

Here $s_{\mathrm{A}+1}^2$ is the residual variance of the quality variable, when it has been adjusted by $\mathbf{t}_{A+1}$. The theoretical motivation for this criterion is shown in Reference [2]. The inequality approximately amounts to saying that the *t*-statistic for the significance of the score vector, $\mathbf{t}_{A+1}$, should be larger than $\sqrt{2}$. If it is smaller, the score vector does not improve the prediction of the response variable.

This criterion is a very weak one, because it only excludes a score vector if Equation (5) is not satisfied. However, for the purposes of the presented cases this criterion was used so that all potential score vectors were included. Afterwards the situation can be analyzed closer, and for example, cross-validation can be used to remove score vectors among those selected ones. There may be some problems in using cross-validation in determining score vectors as pointed out in References [12–14]. But this issue is not considered in detail here.

2.5. Multi-block methods

On-line process data can reveal changes early allowing corrective actions. In industrial processes, such as presented in this case, different types of information can be obtained. The amounts of information between different sources can alter enormously. The number of process control variables is usually relatively limited, maximum couple tens, whereas the variables obtained from optical instruments can be huge, typically over a thousand. When the weight vector is calculated in for example, PLS, the variables in each block share the same importance. This can be a disadvantage, when from the process control point of view it would more lucrative to obtain as much information as possible from the process variables, which would allow early corrective actions and could avert gargantuan economic losses.

In the literature multi-block PLS has been introduced, which can be seen as a similar method to PLS, only exception being that the data are being divided into several blocks. Several variants of the algorithms have been introduced, but the usual approach is to compute weights and score vectors separately for each block so that the covariance between response and the score vectors is maximized. Obtained score vectors can be used to form a super score vector, $\mathbf{t}_S$, as in the case in Reference [3].

In PLS regression the task is to find a weight vector **w** in such a way that the score vector $\mathbf{t} = \mathbf{Xw}$ so that $|\mathbf{Y}^{\mathrm{T}}\mathbf{t}|^2$ is maximized, whereas in the MB method weight vectors, $\mathbf{w}_1$ for $\mathbf{X}_1$ and $\mathbf{w}_2$ for $\mathbf{X}_2$, are found in a manner that the score vectors $\mathbf{t}_1 = \mathbf{X}_1\mathbf{w}_1$ and $\mathbf{t}_2 = \mathbf{X}_2\mathbf{w}_2$ have the property that $|\mathbf{Y}^{\mathrm{T}}(\mathbf{t}_1 + \mathbf{t}_2)|^2$ is maximized. When the score vectors have been found, the **X**-matrices are adjusted for the score vectors as shown in following equations:

$$\mathbf{X}_1 \leftarrow \mathbf{X}_1 - \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}}, \text{ with } \mathbf{p}_1 = \mathbf{X}_1^{\mathrm{T}}\mathbf{t}_1/(\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1) \tag{6}$$

$$\mathbf{X}_2 \leftarrow \mathbf{X}_2 - \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}}, \text{ with } \mathbf{p}_2 = \mathbf{X}_2^{\mathrm{T}}\mathbf{t}_2/(\mathbf{t}_2^{\mathrm{T}}\mathbf{t}_2) \tag{7}$$

The $\mathbf{t}_1$-vectors will be mutually orthogonal and so will the $\mathbf{t}_2$-vectors. But the $\mathbf{t}_1$ vectors will not be orthogonal to the $\mathbf{t}_2$ vectors. Therefore, at each step it is needed to evaluate each score vector to see which contributes to the modelling task. This is done by selecting the score vector that contributes most in terms of fit to the quality variable, the *y*-variable. The quality variable, **y**, is adjusted for this score vector, and a new score vector is found. This selection of score vectors is stopped, if none of the remaining ones satisfy the criterion in Equation (5).

The presented MB methods differ from the ones in the literature. A score vector is computed for each data block, and these score vectors are used in the modelling task. The score vectors of the data blocks allow interpretation of the role of the corresponding data blocks. However, the present method tend to select too many score vectors, because each data block gets one or more score vectors. Thus, the advantage in this method compared to PLS is not in the better predictions, but rather in the improved interpretability and usage of the model.

2.6. Review of methods

In Figure 1, a schematic view is presented for the four methods in this paper. The methods are briefly discussed here.

The Figure 1 illustrates the situation at each iteration or step. In PLS regression the weight vector **w** is found so that $|\mathbf{Y}^{\mathrm{T}}\mathbf{t}|^2 = |\mathbf{Y}^{\mathrm{T}}\mathbf{Xw}|^2$ is maximized. In the CovProc method, the part of **X** that should be used in the analysis is determined. The method can be used as a pre-processing to find good values of **w** in PLS or other regression methods. It can also be used to investigate if there is a redundancy in the data for the method that has been used. In priority regression the weights are restricted to a block of given priority. Other weights are set to zero. In the MB method, here illustrated by two **X**-blocks and two **Y**-blocks, each **X**-block has a role like in PLS regression. The weights $\mathbf{w}_1$ and $\mathbf{w}_2$ are found such that the score vectors $\mathbf{t}_1$ and $\mathbf{t}_2$ are good in describing the response matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

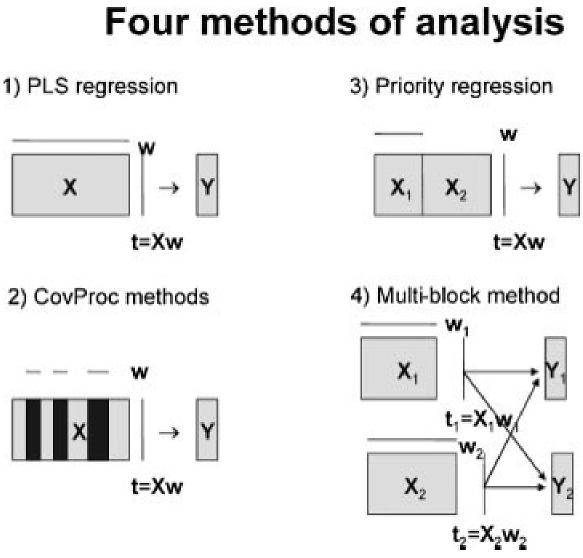

Figure 1. Schematic illustration of the presented four methods [2].

Table I. R^2 of PLS models: regression of 568 samples for the Y. Model (1) NIR data (1397 descriptor variables), and Model (2) NIR and process data (1397 + 8 descriptor variables)

No.	Model 1		Model 2	
	$\Sigma\|\Delta\mathbf{X}\|^2$	$\Sigma\|\Delta\mathbf{Y}\|^2$	$\Sigma\|\Delta\mathbf{X}\|^2$	$\Sigma\|\Delta\mathbf{Y}\|^2$
1	85.669	9.964	85.046	10.228
2	94.520	61.036	94.168	61.271
3	97.279	68.565	96.922	69.371
4	98.015	72.972	97.633	74.374
5	99.223	74.471	98.871	75.938
6	99.525	75.915	99.194	77.576
7	99.667	76.544	99.319	78.754
8	99.745	77.228	99.409	79.726
9	99.764	79.252	99.494	80.454
10	99.801	79.911	99.560	81.243

3. DATA AND THE REFINERY PROCESS

The presented methods are illustrated using a dataset from an oil refinery plant. Oil refining process consists of several different units. They are used to modify the chemical structure of the fractions of crude oil and to improve the product properties. A variety of raw materials, that is, several kinds of crude oils, can be fed to the process and therefore accurate and reliable process control and modelling of the quality of products is a paramount issue. Therefore, the mathematical procedures should be automated and provide the accurate estimates within few seconds for process optimization.

There are seven on-line NIR instruments at the plant, and they provide 1000 spectra daily. Based on the NIR data, over 200 models are used to predict approximately 85 000 different property values, which are used for automatic on-line process control.

In the present example, only one product quality variable **Y** is modelled using data from one NIR instrument and process variables from the unit process at issue. The reference **Y** values are determined in the laboratory, and they are used for calibration. The NIR data consist of measurements over a 6-month period. The NIR spectra are pre-treated: normalized and baseline corrected. The data matrices **X** and **Y** have 568 samples, each of which has eight process variables and 1397 variables from the spectrum. With the CovProc method, however, the original matrix **X** becomes significantly smaller.

4. RESULTS AND DISCUSSION

4.1. Preliminary analysis

PLS method is widely used for predictions based on NIR spectra. A common way to handle situations, where there are different types of variables, is to handle all the variables in a single **X** matrix. If the variables can be, in a reasonable way, divided into different blocks, it is more informative to use a method which handles the blocks separately.

Results from an analysis of the original data using PLS-regression is shown in Table I. Model 1 is found using only NIR data and model 2 is found using both the NIR and process variables with an unfolded PLS. NIR spectra alone do not give satisfactory results. Only less than 80% of the variation of **Y** can be explained. In model 2, process variables give only a slight improvement to this. Cross-validation suggests that there should be ten latent variables in the model. The number is clearly too high, which can be a sign of non-stationarity [10]. The modelling task cannot be considered satisfactory in this way either.

There are ways to improve the results of the modelling task. Analysis of the score vectors, Figure 2, reveals that the first score vector shows an increasing value over the whole time span, negative in the beginning and becoming positive at the middle of the period. A run test for randomness is highly significant for all four score vectors. Especially score vectors 1 and 3 are above and

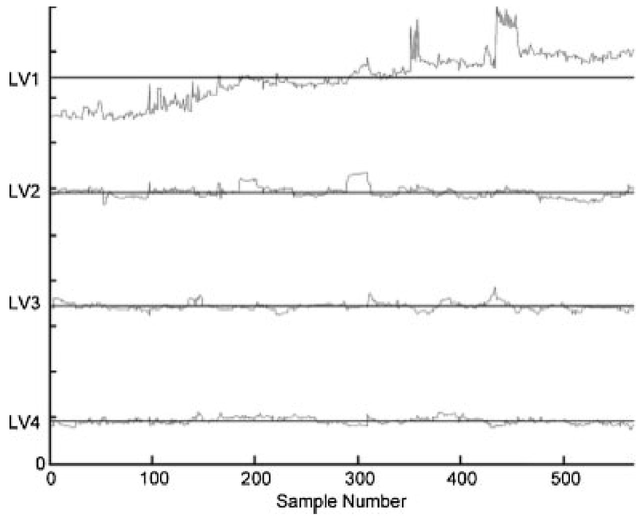

Figure 2. First four score vectors of PLS model having 568 samples and 1397 + 8 descriptor variables (Model 2 in Table I). This figure is available in colour online at www.interscience.wiley.com/journal/cem

Table II. CovProc + PLS regression of 68 most recent samples. Model (3) NIR data (120 descriptor variables), Model (4) process data (8 descriptor variables) and Model (5) NIR and process data (120 + 8 descriptor variables)

No.	Model 3		Model 4		Model 5	
	$\Sigma\lvert\Delta\mathbf{X}\rvert^2$	$\Sigma\lvert\Delta\mathbf{Y}\rvert^2$	$\Sigma\lvert\Delta\mathbf{X}\rvert^2$	$\Sigma\lvert\Delta\mathbf{Y}\rvert^2$	$\Sigma\lvert\Delta\mathbf{X}\rvert^2$	$\Sigma\lvert\Delta\mathbf{Y}\rvert^2$
1	78.961	51.483	45.812	42.910	74.969	51.964
2	91.538	67.559	72.188	74.770	86.786	69.553
3	96.351	76.291	87.284	76.545	91.627	80.643
4	97.942	81.383	93.899	76.944	95.373	85.058
5	98.620	83.900	95.865	77.657	95.919	89.056
6	98.967	85.705	98.202	77.962	97.054	90.050
7	99.205	87.917	99.981	78.029	97.508	91.990
8	99.294	90.472	100.000	78.730	97.990	93.455
9	99.349	92.183			98.667	94.020
10	99.426	92.947			98.896	94.708

Table III. Priority regression: variation extracted from X and Y. Block 1 is the process variables, and block 2 is the NIR spectra

Model 6			
No.	Block	$\Sigma\lvert\Delta\mathbf{X}\rvert^2$	$\Sigma\lvert\Delta\mathbf{Y}\rvert^2$
1	1	38.174	42.910
2	1	51.207	74.770
3	1	61.029	76.545
4	2	78.475	80.152
5	2	90.977	84.114
6	2	97.008	85.880
7	2	97.555	88.482
8	2	98.292	89.610
9	2	98.534	91.593
10	2	98.794	92.670
11	2	98.954	93.257
12	2	99.021	93.969

below the zero line for a longer period of time. It can thus be said that there is a systematic variation in the score vectors over the time range considered. Results may be improved by selecting correct amount of samples and variables into the modelling.

An application of the CovProc method suggests that only some 68 time points (samples) should be used and only around 120 variables of the NIR data. The involved samples are the most recent ones, where the process seems rather stable. The results of modelling after the selection of variables and samples are shown in Table II. If only NIR data are used, 92.95% of the variation of **Y** can be explained. If only the eight process variables are used, 78.73 % is explained. R^2 of 94.71% can be achieved, if all 128 variables are used. The results are still not quite satisfactory. The improvement seen in model 5 is not very significant. As discussed earlier, the effect of the smaller descriptor block can easily get covered by the larger block. In this case the effect of process variables are hidden by the NIR data.

4.2. Priority PLS regression

It can be seen that three score vectors of the process variables could explain 76.55% of the variation of the quality variable. This can be utilized by first carrying out PLS regression for the process variables and then continue with the variables of the NIR data. This is called priority PLS regression. The weight vectors in PLS regression are restricted to be non-zero for the process variables and zero for the NIR variables. This is done for three steps. The fourth score vectors is not significant. At later steps the weight vectors are restricted to be zero for the eight process variables and non-zero for the NIR variables. A successful application of priority regression in multivariate process control is presented in Reference [11].

The results of priority PLS regression is shown in Table III. The method selects three significant score vectors from the process data block and eight from the NIR data block. In comparison with Table I, it can be seen that two additional score vectors are needed in this priority regression model.

However, this approach can reveal more information from the process variables than the unfolded PLS method. The total variation explained by the three PLS score vectors is $R^2 = 76.55\%$. The regression coefficients change when the second block is introduced, but the regression coefficients associated with the first three score vectors are computed using Equation (4). For the original un-scaled data regression equation is shown in Equation (8). This equation can be studied to find out, which variables should be adjusted to give the desirable quality value. The significance of the regression coefficients could be checked with any standard statistical method, but that is not considered here.

$$\begin{aligned} \mathbf{y}_e = & -0.23\mathbf{x}_1 - 5.47\mathbf{x}_2 + 16.66\mathbf{x}_3 - 19.25\mathbf{x}_4 + 7.55\mathbf{x}_5 \\ & + 5.61\mathbf{x}_6 + 53.14\mathbf{x}_7 + 9.34\mathbf{x}_8 \end{aligned} \tag{8}$$

4.3. Multi-block modelling

In MB modelling the score vectors for each block gets equal weights in the modelling task. The data for the process variables are given by $\mathbf{X}_1$, 68 × 8 matrix, and the NIR data, $\mathbf{X}_2$, 68 × 120. The results of the iterations are shown in Table IV.

At the first step the score vector $\mathbf{t}_2$ is more important than $\mathbf{t}_1$. It explains 51.5% of **y**, and it has the squared size of $|\mathbf{t}|^2 = 6.840$. The covariance $(\mathbf{t}_2^T\mathbf{y})$ is 90.9 % of the total covariance. The residual standard deviation after selecting $\mathbf{t}_2$ is $s = 0.086$. The last column reports the *t*-value for the significance of the score vector. Now **y** and $\mathbf{t}_1$ are adjusted for $\mathbf{t}_2$ and then $\mathbf{t}_1$ evaluated. It gives 5.0% reduction in the variation of **y**. Its (reduced) squared size is $|\mathbf{t}_1|^2 = 0.237$, the covariance, $(\mathbf{t}_1^T\mathbf{y})$ (here both are reduced), is 1.13%. After 12 steps no more score vectors appear significant. At steps 3, 4, 7, 9, 10, 11 and 12 only $\mathbf{t}_2$ are selected.

The contribution of $\mathbf{t}_1$-vectors is

$$4.957\% + 4.357\% + 1.079\% + 0.536\% + 0.990\% = 11.92\%$$

The contribution of the $\mathbf{t}_2$-vectors is 84.14%. The score matrices $\mathbf{T}_1$ and $\mathbf{T}_2$ represent the results of the optimisation procedure. They can be studied closer, both their variation and how they relate to the quality variable. If regression is performed on $\mathbf{T}_1$ alone, it can describe 75.7% of the variation of the quality

Table IV. Results from the multi-block modelling

Step no.	Block no.	Score vector	Explained variance	Cumulative explained variance	Squared size	Covariance	Residual standard deviation	*t*-value
			$\|\Delta\mathbf{Y}\|^2$	$\Sigma\|\Delta\mathbf{Y}\|^2$	$\|\mathbf{t}\|^2$	$(\mathbf{t}^T\mathbf{y})$	*s*	
1	2	1	51.483	51.483	6.840	90.864	0.086	8.369
1	1	1	4.957	56.440	0.237	1.131	0.086	2.577
2	2	2	17.638	74.078	1.412	11.298	0.063	6.701
2	1	2	4.357	78.435	0.272	1.697	0.063	3.305
3	2	3	3.949	82.385	0.456	5.253	0.052	3.847
4	2	4	3.436	85.821	0.228	1.506	0.046	3.999
5	2	5	2.163	87.983	0.117	0.630	0.043	3.446
5	1	5	1.079	89.062	0.104	1.006	0.043	2.416
6	2	6	2.254	91.316	0.057	0.145	0.036	4.139
6	1	6	0.536	91.852	0.019	0.064	0.037	2.003
7	2	7	1.206	93.058	0.063	0.325	0.032	3.386
8	1	8	0.990	94.049	0.005	0.002	0.030	3.314
8	2	8	0.423	94.471	0.025	0.151	0.029	2.229
9	2	9	0.580	95.051	0.019	0.063	0.027	2.782
10	2	10	0.459	95.510	0.017	0.059	0.026	2.596
11	2	11	0.291	95.801	0.015	0.081	0.025	2.139
12	2	12	0.260	96.060	0.008	0.025	0.024	2.085

variable. $\mathbf{T}_2$ alone can describe 87.9%. The results of the estimation are shown in Figure 3. Both figures show a fairly clear linear relationship.

It was noted in the beginning that the process variable data, $\mathbf{X}_1$, in a way 'disappear', when used together with $\mathbf{X}_2$. Part of the reason is that $\mathbf{X}_1$ contains 8 variables, while $\mathbf{X}_2$ 120. Each of the eight process variables receive approximately the same weights as each of the 120 NIR variables, and thus will be weighted down. Another reason is that $\mathbf{X}_1$ describes very little of the quality variable, if it has been reduced by what can be described by $\mathbf{X}_2$.

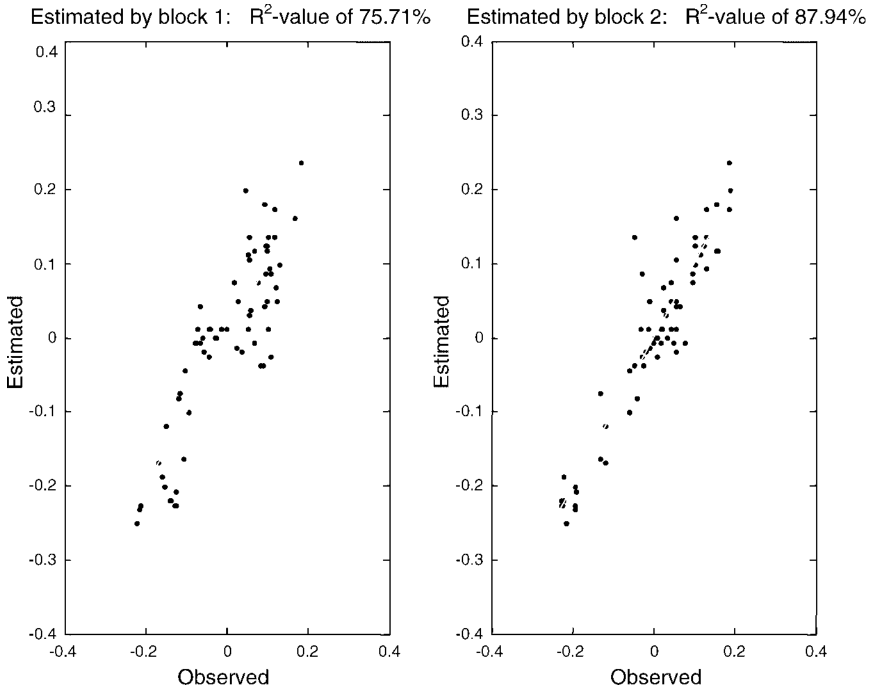

Figure 3. Plots of quality variable *y* (*x*-axis) versus estimated *y*-values. The figure on the left is based on block 1 (process variables) and to the right on block 2 (NIR data). This figure is available in colour online at www.interscience.wiley.com/journal/cem

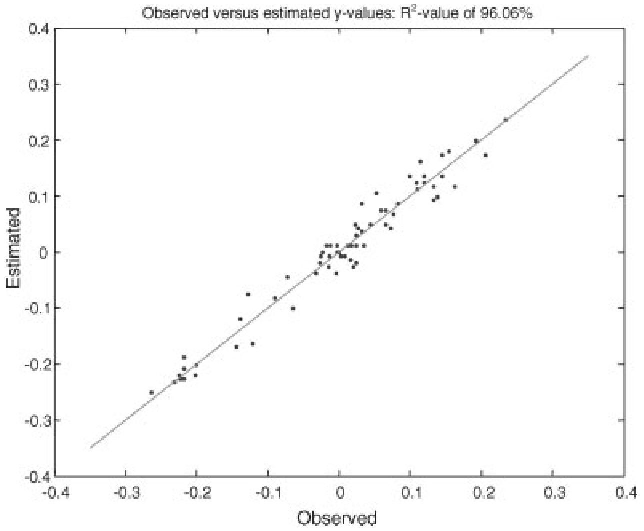

Figure 4. Plot of observed quality variable *y* (*x*-axis) versus the estimated *y*-value based on the multi-block method. This figure is available in colour online at www.interscience.wiley.com/journal/cem

In the MB method $\mathbf{X}_1$ and $\mathbf{X}_2$ play an equal role in the sense that at each step they both provide with a score vector. The two score vectors, $\mathbf{t}_1$ and $\mathbf{t}_2$, are then evaluated. This gives the contribution that each block has to the modelling task. It makes it possible to carry out analysis of the score and loading vectors, and follow how the score vectors relate to the quality variable. It is often the case, that more weight and possibilities to work with the process variables data, $\mathbf{X}_1$, is enabled this way. The results from the multi-block PLS modelling are shown in Figure 4. It can be seen from the figure that the modelling has given satisfactory results.

The disadvantage of MB methods is that there are many score vectors that are suggested and may be found significant. Here a score vector is considered significant, if the *t*-value is significant. This normally gives a slight overfitting. It is probable that the last two $\mathbf{t}_2$-score vectors are not contributing to the modelling task, if evaluated by cross-validation. But this is not considered closer here.

Another important issue is concerning the prediction aspect of the MB method. Constraints are placed on the score vectors, when introducing blocks. The score vectors must come from appropriate data blocks. Furthermore, the MB method tends to select, after cross-validation, more score vectors than for example PLS. In PLS (eventually supplemented by a variables selection method like e.g. the CovProc method) the score vectors are selected freely within **X**. In conclusion, improved interpretability of the MB solution is achieved at the expense of reduced quality of predictions of the **Y**-variables.

5. CONCLUSIONS

Common way in chemometric modelling is to select all **X**-variables into one matrix. This procedure can often lead to a situation where the effect of a sub-process may vanish. This can happen especially, when sub-processes have very different number of variables. In industry as in the nature it is not an unusual situation that data should be divided into different blocks according to their type or to where they are from. If there is, for instance, mechanic data from one process and spectral data from another, it may be impossible to detect the influence of the mechanic sub-process. The loss of information from blocks due to small number of variables could be averted using methods such as priority regression and the multi-block PLS.

In this paper, different methods of modelling are applied to the oil refinery data set. The data consist of NIR data and process control variables. As it is shown, the influence of process control variables is easily distinguished by the proportionally large NIR data. The knowledge of the influence of process control parameters would be beneficial to the process control. The natural division of data is clear in this case, the control variables should be in one block and the NIR data in another. The CovProc method is also applied to NIR data in order to reduce the matrix and to select the most important variables and 'stable' samples. MB method was found to be the most suitable for modelling this type of data. Process variables were utilized in their full extent and thus the information for process control could be maximized.

Acknowledgements

The authors would like to thank the Finnish Funding Agency for Technology and Innovation (TEKES) for financial support.

REFERENCES

1. Kourti T, MacGregor JF. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chem. Int. Lab. Syst.* 1995; **28**: 3–21.
2. Höskuldsson A, Svinning K. Modelling of multi-block data. *J. Chemom.* 2006; **20**: 376–385.
3. Choi SW, Lee I. Multiblock PLS-based localized process diagnosis. *J. Proc. Cont.* 2005; **15**: 295–306.
4. Wold H. Soft modelling. The basic design and some extensions. In Systems Under Indirect Observations vol. II, Jöreskog K-G, Wold H (eds). North-Holland: Amsterdam, 1982.
5. Wangen L, Kowalski B. A multiblock PLS algorithm investigating complex chemical systems. *J. Chemom.* 1989; **5**: 3–20.
6. Eriksson L, Andersson PL, Johansson E, Tysklind M. Megavariate analysis of environmental QSAR data. Part I—A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). *Mol. Divers.* 2006; **10**: 169–186.
7. Brereton RG. Chemometrics: Data Analysis for the Laboratory and Chemical Plant. John Wiley & Sons: United Kingdom, 2006; 298–303.
8. Martens H, Martens M. Multivariate Analysis of Quality: an Introduction. John Wiley & Sons: United Kingdom, 2001.
9. Reinikainen S-P, Höskuldsson A. COVPROC method: strategy in modeling dynamic systems. *J. Chemom.* 2003; **17**: 130–139.
10. Reinikainen S-P, Aaljoki K, Höskuldsson A. On non-stationarity of dynamic systems. *Chemom. Int. Lab. Syst.* 2004; **73**: 119–129.
11. Reinikainen S-P, Höskuldsson A. Multivariate statistical analysis of a multi-step industrial processes. *Anal. Chim. Acta* 2007; **595**: 248–256.
12. Wiklund S, Nilsson D, Eriksson L, Sjöström M, Wold S, Faber K. A randomization test for PLS component selection. *J. Chemom.* 2007; **21**: 427–439.
13. Faber NM, Rajkó R. How to avoid over-fitting in multivariate calibration—the conventional validation approach and an alternative. *Anal. Chim. Acta* 2007; **595**: 98–106.
14. Gómez-Carracedo MP, Andrade JM, Rutledge DN, Faber NM. Selecting the optimum number of partial least squares components for the calibration of attenuated total reflectance-mid-infrared spectra of undesigned kerosene samples. *Anal. Chim. Acta* 2007; **585**: 253–265.