JOURNAL OF CHEMOMETRICS
J. Chemometrics 2006; **20**: 376–385
Published online 27 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/cem.1011

Modelling of multi-block data

Agnar Höskuldsson[1]* and Ketil Svinning[2]
[1]IPL, DTU, Bldg 358, DK_2800 Kgs Lyngby, Denmark
[2]NORCEM A/S, R&D Department, N-3950 Brevik, Norway

Received 23 March 2006; Revised 29 June 2006; Accepted 15 July 2006

Here is presented a unified approach to modelling multi-block regression data. The starting point is a partition of the data X into L data blocks, $X = (X_1, X_2, \ldots, X_L)$, and the data Y into M data-blocks, $Y = (Y_1, Y_2, \ldots, Y_M)$. The methods of linear regression, $X \rightarrow Y$, are extended to the case of a linear relationship between each X_i and Y_j, $X_i \rightarrow Y_j$. A modelling strategy is used to decide if the residual X_i should take part in the modelling of one or more Y_js. At each step the procedure of finding score vectors is based on well-defined optimisation procedures. The principle of optimisation is based on that the score vectors should give the sizes of the resulting Y_js loading vectors as large as possible. The partition of X and Y are independent of each other. The choice of Y_j can be X_j, $Y_i = X_i$, thus including the possibility of modelling $X \rightarrow X_i$, $i = 1, \ldots, L$. It is shown how these methods can be extended to a network of data blocks. Examples of the optimisation procedures in a network are shown. The examples chosen are the ones that are useful to work within industrial production environments. The methods are illustrated by simulated data and data from cement production. Copyright © 2007 John Wiley & Sons, Ltd.

KEYWORDS: multi-block data; regression; optimisation of loadings; network of data blocks; regression in networks

1. INTRODUCTION

Traditionally the chemometricians mainly work with one data matrix **X** or a regression situation where there are given two data matrices, **X** and **Y**. The rows of **X** represent the samples or objects measured and in the regression case the rows of **Y** represent the response values corresponding to the sample, $X \rightarrow Y$. In the empirical work, it is often needed to look at the data in a more structured form and analyse data accordingly. It might be natural to divide **X** and **Y** for instance according to variables (columns), for instance, $X = (X_1, X_2, X_3)$ and $Y = (Y_1, Y_2)$. It can be useful to analyse the data in the light of this structure. May be X_1 is weak to use and contributes only with one score vector, while others may give several score vectors. Furthermore, it might be that Y_1 can be modelled, while Y_2 is difficult to handle. The data may be process measurements, where part of the data is measured before some others. This might result in that a natural description of data is $X_1 \rightarrow X_2 \rightarrow Y$, where samples of X_2 might depend on what was chosen in X_1, and the final result in **Y** depends on both what was selected in X_1 and the derived choice in X_2. An important issue is how to choose the values of intermediate samples (of X_2), such that the prediction of the response sample (of **Y**) is as reliable as possible. This is analysed closer.

In this paper, a unified approach is presented for modelling data of the form $X = (X_1, X_2, \ldots, X_L)$ and $Y = (Y_1, Y_2, \ldots, Y_M)$. The starting point may be modelling from each X_i to every Y_j. A modelling strategy is used to drop a relationship, if a data block X_i does not contribute to modelling of Y_j. The same algorithm and criteria are used in the case there is only an **X** matrix, $X = (X_1, X_2, \ldots, X_L)$. In this case **Y** is chosen as **X**, $Y_i = X_i$. The importance of this approach is due to that the technique of regression analysis is extended to multi-block situation. Cross-validation is carried out in a similar way as in linear regression. If there is no block structure, the algorithm reduces to a linear regression if both **X** and **Y** are present and to Principal Component Analysis (PCA) type of analysis if only **X** is present.

The criteria used are based on maximising the resulting loading vector(s) [1]. The algorithm is presented for the case of three **X**-matrices, $X = (X_1, X_2, X_3)$ and two **Y**-matrices, $Y = (Y_1, Y_2)$, where not all **X**s are used to describe all **Y**s. In Section 6 it is shown what minimal requirements must be satisfied in order that X_i continues to describe Y_j. The algorithm is applied to two types of data, simulated data, where the results are known beforehand, and data from a cement production. Then it is shown how this approach can be extended to more general decomposition of data. The focus of the examples is on data structures that can be found in production environments.

*Correspondence to: A. Höskuldsson, IPL, DTU, Bldg 358, DK_2800 Kgs Lyngby, Denmark.
E-mail: ah@ipl.dtu.dk

In this paper, it is not considered how $\mathbf{X}$ and $\mathbf{Y}$ can be partitioned in order to obtain models of importance. Partitioning can often be found by grouping the variables according to location or activities. For example, variables at one location can be those that generate one of the data blocks, $\mathbf{X}_i$. Partitioning of $\mathbf{X}$ can also be based on the analysis of data. The HELP procedure of Kvalheim [2,3] can be used for this analysis. The HELP procedure can be viewed as follows. Let $\mathbf{w}$ be a weight vector, $\mathbf{t}=\mathbf{X}\mathbf{w}$ the associated score vector. Then, if $|\mathbf{t}|\neq 0$, the reduced matrix $\mathbf{X}_r$ given by

$$\mathbf{X}_r = \mathbf{X} - \mathbf{t}\,\mathbf{p}^T, \quad \text{with } \mathbf{p} = \mathbf{X}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$$

will have a rank which is one less than the that of $\mathbf{X}$. This is independent of the choice of $\mathbf{w}$ as long as $\mathbf{t}$ is not the zero vector. If a plot of column i of $\mathbf{X}$, $\mathbf{x}_i$, against column j, $\mathbf{x}_j$, shows a straight line, it indicates that corresponding columns of $\mathbf{X}_r$ are approximately zeros. The reason is that in this case we are plotting $p_i\mathbf{t}$ against $p_j\mathbf{t}$. Furthermore, small values of p_is indicate that a part of $\mathbf{X}$ is not taking part in the modelling task. This procedure can be extended to weights on both row and columns of $\mathbf{X}$. Let $\mathbf{w}$ be a weight vector for columns and $\mathbf{v}$ for rows, the resulting score vector is $\mathbf{t}=\mathbf{X}\mathbf{w}$ and loading vector is $\mathbf{p}=\mathbf{X}^T\mathbf{v}$. Then, if $\mathbf{v}^T\mathbf{X}\mathbf{w}\neq 0$, the reduced matrix $\mathbf{X}_r$ given by

$$\mathbf{X}_r = \mathbf{X} - d\,\mathbf{t}\,\mathbf{p}^T, \quad \text{with } d = 1/\mathbf{v}^T\mathbf{X}\mathbf{w}$$

will have a rank, which is one less than that of $\mathbf{X}$. This is independent of the choice of $\mathbf{w}$ and $\mathbf{v}$ as long as $\mathbf{v}^T\mathbf{X}\mathbf{w}\neq 0$. In practice it means that, if a plot $\mathbf{x}_i$ against $\mathbf{x}_j$ shows a straight line or a part of a straight line, it indicates that the corresponding part of $\mathbf{X}_r$ is zero. (There should be at least 8 points on a clear straight line.) Same holds if rows are plotted against each other. By studying the data graphically by the HELP methods, appropriate sub-blocks of $\mathbf{X}$ can be identified.

Multi-block methods have been applied in chemometrics [4]. Also some interesting applications in chemometrics have been reported [5]. Multi-block methods have been applied for a long time in psychometrics and social sciences. See Smilde *et al*. [6] for a collection of references. The paper of Smilde *et al*. contains a framework for different methods to analyse multi-block data, where only $\mathbf{X}$ is partitioned, $\mathbf{X}=(\mathbf{X}_1, \mathbf{X}_2,\ldots,\mathbf{X}_L)$ and the modelling task is $\mathbf{X}\rightarrow\mathbf{X}_i$, $i=1,\ldots,L$. Recently, PLS methods have been widely applied in different areas [7]. The paper of M. Vivien *et al*. 2003 [8] has similar objective as the present paper. The differences of the methods in this present paper and those in the literature are discussed in Section 7.

The subject considered in this paper is very large. The emphasis in this paper is on the optimisation step in different model, modelling strategies and data structures, which are useful in production environments. For methods that are not in focus of the present work, we discuss what further analysis may be needed.

2. REGRESSION MODELS OF MULTI-BLOCK DATA

Industrial data are often of different kinds. Some measurement values may come from optical or spectral instruments, others from mechanic or electrical ones. Optical instruments may have many variables, that is, it may provide with 1056 measurement values for each sample that is measured, thus generating 1056 variables. Mechanic instruments, on the other hand, may give only few values as a result of measuring a sample or as results at a given time point. When there are many variables, it is usually most appropriate to model the data by finding a latent structure in data that can do the task that is needed. This is done by weighing the variables to generate the latent structure. If there are many variables and of different kinds, like optical, mechanic, chemical, electrical and so on, it may not be good to treat variables as if they all were of the same kind. It may be necessary to divide the data into data blocks and use the weighing procedures separately for each block.

When working with industrial data, the data are often of very different types. The situation is best illustrated by the schematic example that is shown in Figure 1.

It is supposed here that the instrumental data $\mathbf{X}$ consists of three parts and response data $\mathbf{Y}$ of two parts. We shall now briefly describe a standard modelling procedure and some problems, when there are different types of variables.

When the data are modelled, a weight vector $\mathbf{w}$ is computed, which reflects how well the instrumental data describe $\mathbf{Y}$. The weight vector is used to compute a score vector $\mathbf{t}=\mathbf{X}\mathbf{w}$. The weight vector consists of three parts that correspond to the partition of $\mathbf{X}$, $\mathbf{w}=(\mathbf{w}_1,\mathbf{w}_2,\mathbf{w}_3)$. The score vector is the sum of the corresponding three parts, $\mathbf{t}=\mathbf{X}_1\mathbf{w}_1+\mathbf{X}_2\mathbf{w}_2+\mathbf{X}_3\mathbf{w}_3$. If the chemical measurements are, for example, NIR measurements containing 1056 values, $\mathbf{w}_2$ will contain 1056 values. The engineering measurements might be only 20 values for each sample. That would mean

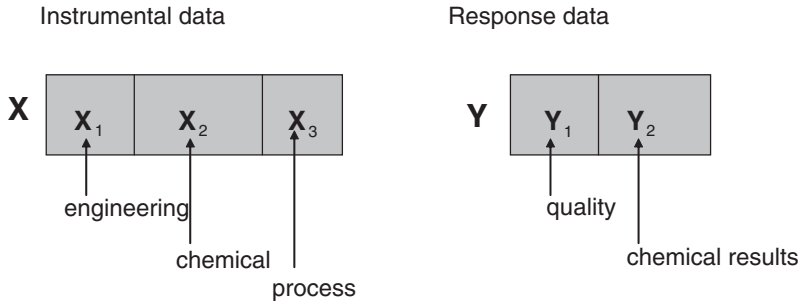

Figure 1. Schematic illustration of instrumental and response data.

DOI: 10.1002/cem

that $\mathbf{w}_1$ contains 20 values. The practical problem is that the weight vector $\mathbf{w}$ is scaled such that it gets the length one. This means that the 20 values in $\mathbf{w}_1$ get the same 'importance' as 20 values of the 1056 ones among $\mathbf{w}_2$. It follows that the importance of $\mathbf{w}_1$ is scaled down. This may be not desirable, if $\mathbf{X}_1$ is in fact good for describing the response data or some of them. There is also a practical problem, when there are many response variables that are of different kinds. This can be illustrated by two sets of response variables. Often there are some quality variables among the response variables, while there may be more variables among them that represent the chemical results in question. There may be 3 quality variables and 10 chemical ones. It may happen that that $\mathbf{X}$ can describe $\mathbf{Y}_1$ well, but there may be difficulties in using $\mathbf{X}$ for describing $\mathbf{Y}_2$. It may be advantageous to divide $\mathbf{Y}$ into the two parts and treat them separately. There are many ways to take into account the partition of $\mathbf{X} = (\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_3)$ and $\mathbf{Y} = (\mathbf{Y}_1,\mathbf{Y}_2)$. If each part of $\mathbf{X}$ is used separately, it may be desirable to have one set of score vectors for each $\mathbf{X}_i$, $i = 1,2,3$. The modelling task with three $\mathbf{X}$-blocks and two $\mathbf{Y}$-blocks is schematically illustrated in Figure 2.

The figure illustrates that the task is to use each $\mathbf{X}_i$, $i = 1,2,3$, to model each $\mathbf{Y}_j$, $j = 1,2$. Each $\mathbf{X}_i$ will have a decomposition of the kind, $\mathbf{X}_i = \mathbf{T}_i\mathbf{P}_i^T + \mathbf{X}_{i0}$, $i = 1,2,3$.

The matrix of score vectors $\mathbf{T}_i = (\mathbf{t}_{i,1},\ldots,\mathbf{t}_{i,Ai})$ will represent the latent structure in $\mathbf{X}_i$ that the description (regression) is based upon. Thus, each $\mathbf{Y}_j$ will be described by $\mathbf{T}_1$, $\mathbf{T}_2$ and $\mathbf{T}_3$. It may be necessary to conclude that for instance $\mathbf{t}_{1,3},\ldots\mathbf{t}_{1,A}$, only contribute to $\mathbf{Y}_1$ but not to $\mathbf{Y}_2$. In this case, only $\mathbf{t}_{1,1}$ and $\mathbf{t}_{1,2}$ contribute to both $\mathbf{Y}_1$ and $\mathbf{Y}_2$, but later score vectors derived from $\mathbf{X}_1$ only contribute to $\mathbf{Y}_1$. If there is only one latent structure $\mathbf{T}_i$ associated with each $\mathbf{X}_i$, it simplifies the interpretation of the latent structure. If there is more than one latent structure for $\mathbf{X}_i$, the interpretation of the results may be difficult. Let us take an example. The engineering measurements may be some initial conditions for the process in question. It may be desirable to find 'good' initial conditions. The latent structure can be used to find these good conditions, which should be used. When working with several response variables, it is an important issue, if one should work with one latent structure or develop one latent structure for each response variable. Experience has shown that we typically need more than one latent structure for obtaining good predictions, when there are several response variables.

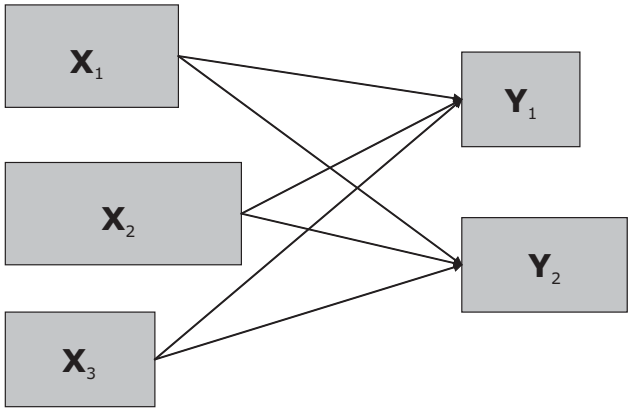

Figure 2. A schematic illustration of the modelling task.

3. TASKS OF MODELLING DATA

When the data is partitioned as described above, there may be as a result many data blocks, both $\mathbf{X}$s and $\mathbf{Y}$s. The user of a modelling task like this is interested in knowing not only how each part is doing the job, but also how it compares to a more overall modelling task. Typical requirements are considered closer.

- Comparison to an overall model, $\mathbf{X}\rightarrow\mathbf{Y}$: If the data is not partitioned at all, specific results can be obtained. Partitioning of data, $\mathbf{X}$ and $\mathbf{Y}$, should improve the modelling task. The user is interested in knowing what improvements there are.
- Significant dimension in each $\mathbf{X}_i\rightarrow\mathbf{Y}_j$: It is important that only score vectors that can describe $\mathbf{Y}_j$ are used for describing $\mathbf{Y}_j$. It may disturb the modelling task, if say 6 score vectors are kept for describing $\mathbf{Y}_2$, when only 2 are needed. But 6 might be needed for $\mathbf{Y}_1$. That $\mathbf{Y}_1$ might need many score vectors should not influence on the modelling of other $\mathbf{Y}$s. This task is considered closer in Section 6.
- Contribution at each step: At each step there are found score vectors from the $\mathbf{X}$s. The user wants to know the contribution that is obtained for each of the $\mathbf{Y}$s.
- Marginal contribution of score vectors: If the score vector say, $\mathbf{t}_{1,2}$ is used for describing both $\mathbf{Y}_1$ and $\mathbf{Y}_2$, the user wants to know the contribution of $\mathbf{t}_{1,2}$ to the task. This would be the marginal contribution of the score vector. The score vector $\mathbf{t}_{1,2}$ contributes together with other score vectors to say, $\mathbf{Y}_1$, but it may be interesting to know how much it would contribute, if it were alone.
- Total contribution of $\mathbf{X}_i$ to $\mathbf{Y}_j$: A part of the score vectors $\mathbf{T}_i$ associated with $\mathbf{X}_i$ contributes to the description of $\mathbf{Y}_j$. It is useful to know the total contribution of $\mathbf{X}_i$ in describing $\mathbf{Y}_j$.
- Separate contribution of $\mathbf{X}_i$ to $\mathbf{Y}_j$: It is natural to check what can be obtained, if only $\mathbf{X}_i$ is used to describe $\mathbf{Y}_j$. This result should be reported for comparison.
- Individual response variable: If it is desired to get best possible predictions for each response variable, they are treated separately. It means that we loop over each response variable and leave the others out. The network of data blocks is then used to estimate the parameters between data blocks. This being carried out for each response variable will show what can be done for the given network of data blocks.

We see that there are many requirements to the modelling task. The important issue is to keep separate, what part is of the multi-block data analysis and what part is a comparison with other views of the modelling task. Note that in the analysis some or all of the $\mathbf{Y}$s can be the $\mathbf{X}$s.

4. THE PRINCIPLE OF OPTIMISATION

4.1. Background

The basic purpose of the mathematical modelling task is to provide with a model that gives good predictions. In order to arrive at such a model it is important to be aware of that the

DOI: 10.1002/cem

modelling task has two independent features. This will be explained in terms of a linear regression model, $\mathbf{X}\rightarrow\mathbf{y}$, (one response variable), the linear least squares solution $\mathbf{b}$, $\mathbf{b}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and normally distributed data. Using standard assumptions, it follows that the residuals, $\mathbf{y}$-$\hat{\mathbf{y}}$, are stochastically independent of the precision, $(\mathbf{X}^T\mathbf{X})^{-1}$. One can show that this implies that the size of improvement in fit due to a score vector $\mathbf{t}$, $|\mathbf{y}^T\mathbf{t}|^2/(\mathbf{t}^T\mathbf{t})$, is independent of the variance of the associated regression coefficients, $\sigma^2/(\mathbf{t}^T\mathbf{t})$. In PLS regression, it is suggested to find $\mathbf{w}$ such that $(\mathbf{y}^T\mathbf{t})$ is maximised. Does this secure that the score vector is large and that the variance is small? The answer is positive, which follows from the Cauchy-Schwarts inequality, $|\mathbf{y}^T\mathbf{t}|\leq|\mathbf{y}|\times|\mathbf{t}|$.

4.2. Approach

In the general case we seek a weight vector $\mathbf{w}$ such that $|\mathbf{q}|^2=|\mathbf{Y}^T\mathbf{t}|^2=|\mathbf{Y}^T\mathbf{X}\mathbf{w}|^2$ is as large as possible. The solution is given by finding the eigen vector associated as a leading eigen value of the set of equations,

$$\mathbf{w}^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}=\lambda\,\mathbf{w} \tag{1}$$

The situation is schematically illustrated in part (a) in Figure 3. The task is to find $\mathbf{w}$ such that the associated score vector generates as large Y-loading vector $\mathbf{q}$ as possible. In part (b) in the figure the task is to find $\mathbf{w}_1$ and $\mathbf{w}_2$ such that the Y-loading vectors, $\mathbf{q}_1$ and $\mathbf{q}_2$, which are generated, become as large as possible. At the optimisation task it is required to find $\mathbf{w}_1$ and $\mathbf{w}_2$ such that $\mathbf{q}=\mathbf{q}_1+\mathbf{q}_2$ is as large as possible. In part (c) the task is to estimate regression coefficients, $\mathbf{B}_x$ and $\mathbf{B}_z$, such that when a new X-sample, $\mathbf{x}$, becomes available, it is possible to use it to estimate a Z-sample $\mathbf{z}_0$, $\mathbf{z}_0=\mathbf{B}_x\mathbf{x}$, and to use $\mathbf{z}_0$ to estimate an Y-sample $\mathbf{y}_0$, $\mathbf{y}_0=\mathbf{B}_z\mathbf{z}_0$. It is desirable to obtain as reliable predictions of Y-samples as possible. Therefore, the optimisation task is to find $\mathbf{w}$ such that the resulting Y-loading vector is as large as possible. The Y-loading vector $\mathbf{q}$ is computed as $\mathbf{q}=\mathbf{Y}^T\mathbf{t}_z=\mathbf{Y}^T\mathbf{Z}\mathbf{q}_z=\mathbf{Y}^T\mathbf{Z}\mathbf{Z}^T\mathbf{t}=\mathbf{Y}^T\mathbf{Z}\mathbf{Z}^T\mathbf{X}\mathbf{w}$. The optimisation task here is to maximise $|\mathbf{q}|^2$.

In summary, the principle behind the optimisation tasks is to maximise the size of the resulting loading vectors that are at the end of the 'network of data blocks'. When this principle is used, it is usually necessary to scale the data, for example, to unit variances within each data block. The issue of scaling is not considered closer here.

5. CRITERIA FOR OPTIMISATION TASKS

It is now considered closer how the criteria for optimisation are formulated for multi-block data analysis. In order to simplify the formulae the situation specified in Figure 4 is chosen as a starting point. The task here is to compute one set of score vectors, one for each of the $\mathbf{X}$s. It is assumed that $\mathbf{X}_1$ only describes $\mathbf{Y}_1$. This can be due to that at previous step it was found that there is no further relationship between $\mathbf{X}_1$ and $\mathbf{Y}_2$. It can also be a part of the model specification that $\mathbf{X}_1$ only models $\mathbf{Y}_1$.

$\mathbf{X}_2$ is used to model both $\mathbf{Y}_1$ and $\mathbf{Y}_2$. But $\mathbf{X}_3$ only models $\mathbf{Y}_2$. This specification is sufficiently simple, and includes all details.

At $\mathbf{Y}_1$ there are two loading vectors, $\mathbf{q}_{11}$ and $\mathbf{q}_{12}$. Following the recommendation above, the total loading $\mathbf{q}_{11}+\mathbf{q}_{12}$ should be as large as possible. Similarly it holds for the loading of $\mathbf{Y}_2$, $\mathbf{q}_{22}$ and $\mathbf{q}_{23}$. Thus, the task is to find $\mathbf{w}_1$, $\mathbf{w}_2$ and $\mathbf{w}_3$ such that the total size of $\mathbf{Y}$-loadings

$$\begin{aligned}&|\mathbf{q}_{11}+\mathbf{q}_{12}|^2+|\mathbf{q}_{22}+\mathbf{q}_{23}|^2\\&=\mathbf{w}_1^T\mathbf{X}_1^T\mathbf{Y}_1\mathbf{Y}_1^T\mathbf{X}_1\mathbf{w}_1+2\mathbf{w}_1^T\mathbf{X}_1^T\mathbf{Y}_1\mathbf{Y}_1^T\mathbf{X}_2\mathbf{w}_2\\&\quad+\mathbf{w}_2^T\mathbf{X}_2^T(\mathbf{Y}_1\mathbf{Y}_1^T+\mathbf{Y}_2\mathbf{Y}_2^T)\mathbf{X}_2\mathbf{w}_2+2\mathbf{w}_2^T\mathbf{X}_2^T\mathbf{Y}_2\mathbf{Y}_2^T\mathbf{X}_3\mathbf{w}_3\\&\quad+\mathbf{w}_3^T\mathbf{X}_3^T\mathbf{Y}_2\mathbf{Y}_2^T\mathbf{X}_3\mathbf{w}_3\end{aligned}$$

becomes as large as possible. Using the Lagrange multiplier technique the terms $\lambda_i(\mathbf{w}_i^T\mathbf{w}_i-1)$, $i=1,2$ and 3, are added to the equation. Differentiating with respect to $\mathbf{w}_i$, the following set of equations are obtained,

$$\begin{aligned}\mathbf{X}_1^T\mathbf{Y}_1\mathbf{Y}_1^T\mathbf{X}_1\mathbf{w}_1 &+\mathbf{X}_1^T\mathbf{Y}_1\mathbf{Y}_1^T\mathbf{X}_2\mathbf{w}_2 &&=\lambda_1\,\mathbf{w}_1\\ \mathbf{X}_2^T(\mathbf{Y}_1\mathbf{Y}_1^T+\mathbf{Y}_2\mathbf{Y}_2^T)\mathbf{X}_2\mathbf{w}_2 &+\mathbf{X}_2^T\mathbf{Y}_2\mathbf{Y}_2^T\mathbf{X}_3\mathbf{w}_3 &&=\lambda_2\,\mathbf{w}_2\\ \mathbf{X}_3^T\mathbf{Y}_2\mathbf{Y}_2^T\mathbf{X}_2\mathbf{w}_2 &+\mathbf{X}_3^T\mathbf{Y}_2\mathbf{Y}_2^T\mathbf{X}_3\mathbf{w}_3 &&=\lambda_3\,\mathbf{w}_3\end{aligned}$$

Consider now the general case, where there are L $\mathbf{X}$-data blocks, $\mathbf{X}=(\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_L)$, and M $\mathbf{Y}$-data blocks, $\mathbf{Y}=(\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_M)$. The set of equations can be viewed as a collection terms of the following type,

$$\mathbf{G}_{ij}=\mathbf{X}_i^T(\delta_{i1}\mathbf{Y}_1\mathbf{Y}_1^T+\delta_{i2}\mathbf{Y}_2\mathbf{Y}_2^T+\ldots+\delta_{iM}\mathbf{Y}_M\mathbf{Y}_M^T)\mathbf{X}_j,$$

$$\text{for } i,j=1,2,\ldots,L$$

Here $\delta_{im}=1$, if $\mathbf{X}_i$ is modelling $\mathbf{Y}_m$ and zero otherwise. If $\mathbf{G}$ is the data matrix containing these terms, $\mathbf{G}=(\mathbf{G}_{ij})$, the equation to solve is given by $\mathbf{G}\mathbf{w}=(\lambda_1\,\mathbf{w}_1,\lambda_2\,\mathbf{w}_2,\ldots,\lambda_L\,\mathbf{w}_L)^T$, with $\mathbf{w}=(\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_L)$. This equation is solved iteratively

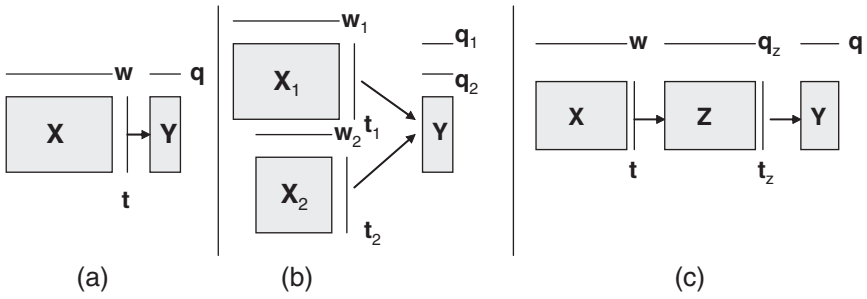

Figure 3. Schematic illustration of the optimisation tasks. (a) $\mathbf{X}\rightarrow\mathbf{Y}$, (b) $\mathbf{X}_1\rightarrow\mathbf{Y}$ and $\mathbf{X}_2\rightarrow\mathbf{Y}$, (c) $\mathbf{X}\rightarrow\mathbf{Z}\rightarrow\mathbf{Y}$.

DOI: 10.1002/cem

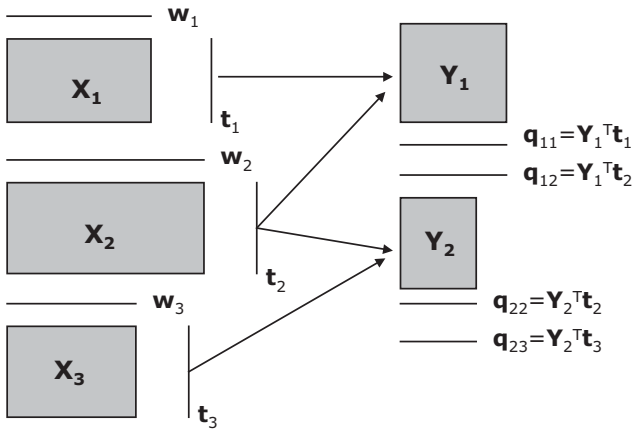

Figure 4. Illustration of the modelling task.

with starting values of $\mathbf{w}_i$ as the eigen vector of the leading eigen value of the $\mathbf{G}_{ii}\mathbf{w}_i = \lambda_i \mathbf{w}_i$. At each iteration all of the weight vectors $\mathbf{w}_i$s must be scaled to unit length. Thus, at each iteration $\mathbf{Gw}$ is partitioned appropriately, and $\mathbf{w}_i$s computed as having unit length. We have observed that the speed of convergence is the same as at the power method of computing the largest eigen value and associated eigen vector. Thus, typically less than 20–30 iterations are necessary to find all weight vectors ($\mathbf{w}_1$, $\mathbf{w}_2$,...,$\mathbf{w}_L$).

When the weight vectors are found, following are computed for i = 1,...,L:

Score vector : $\mathbf{t}_i = \mathbf{X}_i\mathbf{w}_i$

Loading vector : $\mathbf{p}_i = \mathbf{X}_i^T\mathbf{t}_i$

Scaling constant : $d_i = 1/(\mathbf{t}_i^T\mathbf{t}_i)$

Furthermore, the $\mathbf{X}_i$s are adjusted for what has been found:

Adjustment of $\mathbf{X}_i$: $\mathbf{X}_i \leftarrow \mathbf{X}_i - d_i\ \mathbf{t}_i\ \mathbf{p}_i^T$

The adjustment of each $\mathbf{X}_i$ gives orthogonal score vector for $\mathbf{X}_i$, but score vector of one data block $\mathbf{X}_i$ are not orthogonal to score vectors of another data block $\mathbf{X}_j$.

The adjustment of each $\mathbf{Y}_m$ can be carried out as follows. The score vectors that have been found to contribute to $\mathbf{Y}_m$ are collected in a matrix $\mathbf{T}_{a,m}$. Then, linear least squares estimates of the regression coefficients are given by $\mathbf{B}_{am} = (\mathbf{T}a, m^T\mathbf{T}_{a,m})^{-1}\mathbf{T}a, m^T\mathbf{Y}_m$,

adjustment of $\mathbf{Y}_m$: $\mathbf{Y}_m \leftarrow \mathbf{Y}_m - \mathbf{T}_{a,m}\ \mathbf{B}_{am}$

If all of the score vectors that contribute to $\mathbf{Y}_m$ are collected together, $\mathbf{T}_m = (\mathbf{T}_{1,m},...,\mathbf{T}_{A,m})$, the regression coefficients are computed as $\mathbf{B}_m = (\mathbf{T}m^T\mathbf{T}_m)^{-1}\mathbf{T}m^T\mathbf{Y}_m$. The estimated response values are given by $\hat{\mathbf{Y}}_m = \mathbf{T}_m\mathbf{B}_m$.

In the case of many X-blocks, it may be necessary to carry out different adjustments and to compute estimated response values in different ways. When working with an $\mathbf{Y}$, say $\mathbf{Y}_m$, then at each step there have been collected score vectors that are collected in a matrix $\mathbf{T}_{a,m}$. It may be better to work with the PLS solution compared to linear least squares solution. Similarly, when computing final estimates of the response matrix, $\hat{\mathbf{Y}}_m$, there is a collection of score vectors in $\mathbf{T}_m$, each of which has marginally a significant contribution to $\mathbf{Y}_m$. Here also it might be better to use the PLS solution to the linear least squares one.

When we have been working with multi-block methods on industrial data, it has been necessary to identify the part of each $\mathbf{X}_i$ (variables) that should be used. This can be carried out, for example, by studying how $\mathbf{X}_i$ models $\mathbf{Y}$. Variables that do not satisfy Equation (3) below for $\mathbf{t} = \mathbf{x}_i$ (and A = 1) and any Y-variable are automatically excluded. This procedure can be improved, but this is not studied further here.

If we want to study especially how $\mathbf{X}_i$ contributes to the modelling of $\mathbf{Y}_m$, the score vectors associated with $\mathbf{X}_i$ are selected, and we use them to see how $\mathbf{X}_i$ separately contributes to $\mathbf{Y}_m$.

In the practical application of the present multi-block methods there are many practical issues that need to be resolved, like the ones mentioned above. These practical issues are not treated closer in this paper.

The criterion above can take special form depending on the structure of the multi-way blocks. Consider one example, suppose that there is only one $\mathbf{X}$, but the $\mathbf{Y}$s are a part of $\mathbf{X}$, $\mathbf{Y}_i = \mathbf{X}_i$, i = 1,2,...,L. This corresponds to the case, where we want to use all of $\mathbf{X}$ to describe the parts of $\mathbf{X}$. In this case, there is only one equation to solve,

$$\mathbf{X}^T\left(\delta_1\mathbf{X}_1\mathbf{X}_1^T + \delta_2\mathbf{X}_2\mathbf{X}_2^T + \ldots + \delta_L\mathbf{X}_L\mathbf{X}_L^T\right)\mathbf{X}\,\mathbf{w} = \lambda\,\mathbf{w} \qquad (2)$$

Here again $\delta_i = 1$ if $\mathbf{X}_i$ is taking part in the modelling task and zero otherwise. Some further types of models are considered later.

6. MODELLING STRATEGIES

When working with multi-block methods it is essential to determine, if a score vector contributes to the modelling task. We shall consider here the case of one $\mathbf{X}$ matrix, N × K matrix, and a response matrix that consists of only one column, $\mathbf{y}$, associated with one response variable. The score vector $\mathbf{t}$ is computed as $\mathbf{t} = \mathbf{Xw}$ for some weight vector $\mathbf{w}$. When $\mathbf{X}$ is adjusted as described in the previous section, it follows that the score vectors are mutually orthogonal independently of the choice of the weight vectors (as long as the score vectors are not of zero length, $|\mathbf{t}_i| \neq 0$). If the score vectors are $\mathbf{T} = (\mathbf{t}_1,\mathbf{t}_2,...,\mathbf{t}_A)$, $\mathbf{t}_i^T\mathbf{t}_j = 0$, for i≠j, i,j = 1,...,A. The residual variance, when A score vectors have been found, is estimated as

$$s_A^2 = \left[\mathbf{y}^T\mathbf{y} - \left\{(\mathbf{y}^T\mathbf{t}_1)^2/(\mathbf{t}_1^T\mathbf{t}_1) + \ldots + (\mathbf{y}^T\mathbf{t}_A)^2/(\mathbf{t}_A^T\mathbf{t}_A)\right\}\right]/(N - A).$$

Note that the term $(\mathbf{y}^T\mathbf{t}_A)^2/(\mathbf{t}_A^T\mathbf{t}_A)$ is the variation that score vector $\mathbf{t}_A$ explains of $\mathbf{y}$s variation. Consider now the prediction aspect of the modelling task. Assume a standard linear model, $\mathbf{y} \sim N(\mathbf{X\beta}, \sigma^2\mathbf{I})$, which means that the response values are normally distributed with residual variance σ^2. If $\mathbf{x}_0$ is a new sample, the variance of the estimated y-value, $y(\mathbf{x}_0)$, is given by

$$\mathrm{Var}(y(\mathbf{x}_0)) = \sigma^2\left(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{x}_0\right)$$

Suppose that $\mathbf{x}_0$ is the samples (rows) of $\mathbf{X}$, $\mathbf{x}_0 = \mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N$. Summing over the samples we get

$$\sum_1^N \mathrm{Var}(y(\mathbf{x}_i)) = \sigma^2\sum_1^N\left(1 + \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{x}_i\right).$$

DOI: 10.1002/cem

Using that

$$\sum_1^N \left(1 + \mathbf{x}_i^T (\mathbf{X}^T\mathbf{X})^{-1}\, \mathbf{x}_i\right) = \sum_1^N \left(1 + \mathrm{tr}\left\{(\mathbf{X}^T\mathbf{X})^{-1}\, \mathbf{x}_i\, \mathbf{x}_i^T\right\}\right)$$
$$= \mathbf{N} + \mathrm{tr}\left\{(\mathbf{X}^T\mathbf{X})^{-1} \sum_1^N \mathbf{x}_i\, \mathbf{x}_i^T\right\}$$
$$= \mathbf{N} + \mathrm{tr}\{(\mathbf{X}^T\mathbf{X})^{-1}\, \mathbf{X}^T\mathbf{X}\} = \mathrm{N} + \mathrm{K},$$

it follows that

$$\text{Total prediction variance} = \sum_1^N \mathrm{Var}(\mathrm{y}(\mathbf{X}_i)) = \sigma^2\ (\mathrm{N} + \mathrm{K}).$$

When A score vectors have been found, we estimate the total prediction variance as

$$\mathrm{s}_A^2(\mathrm{N} + \mathrm{A}).$$

This estimation presupposes that the bias has become close to zero. It is natural to require that the next score vector, $\mathbf{t}_{A+1}$, should reduce the total prediction variance. It is easy to show that

$$\mathrm{s}_A^2(\mathrm{N} + \mathrm{A}) > \mathrm{s}_{A+1}^2(\mathrm{N} + \mathrm{A} + 1)$$

if and only if

$$(\mathbf{y}^T\mathbf{t}_{A+1})^2/\left(\mathbf{t}_{A+1}^T\mathbf{t}_{A+1}\right) > \mathrm{s}_{A+1}^2(2\mathrm{N} - 1)/(\mathrm{N} + \mathrm{A}). \qquad (3)$$

This follows from the re-writing,

$$\mathrm{s}_A^2(\mathrm{N} + \mathrm{A}) = \left[\mathrm{s}_{A+1}^2(\mathrm{N} - \mathrm{A} - 2) + (\mathbf{y}^T\mathbf{t}_{A+1})^2/\left(\mathbf{t}_{A+1}^T\mathbf{t}_{A+1}\right)\right] \times (\mathrm{N} + \mathrm{A})/(\mathrm{N} - \mathrm{A} - 1),$$

and eliminating the term of the variation explained by $\mathbf{t}_{A+1}$, $(\mathbf{y}^T\mathbf{t}_{A+1})^2/(\mathbf{t}_{A+1}^T\mathbf{t}_{A+1})$. In the examples below we write

$$\text{Variation} = (\mathbf{y}^T\mathbf{t}_{A+1})^2/\left(\mathbf{t}_{A+1}^T\mathbf{t}_{A+1}\right), \qquad (4)$$

$$\text{Minimal requirement} = \mathrm{c} \times \text{residual variance} \qquad (5)$$

with $\mathrm{c} = (2\mathrm{N} - 1)/(\mathrm{N}+\mathrm{A})$ and 'residual variance' $= \mathrm{s}_{A+1}^2$.

It is natural to require that the total (or average) prediction variance should reduce, if a score vector is to be used. It is not a strong criterion. Cross-validation may show in the revision of the model that a score vector may satisfy this criterion,

$$\text{variation} > \mathrm{c} \times \text{residual variance}$$

but does not improve the prediction criterion of the cross-validation. In general, it is safe to discard a score vector, if this criterion is not satisfied.

In the algorithm, we have given a score vector $\mathbf{t}_i$ derived from $\mathbf{X}_i$. In order to see how this score vector works for $\mathbf{Y}_j$, we go through each column of $\mathbf{Y}_j$. If the score vector fails to meet this criterion for all columns of $\mathbf{Y}_j$, the use of $\mathbf{X}_i$ in describing $\mathbf{Y}_j$ is stopped. If the score vector found from $\mathbf{X}_i$ does not contribute to any $\mathbf{Y}_j$, the usage of $\mathbf{X}_i$ is stopped, and a new set of weight vectors is computed for the $\mathbf{X}$-matrices that are taking part in the modelling task. The new set of score vectors found is then evaluated again.

7. COMPARISON TO OTHER CRITERIA IN THE LITERATURE

The paper of Smilde *et al*. 2003 [6] gives a framework for a collection of methods for analysing multi-block data. The paper only treats the case, where $\mathbf{X}$ is divided into blocks, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_L)$. Denote by $\mathbf{S}_i = \mathbf{X}_i\mathbf{X}_i^T$. In the paper, it is shown that many methods amount to finding a vector $\mathbf{t} = \mathbf{t}_{sup}$, such that the equation

$$\sum_1^L \mathrm{c}_i\, \mathbf{S}_i\, \mathbf{t} = \mathrm{d}\, \mathbf{t}$$

is satisfied. The constants $\mathbf{c} = (\mathrm{c}_1, \ldots, \mathrm{c}_L)$ and d depend on the method in question. For some methods this is an eigen value task, while for others it is not. The disadvantage of the non-eigen value methods is that it is not clear what is being carried out, what is being optimised, or what is a good result or not.

In the present approach the set up in reference [6] corresponds to have only one $\mathbf{X}$ and choose $\mathbf{Y}_i$ as $\mathbf{Y}_i = \mathbf{X}_i$. This means that we are looking for a weight vector $\mathbf{w}$ such that the resulting score vector $\mathbf{t} = \mathbf{X}\mathbf{w}$ is good for describing the sub-blocks $\mathbf{X}_i$. The criterion is the eigen value system,

$$\left(\delta_1\mathbf{X}^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{X} + \delta_2\mathbf{X}^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{X} + \ldots + \delta_L\mathbf{X}^T\mathbf{X}_L\mathbf{X}_L^T\mathbf{X}\right)\mathbf{w} = \lambda\ \mathbf{w}$$

where δ_i is 1 if $\mathbf{X}_i$ is participating in the model and zero otherwise. Thus, we see that the present approach is different from the one in [6].

It is also possible to revert the analysis and ask how well the sub-blocks describe the whole $\mathbf{X}$, $\mathbf{X}_i \rightarrow \mathrm{X}$, $\mathrm{i} = 1, \ldots, \mathrm{L}$. For each sub-block there is a computed score vector $\mathbf{t}_i = \mathbf{X}_i\mathbf{w}_i$, and the resulting $\mathbf{X}$-loading vector, $\mathbf{q}_i = \mathbf{X}^T\mathbf{t}_i$. Here it is suggested that the total loading vector $\mathbf{q} = \mathbf{q}_1 + \ldots + \mathbf{q}_L$ should be as large as possible. The solution is the system of equations given by

$$\begin{array}{cccc} \mathbf{X}_1^T\mathbf{X}\mathbf{X}^T\mathbf{X}_1\ \mathbf{w}_1 + & \ldots & +\mathbf{X}_1^T\mathbf{X}\mathbf{X}^T\mathbf{X}_L\ \mathbf{w}_L & = \lambda_1\ \mathbf{w}_1 \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{X}_L^T\mathbf{X}\mathbf{X}^T\mathbf{X}_1\ \mathbf{w}_1 + & \ldots & +\mathbf{X}_L^T\mathbf{X}\mathbf{X}^T\mathbf{X}_L\ \mathbf{w}_L & = \lambda_L\ \mathbf{w}_L \end{array}$$

where the λs and $\mathbf{w}$s are unknown and $\mathbf{w}$s of length 1.

In the reference [8] the task is to model $\mathbf{X}_i \rightarrow \mathrm{Y}_j$, $\mathrm{i} = 1, \ldots, \mathrm{L}$, $\mathrm{j} = 1, \ldots, \mathrm{M}$, like in the present work. In the reference [8] the optimisation task is based on maximising the total covariance. In the reference [9] it is shown that in PLS regression maximising the covariance is equivalent to maximising the size of the Y-loading vector. Thus, in some cases the optimisation task in [8] would lead to the same results as in the present paper. But the approach here is to maximise the loading vectors of the output matrices as suggested in reference [1]. The maximisation of the total covariance does not extend to the multi-block situation, while maximising the size of the output loading vectors does.

In the the reference [7] some 14 optimisation criteria from the literature on multi-block analysis are presented [10–14]. All of them are based on some types of sums. In the present paper, the optimisation criteria are all based on appropriate products of the matrices, which are derived from the path or network specifications.

In the references [4–5] hierarchical modelling procedures are presented. The basic idea of these methods is to compute score vectors for each data block, as it is done here. Then, computed a new score vector $\mathbf{t}_T$, which is a linear combination of these score vectors. This new score vector is used in the modelling task. There are some problems concerning this approach. The score vector $\mathbf{t}_T$ is used to compute the estimated response values and to adjust the X-blocks. It is difficult to relate $\mathbf{t}_T$ to the individual data blocks, which data blocks contribute. Furthermore, when the

DOI: 10.1002/cem

X-blocks are adjusted, a score vector is used that may not be related to the X-values. The score vector may depend mostly on say, chemical values, while when used to adjust NIR data, creates residual X-blocks, which may have no interpretation and may cause difficulties in using the NIR data appropriately.

In summary, the present methods differ from the ones found in the literature. The emphasis of the present methods is to secure optimal prediction of samples, which are derived from the input samples of the path or network.

8. EXAMPLES

Here two examples are considered. One is a situation where there are two X-blocks and two Y-blocks, and the relationships between the blocks are known. The other is a case study.

8.1. A simulated example

$\mathbf{X}_1$ and $\mathbf{X}_2$ are both 10×2 matrices. $\mathbf{Y}_1$ is 10×4 and $\mathbf{Y}_2$ is 10×3. $\mathbf{Y}_1$ is generated such that it depends $\mathbf{X}_1$ and $\mathbf{X}_2$. $\mathbf{Y}_2$ depends only on $\mathbf{X}_1$. In both cases only one score vector from $\mathbf{X}_1$ and $\mathbf{X}_2$ are expected. All data have been auto-scaled (centred and unit variance) before analysis. We want to see if the algorithm identifies the relationships correctly. The modelling results from the first set of score vectors are shown in Table I.

The first task of the algorithm is to find two score vectors $\mathbf{t}_1 = \mathbf{X}_1\mathbf{w}_1$ and $\mathbf{t}_2 = \mathbf{X}_2\mathbf{w}_2$. The score vector $\mathbf{t}_1$ is evaluated for each of the response variables. For the first variable y_1, of the data block $\mathbf{Y}_1$, $\mathbf{t}_1$ explains 52.6218% of the variation of $\mathbf{y}_1$. The minimum requirement to $\mathbf{t}_1$ is 0.093769. Thus, it shows a significant contribution to explaining y_1. The score vector $\mathbf{t}_2$ shows an explanation of 57.8766–59.4852% of the y-variables of $\mathbf{Y}_1$. The score vector $\mathbf{t}_1$ explains from 99.8239% to 99.8787% of the y-variables of $\mathbf{Y}_2$. But the score vector $\mathbf{t}_2$ does not contribute to the variation of $\mathbf{Y}_2$. It fails to pass the values of the minimum requirements around 0.196. The conclusion here is that the algorithm gives correct results and is consistent with the generated models. Note that $\mathbf{t}_1$ explains 52.6218% of the variation of $\mathbf{y}_1$ in data block $\mathbf{Y}_1$, while $\mathbf{t}_2$ explains 58.5040%. The maximum is of course 100%. But they are not independent. Together they explain 99.8% of the variation of $\mathbf{y}_1$.

Table I. Results from one set of score vectors

	c*residual variance	Variation
	$\mathbf{Y}_1$ and $\mathbf{X}_1$	
Dim = 1		
y_1	0.093769	**0.526218**
y_2	0.092595	**0.532150**
y_3	0.095350	**0.518231**
y_4	0.095740	**0.516260**
	$\mathbf{Y}_1$ and $\mathbf{X}_2$	
Dim = 2		
y_1	0.082128	**0.585040**
y_2	0.083369	**0.578766**
y_3	0.080538	**0.593071**
y_4	0.080185	**0.594852**
	$\mathbf{Y}_2$ and $\mathbf{X}_1$	
Dim = 1		
y_1	0.000240	**0.998787**
y_2	0.000327	**0.998349**
y_3	0.000348	**0.998239**
	$\mathbf{Y}_2$ and $\mathbf{X}_2$	
Dim = 2		
y_1	0.196098	0.009189
y_2	0.196313	0.008103
y_3	0.195937	0.010000

8.2. Data from a cement production

Here data from a cement production are considered [15]. There are 163 x-variables and 3 y-variables. Variables are of the following types:

x_1–x_{12}	Chemical variables: mass fraction of SO_3, C, free lime, loss on ignition, SiO_2, Al_2O_3, Fe_2O_3, CaO, MgO, K_2O, Na_2O in cement
x_{13}–x_{110}	Superficial microstructures: mass losses in three regions of temperature up to 946°C from thermo gravimetric analysis (TGA)
x_{111}–x_{135}	Variables describing particle size distribution in 25 size classes
x_{136}–x_{163}	Process variables
y_1	Quality variable related to setting time
y_2	Quality variable related to water content required to achieving standard consistency
y_3	Quality variable related to compressive strength at 1 day

The task of interest here is to model four X-blocks and three Y-blocks, where each Y-block consists of one variable. We are interested in how the X-data blocks model the Y-blocks. Before analysis all variables were auto-scaled (centred and variance 1). The first attempt gave the following results shown in Table II. From the table it can be seen that the data blocks $\mathbf{X}_1$ and $\mathbf{X}_4$ do not contribute to the modelling of the Y-blocks. The first score vector of $\mathbf{X}_3$ explains 10.01% of the variation of $\mathbf{Y}_1$, 18.06% of the variation of $\mathbf{Y}_2$ and 30.69% of the variation of $\mathbf{Y}_3$. If a score vector from $\mathbf{X}_i$ does not contribute to describing $\mathbf{Y}_j$, this modelling connection is stopped. For example, the first 3 score vectors of $\mathbf{X}_3$

Table II. Modelling results for 4 X-blocks and 3 Y-blocks

	$\mathbf{X}_1$		$\mathbf{X}_2$		$\mathbf{X}_3$		$\mathbf{X}_4$	
	1	2	1	2	1	2	1	2
$\mathbf{Y}_1$								
Dim								
1	0.0397	0.0001	0.0359	**0.0953**	0.0367	**0.1001**	0.0341	**0.1406**
2			0.0358	**0.0978**	0.0357	**0.1000**	0.0341	**0.1409**
3			0.0271	0.0191	0.0276	0.0045	0.0273	0.0110
$\mathbf{Y}_2$								
Dim								
1	0.0396	0.0015	0.0303	**0.2371**	0.0325	**0.1806**	0.0397	0.0003
2			0.0303	**0.2352**	0.0325	**0.1807**		
3			0.0244	**0.0470**	0.0263	0.0011		
4			0.0234	**0.0736**				
5			0.0233	0.0030				
$\mathbf{Y}_3$								
Dim								
1	0.0394	0.0055	0.0393	0.0010	0.0275	**0.3069**	0.0345	**0.1308**
2					0.0275	**0.3069**	0.0345	**0.1309**
3					0.0249	0.0139	0.0249	0.0134

1: The minimum requirement to the score vector, see 5.
2: The variation obtained for the score vector, see 4.

DOI: 10.1002/cem

contribute to describing $\mathbf{Y}_1$, but the fourth not. But as the results of the fourth one are shown; it only gives 0.82% explained variation while 2.91% is needed. At dimension 5 only $\mathbf{X}_3$ is modelling $\mathbf{Y}_2$. There can be used several different types of modelling strategies. In Table II the algorithm starts finding the weight vectors $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$ and $\mathbf{w}_4$ as described above. The effects of the resulting score vectors $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$ and $\mathbf{t}_4$ are shown in Table II as the first line for $\mathbf{Y}_1$,$\mathbf{Y}_2$ and $\mathbf{Y}_3$, respectively. Since $\mathbf{t}_1$ and $\mathbf{t}_4$ do not contribute to the modelling of any $\mathbf{Y}$s, we could revise the computation of the weight vectors and only find $\mathbf{w}_2$ and $\mathbf{w}_3$. But it is useful to compute results of Table II and look at how the score vectors together contribute. When there are many variables like here, it is also important to identify non-zero values of a weight vector, that is, to find which variables should be used. If there are many values of the weight vector that are close to zero, which is found here, it may reduce the quality of the model. Therefore, there may be needed at each step some modelling strategies for both the weight vectors and the resulting score vectors. But this is not considered closer here.

In Table III is shown similar results as in Table II, where only $\mathbf{X}_2$ and $\mathbf{X}_3$ are used. The results in Table III are very close to the results of Table II. This is due to that there is very little or no correlation between the X-data blocks.

The conclusion is that only the data block $\mathbf{X}_3$ contributes to $\mathbf{Y}_1$ and $\mathbf{Y}_3$, while $\mathbf{X}_2$ and $\mathbf{X}_3$ contribute to describing $\mathbf{Y}_2$. $\mathbf{X}_1$ and $\mathbf{X}_4$ do not contribute to describing any $\mathbf{Y}$s. We shall not go further into the task of optimising the model closer. There are many results that are interesting to look at. Here it is especially important to select the variables in the data blocks that should be used in the analysis. But it is outside the scope of the present paper.

$\mathbf{X}_1$ and $\mathbf{X}_4$ do not contribute to the modelling task. In this case, the weight vectors associated with $\mathbf{X}_2$ and $\mathbf{X}_3$ are revised, where only these two are used. But this is not shown here because the differences are small. In general, all weight vectors are computed and the associated score vectors evaluated. If there are score vectors that do not contribute to the modelling task, the associated data blocks are removed from analysis and only 'active' data blocks are used to find the appropriate weight vectors.

Table III. Modelling results for 2 X-blocks and 3 Y-blocks

	$\mathbf{X}_2$		$\mathbf{X}_3$	
	1	2	1	2
$\mathbf{Y}_1$				
Dim				
1	0.0383	0.0339	0.0357	**0.1000**
2			0.0328	**0.0744**
3			0.0286	**0.1082**
4			0.0284	0.0082
$\mathbf{Y}_2$				
Dim				
1	0.0267	**0.3279**	0.0325	**0.1806**
2	0.0123	**0.1893**	0.0182	**0.0395**
3	0.0113	**0.0136**	0.0109	**0.0243**
4	0.0096	**0.0196**	0.0100	**0.0119**
5	0.0091	0.0054	0.0088	**0.0127**
$\mathbf{Y}_3$				
Dim				
1	0.0397	0.0001	0.0275	**0.3069**
2			0.0273	0.0052

9. EXTENSIONS IN PRODUCTION ENVIRONMENTS

The present approach can be extended to a network of data blocks. The **X**-matrices would play the role on input data blocks, while the **Y**-matrices are the output matrices. In between there can be any structure of data blocks that are consistent with the **X**- and **Y**-matrices. The task would be to find the weight vector associated with each **X**-matrix, such that the score vector propagated in the network would result in maximal **Y**-loading vectors. Here we shall consider as an example a situation that appears when working with production data. It will be briefly indicated how to find the weight vectors in different situation. The detailed analysis is not shown here, because it is a straightforward extension of the methods presented in previous sections.

9.1. Production data

We shall now consider some extensions to situations that are useful in studying the development of production processes. As a start consider the situation in Figure 5.

At Stage 0 there are available samples, $\mathbf{X}_0$, that are characteristic for the beginning of the production process. At Stage 1 a new set of samples are made available, $\mathbf{X}_1$. The results of Stage 1, quality requirements and other measures, are collected in a matrix $\mathbf{Y}_1$. At Stage 2 there have similarly been collected X-data of process measurements and Y-data of quality, performance and other response data. This setup is sufficient for the present analysis, but methods are the same if more stages are of interest to model.

The data samples can be viewed such that one sample is the result of one production process, that is, one production process generates one sample (row) in $\mathbf{X}_0$, $\mathbf{X}_1$, $\mathbf{Y}_1$, $\mathbf{X}_2$ and $\mathbf{Y}_2$. When a new sample $\mathbf{x}_{0,0}$ is available at Stage 0, it may be needed to estimate the samples $\mathbf{x}_{10}$, $\mathbf{y}_{10}$, $\mathbf{x}_{20}$ and $\mathbf{y}_{20}$, where $\mathbf{x}_{10}$ is a new sample at Stage 1 for $\mathbf{X}_1$ and similarly for the others. At Stage 1, when values of $\mathbf{x}_{10}$ are available, it may be needed to estimate the samples $\mathbf{y}_{10}$, $\mathbf{x}_{20}$ and $\mathbf{y}_{20}$, where $\mathbf{y}_{10}$ is the output of Stage 1, $\mathbf{x}_{20}$ the expected results of the process variables at Stage 2, and $\mathbf{y}_{20}$ the output of Stage 2. Finally, when $\mathbf{x}_{20}$ has become available at Stage 2, it may be needed to estimate the output, $\mathbf{y}_{20}$.

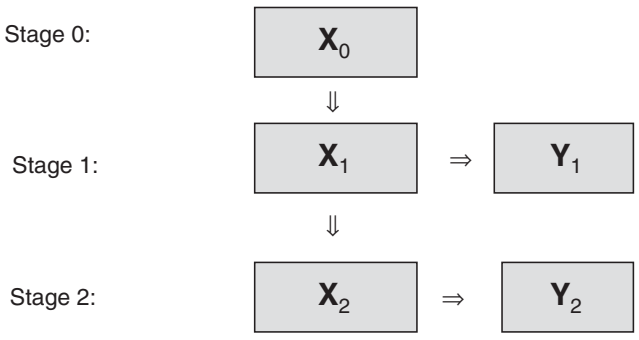

Figure 5. Schematic illustration of three stages in a production process.

DOI: 10.1002/cem

9.2. Modelling seen from Stage 0

Figure 6 illustrates schematically the available data and the task of modelling.

The estimation of each of $\mathbf{x}_{10}$, $\mathbf{y}_{10}$, $\mathbf{x}_{20}$ and $\mathbf{y}_{20}$, when $\mathbf{x}_{00}$ is known can be carried out by a standard regression analysis. We could carry out the four regressions $\mathbf{X}_0 \rightarrow \mathbf{X}_1$, $\mathbf{X}_0 \rightarrow \mathbf{Y}_1$, $\mathbf{X}_0 \rightarrow \mathbf{X}_2$ and $\mathbf{X}_0 \rightarrow \mathbf{Y}_2$. But this may not be the best approach. The objective of modelling is to provide with as good predictions of $\mathbf{y}_{10}$ and $\mathbf{y}_{20}$ as possible. For that purpose good values of $\mathbf{x}_{10}$ and $\mathbf{x}_{20}$ are needed, which give good estimates of $\mathbf{y}_{10}$ and $\mathbf{y}_{20}$. Thus, the best approach need not be to carry out the regressions $\mathbf{X}_0 \rightarrow \mathbf{X}_1$ and $\mathbf{X}_0 \rightarrow \mathbf{X}_2$, but to model the path in question. The task is to start with a loading vector $\mathbf{w}_0$ for $\mathbf{X}_0$, compute the score vector $\mathbf{t}_0 = \mathbf{X}_0\mathbf{w}_0$, and then compute the loading and score vectors that propagate further in the network. The optimisation task is to find $\mathbf{w}_0$ that maximises the total size of the loading vectors for the output matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$. The optimisation task is not shown here, but in next section it is shown for the next modelling stage. When the weight vector $\mathbf{w}_0$ has been found, score and loading vectors for the later matrices are found, and they are used to compute the regression coefficients between data blocks as shown previously.

9.3. Modelling at Stage 1

At Stage 1 the results of the samples $\mathbf{x}_{00}$ and $\mathbf{x}_{10}$ are known. The task of modelling is schematically illustrated in Figure 7.

$\mathbf{X}_0$ and $\mathbf{X}_1$ are the input data blocks. We want to use the model to estimate $\mathbf{y}_{10}$, $\mathbf{x}_{20}$ and $\mathbf{y}_{20}$, when the samples $\mathbf{x}_{00}$ and $\mathbf{x}_{10}$ have become available. The task here is to find a weight vector $\mathbf{w}_0$ for $\mathbf{X}_0$ and a weight vector $\mathbf{w}_1$ for $\mathbf{X}_1$ such that the resulting loading vectors for $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are as large as possible. There are two loading vectors computed for both $\mathbf{Y}_1$ and $\mathbf{Y}_2$. The loading vectors for $\mathbf{Y}_1$ are computed as $\mathbf{q}_{10} = \mathbf{Y}_1^{\mathrm{T}}\mathbf{t}_0 = \mathbf{Y}_1^{\mathrm{T}}\mathbf{X}_0\mathbf{w}_0$ and $\mathbf{q}_{11} = \mathbf{Y}_1^{\mathrm{T}}\mathbf{t}_1 = \mathbf{Y}_1^{\mathrm{T}}\mathbf{X}_1\mathbf{w}_1$. Those of $\mathbf{Y}_2$ are similarly computed as $\mathbf{q}_{20} = \mathbf{Y}_2^{\mathrm{T}}\mathbf{t}_{20} = \mathbf{Y}_2^{\mathrm{T}}\mathbf{X}_2\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_0\mathbf{w}_0$ and $\mathbf{q}_{21} = \mathbf{Y}_2^{\mathrm{T}}\mathbf{t}_{21} = \mathbf{Y}_2^{\mathrm{T}}\mathbf{X}_2\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\mathbf{w}_1$. For $\mathbf{q}_1 = \mathbf{q}_{10} + \mathbf{q}_{11}$ and $\mathbf{q}_2 = \mathbf{q}_{20} + \mathbf{q}_{21}$ it is desired to make $|\mathbf{q}_1|^2 + |\mathbf{q}_2|^2$ as large as possible. By expressing $|\mathbf{q}_1|^2 + |\mathbf{q}_2|^2$ in terms of $\mathbf{w}_0$ and $\mathbf{w}_1$, we arrive at

$$
\begin{aligned}
&|\mathbf{q}_1|^2 + |\mathbf{q}_2|^2 \\
&\quad = \mathbf{w}_0^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{E}\mathbf{X}_0\mathbf{w}_0 + 2\,\mathbf{w}_0^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{E}\mathbf{X}_1\mathbf{w}_1 + \mathbf{w}_1^{\mathrm{T}}\mathbf{X}_1^{\mathrm{T}}\mathbf{E}\mathbf{X}_1\mathbf{w}_1 \\
&\qquad + \mathbf{w}_0^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{F}\mathbf{X}_0\mathbf{w}_0 + 2\,\mathbf{w}_0^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{F}\mathbf{X}_1\mathbf{w}_1 + \mathbf{w}_1^{\mathrm{T}}\mathbf{X}_1^{\mathrm{T}}\mathbf{F}\mathbf{X}_1\mathbf{w}_1,
\end{aligned}
$$

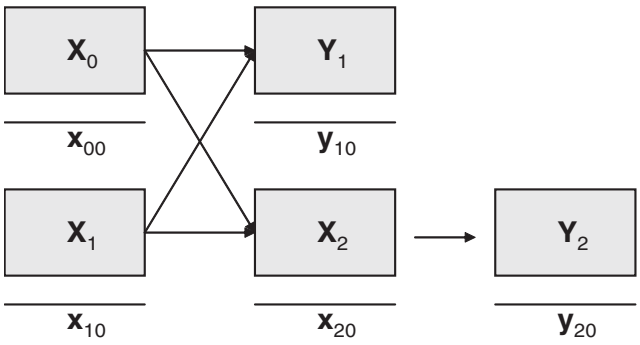

Figure 7. Schematic illustration of the modelling task at Stage 1.

where $\mathbf{E} = \mathbf{Y}_1\mathbf{Y}_1^{\mathrm{T}}$ and $\mathbf{F} = \mathbf{X}_2\mathbf{X}_2^{\mathrm{T}}\mathbf{Y}_2\mathbf{Y}_2^{\mathrm{T}}\mathbf{X}_2\mathbf{X}_2^{\mathrm{T}}$. Adding the side conditions $\lambda_i(\mathbf{w}_i - 1)$ for $i = 0,1$, the Lagrange techniques results in the set of equations

$$
\begin{aligned}
\mathbf{X}_0^{\mathrm{T}}\mathbf{H}\,\mathbf{X}_0\,\mathbf{w}_0 \;+\; \mathbf{X}_0^{\mathrm{T}}\mathbf{H}\,\mathbf{X}_1\,\mathbf{w}_1 &= \lambda_0\,\mathbf{w}_0 \\
\mathbf{X}_1^{\mathrm{T}}\mathbf{H}\,\mathbf{X}_0\,\mathbf{w}_0 \;+\; \mathbf{X}_1^{\mathrm{T}}\mathbf{H}\,\mathbf{X}_1\,\mathbf{w}_1 &= \lambda_1\,\mathbf{w}_1
\end{aligned}
$$

Here $\mathbf{H} = \mathbf{E} + \mathbf{F}$. These equations show how the set of equations are defined for a larger network of data blocks. There is one equation for each input data matrix. The matrix $\mathbf{H}$ collects the matrices associated with the 'paths' from input matrices to the output ones. When the weight vectors $\mathbf{w}_0$ and $\mathbf{w}_1$ have been found, the score and loading vectors of later data blocks are determined, and used to compute the regression coefficients between the data blocks as described previously in this paper.

10. CONCLUSION

Here we have presented a unified approach to model a network of data blocks. The methods of standard regression analysis are extended to model relationships between data blocks. The emphasis has been on to show how to find weight vectors that generate the score vectors that the regressions are based upon.

The first part considers a set of input matrices ($\mathbf{X}_1$, $\mathbf{X}_2$,...,$\mathbf{X}_L$) and a set of output matrices ($\mathbf{Y}_1$, $\mathbf{Y}_2$,...,$\mathbf{Y}_M$). It is shown how we proceed, when each $\mathbf{X}_i$ is supposed to model each $\mathbf{Y}_j$, $\mathbf{X}_i \rightarrow \mathbf{Y}_j$, $i = 1,...L$, $j = 1,..,M$. Using the structure of

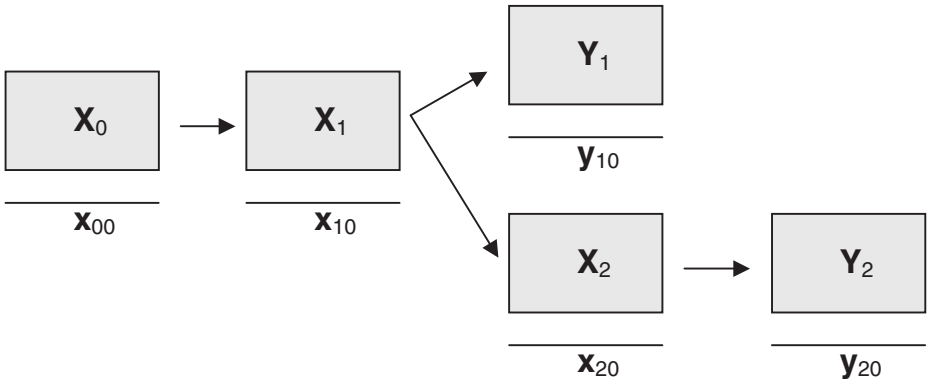

Figure 6. Schematic illustration of the modelling task as seen from Stage 0.

DOI: 10.1002/cem

production data as a prototype, it is shown how these methods extend to a general network of matrices.

The validation of the model can be carried out by a typical cross-validation, where a certain amount of samples are excluded. The regression equations between the data blocks can be used to estimate how original input sample propagates in the network or path in an analogous way as in ordinary regression analysis.

Graphic analysis can be carried out in a similar way as in regression analysis. There are more choices here. There are score vectors of one data block, which can be related to score vectors later in the network. The important ones are the score vectors of the output matrices and the score vectors earlier in the network.

The methods presented here allow a detailed analysis of parts of data. It can in general be recommended to try out some subdivisions of data even if the final analysis may result in one or few regression analysis only. By working with regression analysis between different data blocks one learns about the dependencies between parts of data, which may give useful knowledge about data.

The basic idea of these methods is to obtain maximal size of the resulting loading vectors at the output matrices. Before analysis it is necessary to scale all data matrices, that is, such that each variable has unit variance. After the analysis the data are scaled back like in standard regression analysis. This topic may need further analysis for specific sets of data blocks. But this is not considered closely here.

A proper understanding of the methods presented is important. Let us consider an example. Suppose that there are three matrices in a path, $\mathbf{X}_1 \rightarrow \mathbf{X}_2 \rightarrow \mathbf{X}_3$. One sequence always identifies three score vectors $\mathbf{t}_1$, $\mathbf{t}_2$ and $\mathbf{t}_3$. If the X-blocks are equal, the score vectors will also be equal. Therefore, the methods are useful to detect changes in a path. There can be as many sequences as the rank of $\mathbf{X}_1$. At some stage (reduced) $\mathbf{X}_2$ does not support further modelling, because $\mathbf{t}_2$ may not contribute to the predictions of $\mathbf{X}_3$-samples. In this case, the modelling task (of this part of the path) stops. In general, the score vectors can be viewed as 'supporting pillars' of the modelling along the path or network.

The methods presented are straightforward extensions of standard linear regression analysis. If there is a given procedure for allowing missing values in the data, it should not be difficult to implement it in the present framework.

When working with many data blocks, we sometimes find non-linearity in some of the data blocks. The non-linearity found is usually in a form of a curvature in the score space. This can be modelled by the methods presented in reference [16]. In a sequel paper, it will be shown how a smooth surface in low dimensions (typically of second or third order in the score vectors) can be determined within the framework of multi-block modelling.

In the applications of these methods it has been found necessary to identify the variables, which should be used in each data block. This is a large topic, which is also not considered here.

Finally, also not studied here, the results of analysis will depend on how the intermediate data blocks are adjusted before a new iteration, where a new set of score and loading vectors are computed.

These three topics, scaling data, variable selection and adjustment of data blocks, are all important topics to secure optimal predictions along the path of data blocks. Each of these topics is large and there are many options that need to be checked, when modelling a path of data blocks. These topics are out of the scope of the present paper and therefore not considered here.

REFERENCES

1. Höskuldsson A. Causal and path modelling. *Chemom. Intell. Lab. Syst.* 2001; **58**: 287–311.
2. Kvalheim OM, Liang YZ. Heuristic evolving latent projections: resolving two-way multicomponent data. *Anal. Chem.* 1992; **64**: 936–946.
3. Liang YZ, Kvalheim OM. Heuristic evolving latent projections: resolving hyphenated chromatographic profiles by component stripping. *Chemom. Intell. Lab. Syst.* 1993; **20**: 115–125.
4. Wold S, Kettaneh N, Tjessem K. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *J. Chemom.* 1996; **10**: 463–482.
5. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.* 1998; **12**: 301–321.
6. Smilde AK, Westerhuis JA, de Jong S. A framework for sequential multiblock component methods. *J. Chemom.* 2003; **17**: 323–337.
7. Michel Tenenhaus, Vincenzo Esposito Vinzi. PLS regression, PLS path modeling and generalized Procrustean analysis: a combined approach for multiblock analysis. *J. Chemom.* 2005; **19**: 145–153.
8. Myrtille Vivien , Robert Sabatier. Generalized orthogonal multiple co-inertia analysis (-PLS): new multiblock coponent and regression methods. *J. Chemom.* 2003; **17**: 287–301.
9. Höskuldsson A. PLS Regression Methods. *J. Chemom.* 1988; **2**: 211–228.
10. Horst P. Relations among m sets of variables. *Psychometrika* 1961; **26**: 126–149.
11. Carroll JD. A generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th Annual Convention of the American psychological Association.* 1968; 227–228.
12. Kettenring JR. Canonical analysis of several sets of variables. *Biometrika* 1971; **58**: 433–451.
13. Van de Geer JP. Linear relations among K sets of variables. *Psychometrika* 1984; **49**: 79–94.
14. Ten Berge JMF. Generalised approaches to the MAXBET problem and the MAXDIFF problem with applications to canonical correlations. *Psychometrika* 1988; **53**(4): 487–494.
15. Svinning K, Datu KA. Prediction of microstructure and properties of Portland Cement from production condition in cement mill. Part I: Evaluation of prediction models, *11th International congress on the chemistry of cement, Durban* 2003.
16. Höskuldsson A. The Heisenberg modelling procedure and application to non-linear modelling. *Chemom. Intell. Lab. Syst.* 1998; **44**: 15–30.

DOI: 10.1002/cem