Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 88 (2007) 84–99
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Path modeling and process control

Agnar Höskuldsson [a,*], Oxana Rodionova [b], Alexey Pomerantsev [b]

[a] *Technical University of Denmark, Building 358, 2800 Lyngby, Denmark*
[b] *Institute of Chemical Physics, Kosygin Str. 4, Moscow, 119991, Russia*

Received 18 April 2006; received in revised form 30 August 2006; accepted 28 September 2006
Available online 21 November 2006

Abstract

Many production processes are carried out in stages. At the end of each stage, the production engineer can analyze the intermediate results and correct process parameters (variables) of the next stage. Both analysis of the process and correction to process parameters at next stage should be performed regarding the foreseeable output property *y*, and with respect to an admissible range of correcting actions for the parameters of the next stage. In this paper the basic principles of path modeling is presented. The mathematics is presented for processes having only one stage, having two stages and having three or more stages. The methods are applied to a process control of a multi-stage production process having 25 variables and one output variable. When moving along the process, variables change their roles. It is shown how the methods of path modeling can be applied to estimate variables of the next stage with the purpose of obtaining optimal or almost optimal quality of the output variable. An important aspect of the methods presented is the possibility of extensive graphic analysis of data that can provide the engineer with a detailed view of the multi-variate variation in data.

Keywords: Path modeling; Process control; H-principle; PLS regression; Latent structure

1. Introduction

In industry there is great interest for numerical methods that can be used in process control of the production. There have been developed numerous methods for process control. But most of them are designed for situations, where there are relatively few variables that are measured during the production process. But conditions are changing. Now the companies are measuring numerous variables and using different types of measurement instruments. For instance, if an NIR (Near Infra-Red) instrument is being used, it automatically generates typically 1056 values for each sample measured. A company using NIR instruments for process control may have hundreds of them.

An important trend today is that authorities are providing recommendations or requirements concerning the process control with the purpose of securing the quality of the final product. An example is FDA in USA that is setting up prescriptions for process control in order to secure quality of food and medical products.

In recent years it has been successful to implement *predictive control*. It means that models are used to predict some future measurements. When the actual measurement is carried out, the measured value is compared with the predicted one. If there are too large differences between the measured and the predicted ones, it indicates that something is wrong. When the operators can see the predicted values for the future development, they may decide that the values are not good and change the conditions for the future values. The methods presented here are provided with models that can be used for effective predictive control.

Linear regression is a method that is often used, when predictions are needed. Instrumental data are collected in a matrix **X**, and the output (quality or result-data) are collected in a matrix **Y**. These data are used to estimate the parameters **B** in a lines model, $\mathbf{Y}^{*}=\mathbf{XB}$. When a new sample is available, $\mathbf{x}_0$, the predictions are computed from the model $\mathbf{y}_0^{*}=\mathbf{B}^{\mathrm{T}}\mathbf{x}_0$. Although this approach is often good, it is frequently not adequate. There may be many reasons for this. The parameters **B** may change with time or the linear model may not be adequate to provide

* Corresponding author.
E-mail address: ah@ipl.dk (A. Höskuldsson).

0169-7439/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.chemolab.2006.09.010

good predictions. What is often needed is a more detailed description of the production processes and models that reflect the description.

An example is where the production is organized as batch processes. Each batch can be organized in stages, where at each stage a certain amount of variables is measured. The variables at later stages are measured at a later time than at previous stages. Thus, there are typically two time concepts. One is from sample to sample and another is within a sample (a batch). This is illustrated schematically in Fig. 1. The data matrix $\mathbf{X}$ can be partitioned in parts as $\mathbf{X}=(\mathbf{X}_1 |\mathbf{X}_2|\ldots|\mathbf{X}_n)$. The measurements in $\mathbf{X}_1$ are the ones found at stage 1 and similarly for the other parts. When new values, $\mathbf{x}_{10}$, from the first stage are obtained, one would like to know if these are good values. Can values at later stages, $\mathbf{x}_{20},\ldots,\mathbf{x}_{n0}$, be adequately estimated from $\mathbf{x}_{10}$, and how good can we expect the output results, $\mathbf{y}_0$, for this batch to be? The methods presented here are designed to deal with this way of structuring data.

The methods in this paper are based on the path modeling methods developed in Ref. [1]. In this paper the basic methods are shown. Thus the presentation in this paper supplements the ones in Ref. [1]. The basic methods of path modeling are concerned with a network of data blocks, where certain matrices (data blocks) are defined as the instrumental data as input data and some other as output data blocks, the **Y**'s. Between the input and output data blocks there can be any amount of data blocks that make up the network. This paper is concerned with a simple network, where a linear regression, $\mathbf{X} \Rightarrow \mathbf{Y}$, is extended to $\mathbf{X} \Rightarrow \mathbf{Z}_1 \Rightarrow \mathbf{Z}_2 \Rightarrow \ldots \Rightarrow \mathbf{Z}_n \Rightarrow \mathbf{Y}$. Regression coefficients for this is developed. If $\mathbf{Y}_e, \mathbf{Z}_{ne},\ldots,\mathbf{Z}_{1e}$ denote the computed values, the regression coefficients give $\mathbf{Y}_e=\mathbf{Z}_n\mathbf{B}_y$, $\mathbf{Z}_{ne}=\mathbf{Z}_{n-1}\mathbf{B}_n,\ldots$, $\mathbf{Z}_{2e}=\mathbf{Z}_1\mathbf{B}_2$ and $\mathbf{Z}_{1e}=\mathbf{X}\mathbf{B}_1$, This means that when a new **X**-sample, $\mathbf{x}_0$, is available, these regression coefficients are used to compute the estimate for a $\mathbf{Z}_1$-sample, $\mathbf{z}_{10}$, $\mathbf{Z}_2$-sample, $\mathbf{z}_{20},\ldots$, $\mathbf{Z}_n$-sample, $\mathbf{z}_{n0}$ and an **Y**-sample, $\mathbf{y}_0$. In Section 2 the ideas of path modeling are considered more closely.

In Section 3 it is shown how modeling data that are collected in batch processes fit within the present framework. It is emphasized how the estimation procedure should be carried out, what an overall criterion of the estimation should be and what results are important for the analysis.

In Section 4 the case of only one data matrix **X** is considered. The results and interpretation in this case are similar to the ones in more general context. Also the basic formulae are the same as in a general network of data blocks.

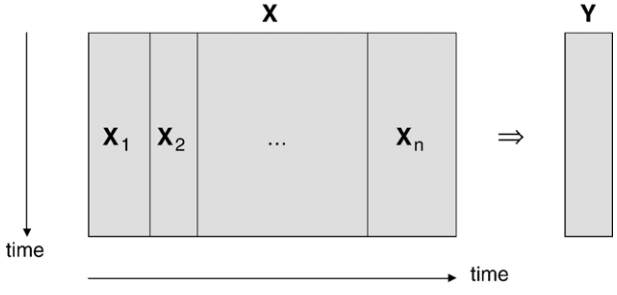

Fig. 1. Schematic illustration of two time concepts in batch processes.

In Section 5 the case of linear regression, $\mathbf{X} \Rightarrow \mathbf{Y}$, is treated. It is shown how the optimization for one matrix extends to the case of two matrices. Furthermore, it is shown how the results of the optimization task are illustrated graphically.

The methods presented are of mathematical type. The same algorithm is used for any number of matrices in a path. Therefore, it has been chosen to present the basic formulae for one and two data matrices. Once the algorithm has been understood for these two cases it immediately follows the general formulae for the arbitrary number of data matrices.

In Section 6 the case of three data matrices, **X**, **Z** and **Y**, is considered, where the modeling task is $\mathbf{X} \Rightarrow \mathbf{Z} \Rightarrow \mathbf{Y}$. It is shown how the optimization task for two matrices extends naturally to the case of three or more matrices. It is shown how the regression coefficients are estimated and used to provide predictions for later stages of the process.

The methods presented are based on the H-principle of mathematical modeling [2]. It suggests that the modeling task should be carried out in steps, where at each step the estimation and precision aspect of the model should be balanced with the purpose of obtaining optimal predictions. It is an important aspect of the method presented that the same algorithm is used for one, two, three or more matrices. In the case of one matrix **X** the interpretation is based on $\mathbf{X} \Rightarrow \mathbf{X}$.

In Section 7 the methods are applied to the process data that are presented in Section 3. It is shown how the modeling task of $\mathbf{X} \Rightarrow \mathbf{Z} \Rightarrow \mathbf{Y}$ can be restructured in different ways depending on which stage we are at. It is shown how graphic procedures can be used to illustrate different features in the data.

In Ref. [1] the theory of path modeling is presented, which is based on the H-principle. The present paper presents the theory from a different point of view. The path modeling here is viewed as extensions of modeling one data or two data blocks. It is started with one data block and all measures, regression coefficients and graphic procedures are developed. The regression task is viewed as deriving the data **X** from itself. This is extended to two data blocks, $\mathbf{X} \Rightarrow \mathbf{Y}$, such that the same procedures are used. If **Y** is replaced by **X**, the results of one data block are obtained. This approach makes it natural to extend the procedures to three or more data blocks. If the data blocks are organized in a path, for instance as $\mathbf{Z}_1 \Rightarrow \mathbf{Z}_2 \Rightarrow \ldots \Rightarrow \mathbf{Z}_n$, the procedures reduce to linear regression if $n=2$, and to one block if $n=1$. Here it is argued that the data at different stages in a batch process adequately can be viewed as data blocks in a path. Furthermore, it is argued that the path modeling approach is the natural approach to modeling the stages of the batch, where the purpose of the modeling is the prediction of the final quality.

The theory of path modeling is extensive, see Ref. [3]. The presentation of path modeling in this paper is different from what is found in the literature. Most of the theory of path modeling in the literature can be viewed as the analysis of the correlation structure among the data blocks. It is also typically only concerned with finding one latent vector (latent variable) for each data block, see Refs. [3–7]. The papers in the literature are very much focused on the estimation aspect in these types of models, see Ref. [8]. Path modeling can be viewed as multi-block modeling. In Ref. [9] there is a review and a framework for different

methods to analyse multi-block data. It also contains extensive references to literature on multi-block methods. The methods presented here are different from the ones in the literature. The focus here is to use the H-principle to secure optimal predictions along the path, from the starting input data block to the propagated values in the path.

2. Path modeling

Here some basic ideas of path modeling are reviewed. It is assumed that there are given a collection of data matrices that make a network of interconnected data. In a linear regression there are given two data matrices, $\mathbf{X} \Rightarrow \mathbf{Y}$. When a new X-sample, $\mathbf{x}_0$, is available, the linear model is used to estimate the Y-sample, $\mathbf{y}_0 = \mathbf{B}^{\mathrm{T}}\mathbf{x}_0$. In case of three data blocks, $\mathbf{X} \Rightarrow \mathbf{Z} \Rightarrow \mathbf{Y}$, prediction is wanted for a Z-sample and a Y-sample, when X-sample $\mathbf{x}_0$ is available, $\mathbf{z}_0^* = \mathbf{B}_x^{\mathrm{T}}\mathbf{x}_0$ and $\mathbf{y}_0^* = \mathbf{B}_z^{\mathrm{T}}\mathbf{z}_0^*$. Path modeling is concerned with a directed network of data blocks, which means that each data block is connected to one or more data blocks later in the network. There are given a number of input data matrices $\mathbf{X}_i$ and some output matrices $\mathbf{Y}_j$. There is direction in the network in the sense that each data matrix is supposed to explain one or more data matrices later in the network. The starting point is the input data matrices. The task is to establish regression relationship between the data blocks such that when the input samples become available, $\mathbf{x}_{i0}$, $i = 1,\ldots,I$, intermediate and output samples can be estimated by regression relationships. In this paper there is one input matrix, $\mathbf{X}$, one output matrix, $\mathbf{Y}$, and a set of intermediate data blocks reflecting the time sequence of the measurements, $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$. The path considered is $\mathbf{X} \Rightarrow \mathbf{Z}_1 \Rightarrow \ldots \Rightarrow \mathbf{Z}_n \Rightarrow \mathbf{Y}$.

The modeling follows the recommendations of the H-principle. It suggests that the modeling should be carried out in steps. At each step there should be applied a weighing procedure that reflects the emphasis of the analysis. In the present case at each step there should be found a weight vector $\mathbf{w}$ such that the score vector $\mathbf{t} = \mathbf{X}\mathbf{w}$ has some optimal properties. The H-principle suggests that the estimation and the precision aspect of the model should be determined such that the prediction is optimized. In the path considered the score vector $\mathbf{t}$ generates loading and score vector at later stages as given by

$$\begin{aligned}\mathbf{p}_1 &= \mathbf{Z}_1^{\mathrm{T}}\mathbf{t}, \mathbf{t}_1 = \mathbf{Z}_1\mathbf{p}_1, \mathbf{p}_2 = \mathbf{Z}_2^{\mathrm{T}}\mathbf{t}_1, \mathbf{t}_2 = \mathbf{Z}_2\mathbf{p}_2, \ldots, \mathbf{p}_n \\ &= \mathbf{Z}_n^{\mathrm{T}}\mathbf{t}_{n-1}, \mathbf{t}_n = \mathbf{Z}_n\mathbf{p}_n, \mathbf{q} = \mathbf{Y}^{\mathrm{T}}\mathbf{t}_n, \mathbf{u} = \mathbf{Y}\mathbf{q}. \end{aligned} \quad (1)$$

This sequence of vectors is schematically illustrated in Fig. 2. Each of the vectors has an interpretation as 'the effect of'. For instance, the projection of $\mathbf{Z}_1$ onto $\mathbf{t}$ is $\mathbf{t}\mathbf{p}_1^{\mathrm{T}}/(\mathbf{t}^{\mathrm{T}}\mathbf{t})$.

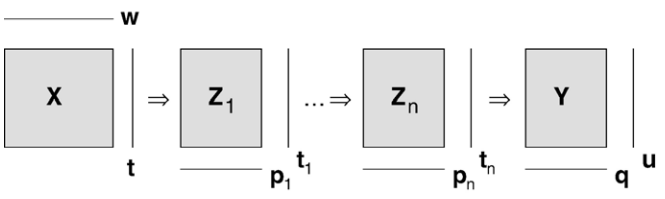

Fig. 2. Schematic illustration of the sequence of vectors of Eq. (1).

Thus $\mathbf{p}_1$ is proportional to coefficients of projecting $\mathbf{Z}_1$ onto $\mathbf{t}$. The H-principle suggests that the loading vector of the data matrix, which is to be predicted, should be maximized. Here the primary concern is the $\mathbf{Y}$-matrix. Thus a weight vector $\mathbf{w}$ should be found such that the size of the loading vector $\mathbf{q}$ is maximized, max $|\mathbf{q}|$, subject to $|\mathbf{w}| = 1$. A weight vector $\mathbf{w}$ chosen in this way secures best possible predictions of new Y-samples in the present path. The predictions obtained for the $\mathbf{Z}_i$-samples, $\mathbf{z}_{i0}$, may not be optimal for predicting Z_i-samples. But it is 'the best' for providing with good predictions for the Y-samples. When analyzing paths it may be important to detect if some of the $\mathbf{Z}_i$s are not functioning well. It might be that for instance $\mathbf{Z}_2$ is not good, either that $\mathbf{p}_2 = \mathbf{Z}_2^{\mathrm{T}}\mathbf{t}_1$ is close to zero or $\mathbf{p}_3 = \mathbf{Z}_3^{\mathrm{T}}\mathbf{t}_2$ is zero. If $\mathbf{p}_2$ is zero, $\mathbf{Z}_1$ cannot describe $\mathbf{Z}_2$, and if $\mathbf{p}_3$ is zero $\mathbf{Z}_2$ cannot describe $\mathbf{Z}_3$.

It is important to understand the task of path modeling. We are looking for a weight vector $\mathbf{w}$ such that the predictions of the $\mathbf{Z}_i$-samples 'fit' in the path and give good predictions of the $\mathbf{Y}$-samples. The operators get the 'best' estimates of the variables at the intermediate stages and an optimal linear prediction of the output. Note that it might be possible to get more precise predictions of $\mathbf{Y}$-samples, if only the $\mathbf{X}$-data are used. The more detailed path that is specified the more requirements are desired for the data in order to get good predictions.

The sequence (1) is generated for each step. When the sequence of vectors has been computed, the matrices ($\mathbf{X}$, $\mathbf{Z}_1$,.., $\mathbf{Z}_n$ and $\mathbf{Y}$) are adjusted appropriately before a new sequence is found. An important aspect of the analysis is the possibility of graphic analysis of the path. Plots of score vectors show the changes of the samples along the path, and plot loadings the changes in the correlation structure. If all matrices are equal, $\mathbf{X} = \mathbf{Z}_1 = \ldots = \mathbf{Z}_n = \mathbf{Y}$, the score vectors (and loading vectors) will be equal apart from a scaling constant. In the general case the plot of score vectors can reveal changes that have occurred in the data path.

The methods of this paper require appropriate scaling of all data blocks. The reason is that the strength of relationship is measured by the covariance matrices. If data are not scaled, the results may depend on some variables that have large variance but not very good modeling power. Another aspect is the case of many variables. There it is necessary to scale the data in order to secure numerical stability of the algorithm. When working with industrial data we typically use reduced rank solutions because they show better prediction ability than full rank solutions. Therefore, the procedure of scaling data is considered here briefly.

The scaling of variables of data ($\mathbf{X}$,$\mathbf{Y}$) amounts to multiplying by diagonal matrices. The scaled data are $\mathbf{X}_1 = \mathbf{X}\mathbf{C}_1$ and $\mathbf{Y}_1 = \mathbf{Y}\mathbf{D}_1$. The linear least squares solution is given by $\mathbf{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. If it is computed for the scaled data, the result is $\mathbf{B}_1 = (\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{Y}_1 = \mathbf{C}_1^{-1}\mathbf{B}\mathbf{D}_1$. It gives $\mathbf{B} = \mathbf{C}_1\mathbf{B}_1\mathbf{D}_1^{-1}$. This shows that scaling of data can be viewed as a way to obtain a stable least squares solution. In the procedures the solution is computed in steps, where at each step a rank one contribution to the solution is computed. In the analysis only A terms are used in the solution. The same scaling is used for the rank A solution, $\mathbf{B}_A = \mathbf{C}_1\mathbf{B}_{1,A}\mathbf{D}_1^{-1}$. All measures are computed for the scaled data. Also, the graphic procedures are based on the scaled data.

3. Batch process data

We consider a multi-stage continuous technological process that is represented by 25 process variables $\boldsymbol{x}$ and one output variable y, which is the final quality of the end-product. The data represent the production of a well-known Russian strong drink. The production cycle (see Fig. 3) is divided into seven stages numbered by Roman numerals. At each stage there certain variables are measured. Variables used on all previous stages are treated as the input fixed variables, current variables are the controlled ones, and variables that characterize next stages are out of scope on the current stage. Moving along the process, variables change their roles.

The first stage (I) is represented by six input variables (W1, W2, W3 and S1, S2, S3) that stand for the properties of the raw components S and W. At the second stage (II) component W is refined and variables WR1 and WR2 characterize this process. Variables CW1, CW2, and CW3 (Stage III) represent the properties of the outcome product CW. The next stage (IV) is the mixing of the raw component S and the refined component CW. The result M is characterized by variables M1, M2, and M3. Afterward, blend M is also refined (Stage V) with the process characteristics MR1 and MR2, and the properties of outcome CM are presented by variables CM1, CM2, and CM3 (Stage VI). The last stage (VII) stands for the ultimate amendments, which are done with additives A1,..., A6. The output variable (P=y) is the final product quality. The data used are from 154 batches.

At the end of each stage, the production engineer could analyze the intermediate results and correct process parameters (variables) of the next stage. Both analysis and correction should be performed regarding the foreseeable output property y, and with respect to the admissible range of correcting actions for the next stage.

4. One data block

Assume that there is given one data block, $\mathbf{X}$ an $N \times K$ matrix, $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_K)$. $\mathbf{x}_k$ is the kth column of $\mathbf{X}$ and $\mathbf{x}^n$ the nth row of $\mathbf{X}$. We shall start by looking at the decomposition of $\mathbf{X}$. Then it is considered how we can report from the decomposition procedure.

4.1. Decomposition of $\mathbf{X}$. Weighing variables

4.1.1. General decomposition procedure

The H-principle suggests that the modeling task should be carried out in steps. The reason for this is that in practice data have their own 'identity'. There may be reduced dimension in data; there may appear unexpected non-linearity or unforeseen grouping of the samples. It is important to detect, when data says 'stop'. At each step we are looking for a good score vector $\mathbf{t}$, which is computed by

$$\mathbf{t} = \mathbf{X}w = w_1\mathbf{x}_1 + \ldots + w_K\mathbf{x}_K. \tag{2}$$

Between steps the matrix $\mathbf{X}$ is reduced. The reduction is carried out by

$$\mathbf{X} \leftarrow \mathbf{X} - d\,\mathbf{t}\,\mathbf{p}^{\mathrm{T}}, \text{ with } \mathbf{p} = \mathbf{X}^{\mathrm{T}}\mathbf{t} \text{ and } d = 1/(\mathbf{t}^{\mathrm{T}}\mathbf{t}) \tag{3}$$

The vectors involved at each step are schematically illustrated in Fig. 4. The matrix $\mathbf{X}$ in Eq. (2) will typically be the reduced one, $\mathbf{X}=\mathbf{X}_{(A)}$,

$$\mathbf{X}_{(\mathrm{A})} = \mathbf{X} - \left(d_1\,\mathbf{t}_1\,\mathbf{p}_1^{\mathrm{T}} + \ldots + d_A\,\mathbf{t}_A\,\mathbf{p}_A^{\mathrm{T}}\right). \tag{4}$$

The index A refers to the numerical step in the decomposition and A can be any integer between 1 and min(N,K). Note that ($d\,\mathbf{t}\,\mathbf{p}^{\mathrm{T}}$) in Eq. (3) can be viewed as the projection of $\mathbf{X}$ onto $\mathbf{t}$. Thus the score vectors will be orthogonal independent of the choice of weight vectors, $\mathbf{w}$'s. In order to simplify the notation the index is often dropped, like in Eq. (3), when there is no risk of misunderstanding. At each step there are generated vectors, $\mathbf{w}$, $\mathbf{t}$ and $\mathbf{p}$. These vectors are collected in matrices $\mathbf{W}=(\mathbf{w}_1,\ldots,\mathbf{w}_A)$, $\mathbf{T}=(\mathbf{t}_1,\ldots,\mathbf{t}_A)$, $\mathbf{P}=(\mathbf{p}_1,\ldots,$

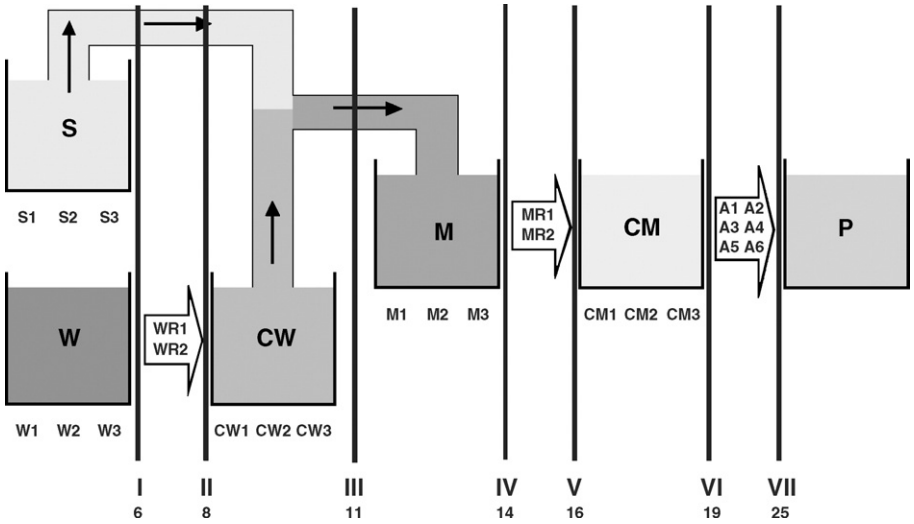

Fig. 3. Production cycle.

$\mathbf{p}_A$). Besides these vectors there are needed the **r**-vectors. They are generated by the formula,

$$\mathbf{r}_A = \mathbf{w}_A - \left[\left(\mathbf{w}_A^{\mathrm{T}}\mathbf{p}_1\right)\mathbf{r}_1 + \ldots \left(\mathbf{w}_A^{\mathrm{T}}\mathbf{p}_{A-1}\right)\mathbf{r}_{A-1}\right] \quad (\mathbf{r}_1 = \mathbf{w}_1). \tag{5}$$

They are generated from the property,

$$\mathbf{t}_A = \mathbf{X}\mathbf{r}_A, \tag{6}$$

where **X** is the original data matrix. The matrix $\mathbf{R}=(\mathbf{r}_1,\ldots,\mathbf{r}_A)$ satisfies

$$\mathbf{R}^{\mathrm{T}}\mathbf{P} = \mathbf{D}^{-1}, \tag{7}$$

where **D** is the diagonal matrix with d_A in the diagonal, see Ref. [2]. Eq. (5) uses $\mathbf{X}_{(A)}=\mathbf{X}_{(A-1)}(\mathbf{I} - d_A\mathbf{w}_A\mathbf{p}_A^{\mathrm{T}})=\mathbf{X}(\mathbf{I} - d_1\mathbf{w}_1\mathbf{p}_1^{\mathrm{T}}) \times \ldots \times (\mathbf{I} - d_A\mathbf{w}_A\mathbf{p}_A^{\mathrm{T}})$. Thus it is easy to show that Eq. (5) does not depend on the way the **p**-vectors are given. This property of Eq. (5) is used in later sections, when computing the regression coefficients between data blocks.

4.1.2. Interpretation of the vectors

4.1.2.1. $\mathbf{w}_A$*, the weight vector.* The weight vector is found from the task in question. There is no restriction on $\mathbf{w}_A$ except that the resulting score vector **t** may not be the zero vector. For any such choice of $\mathbf{w}_A$, Eqs. (2)–(7) are valid. Usually, the weight vector is scaled to unit length.

4.1.2.2. $\mathbf{t}_A$*, the score vector.* It is the result of weighing of the variables, which represent the 'profile' in data at this step. It is sometimes said that the score matrix **T** represent the latent structure in data. Often the aim of the analysis is to obtain a score vector having some properties. E.g., in Principal Component Analysis, PCA, the task is to find a score vector $\mathbf{w}_A$ such that the resulting score vector $\mathbf{t}_A$ has as large size as possible. The reduction of **X**, Eq. (3), at each step implies that the score vectors are always mutually orthogonal whatever choices of the weight vectors there have been made.

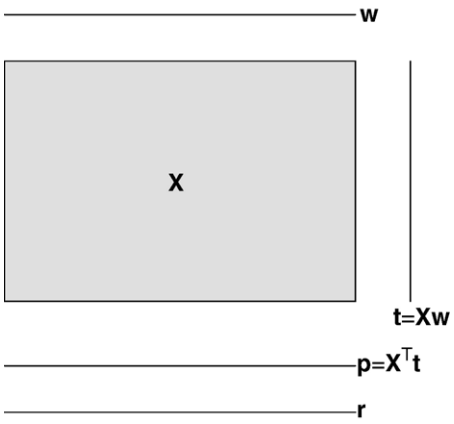

Fig. 4. Schematic illustration of vectors.

4.1.2.3. $\mathbf{p}_A$*, the loading vector.* It has the interpretation of being proportional to the correlation coefficients between the *Kx*-variables and the score vector $\mathbf{t}_A$. Therefore, it is often useful to study scatter plots of two loading vectors, because these plots show how the *x*-variables 'contribute' to the score vectors.

4.1.2.4. $\mathbf{r}_A$*, the transformation vector.* It has the interpretation (Eq. (6)) or $\mathbf{T}=\mathbf{XR}$, or $\mathbf{P}=\mathbf{X}^{\mathrm{T}}\mathbf{XR}$, where **X** is the original data matrix. The $\mathbf{r}_A$-vectors are used to study how the values of the score vectors are derived from the original data. They are also used to study how the loading vectors are derived from the correlation matrix.

4.1.2.5. $\mathbf{d}_A$*, the scaling constant.* It has an interpretation of a variance. In case of PCA, $d_A = 1 / \lambda_A$, where λ_A is the eigen value from $\mathbf{X}^{\mathrm{T}}\mathbf{Xw}=\lambda\mathbf{w}$. The scaling constants can be absorbed into either **T** or **P**. They are computed separately to secure numerical stability.

4.1.3. The mathematical decomposition of X and $\mathbf{X}^+$

Assume that there are more samples than variables, $K < N$. The decomposition of **X** given by Eqs. (2)–(5) can be written as

$$\mathbf{X} = d_1\mathbf{t}_1\mathbf{p}_1\mathrm{T} + \ldots + d_A\mathbf{t}_A\mathbf{p}_A^{\mathrm{T}} + \ldots + d_K\mathbf{t}_K\mathbf{p}_K T = \mathbf{TDP}^T \tag{8}$$

$$\mathbf{X}^+ = d_1\mathbf{r}_1\mathbf{t}_1^{\mathrm{T}} + \ldots + d_A\mathbf{r}_A\mathbf{t}_A^{\mathrm{T}} + \ldots + d_K\mathbf{r}_K\mathbf{t}_K^{\mathrm{T}} = \mathbf{RDT}^{\mathrm{T}}. \tag{9}$$

The matrix $\mathbf{X}^+$ is the generalized inverse of **X**, $\mathbf{X}=\mathbf{XX}^+\mathbf{X}$. This follows from (7). If $\mathbf{X}_A$ and $\mathbf{X}_A^+$ are the truncated versions, $\mathbf{X}_A=d_1\mathbf{t}_1\mathbf{p}_1^{\mathrm{T}}+\ldots+d_A\mathbf{t}_A\mathbf{p}_A^{\mathrm{T}}$, $\mathbf{X}_A^+$ is also the generalized inverse of $\mathbf{X}_A$, $\mathbf{X}_A=\mathbf{X}_A\mathbf{X}_A^+\mathbf{X}_A$. Usually only *A* terms are used in Eqs. (8) and (9) since further terms do not improve the modeling of **X**.

4.1.4. Regression coefficients

It is useful to apply the notation and concepts of the well-known regression analysis. If

$$\mathbf{B}_A = d_1\mathbf{r}_1\mathbf{p}_1^{\mathrm{T}} + \ldots + d_A\mathbf{r}_A\mathbf{p}_A^{\mathrm{T}} = \mathbf{R}_A\mathbf{D}_A\mathbf{P}_A^{\mathrm{T}}, \tag{10}$$

we can write $\mathbf{X}_A=\mathbf{XB}_A$. The estimate of **X** is given by $\mathbf{X}_A$ and the regression equation by $\mathbf{XB}_A$.

4.2. Decomposition of X. Weighing samples

The above shows the results of decomposing **X**, when the variables have been weighted. This approach is appropriate, when the rows are repeated samples or objects, where the same type of measurements is carried out. It is also common to see data where this interpretation is not suitable. An example is the situation of rating food by tasting samples of food. The variables of columns of **X** may represent the types of food,

while the rows are the persons tasting the food. The persons may be different and their rating may not be considered as repeated samples. In fact it may be natural to consider the persons as variables. The decomposition above can be applied to the case of weighing rows (samples) of **X** by considering the transpose of the formulae above. But for later purpose it will be natural in this case to change the notation. Let **v** be the weight vector for the rows. It is used to compute the loading vector **p**,

$$\mathbf{p} = \mathbf{X}^{\mathrm{T}}\mathbf{v} = v_1\mathbf{x}^1 + \ldots + v_N\mathbf{x}^N.$$

Here $\mathbf{x}^1$ is the first row of **X** and similarly for the other rows. The adjustment of **X** at each step is carried out as

$$\mathbf{X} \leftarrow \mathbf{X} - d\,\mathbf{t}\,\mathbf{p}^{\mathrm{T}}, \text{ with } \mathbf{t} = \mathbf{X}\mathbf{p} \text{ and } d = 1/\left(\mathbf{p}^{\mathrm{T}}\mathbf{p}\right).$$

This way of adjusting **X** secures that the matrix $\mathbf{P}_A=(\mathbf{p}_1,\ldots,\mathbf{p}_A)$ will have orthogonal columns. Like above (in the case of **r**-vectors) there are needed the **s**-vectors. They are generated by the formula,

$$\mathbf{s}_A = \mathbf{v}_A - \left[\left(\mathbf{v}_A^{\mathrm{T}}\mathbf{t}_1\right)\mathbf{s}_1 + \ldots + \left(v_A^{\mathrm{T}}\mathbf{t}_{A-1}\right)\mathbf{s}_{A-1}\right] \quad (11)$$

The matrix $\mathbf{S}_A=(\mathbf{s}_1,\ldots,\mathbf{s}_A)$ is computed such that

$$\mathbf{P}_A = \mathbf{X}^{\mathrm{T}}\mathbf{S}_A, \quad (12)$$

where **X** here is the original matrix. This corresponds to Eq. (6). Similar to Eq. (7) it can be shown that $\mathbf{S}_A^{\mathrm{T}}\mathbf{T}_A=\mathbf{D}_A^{-1}$. Eq. (8) will look the same, while for Eq. (9), we now have

$$\mathbf{X}^{+} = d_1\mathbf{p}_1\mathbf{s}_1^{\mathrm{T}} + \ldots + d_K\mathbf{p}_K\mathbf{s}_K^{\mathrm{T}} = \mathbf{P}\mathbf{D}\mathbf{S}^{\mathrm{T}} \quad (13)$$

Both $\mathbf{X}^{+}$ and the truncated version $\mathbf{X}_A^{+}$ will be the generalized inverses of **X** and $\mathbf{X}_A$ resp. The vectors involved, when decomposing **X** according to rows, are schematically shown in Fig. 5.

4.3. Decomposition of X. Weighing both variables and samples

It may be natural to weigh both columns and rows in cases like mentioned, where both rows and columns can be viewed as variables. In this case the there are weight vectors for columns, **w**, and rows, **v**, that give

$$\mathbf{t} = \mathbf{X}\mathbf{w} = w_1\mathbf{x}_1 + \ldots + w_K\mathbf{x}_K$$
$$\mathbf{p} = \mathbf{X}^{\mathrm{T}}\mathbf{v} = v_1\mathbf{x}^1 + \ldots + v_N\mathbf{x}^N.$$

The adjustment of **X** at each step is now carried out as

$$\mathbf{X} \leftarrow \mathbf{X} - d\,\mathbf{t}\,\mathbf{p}^{\mathrm{T}}, \text{ with } d = 1/\left(\mathbf{w}^{\mathrm{T}}\,\mathbf{X}^{\mathrm{T}}\mathbf{v}\right).$$

Neither the **t**'s nor the **p**'s will in general be orthogonal. But the adjustment will always give a rank one reduction of **X**. Thus there will be at most min(*N*,*K*) number of steps in the modeling of **X**. The **r**-vectors are generated by Eq. (5) and the **s**-vectors by Eq. (11). These vectors are generated to satisfy

$$\mathbf{T} = \mathbf{X}\mathbf{R} \quad \text{and} \quad \mathbf{P} = \mathbf{X}^{\mathrm{T}}\mathbf{S},$$

with **X** as the original matrix. They have the property

$$\mathbf{R}^{\mathrm{T}}\mathbf{P} = \mathbf{D}^{-1} \quad \text{and} \quad \mathbf{S}^{\mathrm{T}}\mathbf{T} = \mathbf{D}^{-1}.$$

The decompositions of **X** and $\mathbf{X}^{+}$ have now the form,

$$\mathbf{X} = d_1\mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \ldots + d_A\mathbf{t}_A\mathbf{p}_A^{\mathrm{T}} + \ldots + d_K\mathbf{t}_K\mathbf{p}_K^{\mathrm{T}} = \mathbf{T}\mathbf{D}\mathbf{P}^{\mathrm{T}}$$
$$\mathbf{X}^{+} = d_1\mathbf{r}_1\mathbf{s}_1^{\mathrm{T}} + \ldots + d_A\mathbf{r}_A\mathbf{s}_A^{\mathrm{T}} + \ldots + d_K\mathbf{r}_K\mathbf{s}_K^{\mathrm{T}} = \mathbf{R}\mathbf{D}\mathbf{S}^{\mathrm{T}}.$$

Both $\mathbf{X}^{+}$ and its truncated version, $\mathbf{X}_A^{+}$, are generalized inverses of **X** and $\mathbf{X}_A$, resp. The vectors involved, when *X* is decomposed according to both variables and samples, are schematically illustrated in Fig. 6.

4.3.1. Algorithmic considerations

The same algorithm is used in all three cases treated above, weighing columns, weighing rows and weighing both columns and rows. If only columns are weighted, the weight vector **v** is chosen as scaled **t**, $\mathbf{v}=\mathbf{t}/|\mathbf{t}|$. In this case **s** becomes **t**, $\mathbf{s}=\mathbf{t}$. If only rows are weighted, **w** is chosen as scaled **p**, $\mathbf{w}=\mathbf{p}/|\mathbf{p}|$, and **r** reduces to **p**, $\mathbf{r}=\mathbf{p}$. If both rows and columns are weighted, there is the risk that the value of $\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{v}$ will become close to zero, or negative, which indicates a conflict between the way we want to look at the columns and at the rows. This is often reflected by that the reduced **X** increases in size although the rank is diminished by one. In this case it may be necessary to switch to either weighing only columns or only rows, depending on which is giving 'most'.

4.4. Interpretation and application of PCA

Here we shall consider an application of Principal Component Analysis (PCA). It is an important aspect of PCA that it can be considered as weighing columns, weighing rows or as weighing both rows and columns. In fact PCA can be viewed as a stepwise procedure, where at each step one of the following equivalent tasks is solved.

a) Maximize $|\mathbf{X}\mathbf{w}|^2$, subject to $|\mathbf{w}|=1$,
b) Maximize $|\mathbf{X}^{\mathrm{T}}\mathbf{v}|^2$, subject to $|\mathbf{v}|=1$,
c) Maximize $(\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{v})$, subject to $|\mathbf{w}|=1$ and $|\mathbf{v}|=1$.

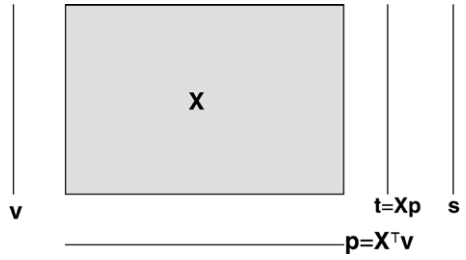

Fig. 5. Schematic illustration of vector, weighing of rows of *X*.

One can use this property of PCA to be inspired to formulate different criteria that better fits to the data that are at present. This property also motivates to view the results from different decompositions in a similar way as in the case of PCA.

The weight vector **w** is found as the eigen vector associated with the largest eigen value of $\mathbf{X}^T\mathbf{X}\mathbf{w}=\lambda\mathbf{w}$. The set of score vectors are orthogonal (because of a)), and so is the set of loading vectors (because of b)). This property characterizes PCA apart from rotation of score and loading vectors.

4.4.1. Batch process data

These are process data. In case of process data we often find reduced rank. Here there are 25 *x*-variables. Thus $\mathbf{X}^T\mathbf{X}$ is a 25 × 25 matrix. The data are here auto-scaled. In case of PCA we find a weight vector **w** such that the score vector **t** has maximal size, max |**t**| for |**w**| = 1. Therefore it is natural to plot the size of $|\mathbf{t}|=\sqrt{\lambda}$ against the dimension. This is done in Fig. 7. It shows that the size has become zero at dimension 12 or so. The next question is: What dimension should be used? In analogy with regression analysis it is natural to consider the expression.

$$\begin{aligned} f(A) &= |\mathbf{X}-\mathbf{X}_A|^2\left(1+|\mathbf{X}_A^+|^2\right) \\ &= |\mathbf{X}-\mathbf{X}_A|^2(1+d_1+\ldots+d_A) \\ &= (\lambda_{A+1}+\ldots+\lambda_{25})(1+1/\lambda_1+\ldots+1/\lambda_A). \end{aligned} \quad (14)$$

This expression can be viewed as the variance of a prediction of a new sample. Here it is the ability of the regression matrix $\mathbf{B}_A$ to regenerate the sample.

In Fig. 8 the values of *f*(*A*) are shown. The way we look at the figure is that we look for a minimum value or when *f*(*A*) reaches zero. The figure suggests that dimension of 8 should be used. The *x*-axis only goes to 11, because the score vectors are almost zero for dimension 12 and later.

The score vector **t** is to 'describe' **X** as well as possible. Following the H-principle we should look at the size of $|\mathbf{t}^T\mathbf{X}|^2$. In the case of PCA there is the same information in $|\mathbf{t}|^2=\lambda$ as in $|\mathbf{t}^T\mathbf{X}|^2=\lambda^2$. Therefore it is not shown here. The term $|\mathbf{t}^T\mathbf{X}|^2$ can be written as $|\mathbf{t}^T\mathbf{X}|^2=\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{w}$. It shows that the matrix

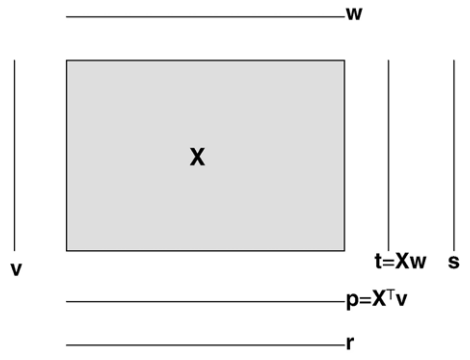

Fig. 6. Schematc illustration of vectors, weighing both columns and rows.

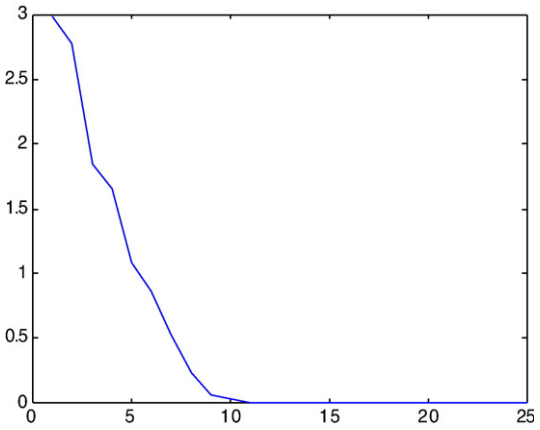

Fig. 7. Plot size of score vector, |**t**|, versus the dimension.

$\mathbf{C}=\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X}$ is important in describing the modeling steps. The diagonal elements of this matrix is

$$C_{ii} = \left(\mathbf{x}_i^T\mathbf{x}_1\right)^2+\ldots+\left(\mathbf{x}_i^T\mathbf{x}_{25}\right)^2, \quad i=1,\ldots,25.$$

It is useful to look at the values of C_{ii} before the analysis and the results after dimension 8. In Fig. 9 the values of $\sqrt{C_{ii}}$ are shown both before the analysis and after step 8. The figure shows that the values of $\sqrt{C_{ii}}$ are practically zero after step 8. Thus very little of the covariation in data is left, when step no. 8 has been completed. It can also be seen from the figure that the last 6 variables show smaller covariance with the other variables than the first 18 variables.

5. Two data blocks. Regression analysis

It is supposed here that there are given two data blocks **X**, an *N* × *K* matrix, and **Y**, an *N* × *M* matrix.

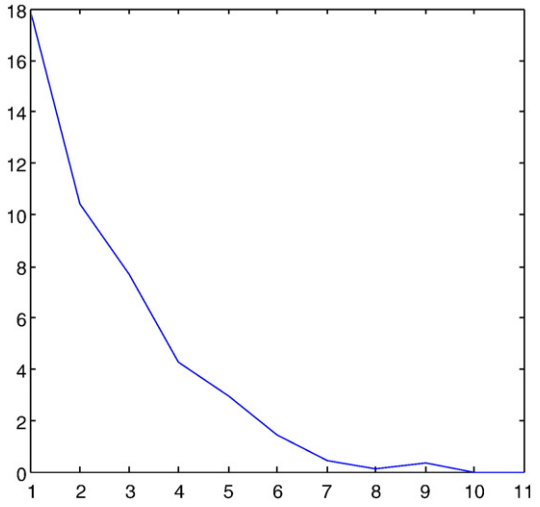

Fig. 8. Plot of *f*(*A*), Eq. (14), versus dimension.

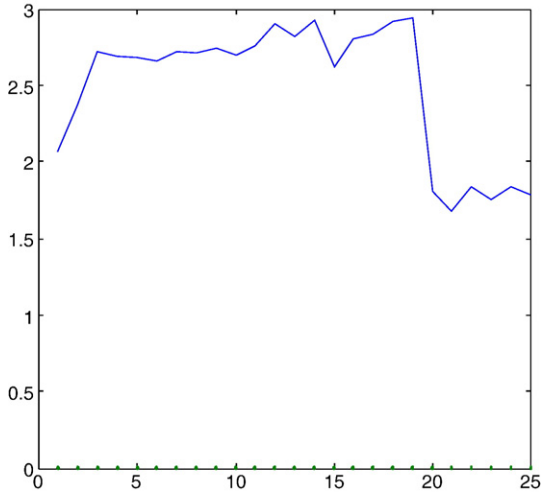

Fig. 9. Plot of $\sqrt{C_{ii}}$ versus variable number. Line: before analysis, ...: after dimension 8.

5.1. Decomposition of X

When the modeling task is carried out in steps, we are looking for a weight vector **w** such that the resulting score vector, $\mathbf{t}=\mathbf{Xw}$, is good in describing the **Y**-data. There are many ways to find **w**. In fact most standard regression analysis can be formulated as special choices of **w**.

Having found the weight vector **w**, the score vector $\mathbf{t}=\mathbf{Xw}$ is computed and the matrices **X** and **Y** are adjusted for what has been found,

$$\mathbf{X}\leftarrow\mathbf{X}-d\,\mathbf{t}\,\mathbf{p}^{\mathrm{T}},\ \text{with}\ \mathbf{p}=\mathbf{X}^{\mathrm{T}}\mathbf{t}\ \text{and}\ d=1/\left(\mathbf{t}^{\mathrm{T}}\mathbf{t}\right).$$

$$\mathbf{Y}\leftarrow\mathbf{Y}-d\,\mathbf{t}\,\mathbf{q}^{\mathrm{T}},\ \text{with}\ \mathbf{q}=\mathbf{Y}^{\mathrm{T}}\mathbf{t}.$$

This procedure is then repeated for the reduced matrices. The steps are continued as long as the predictions derived from the model are improved. Like in Eq. (11), **X** is projected onto **t**, and the result is subtracted from **X**. Therefore, the score vectors will be mutually orthogonal independently of the weight vectors **w**'s. Note that in a linear regression case, $\mathbf{X}\Rightarrow\mathbf{Y}$, both **X** and **Y** are adjusted by the score vectors that have been found.

The results of the regression analysis are the following decompositions of **X** and **Y**,

$$\mathbf{X}=d_1\mathbf{t}_1\mathbf{p}_1^{\mathrm{T}}+\ldots+d_A\mathbf{t}_A\mathbf{p}_A^{\mathrm{T}}+\mathbf{X}_0=\mathbf{T}_A\mathbf{D}_A\mathbf{P}_A^{\mathrm{T}}+\mathbf{X}_0 \tag{15}$$

$$\mathbf{Y}=d_1\mathbf{t}_1\mathbf{q}_1^{\mathrm{T}}+\ldots+d_A\mathbf{t}_A\mathbf{q}_A^{\mathrm{T}}+\mathbf{Y}_0=\mathbf{T}_A\mathbf{D}_A\mathbf{Q}_A^{\mathrm{T}}+\mathbf{Y}_0 \tag{16}$$

where $\mathbf{Q}_A=(\mathbf{q}_1,\ldots,\mathbf{q}_A)$ and $\mathbf{q}_1=\mathbf{Y}^{\mathrm{T}}\mathbf{t}_1,\ldots,\mathbf{q}_A=\mathbf{Y}^{\mathrm{T}}\mathbf{t}_A$. $\mathbf{X}_0$ is the part of **X** that is not used, and $\mathbf{Y}_0$ is the unexplained part of **Y**. The regression coefficients are given by

$$\mathbf{B}_A=d_1\mathbf{r}_1\mathbf{q}_1^{\mathrm{T}}+\ldots+d_A\mathbf{r}_A\mathbf{q}_A^{\mathrm{T}}=\mathbf{R}_A\mathbf{D}_A\mathbf{Q}_A^{\mathrm{T}}, \tag{17}$$

with $\mathbf{R}_{\mathrm{A}}$ generated from Eqs. (5) to (6). It gives, with $\mathbf{Y}_A=\mathbf{XB}_A$ the estimated response values,

$$\mathbf{Y}=\mathbf{XB}_A+\mathbf{Y}_0=\mathbf{Y}_A+\mathbf{Y}_0.$$

5.2. Optimal choice of weight vector w

It is an important issue how the weight vectors should be chosen. The primary interest in industry is the variance or uncertainty of predictions. If $\mathbf{x}_0$ is a new sample, the corresponding response value, $\mathbf{y}_0$, is in linear regression computed as $\mathbf{y}_0=\mathbf{b}^{\mathrm{T}}\mathbf{x}_0$. In order to simplify the notation assume that there is only one response variables, and the data can be described by a normal distribution, $\mathbf{y}\sim N(\mathbf{X\beta},\sigma^2)$. Then the variance of y_0 is given by

$$\begin{aligned}\mathrm{Var}(y_0)&\cong s^2\left(1+\mathbf{x}_0^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}_0\right)\\&=\left[\mathbf{y}^{\mathrm{T}}\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right)\mathbf{y}\right]\left(1+\mathbf{x}_0^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}_0\right)/(N-K).\end{aligned}$$

Here the full model is assumed. We would like this variance to be as small as possible. When the model is expanded and new score vector is selected, one can show that

a) the residual variation, $\mathbf{y}^{\mathrm{T}}(\mathbf{I}-\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{y}$, always decreases
b) the model variation, $\mathbf{x}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_0$, always increases.

The H-principle looks closer at these terms and suggests optimizing a balance between the two terms. The conclusion is that the size of $\mathbf{Y}^{\mathrm{T}}\mathbf{t}$ should be maximized, or

$$\text{maximize}\ |\mathbf{q}|^2=|\mathbf{Y}^{\mathrm{T}}\mathbf{Xw}|^2,\ \text{for}\ \mathbf{w}\ \text{subject to}\ |\mathbf{w}|=1. \tag{18}$$

The solution to this task is to choose **w** as the eigen vector associated with the largest eigen value of

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{Xw}=\lambda\mathbf{w}. \tag{19}$$

This choice of **w** gives PLS regression. The importance of the H-principle is due to that, assuming normally distributed data, the residual variation, $\mathbf{Y}^{\mathrm{T}}(\mathbf{I}-\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{Y}$, and the precision, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, are stochastically independent. Therefore both must be modeled in order to secure small variance of predictions. This theory assumes that it is appropriate to use all of **X** in the modeling task. Usually it is necessary to find the part of **X** that should be used. But this is not considered closer here, because this is an extensive topic.

Like in the case of PCA there are some equivalent forms of maximization that can be used. An important one is to find a weight vector **w** for **X** and a weight vector **q** for **Y** such that the resulting score vectors have maximal covariance,

$$\begin{aligned}&\text{maximize}\left(\mathbf{t}^{\mathrm{T}}\mathbf{u}\right),\ \text{for}\ \mathbf{t}=\mathbf{Xw}\ \text{and}\ \mathbf{u}\\&=\mathbf{Yq},\ \text{for}\ \mathbf{w}\ \text{and}\ \mathbf{q}\ \text{subject to}\ |\mathbf{w}|=1\ \text{and}\ |\mathbf{q}|=1.\end{aligned} \tag{20}$$

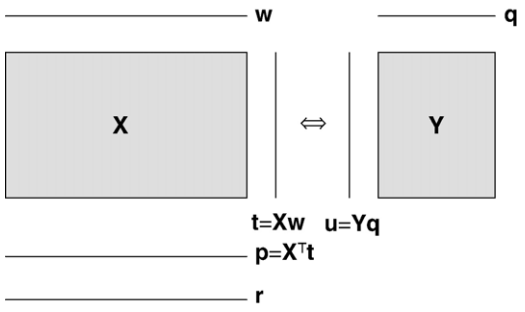

Fig. 10. Vectors at each step in a regression analysis.

This interpretation is important. It indicates theoretical closeness with Canonical Correlation Analysis. Also, the results can be analyzed graphically by studying the scatter plot of the **u**-vector against the **t**-vector. This plot should show a scatter that approximately follows a straight line. If this is not the case, it is time to stop modeling data.

Fig. 10 gives a schematic illustration of the vectors that are computed in the regression analysis. Note, that **u** is only used for graphical analysis and statistical testing for the significance of the step.

5.3. Graphical analysis

The maximal covariance, $\max(\mathbf{t}^T\mathbf{u})$, obtainable at each step is the largest eigen value λ in Eq. (19). It can be written as $\lambda=\mathbf{t}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t}=|\mathbf{Y}^T\mathbf{t}|^2$. It is useful to report this value at each iteration in order to see how much it decreases. This is shown in Fig. 11 for the process data using all 25 variables. Data have been auto-scaled before analysis. It shows that λ decreases rapidly. From the figure we would consider the dimension to be around 8. In Fig. 12 are shown the values of $f(A)$ in Eq. (14) for these data.

Fig. 12 suggests that the dimension should be 7 or 8. The analysis of the dimension can be supplemented in different ways, but that is not considered here.

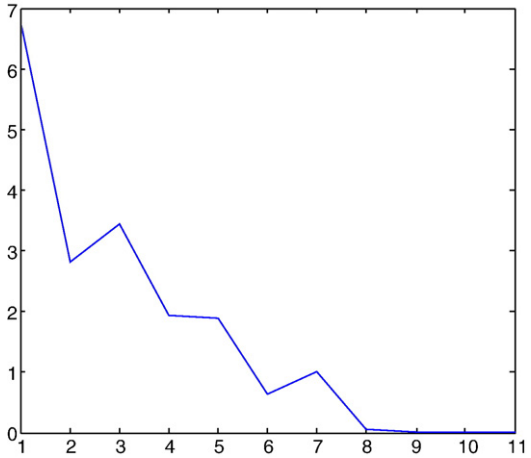

Fig. 11. Plot of λ, Eq. (18), versus dimension.

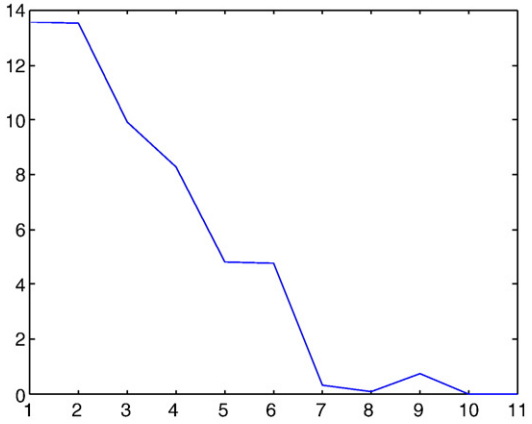

Fig. 12. Plot of f(A), (14), versus dimension.

It is useful to look at the diagonal elements of $\mathbf{C}=\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$, the matrix in Eq. (19),

$$\mathbf{C}_{kk} = \left(\mathbf{x}_k^T\mathbf{y}_1\right)^2+\ldots+\left(\mathbf{x}_k^T\mathbf{y}_M\right)^2 = \left(\mathbf{x}_k^T\mathbf{y}\right)^2, \quad k=1,\ldots,K, \quad (M=1).$$

A plot of the $\sqrt{C_{kk}}$ values for the first 8 dimensions should be looked at. There are two things that one is looking for. The first is the figure, where all values of $\sqrt{C_{kk}}$ seem to be zeros. If they seem to be almost zeros on a graph, it indicates that the modeling should stop at this dimension. The second thing one is looking for is if there are many variables that do not show covariance with the y-variables. If this is the case it is important to remove them from the analysis. This is an important topic because successful modeling is based on that the x-variables contribute to the modeling task. But this is not considered closer here.

There are two types of graphs that one should always look at. The first type is the plot of observed y-values against the computed y-values. There should also be a plot where the y-values are computed by cross-validation. The other types of plots are plots of the u-vector versus the corresponding t-vector. This should be done for each step or dimension. These plots are also important in the path modeling task. We shall look at them as illustrations of the results from path modeling.

6. Three data blocks

Consider now the task of modeling three data blocks. The situation is illustrated in Fig. 13. The variables in **X** and **Z** may be process data and those in **Y** are quality data. The primary

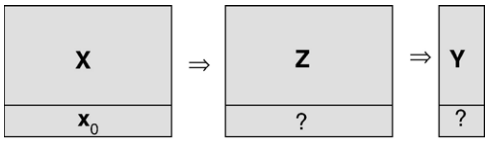

Fig. 13. Modeling three data blocks.

objective may be to provide with predictions for new values of *Y*-variables. But part of the process variables is available. This is indicated in Fig. 13. The values associated with **X** may be available, and it may be needed to predict both the *Z*-values and the final quality, the *Y*-values. What are needed are 'good' *Z*-values, $\mathbf{z}_0$, such that $\mathbf{z}_0$ give good predictions for the *Y*-values, $\mathbf{y}_0$. Thus, we are not looking for good predictions of $\mathbf{z}_0$ per se, but we want to estimate the values of $\mathbf{z}_0$ such that they are good values for entering predictions of $\mathbf{y}_0$. This procedure is sometimes called 'path modeling'.

6.1. Criteria for modeling three data blocks as a path

It shall now be considered, how appropriate score and loading vectors can be defined. It will also be shown how the criteria (18) extend naturally to this setup.

Consider Fig. 14 closer. The task is to find a weight vector **w** such that the derived score vector, $\mathbf{t}=\mathbf{X}\mathbf{w}$ is good. The score vector **t** generates a *Z*-loading vector **p**, $\mathbf{p}=\mathbf{Z}^{\mathrm{T}}\mathbf{t}$, which is used to produce a score vector $\mathbf{s}=\mathbf{Z}\mathbf{p}$ that is expected to be good in predicting samples of **Y**. A score vector **s** is good if the resulting loading vector $\mathbf{q}=\mathbf{Y}^{\mathrm{T}}\mathbf{s}$ is large, cf Eq. (18).

Thus, the optimization task is

$$\text{maximize } |\mathbf{q}|^2, \text{ for } \mathbf{w} \text{ subject to } |\mathbf{w}| = 1.$$

If the expression for **q** is expanded, the result is $\mathbf{q}=\mathbf{Y}^{\mathrm{T}}\mathbf{s}=\mathbf{Y}^{\mathrm{T}}\mathbf{Z}\mathbf{p}=\mathbf{Y}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{t}=\mathbf{Y}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{X}\mathbf{w}$. The task is therefore,

$$\text{maximize } |\mathbf{Y}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{X}\mathbf{w}|^2, \text{ for } \mathbf{w} \text{ subject to } |\mathbf{w}| = 1.$$

Using Lagrange multiplier technique it can be shown that **w** should be chosen as the eigen vector associated with the largest eigen value of

$$\mathbf{X}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\,\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \lambda\mathbf{w}. \tag{21}$$

When **w** has been found, **t** is computed as $\mathbf{t}=\mathbf{X}\mathbf{w}$, **p** as $\mathbf{p}=\mathbf{Z}^{\mathrm{T}}\mathbf{t}$, **s** as $\mathbf{s}=\mathbf{Z}\mathbf{p}$, **q** as $\mathbf{q}=\mathbf{Y}^{\mathrm{T}}\mathbf{s}$ and **u** as $\mathbf{u}=\mathbf{Y}\mathbf{q}$. The success of this estimation is studied by looking at the scatter-plot of **s** against **u**.

It should be emphasized that the task here is the modeling between **Z** and **Y** with weights generated from **X**. This type of modeling may be unsuccessful either because **Z** cannot model **Y** or because **X** cannot adequately describe **Z**. In both cases it might be that **X** is good in describing **Y**.

The procedure above shows how **w** is found, which generates the further score and loading vectors, **t**, **p**, **s**, **q** and **u**. The important question is: How should the next weight vector **w** be

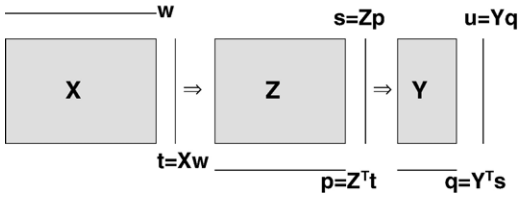

Fig. 14. Schematic illustration of score and loading vectors in a path.

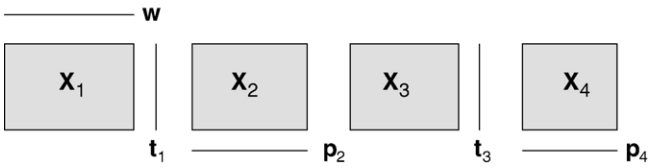

Fig. 15. Schematic illustration of matrices and vectors.

found and how should the score and loading vectors be determined? In many industrial and business applications this is not carried out. The path containing latent variables is often only represented by one set of latent variables, one latent variable for each data block. This is not satisfactory, because more latent variables are usually needed in order to obtain satisfactory modeling results.

In linear regression, $\mathbf{X} \Rightarrow \mathbf{Y}$, both **X** and **Y** are adjusted by the score vectors derived from **X**. When there is a given path, $\mathbf{X} \Rightarrow \mathbf{Z} \Rightarrow \mathbf{Y}$, **X** and **Z** are adjusted by score vectors derived from **X**. The pair of matrices, **Z** and **Y**, needs to be adjusted similarly. This is explained in the following.

The weight vector **w** is found by maximizing the size of $\mathbf{Y}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{X}\mathbf{w}$. If the matrix product is split into parts, we get

$$\mathbf{t} = \mathbf{X}\mathbf{w} \rightarrow \mathbf{Z},\ \mathbf{p} = \mathbf{Z}^{\mathrm{T}}\mathbf{t} \rightarrow \mathbf{Z}^{\mathrm{T}},\ \mathbf{s} = \mathbf{Z}\mathbf{p} \rightarrow \mathbf{Y}.$$

Here the arrow → marks that the vector is to describe the matrix, e.g., **t** is to describe **Z**. From an algorithmic point of view it is natural to work with four matrices,

$$\mathbf{X}_1 = \mathbf{X}, \mathbf{X}_2 = \mathbf{Z}, \mathbf{X}_3 = \mathbf{Z} \text{ and } \mathbf{X}_4 = \mathbf{Y}$$

The reason is that in a regression model, $\mathbf{X} \Rightarrow \mathbf{Y}$, both **X** and **Y** are adjusted by the *X*-score vectors. In the path case there are two sets of regressions, $\mathbf{X} \Rightarrow \mathbf{Z}$ and $\mathbf{Z} \Rightarrow \mathbf{Y}$. **X** and **Z** are adjusted by the *X*-score vectors, and **Z** and **Y** are adjusted by the derived vectors. The notation is changed slightly to conform to the four matrices,

$$\mathbf{t}_1 = \mathbf{X}_1\mathbf{w},\ \mathbf{p}_2 = \mathbf{X}_2^{\mathrm{T}}\mathbf{t}_1,\ \mathbf{t}_3 = \mathbf{X}_3\mathbf{p}_2, \text{ and } \mathbf{p}_4 = \mathbf{X}_4^{\mathrm{T}}\mathbf{t}_3.$$

These vectors are schematically illustrated in Fig. 15. Note that the vector **u** (= $\mathbf{X}_4\mathbf{p}_4$), like in the case of linear regression, is only used for studying the results of the modeling task. The figure emphasizes that the vectors describe the succeeding matrices,

$$\mathbf{t}_1 \rightarrow \mathbf{X}_2,\ \mathbf{p}_2 \rightarrow \mathbf{X}_3^{\mathrm{T}},\ \mathbf{t}_3 \rightarrow \mathbf{X}_4.$$

The adjustment now follows the rule of linear regression,

$$\begin{aligned}
&\mathbf{X}_1 \leftarrow \mathbf{X}_1 - d_1\mathbf{t}_1\mathbf{p}_1^{\mathrm{T}}, \text{with } \mathbf{p}_1 = \mathbf{X}_1^{\mathrm{T}}\mathbf{t}_1,\ d_1 = 1/\left(\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1\right)\\
&\mathbf{X}_2 \leftarrow \mathbf{X}_2 - d_1\mathbf{t}_1\mathbf{p}_2^{\mathrm{T}}, \text{with } \mathbf{p}_2 = \mathbf{X}_2^{\mathrm{T}}\mathbf{t}_1,\ d_1 = 1/\left(\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1\right)\\
&\mathbf{X}_3 \leftarrow \mathbf{X}_3 - d_2\mathbf{t}_3\mathbf{p}_2^{\mathrm{T}}, \text{with } \mathbf{t}_3 = \mathbf{X}_3\mathbf{p}_2,\ d_2 = 1/\left(\mathbf{p}_2^{\mathrm{T}}\mathbf{p}_2\right)\\
&\mathbf{X}_4 \leftarrow \mathbf{X}_4 - d_3\mathbf{t}_3\mathbf{p}_4^{\mathrm{T}}, \text{with } \mathbf{p}_4 = \mathbf{X}_4^{\mathrm{T}}\mathbf{t}_3,\ d_3 = 1/\left(\mathbf{t}_3^{\mathrm{T}}\mathbf{t}_3\right).
\end{aligned}$$

Note that $\mathbf{t}_1$-vectors are orthogonal, but this does not hold for the other set of vectors. Thus, $\mathbf{X}_1$ is reduced by rank 1, while the

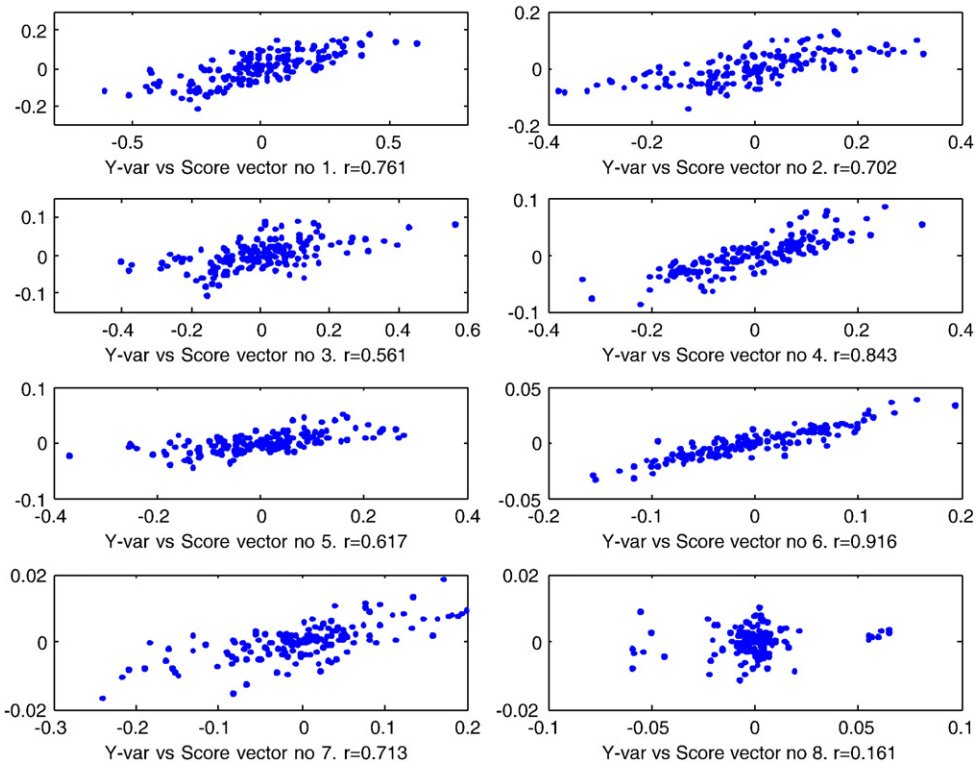

Fig. 16. Pair-wise plots of the first 8 *X*-score vectors, $\mathbf{t}_A$, and *Y*-score vectors, $\mathbf{u}_A$. The first pair is the upper most to the left, the next is the upper most to the right and so on.

others are not. Using the reduced matrices a new weight vector is determined from Eq. (21), which is now

$$\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2\mathbf{X}_3^{\mathrm{T}}\,\mathbf{X}_4\mathbf{X}_4^{\mathrm{T}}\mathbf{X}_3\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\mathbf{w} = \lambda\mathbf{w}.$$

When $\mathbf{w}$ has been found, the score vectors $\mathbf{t}_1$, $\mathbf{t}_3$, … and the loading vectors $\mathbf{p}_2$, $\mathbf{p}_4$, … are computed and the $\mathbf{X}_i$-matrices adjusted. By appropriate identification of score and loading vectors, we can write the part of the data block $\mathbf{X}_i$ that is used, here $\mathbf{X}_{i,A}$, as

$$\mathbf{X}_{i,A} = d_{i,1}\,\mathbf{t}_{i,1}\mathbf{p}_{i,1}^{\mathrm{T}} + \ldots + d_{i,A}\,\mathbf{t}_{i,A}\mathbf{p}_{i,A}^{\mathrm{T}} = \mathbf{T}_{i,A}\mathbf{D}_{i,A}\mathbf{P}_{i,A}^{\mathrm{T}}.$$

In order to compute the regression coefficients between the data blocks, we need the **R**-matrices. They are computed using the Eq. (5) or (11). $\mathbf{R}_i$ is computed by Eq. (11) for i=1,3, … and by Eq. (5) for i=2,4, …. They are computed such that they satisfy

$$\begin{aligned}\mathbf{P}_i &= \mathbf{X}_{i-1}^{\mathrm{T}}\mathbf{R}_i,\ i = 1,3,\ldots(\mathbf{X}_0 = \mathbf{X}_1)\\ \mathbf{T}_i &= \mathbf{X}_{i-1}\,\mathbf{R}_i,\ i = 2,4,\ldots\end{aligned}.$$

In fact only Eq. (5) is used, because Eq. (11) is similar with exchange of loading and score vectors. The matrices $\mathbf{T}_i$, $\mathbf{P}_i$, $\mathbf{D}_i$ and $\mathbf{R}_i$ each contains A vectors. Consider now the regression coefficients between the data blocks. They are computed as shown in Eq. (10) or (17), as

$$\begin{aligned}\mathbf{B}_i &= \mathbf{T}_i\mathbf{D}_i\mathbf{R}_i^{\mathrm{T}},\quad i = 1,3,\ldots\\ \mathbf{B}_i &= \mathbf{R}_i\mathbf{D}_i\mathbf{P}_i^{\mathrm{T}},\quad i = 2,4,\ldots\end{aligned}.$$

For the regression between the data blocks we can write

$$\mathbf{X}_{i,A} = \mathbf{B}_i\mathbf{X}_{i-1},\quad i = 1,3,\ldots(\mathbf{X}_0 = \mathbf{X}_1) \tag{22}$$

$$\mathbf{X}_{i,A} = \mathbf{X}_{i-1}\mathbf{B}_i,\quad i = 2,4,\ldots. \tag{23}$$

In terms of the original matrices (23) is

$$\mathbf{Z}_A = \mathbf{X}\mathbf{B}_2,\ \mathbf{Y}_A = \mathbf{Z}\mathbf{B}_4. \tag{24}$$

Consider now how these regressions are used for predictions. Here $\mathbf{Z}_A$ is the estimate of $\mathbf{Z}$ based on A dimensions and similarly $\mathbf{Y}_A$ is an estimate of $\mathbf{Y}$. In terms of the samples Eq. (24) can be written as

$$\left(\mathbf{z}_A^i\right)^{\mathrm{T}} = \left(\mathbf{x}^i\right)^{\mathrm{T}}\mathbf{B}_2,\ \text{and}\ \left(\mathbf{y}_A^i\right)^{\mathrm{T}} = \left(\mathbf{z}^i\right)^{\mathrm{T}}\mathbf{B}_4,\ i = 1,2,\ldots,N.$$

Here $\mathbf{x}^i$ is the ith row of $\mathbf{X}$, but treated as a column vector. Similarly, $\mathbf{z}_A^i$ and $\mathbf{y}_A^i$ are the ith rows of $\mathbf{Z}_A$ and $\mathbf{Y}_A$, resp. When a

Table 1
Cumulative percentage of **X**, Cum(**X**), and cumulative percentage of **Y**, Cum(**Y**), that is selected at each dimension

Dimension	1	2	3	4	5	6	7	8	9	10	11
Cum(**X**)	31.15	46.91	72.78	81.76	92.78	95.695	99.757	99.979	99.992	100.000	100.000
Cum(**Y**)	57.86	78.61	85.33	95.75	97.37	99.574	99.790	99.796	99.797	99.798	99.800

new sample is available for **X**, $\mathbf{x}_0$, the estimated **Z**-sample is computed as $\mathbf{z}_0^* = \mathbf{B}_2^T \mathbf{x}_0$. The prediction value for the **Y**-sample, $\mathbf{y}_0^*$, is computed as $\mathbf{y}_0^* = \mathbf{B}_4^T \mathbf{z}_0^*$, where $\mathbf{z}_0^*$ is the estimated **Z**-sample.

The regression Eq. (22) is used to specify how much of a data block is being used in the modeling task, and how the variables contribute to the modeling task. In terms of the original matrices we have

$$\mathbf{X}_A = \mathbf{B}_1 \mathbf{X},\ \mathbf{Z}_A = \mathbf{B}_3 \mathbf{Z}.$$

In terms of the variables we get

$$\mathbf{x}_{i,A} = \mathbf{B}_1 \mathbf{x}_i,\ i = 1,\ldots,\mathbf{K},\ \text{and}\ \mathbf{z}_{j,A} = \mathbf{B}_3 \mathbf{z}_j,\ j = 1,\ldots,J.$$

These equations can be used to relate the data for the original variable, $\mathbf{x}_i$, with the values actually used in the computations, $\mathbf{x}_{i,A}$. Similarly for the z-variables.

7. Case study

The theory shall now be applied to the Batch process data of Section 3. We start with an overall analysis of the data. Then the results after stage 4 are analyzed closer. The results from the path modeling procedure are then studied closer.

7.1. An overall analysis

In Figs. 11 and 12 the plots of λ_A and $f(A)$ were shown. Fig. 12 suggested that dimension 7 should be used. In order to illustrate this closer it is useful to look at the plots of the score vectors, $\mathbf{t}_A$, versus the Y-score vectors, $\mathbf{u}_A$. When there is only one response variable, like here, the u-vectors are the reduced y-vector. The equation is

$$\mathbf{u}_1 = \mathbf{y}_1,\quad \mathbf{u}_A = \mathbf{u}_{A-1} - (\mathbf{y}^T\mathbf{t}_A)/(\mathbf{t}_A^T\mathbf{t}_A)\mathbf{t}_A,\ A = 2,3,\ldots.$$

A plot showing a pair-wise scatter plot of $\mathbf{u}_A$ versus $\mathbf{t}_A$ for $A = 1,\ldots,8$ is shown in Fig. 16.

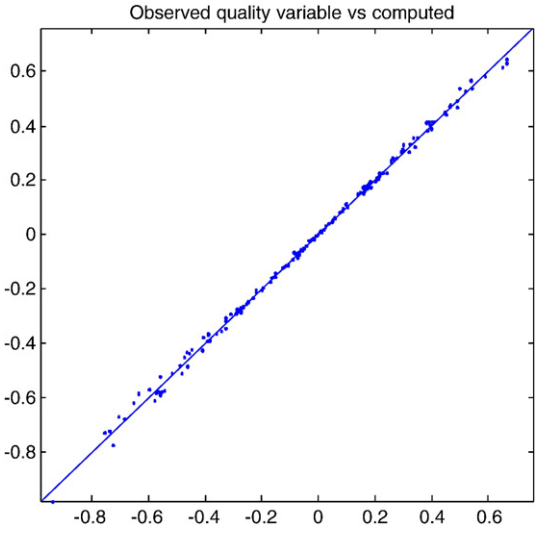

Fig. 17. Plot of observed versus computed response values after step 7.

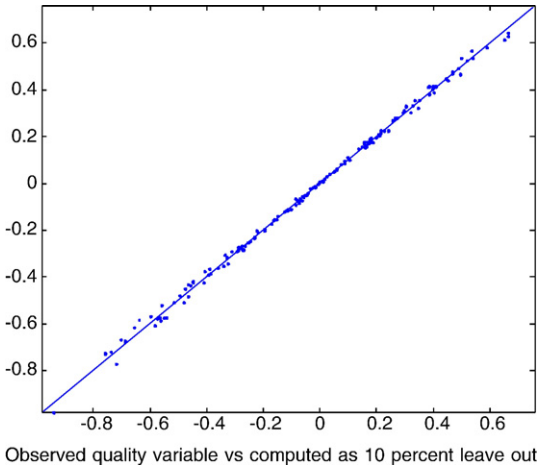

Fig. 18. Plot of observed versus cross-validated response values after step 7.

The figure confirms that the dimension should be 7. The correlation coefficient between $\mathbf{t}_7$ and $\mathbf{u}_7$ is 0.713, which is very significant. But it should be observed that the response variable varies between − 0.02 and 0.02. If these values are below the precision of the y-values, the dimension 6 should be chosen. The following table shows how much variation is selected at each step or dimension.

From the Table 1 it can be seen that at dimension 7 there has been selected 99.757% of **X** and 99.790% of **Y**. The 7th component is selecting 4.1% of **X** and 0.2% of **Y**. This is not very much, but it is from a numerical point of view a significant improvement compared to selecting only 6 dimensions. At dimension 6 there is 2.9% of **X** and 2.2% of **Y** selected. This is not very much. When we look at Fig. 16 it can be seen that the simple correlation coefficient between $\mathbf{t}_6$ and $\mathbf{u}_6$ is 0.916. It is a disadvantage with PLS regression that a very high correlation can be found at a step, where relatively little is being selected. It indicates that an improvement can be made by identifying the part of **X** that should be used in the modeling task. But it is not the aim of this paper to identify optimal parts of **X** that should be used.

It is useful to look closer at the observed and computed response values, when the dimension is 7. This is shown in Figs. 17 and 18.

Fig. 17 shows that the points are located close to a line with a slope of 45°. The explained variation is $R^2 = 99.790\%$ and the residual standard deviation is $s = 0.0166$. The cross-validated response values are computed as follows. 10% of the samples, here 15, are excluded from the analysis. The parameters are estimated using 90%, here 139 samples. These parameters are then used to estimate the 15 y-values that were excluded. This is repeated 10 times, each time a different set of samples is excluded. At the end of the analysis each cross-validated y-value has been computed by using 90% of the samples. Thus the cross-validated y-values represent the typical results, when the model is used to estimate the response value of a new sample. The simple correlation coefficient, r, between the observed and the cross-validated y-value is $r = 0.9989$, and the $Q^2 = (r)^2 = 0.9977$. Further, $s_c = [\Sigma(y_i - y_{ic})^2 / N]^{1/2} = 0.017$. Here (y_i) is defined as the original response values and (y_{ic}) the estimated cross-validated response values. It shows that after the 7th stage the

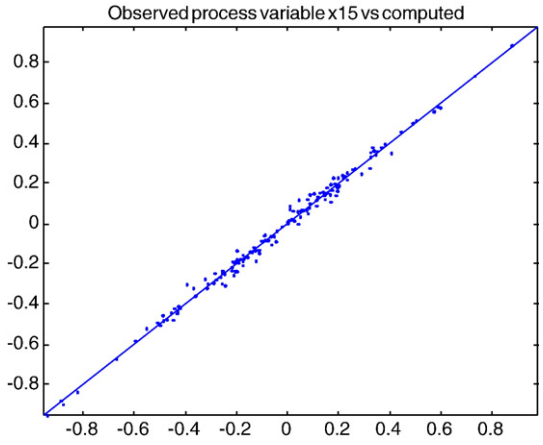

Fig. 19. Observed values of the process variable $\mathbf{x}_{15}$ versus the computed value.

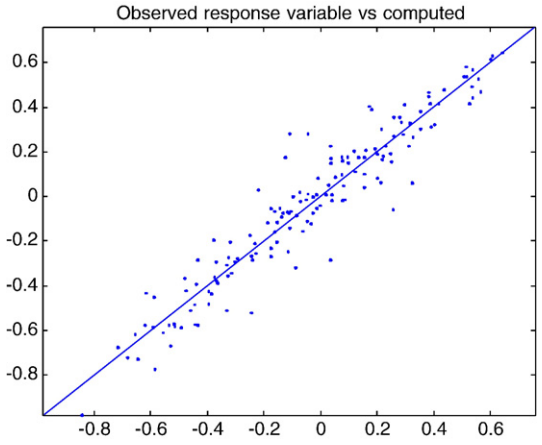

Fig. 20. Plot of observed versus computed response values after stage 4, based on model b).

response values can be estimated with an uncertainty that in 95% of the cases is smaller than around $\pm 2 \times 0.017 = \pm 0.034$.

7.2. Actual models at stage 4

Here we shall consider the situation after stage 4 of the batch process has been completed. The other stages are treated in the same way. Note that there is only one response variable, **y**. If the process has reached stage 4, there are several models of interest:

a) $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4) = (\mathbf{x}_1,\ldots,\mathbf{x}_{14}) \Rightarrow \mathbf{X}_5 = (\mathbf{x}_{15},\mathbf{x}_{16})$
How well can we predict the process variables at stage 5?

b) $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4) \Rightarrow \mathbf{y}$
How well can the response variable be predicted from stage 4?

c) $\mathbf{X}_5 \Rightarrow \mathbf{y}$
How well do process variables at stage 5, which only describes **y**?

d) $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4) \Rightarrow \mathbf{X}_5 \Rightarrow \mathbf{y}$
What would be a natural choice of process variables at stage 5 if the fit of expected **y** is as good as possible?

e) $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5) \Rightarrow \mathbf{y}$

From a) and c) 'good' guesses of process variables at stage 5 are known. How do these values fit to an overall model?

The results of these models shall be considered now.

a) $(\mathbf{x}_1,\ldots,\mathbf{x}_{14}) \Rightarrow \mathbf{X}_5 = (\mathbf{x}_{15},\mathbf{x}_{16})$

In Fig. 19 it is shown the observed values of the process variable $\mathbf{x}_{15}$ versus the computed value. The method used is PLS regression with 7 components. It shows a relatively good fit, $R^2 = (0.9965)^2 = 99.21\%$. A cross-validation that selects 90% of the samples uniformly out of the 154 ones and is repeated 10 times gives a residual standard deviation of $s_c = 0.045$.

The results for $\mathbf{x}_{16}$ are not shown here, because they are almost identical to the one for $\mathbf{x}_{15}$.

Table 2
R^2 values and cross-validated residual standard deviation s_c

Up to stage	1	2	3	4	5	6	7
R^2	81.00%	87.32%	86.88%	90.90%	98.48%	99.62%	99.80%
s_c	0.156	0.128	0.130	0.111	0.045	0.022	0.017

b) $(\mathbf{x}_1,\ldots,\mathbf{x}_{14}) \Rightarrow \mathbf{Y}$

Table 2 shows the R^2 values and the cross-validated residual standard deviation s_c for modeling the response values after each of the seven stages.

All analysis is based on PLS regression with 7 components (the first stage uses 6). The table shows that it is first at stage 6 that a satisfactory prediction of the response variables is obtained. Looking at s_c the results of stage 3 do not improve the results compared with stage 2. Fig. 20 shows the observed values of the response variable versus the computed values at stage 4. It shows a fairly good linear relationship, but a rather large variation around the line of 45°. The variation seems to be relatively homogenous around the line. Thus the model can be used to estimate the response values, but there will be relatively large uncertainty in the prediction, $\pm 2 \times 0.111 = \pm 0.222$.

c) $\mathbf{X}_5 = (\mathbf{x}_{15},\mathbf{x}_{16}) \Rightarrow \mathbf{Y}$

In Fig. 21 it is shown how well the two variables at stage 5 describe **Y**. The results are not very good, $R^2 = 59.67\%$ and $s_c = 0.228$.

d) $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4) \Rightarrow \mathbf{X}_5 \Rightarrow \mathbf{y}$

Fig. 22 shows the plot of the values of the response variable **Y** against the first two score vectors from $\mathbf{X}_5$. The both show a rather week relationship. In Fig. 23 is shown the plot of observed versus computed response values from the path model. The results are very close to the ones found in the previous model, see Fig. 21. Here $R^2 = 57.41\%$ and $s_c = 0.232$. A further perspective of this model is discussed in the summary below.

e) $(\mathbf{x}_1,\ldots,\mathbf{x}_{16}) \Rightarrow \mathbf{y}$

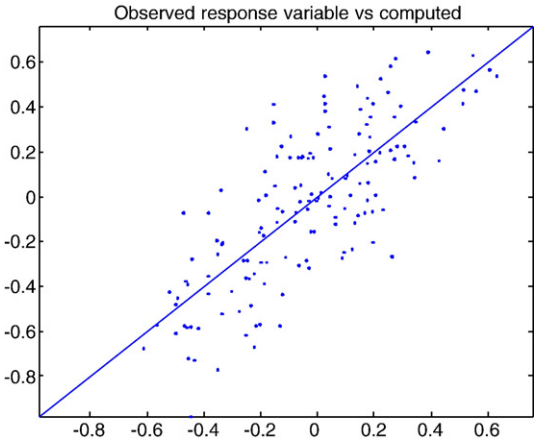

Fig. 21. Plot of observed response values versus computed, using variables $\mathbf{x}_{15}$, and $\mathbf{x}_{16}$, based on model c).

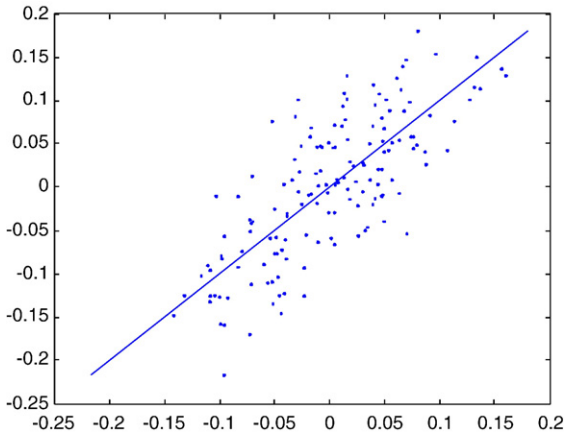

Fig. 23. Plot of observed versus computed response values resulted from model d).

In Fig. 24 it is shown the results of modeling after stage 5, where the values of the first 16 variables are available. Here $R^2=98.48\%$ and $s_c=0.045$. Here the measured values of ($\mathbf{x}_{15}$, $\mathbf{x}_{16}$) are used. But other values for ($\mathbf{x}_{15}$, $\mathbf{x}_{16}$) can be used, as discussed below.

7.2.1. Summary

At the end of stage 4 the operator has from the operating procedure certain values for the variables at stage 5, $\mathbf{x}_{15}$ and $\mathbf{x}_{16}$. The values of $\mathbf{x}_{15}$ and $\mathbf{x}_{16}$ can be fairly well predicted from the values known for $\mathbf{x}_1,\ldots,\mathbf{x}_{14}$ as shown in Fig. 19. The predicted values and the operating procedure values can be compared. The values of $\mathbf{x}_{15}$ and $\mathbf{x}_{16}$ of stage 5 have a rather weak relationship to the response variable. For a given value of $\mathbf{x}_1,\ldots,\mathbf{x}_{14}$ the operator can look at Fig. 20 to identify where the quality variable is now, and Fig. 24 of what may be the expected results after stage 5, where either the predicted or operating procedure values of $\mathbf{x}_{15}$ and $\mathbf{x}_{16}$ have been used. The regression coefficients from model d) can be used to compute the estimated value to adjust the values of $\mathbf{x}_{15}$ and $\mathbf{x}_{16}$ in order to get improved results for the quality variable. From a data analysis point of view it is important that the operator selects values for $\mathbf{x}_{15}$ and $\mathbf{x}_{16}$ that are consistent with model a), because these values are dependent on the previous values as shown in Fig. 19. Thus, the recommendation is to adjust the values of $\mathbf{x}_{15}$ and $\mathbf{x}_{16}$ using model d), but securing that the values are consistent with model a). Model e) will show the expected results on the quality variable. From data analysis point of view any adjustment of $\mathbf{x}_{15}$ and $\mathbf{x}_{16}$ can be made as long as the residual obtained from e) is within the limits shown in Fig. 24. Thus we can use the residual standard deviation $s_c=0.045$ to decide how much the two variables can be adjusted. The adjusted values should give in model e) residual within $2\times s_c=0.09$.

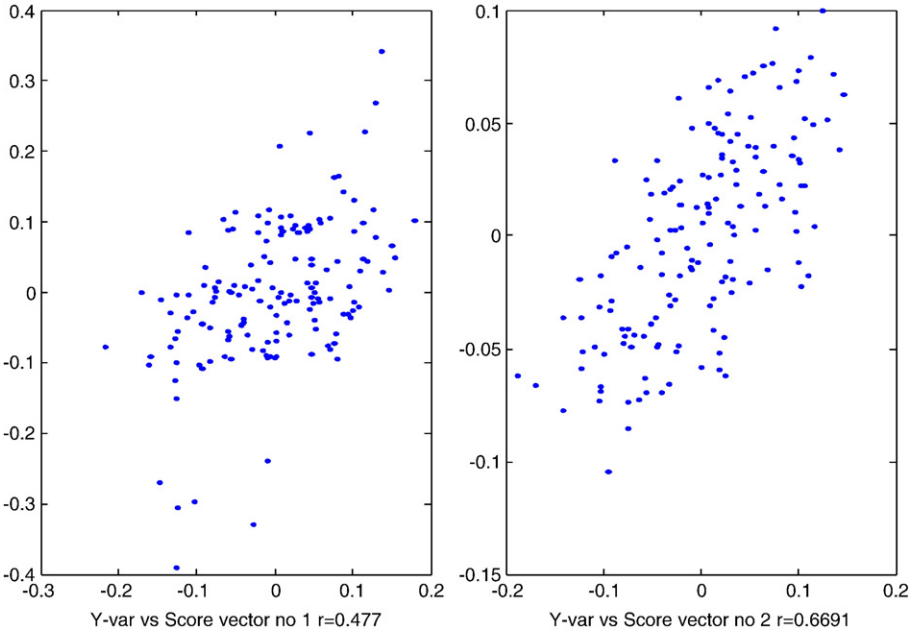

Fig. 22. Plot of observed response variable versus the first two score vectors from $\mathbf{X}_5$.

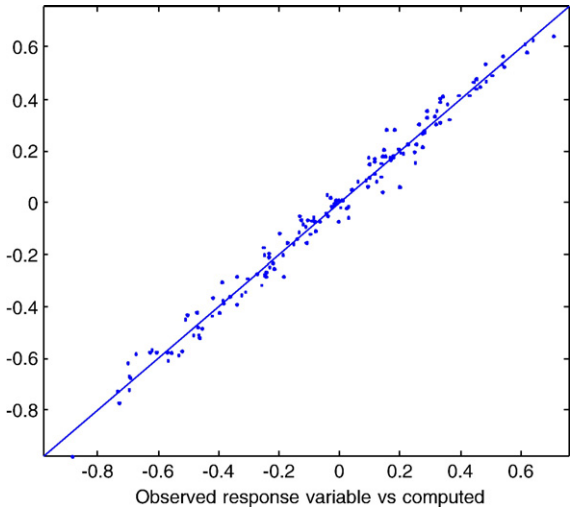

Fig. 24. Plot of observed response values versus computed from model e).

Before the start of stage 5 there is proposal, $\mathbf{x}_{\text{proposal}}=(\mathbf{x}_{15}, \mathbf{x}_{16})$, which the process engineer has from the operating manual. Using model a) the engineer can estimate the values, $\mathbf{x}_{\text{regression}}=(\mathbf{x}_{15},\mathbf{x}_{16})$, based on the previous of the process variables, $\mathbf{x}_1-\mathbf{x}_{14}$, for this batch. Using the model d) the engineer can use the regression coefficients shown in Eq. (24) to estimate a new set of values, $\mathbf{x}_{\text{path}}=(\mathbf{x}_{15},\mathbf{x}_{16})$, for the next stage. These values are the ones, which for given values of $\mathbf{x}_1-\mathbf{x}_{14}$ give the best prediction of the response variable. In model c) we can study how $(\mathbf{x}_{15}, \mathbf{x}_{16})$ predict the response variable. In d) there is placed a restriction on the model in terms of that there is given the past values of $\mathbf{x}_1-\mathbf{x}_{14}$. The different suggested values for stage 5, $\mathbf{x}_{\text{proposal}}$, $\mathbf{x}_{\text{regression}}$ and $\mathbf{x}_{\text{path}}$, can be combined with the values of $\mathbf{x}_1-\mathbf{x}_{14}$ of the present batch and the model e) can be used to estimate the response value. In this example the response variable is a quality variable, where the higher value the better product. The engineer can choose among the three sets of values for $(\mathbf{x}_{15}, \mathbf{x}_{16})$, which gives the best quality.

The path model gives a proposal for the intermediate process variables, when there already has been observed a part of the batch. The proposal is aimed at predicting the response variable as well as possible in the light of the values already obtained for the process. This can be used to improve the process already after the first stage. It is a basic problem for the process engineer to improve the process during the batch. The path models give proposals that the engineer can evaluate.

There is one practical issue, when path models are applied. The predictions derived from $X \Rightarrow Z \Rightarrow Y$ will always be worse than ones derived from $Z \Rightarrow Y$. Therefore, when path models are applied, it may be useful to select Z carefully. In the present case it was chosen as stage 5. It may be better to enlarge the intermediate period by both stage 5 and 6, because stage 5 is not good in predicting the response values. But this is not considered closer here.

The data is divided into blocks according to stages. Thus, we can look at the data as multi-block data, which are used to predict the quality variable. The first author has developed new methods for multi-block data [10], which are natural extensions of regression analysis for one X-block. These methods are good to detect the role of each data block (stage) in predicting the quality variable. The results of a multi-block regression analysis are how the individual data blocks contribute to the modeling task. This is different from path modeling, where the aim is to estimate intermediate values, such that they, together with known values, give as good predictions as possible.

The aim of the models presented here is to use the data to give good regression models, which can predict the quality values with sufficient precision. The residuals and s_c-values from the models show how good the models are. Adjustments of individual process parameters can be done within the variation of the residuals. For instance, change of variable $\mathbf{x}_{12}$ can be made such that the variation of $b_{12}\mathbf{x}_{12}$ is within the variation of the residuals. More active control of the processes usually requires the knowledge of the possible variation of each process parameters. This can be carried out. In Ref. [11] both passive and active ways of control are considered in details. The advantage of using the models of the type presented here is that the models fit well to the historical batches and they provide us with tools to improve the process. Often the chemical processes are so complicated that the company may be more interested in tools that assist the engineer in improving the process than in methods that seek to actively control the process, because this often requires detailed knowledge of the chemistry and processes.

When the operators get training in using these models there may appear new types of models that are interesting for the operators. E.g., it might be useful to look at stages 5 and 6 as one stage, or to estimate all remaining process parameters, when stage 1 has been completed.

8. Conclusion

The methods of path modeling have been presented. The importance of the methods is due to that they are natural extensions of one data block (PCA type of analysis) and two data blocks (regression type of analysis) to a sequence of data blocks. An important aspect of these methods is the graphic procedures. The same graphic procedures are used for analyzing $Z_{i-1} \Rightarrow Z_i$ in a path $\ldots \Rightarrow Z_{i-1} \Rightarrow Z_i \Rightarrow \ldots$ as is done in a linear regression analysis, $X \Rightarrow Y$.

It is suggested to use Eqs. (19) and (21) (and their extension to multiple data blocks) as a criterion for finding good weight vectors. But other criteria could be used, which better may reflect a specific purpose of analysis.

Here only three stages have been modeled. It may have some interest to model more stages. When the results from the first stage are available, a path model for all 7 stages would give regression coefficients between stages. Thus we could get estimates of the remaining variables and the response variable. If this is done at each stage, it might give the operator training in adjusting variables at the 'next' stage with the purpose of improving the result of the quality variable.

As the operator gets more experience with the models, he may ask for more extensive models and graphs. This may assist him in finding at what stages it is important to get models for the future stages.

In this paper there has been presented a unified approach to path modeling. The aim of the methods is to find best predictions along a path. This theory can be applied to any situations, where there are given a sequence of data blocks, where each data block represents the values of a given number of variables. The methods are based on well-defined optimization procedure to find the weight vectors of the initial (input) data block. Other optimization procedures can be used to find the weight vectors, but the remaining computations would be the same.

Standard procedures in linear regression like e.g., cross-validation, sensitivity analysis and others have natural extensions to this way of carrying out path analysis. For instance, in the case of cross-validation the corresponding samples are deleted from all the data blocks. Using the regression coefficients the values of the last data block are estimated. When this is repeated one can find out how well the model performs on the present data.

In industry there are often measured many more variables than are appropriate to use. In these cases it may be necessary to select the variables that should be used in each data block. This is easy to implement in the procedures presented here. But it may be time consuming on the computer if all combination of variables is investigated. It may be preferable to start with the second to the last data block and eliminate the variables that cannot describe the variables of the last data block then to eliminate the variables in the third last data that do not contribute to the modeling task that starts at the third last data block. In this way a manageable procedure can be established for data blocks that have many variables, like those that are obtained from NIR instruments.

The path of data blocks considered here consists of serially organized ones. There is one starting data block, input data block, serially connected interconnected data blocks and one ending data block, the output data block. The criterion (21) has been extended to a network of data blocks with several input data blocks and many output data blocks. In between the input data blocks and output data blocks there can be any network of data blocks. The regression Eqs. (22) and (23) have been extended to show how any data block are depending on data blocks that lead to the given data block. By using the H-principle the prediction aspect of the modeling task is optimized, when estimating parameters. The importance of these new methods is due to that they extend the standard methods of linear regression analysis to an arbitrary network of data blocks with a given set of input and output data blocks. The same algorithm handles any number of data blocks. By using the software to work with one, two or three data blocks one gets good training in working with a network of data blocks.

References

[1] A. Höskuldsson, Causal and path modelling, J. Chemom. 58 (2001) 287–311.
[2] A. Höskuldsson, Prediction Methods in Science and Technology. vol. 1. Basic Theory, Thor Publishing, Copenhagen, 1996 87–985941–0-9.
[3] K.A. Bollen, Structural Equations with Latent Variables, Wiley, New York, 1988.
[4] H. Wold, Soft modelling. The basic design and some extensions, in: Jöreskog og Wold (Ed.), System Under Indirect observations. Causality-Structure-Prediction, vol. 1, North Holland Pub. Co., Amsterdam, 1982, pp. 263–271.
[5] H. Wold, Partial Least Squares, Encyclopaedia of Statistical Sciences, vol. 6, Wiley, New York, 1985.
[6] J.-B. Lohmöller, Latent Variables Path Modelling with Partial Least Squares, Physica-Verlag, Heidelberg, 1989.
[7] S. Wold, et al., Multivariate analysis in Chemistry, in: B.R. Kowalski (Ed.), Chemometrics: Mathematics and Statistics in Chemistry, Reidel, Dordrecht, 1984, pp. 17–95.
[8] Michel Tenenhaus, Vincenzo EspositoVinzi, PLS regression, PLS path modeling and generalized Procrustean analysis: a combined approach for multiblock analysis, J. Chemom. 19 (2005) 145–153.
[9] A.K. Smilde, J.A. Westerhuis, S. de Jong, A framework for sequential multiblock component methods, J. Chemom. 17 (2003) 323–337.
[10] A. Höskuldsson, K. Svinning, in press. Modelling of Multi-Block Data, J. Chemometrics.
[11] A.L. Pomerantsev, O.Ye. Rodionova, A. Höskuldsson, Process control and optimization with simple interval calculation method, Chemometr. Intell. Lab. Syst. 18 (2) (2006) 165–179.