1 Partitioning of data

Industrial companies carry out many types of measurements in order to secure the quality of their products. Measurements may be carried out online or sampled at specific time intervals for further analysis.

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, i.e., it may provide with say, 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.

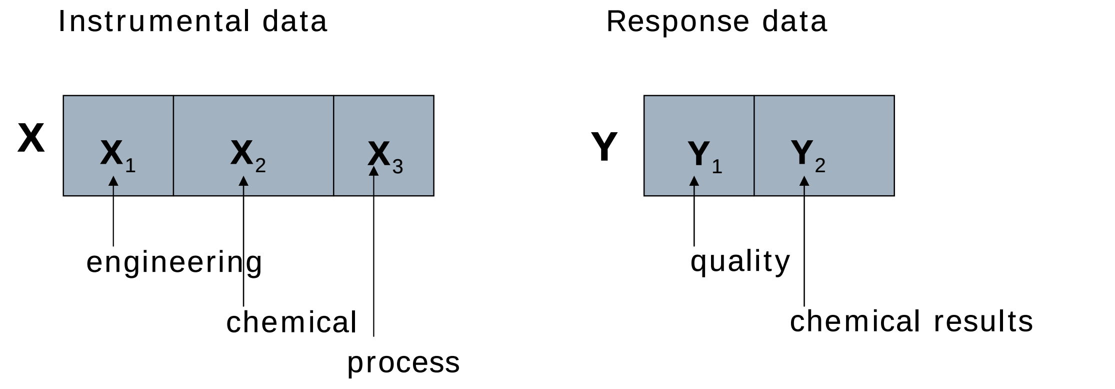


Figure 1. Schematic illustration of instrumental and response data.

It is supposed here that the instrumental data **X** consists of three parts and response data **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 **w** is computed, which reflects how well the instrumental data describe **Y**. The weight vector is used to compute a score vector **t**=**Xw**. The weight vector consists of three parts that correspond to the partition of **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 e.g., NIR measurements containing 1056 values, $\mathbf{w}_2$ will contain 1056 values. The engineering measurements might be only say, 20 values for each sample. That would mean 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 few quality variables among the response variables, while there may be more variables among them that represent the chemical results in question. There may be say, 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 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.

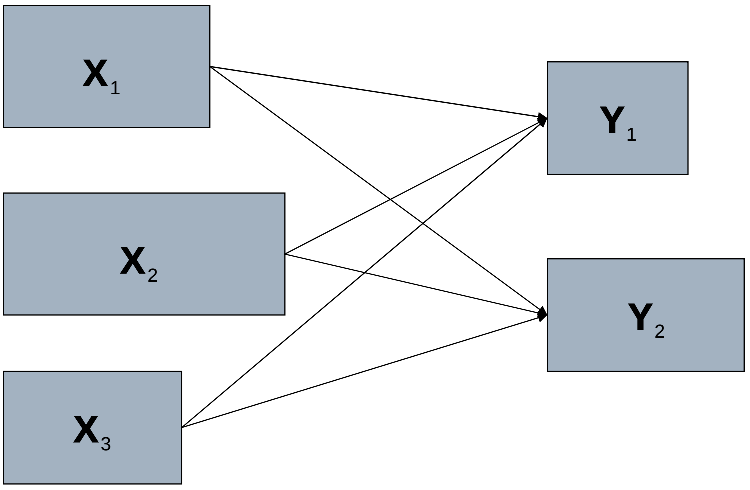


Figure 2. A schematic illustration of the modelling task.

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 $\mathbf{T}_i=(t_{i,1},\ldots,\mathbf{t}_{i,Ai})$ 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 $\mathbf{t}_{1,3}$,…$\mathbf{t}_{1,A}$, 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.

2 Tasks of modelling data

When the data is partitioned as described above, there may be as a result many data blocks, both **X**'s and **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, X→Y**. If the data is not partitioned at all, specific results can be obtained. Partitioning of data, **X** and **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 **Y**'s. This task is considered closer in section 6.
- **Contribution at each step**. At each step there are found score vectors from the **X**'s. The user wants to know the contribution that is obtained for each of the **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 contribute, if it was 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 are 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 is a part 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 **Y**'s can be the **X**'s.

3 The principle of optimisation

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 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 thus that the variance is small? The answer is positive, which follows from the Cauchy-Schwarts inequality, $|\mathbf{y}^T\mathbf{t}| \le |\mathbf{y}|\times|\mathbf{t}|$.

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,

$$(1) \qquad \mathbf{w}^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w} = \lambda\ \mathbf{w}$$

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 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 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 is here to maximize $|\mathbf{q}|^2$.

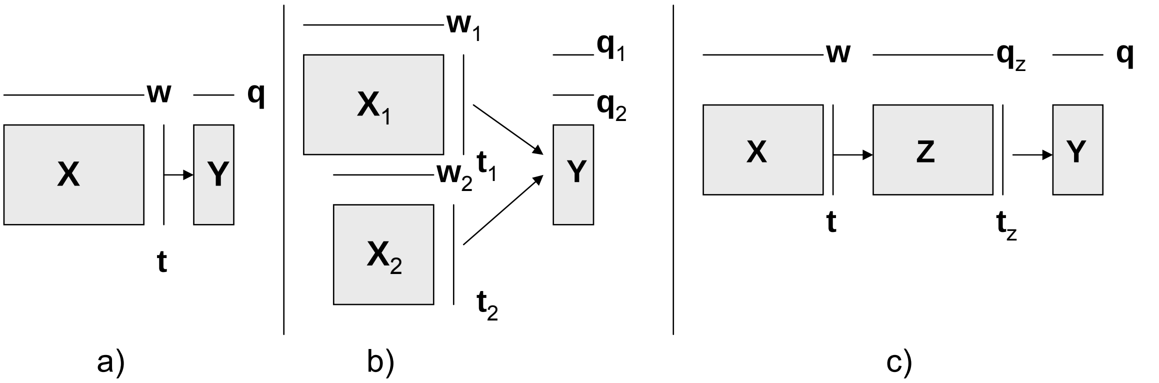
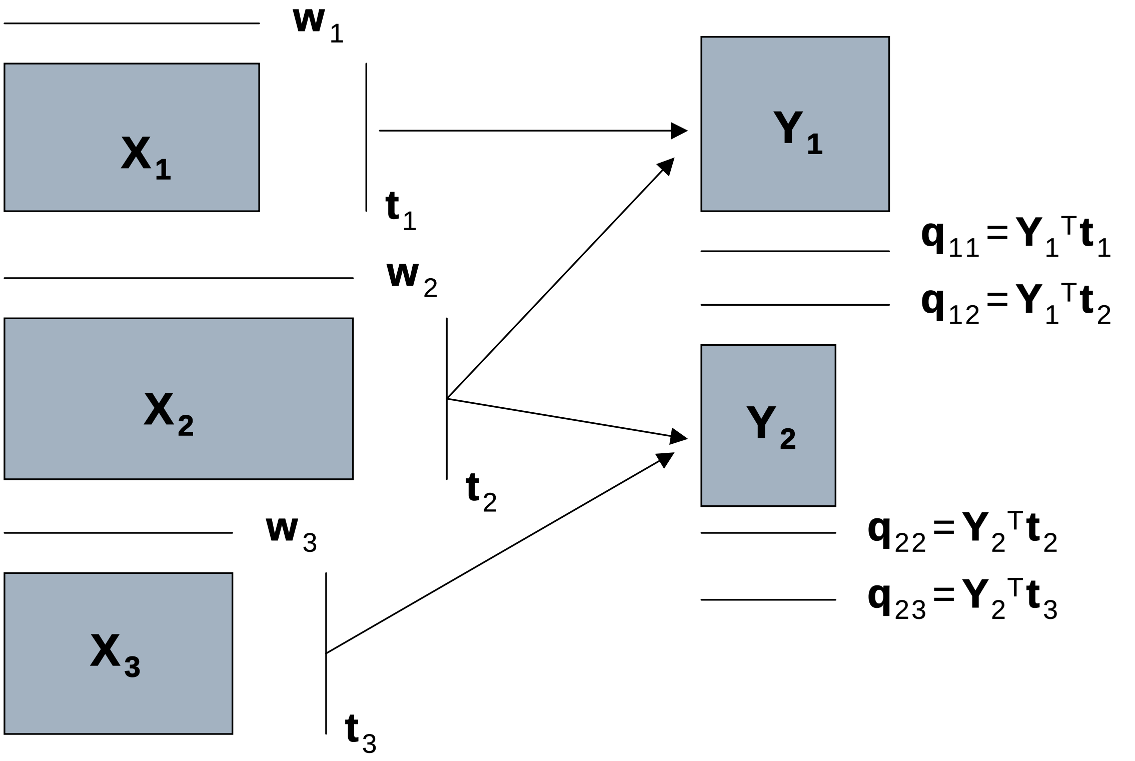


Figure 3. Schematic illustration of the optimization tasks. a) $\mathbf{X}\rightarrow\mathbf{Y}$, b) $\mathbf{X}_1\rightarrow\mathbf{Y}$ and $\mathbf{X}_2\rightarrow\mathbf{Y}$, c) $\mathbf{X}\rightarrow\mathbf{Z}\rightarrow\mathbf{Y}$.

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 e.g., to unit variances within each data block. The issue of scaling is not considered closer here.

4 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 **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 $\mathbf{Y}_1$.



$\mathbf{w}_1$
$\mathbf{X}_1$
$\mathbf{t}_1$
$\mathbf{Y}_1$
$\mathbf{q}_{11}=\mathbf{Y}_1^T\mathbf{t}_1$
$\mathbf{q}_{12}=\mathbf{Y}_1^T\mathbf{t}_2$
$\mathbf{w}_2$
$\mathbf{X}_2$
$\mathbf{Y}_2$
$\mathbf{t}_2$
$\mathbf{w}_3$
$\mathbf{q}_{22}=\mathbf{Y}_2^T\mathbf{t}_2$
$\mathbf{q}_{23}=\mathbf{Y}_2^T\mathbf{t}_3$
$\mathbf{X}_3$
$\mathbf{t}_3$

Figure 4. Illustration of the modeling task.

$\mathbf{X}_2$ 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. Similar 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 **Y**-loadings

$$|\mathbf{q}_{11} + \mathbf{q}_{12}|^2 + |\mathbf{q}_{22} + \mathbf{q}_{23}|^2 = \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{Y}_1\mathbf{Y}_1^T\mathbf{X}_1\mathbf{w}_1 + 2\mathbf{w}_1^T\mathbf{X}_1^T\mathbf{Y}_1\mathbf{Y}_1^T\mathbf{X}_2\mathbf{w}_2 +$$

$$\mathbf{w}_2^T\mathbf{X}_2^T(\mathbf{Y}_1\mathbf{Y}_1^T+\mathbf{Y}_2\mathbf{Y}_2^T)\mathbf{X}_2\mathbf{w}_2 + 2\mathbf{w}_2^T\mathbf{X}_2^T\mathbf{Y}_2\mathbf{Y}_2^T\mathbf{X}_3\mathbf{w}_3 + \mathbf{w}_3^T\mathbf{X}_3^T\mathbf{Y}_2\mathbf{Y}_2^T\mathbf{X}_3\mathbf{w}_3$$

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,

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

Consider now the general case, where there are L **X**-data blocks, $\mathbf{X}=(\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_L)$, and M **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,

$$\mathbf{G}_{ij} = \mathbf{X}_i^T(\delta_{i1}\mathbf{Y}_1\mathbf{Y}_1^T+\delta_{i2}\mathbf{Y}_2\mathbf{Y}_2^T+\ldots+\delta_{iM}\mathbf{Y}_M\mathbf{Y}_M^T)\mathbf{X}_j, \qquad \text{for } i,j=1,2,\ldots,L.$$

Here $\delta_{im}=1$, if $\mathbf{X}_i$ is modelling $\mathbf{Y}_m$ and zero otherwise. If **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 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 **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 $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_L)$.

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, $\mathbf{T}_m=(\mathbf{T}_{1,m},\ldots,\mathbf{T}_{A,m})$, 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 **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, e.g., by studying how $\mathbf{X}_i$ models $\mathbf{Y}$. Variables that do not satisfy (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 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 **X**, but the **Y**'s are a part of **X**, $\mathbf{Y}_i=\mathbf{X}_i$, i=1,2,…,L. This corresponds to the case, where we want to use all of **X** to describe the parts of **X**. In this case there is only one equation to solve,

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

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.

5 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 straight forward extension of the methods presented in previous sections.

5.1 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.

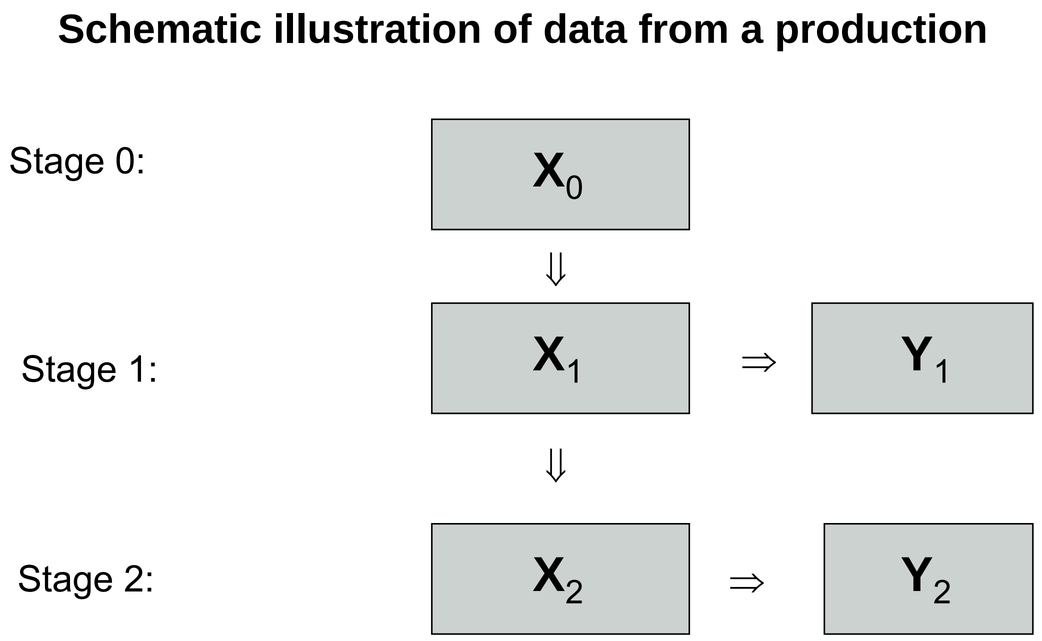


Figure 5. Schematic illustration of three stages in a production process.

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, i.e., 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, $\mathbf{y}_{20}$.

5.2 Modelling seen from Stage 0

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

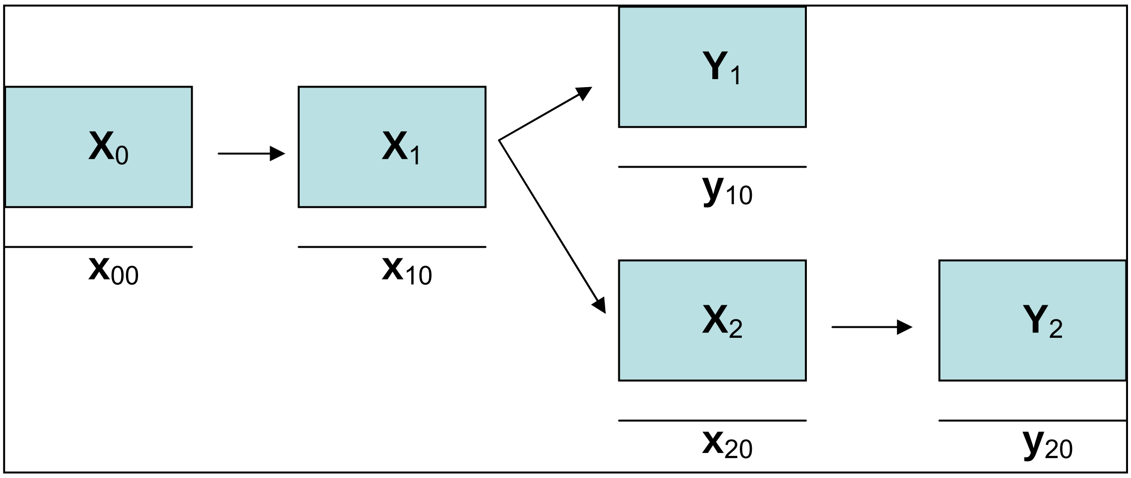


Figure 6. Schematic illustration of the modelling task as seen from Stage 0.

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 maximizes 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.

5.3 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.

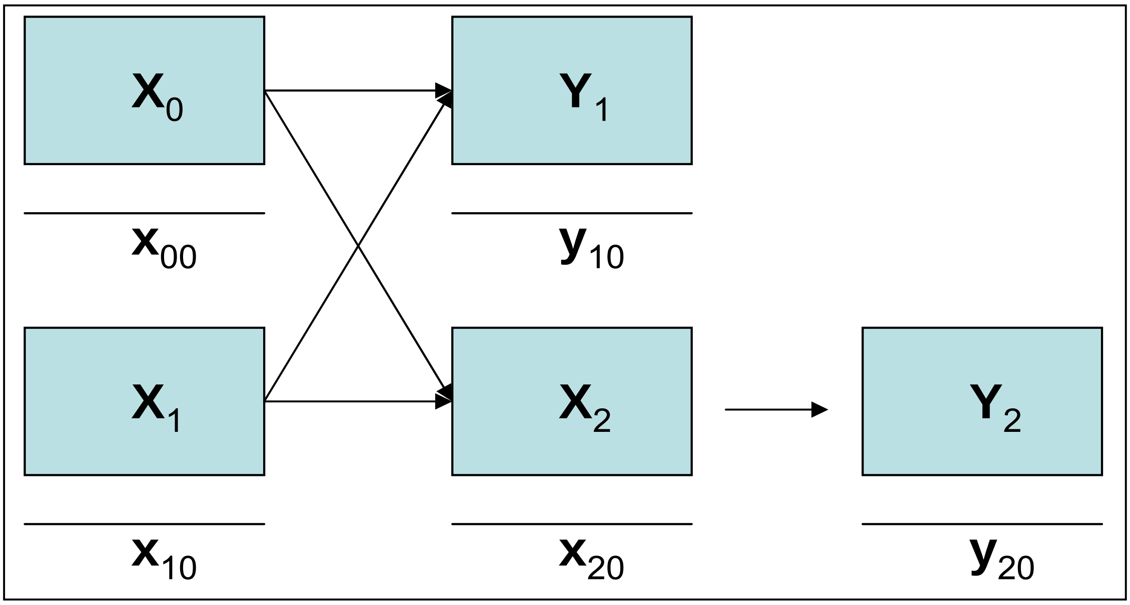


Figure 7. Schematic illustration of the modelling task at stage 1.

$\mathbf{X}_0$ 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 vector computed for both $\mathbf{Y}_1$ and $\mathbf{Y}_2$. The loading vectors for $\mathbf{Y}_1$ are computed as $\mathbf{q}_{10}=\mathbf{Y}_1^T\mathbf{t}_0= \mathbf{Y}_1^T\mathbf{X}_0\mathbf{w}_0$ and $\mathbf{q}_{11}=\mathbf{Y}_1^T\mathbf{t}_1= \mathbf{Y}_1^T\mathbf{X}_1\mathbf{w}_1$. Those of $\mathbf{Y}_2$ are similarly computed as $\mathbf{q}_{20}=\mathbf{Y}_2^T\mathbf{t}_{20}= \mathbf{Y}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{X}_0\mathbf{w}_0$ and $\mathbf{q}_{21}=\mathbf{Y}_2^T\mathbf{t}_{21}= \mathbf{Y}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{X}_1\mathbf{w}_1$. 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

$$|\mathbf{q}_1|^2+|\mathbf{q}_2|^2 = \mathbf{w}_0^T\mathbf{X}_0^T\mathbf{E}\mathbf{X}_0\mathbf{w}_0 + 2\,\mathbf{w}_0^T\mathbf{X}_0^T\mathbf{E}\mathbf{X}_1\mathbf{w}_1 + \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{E}\mathbf{X}_1\mathbf{w}_1$$

$$+\, \mathbf{w}_0^T\mathbf{X}_0^T\mathbf{F}\mathbf{X}_0\mathbf{w}_0 + 2\,\mathbf{w}_0^T\mathbf{X}_0^T\mathbf{F}\mathbf{X}_1\mathbf{w}_1 + \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{F}\mathbf{X}_1\mathbf{w}_1,$$

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

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

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.