1 New horizons in industrial mathematics

During the last decades a new form for mathematics has gained success in analysing industrial data. The basic idea has been to build the solutions to the mathematical models in steps in order to optimize the predictions ability of the final solutions. At each step the improvement of the solution is judged in two ways, both in the light of the present task (fit, optimization etc) and also, how precise the solution is for the given uncertain data. The developed methods have business success both at institutions and industry. The success is partly due to that the new methods secure better predictions than traditional methods and partly that they provide with graphical analysis of data that effectively shows the inherent variation in data. The tools presented here are now working tools within the field of chemometrics.

1.1 Large sets of data

In the industry there are for the time being large developments within data collection. There are two circumstances that are especially prevailing. The first one is that it is necessary to collect data from the whole production in order to secure the necessary level of quality and security of operations. Here the automobile factories have shown the way by extensive investments during the last decade. The second is the new measurement instruments that have been developed. Especially within the fields of chemistry there have appeared advanced measurement instruments, e.g. the modern spectroscopic instruments. These instruments cover broad area of wavelengths and therefore give many measurement values for each sample. E.g., a NIR (Near Infra-Red) instrument typically gives 1050 values for each sample. Another instrument, a Raman spectroscopy, can give 3300 values for each sample. The new mathematics has proved to be efficient in handling mathematical models having many variables (hundreds, thousands or tenths of thousands).

1.2 Procedures of natural sciences

When there are given data that should be modelled, it is recommended to setup a mathematical model, which one can argue for by theoretical arguments, and where it is reasonable to assume the model for the present data. An example could be the use of the least squares method, which is summarized in Box 1. This method is popular in statistics, and is used as a starting point in many types of analysis, like e.g., regression analysis. When we work with industrial data, it is usually difficult to argue theoretically for a 'correct model'. If we take the example of a NIR instrument, that gives 1050 data values for each sample, it is difficult or impossible to give precise physical explanations for what the individual data values stand for. The technicians can at most say that they expect a material to give reflections in an interval, say from 400 to 500. There is in general not available knowledge about the details of the reflections. And it is not necessary to have this knowledge in order to obtain satisfactory modelling results. The way natural sciences handle mathematical models, which are implemented in many program packages, have three basic problems involved.

1. *Great expectation to data*. There are typically placed expectations to the available data, which the data can not meet. The variables are often highly correlated and there can be considerable uncertainties in the measurements of the individual values. Data are thus very much different from what we expect, when the model is formulated.
2. *Special conditions in data*. There are often special conditions in data, like e.g., groupings of data values, which should be taken into account when formulating the model. It means that working with the data often leads to the models that are finally applied.
3. *Too large models*. The natural science background typically suggests a large model, because there are many variables, and knowledge suggests a detailed model. An example is data from a NIR instrument. It gives 1050 variables. Estimation in linear models that contain all the variables (it can usually be done, if there are more samples than variables), results in models that give good fit, but will be not adequate for prediction.

The conclusion from extensive comparisons is that mathematical methods that apply exact or (numerically) optimal solutions, do not give satisfactory results, when they are applied to industrial data.

Box 1.1. Least squares method

There are given corresponding data $\mathbf{X}$ (N×K), and $\mathbf{Y}$ (N×M) matrix.

It is supposed a linear model:
$\mathbf{y}=\mathbf{b}_1x_1 + \ldots + \mathbf{b}_Kx_K$,
where the parameters $\mathbf{B}= (\mathbf{b}_1,\ldots,\mathbf{b}_K)$ are unknown. They are found by minimising $|\mathbf{Y}-\mathbf{XB}|^2$ with respect to $\mathbf{B}$.

The solution given by the least squares method is $\mathbf{B}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

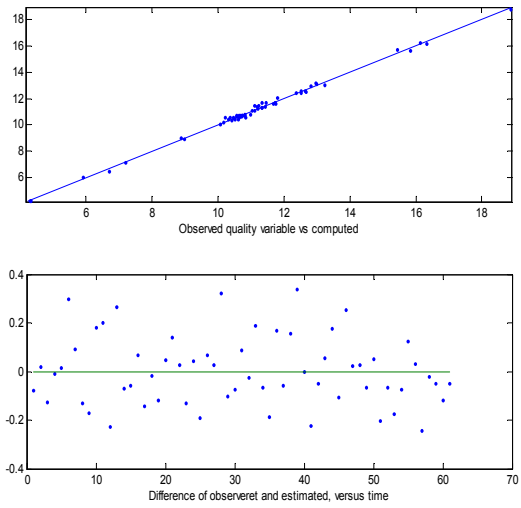

Figure 1.1. Results of PLS regression

1.3 The new procedures

The basic idea is to build the solution by using effective building blocks[1]. The procedure is illustrated here by an example, the Beer data. There are present 61 samples of NIR data from a brewery production of beer. Furthermore, there is given 61 values of a quality parameter for beer. Thus there are given 61 sets, where each set contains 1050 values from a NIR instrument and the corresponding value of the quality parameter. The task is to investigate if it is possible to use the NIR measurements to predict the quality of the final beer. A closer analysis of data shows that it is best only to use variables in the (frequency) interval from 401 to 440. The first step is to find a score vector, a profile vector, that is as reliable as possible and as good as possible in describing the quality parameter. When such a score vector has been found, the data are adjusted for the score vector found, and the analysis starts over again. Altogether four score vectors were found. The results are shown in Figure 1. The upper part shows the observed value of the quality parameter on the y-axis and the corresponding computed value on the x-axis. A line with slope 45^{o} is drawn in the figure in order to better be able to judge the results. It can be concluded that the four score vectors can be used to predict the quality parameter for beer with the desired precision. The lower part of the figure shows the difference of the observed and computed value of the quality parameter on the y-axis and the process time on the x-axis. It shows that the quality parameter can be determined with an uncertainty of around ca. ±0.35. The analysis carried out here is a standard PLS regression. Confidence interval and significance testing of parameters can be carried out by Bootstrapping and Jacknifing[2].

1.4 The H-principle of mathematical modelling – A review

The H-principle or the H-method is a recommendation of how the modelling procedure for any mathematical model should be carried out.. A summary of the ideas is presented in Box 1.2.

A large collection of weighing schemes has been developed to adapt to specific situations. The weighing schemes have been expanded to multi-linear algebra (multi-way data). Regression methods have been extended to path modelling. Methods have been extended to non-linear analysis, dynamic systems, multi-block methods, and other fields of applied mathematics.

1.5 Interpretation of the H-principle

The basic idea is to build up the solution to the mathematical model in steps, where at each step an optimal balance between the prediction ability of the solution and the numerical improvement (in terms of fit, optimality etc). The derivation of the solution is halted, if further steps in deriving the solution do not improve the prediction ability of the solution. The solution to the mathematical model in question obtained in this way has the property that in the light of the given data it provides with almost optimal predictions (the optimisation of predictions is carried out at each step). This way of carrying out model validation at each step, when finding solutions to mathematical models, has been called the H-principle (or the H-method) of mathematical modelling because of the close analogy to the Heisenberg uncertainty principle. The mathematical model is the 'tools' used. The more detailed model that is obtained, the more uncertainty is introduced due to the tools and the variation in data. The H-principle finds the solution that gives a balance between improving the solution and the 'price' in terms of predictions, when the model is applied.

This approach is especially important when analysing industrial data. This is due to the relatively low rank that we typically find in industrial data. The traditional approach, like e.g. we find the program packages (SAS and others), where we first specify and estimate the full model and then carry out a significance testing on parameters is not satisfactory, because the residual values are not reliable values.

It is recommended to build up the solution in terms of elementary parts. An expression should be computed for the possible improvements of the solution in terms of fit or other some criteria. Also, compute an expression for the precision of the possible solutions. These two expressions should be balanced as shown in 3. The optimisation of 3. gives a solution where equal weight is given to the (absolute value of) relative increase in these terms. The improvement is evaluated as suggested in 4.	**Steps of the H-principle** 1.Carry out the modelling in steps. You specify how you want to look at the data at this step by formulating how the weights are computed. 2.At each step compute expressions for a) improvement in fit, ΔFit b) the associated prediction, ΔPrecision 3. Compute the solution that maximizes the product $\Delta\text{Fit} \times \Delta\text{Precision}$ 4. In case the computed solution improves the prediction abilities of the model, the solution is accepted. If the solution does not provide this improvement, we stop modelling. 5. The data is adjusted for what has been selected and start again at 1).
The main motivation for this approach is the prediction aspect of the model. The prediction variance for a standard regression model is, $\text{Var}(\mathbf{Y}(\mathbf{x}_0)) =$ $[\mathbf{Y}^T(\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y}]\times\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0/(\text{N-K})$ Assuming normal distribution, $(\mathbf{X}^T\mathbf{X})^{-1}$ and $[\mathbf{Y}^T(\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y}]$ are stochastically independent, hence both need to be modelled.	**Application to linear regression** It is desired to find a score vector $\mathbf{t}$, $\mathbf{t}=\mathbf{Xw}$. There are two basic aspects of the task of the score vector: a) the size of the improvement in fit, $\lvert\mathbf{Y}^T\mathbf{t}\rvert^2/(\mathbf{t}^T\mathbf{t})$ b) the associated precision in terms of the variance, $\sigma^2/(\mathbf{t}^T\mathbf{t})$ Balancing these two terms as suggested in 3. we get maximise $[\lvert\mathbf{Y}^T\mathbf{t}\rvert^2/(\mathbf{t}^T\mathbf{t})]\times[1/(\sigma^2/(\mathbf{t}^T\mathbf{t}))]$=maximise $\lvert\mathbf{Y}^T\mathbf{t}\rvert^2/\sigma^2$ =maximise $\mathbf{w}^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}/\sigma^2$ for $\lvert\mathbf{w}\rvert=1$. Assuming σ^2 constant the solution is to choose $\mathbf{w}$ as the eigen vector associated with the leading eigen value to $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}=\lambda\mathbf{w}$. This criterion leads to PLS regression.
For $\mathbf{S}=\mathbf{X}^T\mathbf{X}+\mathbf{U}$, with $\mathbf{U}$ positive semi-definite, algorithms have been developed of approximating the exact solution $\mathbf{B}$. In the analysis each term of the decompositions is evaluated. *A* terms are used, if it is judged that further terms do not improve the prediction ability of the model.	**An example of decompositions in linear regression** $\mathbf{S} = d_1\mathbf{p}_1\mathbf{p}_1^T + \ldots + d_A\mathbf{p}_A\mathbf{p}_A^T + \ldots + d_K\mathbf{p}_K\mathbf{p}_K^T = \mathbf{PDP}^T$ $\mathbf{S}^{-1} = d_1\mathbf{v}_1\mathbf{v}_1^T + \ldots + d_A\mathbf{v}_A\mathbf{v}_A^T + \ldots + d_K\mathbf{v}_K\mathbf{v}_K^T = \mathbf{VDV}^T$ $\mathbf{X} = d_1\mathbf{t}_1\mathbf{p}_1^T + \ldots + d_A\mathbf{t}_A\mathbf{p}_A^T + \ldots + d_K\mathbf{t}_K\mathbf{p}_K^T = \mathbf{TDP}^T$ $\mathbf{X}^T\mathbf{Y} = d_1\mathbf{p}_1\mathbf{q}_1^T + \ldots + d_A\mathbf{p}_A\mathbf{q}_A^T + \ldots + d_K\mathbf{p}_K\mathbf{q}_K^T = \mathbf{PDQ}^T$ $\mathbf{B} = d_1\mathbf{v}_1\mathbf{q}_1^T + \ldots + d_A\mathbf{v}_A\mathbf{q}_A^T + \ldots + d_K\mathbf{v}_K\mathbf{q}_K^T = \mathbf{VDQ}^T$ $\hat{\mathbf{Y}}=\mathbf{XB} = d_1\mathbf{t}_1\mathbf{q}_1^T + \ldots + d_A\mathbf{t}_A\mathbf{q}_A^T + \ldots + d_K\mathbf{t}_K\mathbf{q}_K^T = \mathbf{TDQ}^T$

Box 1.2. Summary of ideas of the H-principle

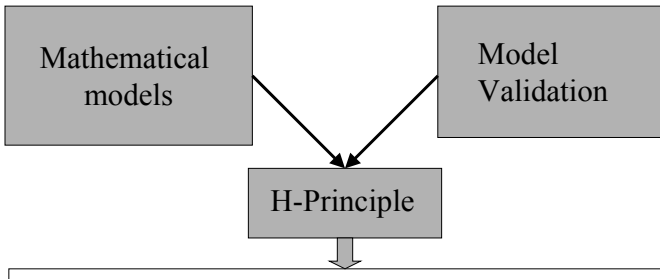

Figure 1.2 The role of the H-principle

E.g., NIR data may contain 1050 variables, but the rank (obtained by the H-principle) may be 10 or less. Better and more reliable significance testing is obtained, when it is based on solution found by the H-principle.

Many mathematical methods are looking for solutions that have full rank, possibly after a regularisation (Ridge or Marquardt) of the covariance matrix. Examples of such methods are found, when solving non-linear models. It will in general be better to use the approach suggested by the H-principle and use low rank solutions at the iterations.

The methods developed provide with a conceptual basis of some methods used in chemometrics.

I believe that these methods will be the ones that future statistical methods will be based upon. The reason is the advance of computers. Computers are so fast that an extensive study of the prediction ability of the model as suggested by the H-principle can be carried out in a very short time. Inference in mathematical models is more reliable, when the

solution has been derived by this approach than by traditional ones.

1.6 Changes in the solution by increasing dimension

For many scientists it is natural to base their methods on a certain criterion like e.g., the linear least squares, where the sizes of the residuals are minimized. In this method it is required to estimate the regression coefficients **B** such that the sizes of the residuals, difference **Y**-**XB**, are as small as possible. This is perfectly normal requirement to the regression coefficients.

In the literature it is common to see some optimization criteria that are used as basis for the estimation of the unknown parameters. These criteria are natural to use. But there is a price to be paid that people may not be aware of. The issues are illustrated in the light of an example.

In Figure 1.3 is shown the variation of the solution vector for the regression analysis of the Beer data that was used in Figure 1.1. Here the wavelengths in the range 400-480 are used. The data matrix **X** is thus 61 times 81. It is possible to compute the dimension up to 61, but only the first 25 is shown. By using double precision it is not a numerical problem to compute the full linear least squares solution. In the figure there are 81 curves associated with the 81 x-variables. When the data matrix **X** is of low rank, curves will fluctuate more by increasing dimension than shown here. The regression coefficients are shown for auto-scaled data. When the regression coefficients are presented in this way, the curves typically will be located within a cone. The cone can be very wide. A regression coefficient b_i, may be positive in the beginning, but ending by having a large negative value. It may be small in the beginning, get large negative value and end with large positive value.

There are typically three values of the dimension that are important. There is a value of the dimension, above which we are beyond the numerical precision of the data. For this and higher dimension, the residuals **Y**-**XB**, are smaller than the numerical precision given in values of **Y**. The second is the dimension that was found significant by traditional significance tests. The third is the dimension that is optimal in the sense that the best prediction is obtained using this dimension. One can ask: Why is the significant dimension often a lot larger than the one found at the best prediction? The answer is that when

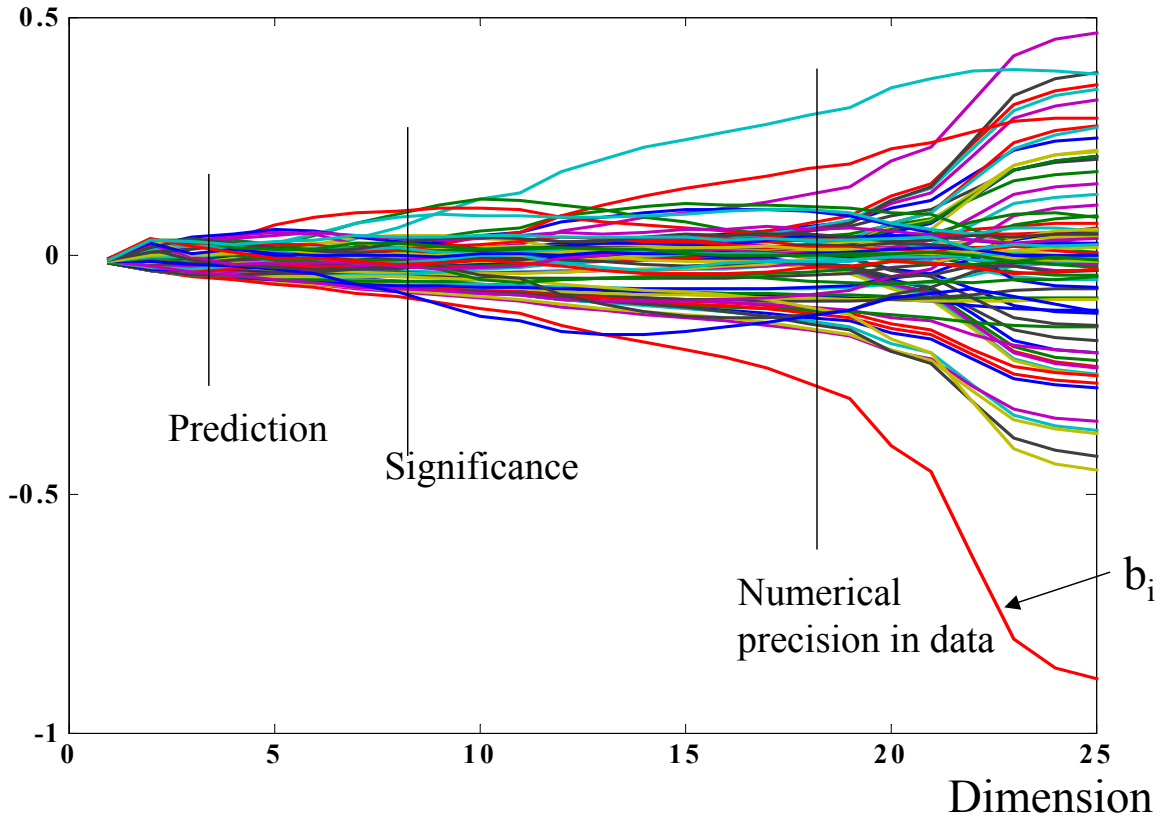

Figure 1.3 Variation of the solution vector with increasing dimension

searching among many correlation coefficients, there will automatically be found several coefficients that exhibit significance. For 81 x-variables, like here, we are working in an 81-dimensional space. When searching among 81 x-variables, we are bound to get spurious effects. This shows normally in small score vectors (a score vector is the vector showing the additional values from **X** compared to the previous ones). It is a fundamental problem in using tests based on the correlation coefficient, or equivalent ones, that they are invariant to the size of the score vector involved.

When scientists use the exact solution based on some optimization criterion (that theoretically might provide with optimal (unbiased, minimum variance etc) solution), the solution vector will typically exhibit behaviour as shown in Figure 1.3, when working with industrial data. Why do the optimization criteria give solutions that provide with bad predictions, when applied to industrial data? The reason is that the criteria assume that the data fit appropriately to the specified data. E.g., the linear least squares method gives unbiased and minimum variance estimates of regression coefficients if the model is correct for the present data. But typically the model formulated is not correct.

It is a fundamental problem for the program packages for statistical analysis that they base their significance testing on the full rank solution. The full solution may be fluctuating wildly, and a significance testing may give highly unreliable results.

Companies working with NIR data that use the methods presented here, measure 1050 wavelengths, variables, select typically 30-60 wavelengths by the methods presented here. Also, the dimension used is typically between 3 and 6. Thus, optimal predictions are obtained by using 3 to 6 score vectors that are derived from an **X** matrix with the number of columns between 30 and 60. The methods used are the ones presented here or are basically equivalent to them in the sense that the same results can be obtained by the ones presented here.

1.7 Standard plots associated with data analysis

When carrying out data analysis it can be recommended to look at following plots to study the results of the analysis:

- **Observed versus computed Y-values**. The columns of **Y** are drawn against the corresponding columns of $\hat{\mathbf{Y}}=\mathbf{XB}_A$. The graphs are supplied by different measures of how good a fit has been obtained.
- **Y-values against the score vectors**. These graphs show the quality of the fit at each step of the computations. In case of Stepwise Regression analysis the weight vectors are given as $\mathbf{w}_a$ =(0,0,…,0,1,0,…), where 1 corresponds to the variable selected. In this case the graphs are called 'Added variable plots'.
- **Y-values against the residuals**. The residuals are given as $\mathbf{E}=\mathbf{Y}-\mathbf{XB}_A$. If the plots of the columns of **Y** against the corresponding column of **E** show systematic variations, it indicates that the modelling task has not been successful.
- **The Y-residuals**. The columns of the residual matrix **E** are to exhibit random behaviour. Therefore, plots, where the y-axis is a column of **E** and x-axis is e.g., the sample number or a score vector, should show random scatter of points.

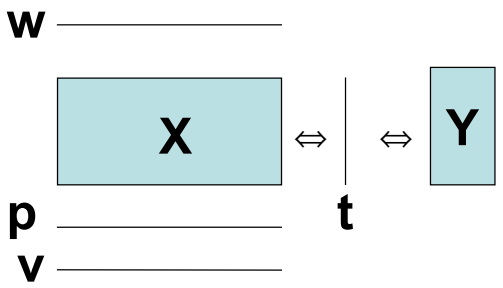

Figure 1.4. Schematic illustration of vectors

In the numerical computations associated with using the algorithms of the H-principle there are computed four sets of vectors, $\mathbf{w}_a$, $\mathbf{p}_a$, $\mathbf{v}_a$, and $\mathbf{t}_a$, at each step. These vectors are schematically illustrated in Figure 1.4. In the following it is described how we look at these vectors and how they can be used in different types of plots.

- $\mathbf{w}_a$, **the weight vector**. It reflects the emphasis of the analysis. Different weights give different regression analysis. In PLS-regression or PLS-regression type of analysis they are computed as shown in Box 1.2, where **X** is the reduced **X**-matrix, $\mathbf{X}=\mathbf{X}_{a-1}$. In the plots of the vectors we look for if one or more variables get small weights for all weight vectors. If we can see that one or more of them get small weights, it is investigated if they should be removed from analysis.
- $\mathbf{t}_a$**, the score vector**. It is computed as $\mathbf{t}_a=\mathbf{X}_{a-1}\mathbf{w}_a$ or $\mathbf{t}_a=\mathbf{X}\mathbf{v}_a$. They define the latent structure. They show what has been used of **X** and how we can describe **Y**. Pair wise plots of the score vectors show the variation in the part of data that is being used.

- $\mathbf{p}_a$, **the loading vector**. It is computed as $\mathbf{p}_a = \mathbf{S}_{a-1}\mathbf{w}_a$. If $\mathbf{S}=\mathbf{X}^T\mathbf{X}$, then $\mathbf{p}_a = \mathbf{X}^T\mathbf{t}_a$. If the $\mathbf{X}$ matrix has been auto-scaled, and $\mathbf{t}_a$ scaled to unit length, the loading vector $\mathbf{p}_a$ can be viewed as the correlation coefficients between the original variables and the a^{th} score variable. In the general case where $\mathbf{S}$ is any positive definite matrix, a similar interpretation is used. Pair wise plots of the loading vectors show the correlation structure in data.
- $\mathbf{v}_a$, **the loading weight vector**. It is given by $\mathbf{p}_a=\mathbf{S}\mathbf{v}_a$. They show how $\mathbf{p}_a$ is derived from the correlations of the original X-variables. Since $\mathbf{S}_0=\mathbf{S}$, we have $\mathbf{v}_1=\mathbf{w}_1$. We study the loading weight vector in order to know how the original variables generate the latent structure.

Figure 1.4 shows that the vectors $\mathbf{w}_a$, $\mathbf{p}_a$, and $\mathbf{v}_a$ are of the same size. It also emphasizes that the score vectors $\mathbf{t}_a$ are used for describing both $\mathbf{X}$ and $\mathbf{Y}$, although the primary purpose with the analysis is to describe $\mathbf{Y}$. In the applied work much time is spent on analyzing how the score vectors describe $\mathbf{X}$. Note that all the graphical analysis above can be done for any choices of the weight vectors $\mathbf{w}_a$ that have been selected and any covariance matrix $\mathbf{S}$.

Example. Score plots from PLS Regression

The McMaster University in Toronto, Canada, has established in cooperation with 16 large companies a centre, McMaster Advanced Control Consortium[3]. The purpose of this centre is to apply these techniques on the 'factory floor'. It has given the companies essential competition advantages to apply these techniques[4]. Furthermore, it has had positive influence on the staff to be trained in reading these plots in order to study the operating conditions. We shall here look at an example from an introductory analysis of the operations. There have been measured 12 process variables. $\mathbf{Y}$ contains the values of the quality variables. The measurements are 289 hourly ones. Thus $\mathbf{X}$ is a 289×12 matrix. In the practical work much time is spent on how the score

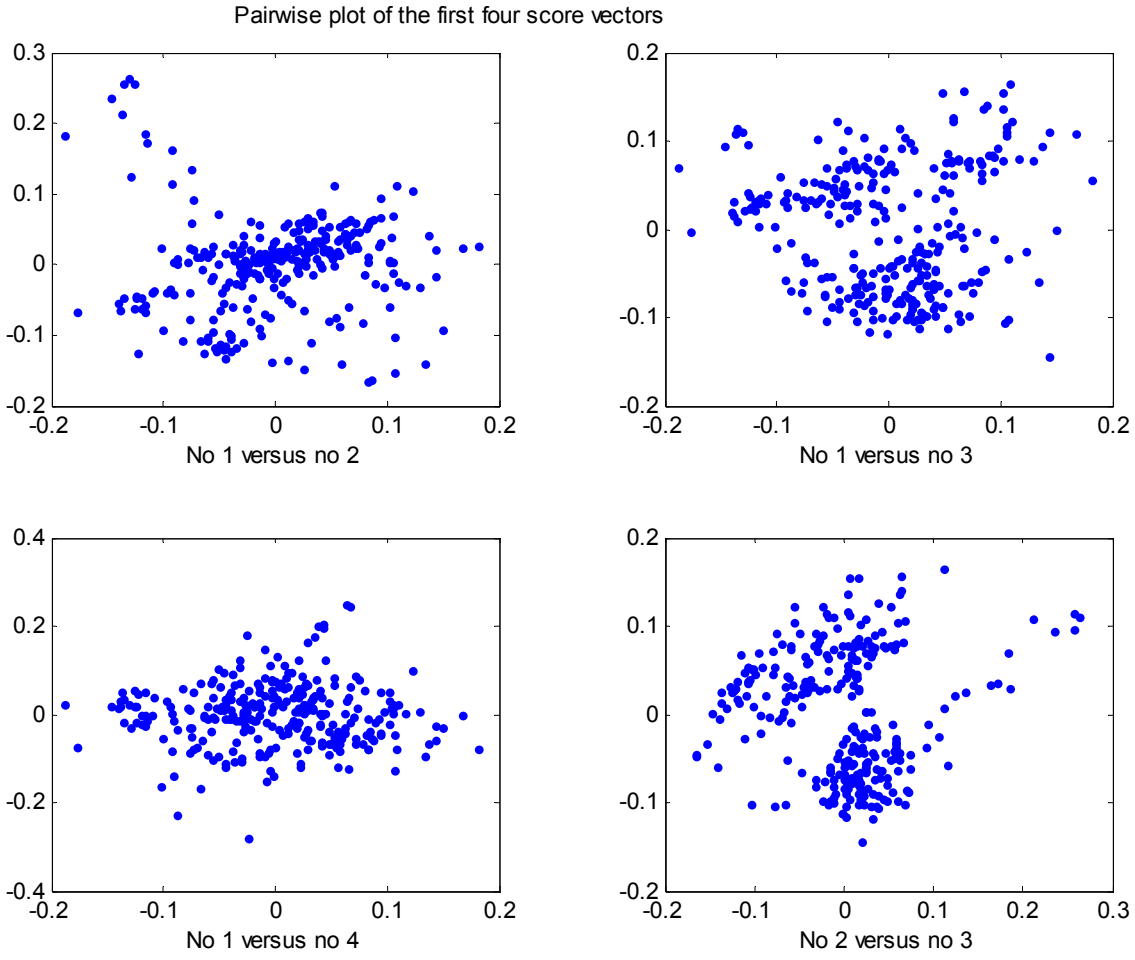

Figure 1.5. Pair wise plot of the first four score vectors. 298 points on each graph.

vectors relate to **X**. An example is illustrated in Figure 1.5, which shows the pair-wise plot of the first four score vectors. When looking at the plot of score vector no. 2 versus no. 3, we can see that the points fall into two groups. A closer analysis shows that one group corresponds to the samples for the first 144 hours, while the second group the last 154 hours. By modelling **X** also, we are able to detect quickly that there have been changes in the operating condition almost half ways of the process.

1.8 New methods and new mathematics

The criteria that the H-principle suggests are easy to compute and fast to implement. This has led to new methods in modelling data. Here we shall briefly describe briefly one type of methods, Path Modelling, that is used for analyzing a network of data blocks. In a regression analysis there are given two data blocks, **X** and **Y**, where the rows of X and Y are corresponding samples. We write **X**⇒**Y** to show that a new sample in **X** is used to estimate a new sample in **Y**. Here **X** is an *input* data block and **Y** an *output* data block. For three data blocks we can write **X**⇒**Y**⇒**Z**, to indicate that a sample in **X** is used to estimate a sample in **Y**, and this estimate is used to estimate a new sample in **Z**. Here **X** is an input data block and **Z** an output data block. The numerical methods do not distinguish rows from columns. Thus, they can be used to project along rows, **X** ⇑ **Y**. And these can be mixed in different ways, **X**⇑**Y**⇒**Z**. The H-principle and these simple ideas have been used to formulate methods to analyze an arbitrary large network of data blocks, with severalinput and output data blocks. These methods have great possibilities in the industry, because we can view one data block as the data from one production place. These methods provide with possibilities to estimate from production places that are located somewhere in the network, the production data that are located later in the production.

The procedures suggested by the H-method are based on weighing the variable such that the results are reliable and good to use. Similar weighing procedures can be used on the rows of **X**. These two types of weighing can be combined to satisfy some special modelling tasks. In fact this weighing procedure can be extended to multi-linear algebra, where data matrices have multiple indices. This has led to a new type of mathematics for multi-linear algebra. Using this approach the concepts like rank, inverse and others become uniquely defined. The importance of this approach is due that these methods are natural extensions of weighing variables to two-ways weighing two-way data and multi-way weighing of multi-way data. These multi-way methods reduce to standard methods of e.g. regression, if only one type of weighing is used or only one mode gives significant variation.

1.9 Industrial success

The methods described here have obtained success at many companies and institutions. At the department of Dairy and Food Science, KVL, Copenhagen, there are around 30 employees working daily with these methods[5]. The company Foss-Electric[6] in Hillerød, Denmark, produces different types of measurement instruments based on the NIR technology. They use these methods and the NIR technology to estimate the chemical contents of the samples (diary product, corn, oil, wine and other products). They have yearly sales of these measurement instruments of around 300 mio euros.

Further reading

1. Höskuldsson, A.: *Prediction Methods in Science and Technology. Vol 1. Basic Teory*. 1996.
Thor Publishing. København. ISBN 87-985941-0-9
2. Martens, H., M. Martens: Multivariate analysis of quality – an introduction. J. Wiley, 2001
3.
http://www.chemeng.mcmaster.ca/MACC/default.htm
4. P. Nomikos & J.F. MacGregor: Multivariate SPC charts for Monitoring Batch Processes. *Technometrics*, 37, 41-59, 1995.
5. Munck L., L. Nørgaard, S. B. Engelsen, R. Bro, C. A. Andersson: Chemometrics in food science – a demonstration of a highly exploratory, inductive evaluation strategy of fundamental scientific importance. *Chemometrics and Intelligent Laboratory Systems*, 44 (1998) 31-60.
6. www.Foss-Electric.dk