Appl. Math. and Comp. Sci., 1998, Vol.8, No.4, 755–773

THE HEISENBERG MODELLING PROCEDURE AND APPLICATIONS TO PROCESS CONTROL

Agnar HÖSKULDSSON*

The Heisenberg modelling procedure is presented. The underlying idea is to define the *basic modelling units* that describe the modelling task. The associated algorithm gives a simultaneous expansion of the data matrix $\boldsymbol{X}$ and its generalised inverse $\boldsymbol{X}^+$. The procedure is applied to linear regression, time series analysis, Kalman filtering and nonlinear modelling. Some applications to Statistical Process Control (SPC) are described. Applications of the Heisenberg modelling procedure give new algorithms in different fields. The advantage of the procedure is that it gives stable estimates of parameters.

1. Introduction

Modern measurement equipment is advanced in many ways. In chemistry, e.g., if a sample is measured we get a 'landscape' of measurement values that depict different aspects of the sample. It is common to get, say, 300 values as a result of the measurement process. If 50 samples are measured the 50 times 300 values can be organized in a matrix $\boldsymbol{X}$ of size 50 times 300, where each row of $\boldsymbol{X}$ contains the values of the respective sample. In manufacturing industries automatic inspection procedures are often very cheap to install. This makes it relatively easy to measure many parameters on each unit manufactured or to measure some process characteristics. If the measurement of each unit or at a certain time are placed in a row of $\boldsymbol{X}$, the size of $\boldsymbol{X}$ tends to be very large. If all measurements are selected, the matrix $\boldsymbol{X}$ may easily contain thousands of columns. On the other hand, if quality characteristics, process performance measures, economic magnitudes or similar measurements are organized similarly in a matrix $\boldsymbol{Y}$, we often get much smaller matrices in terms of number of columns. The Heisenberg modelling procedure has been developed in the light of these large data matrices. The basic consideration is to define a weight vector $\boldsymbol{w}$, which reflects the columns of $\boldsymbol{X}$, that is used to compute a score vector $\boldsymbol{t}$, $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$. The score vector $\boldsymbol{t}$ is a representative for the columns (variables) of $\boldsymbol{X}$. Similarly, a weight vector $\boldsymbol{v}$ is defined in some way that is used to compute a loading vector $\boldsymbol{p}$, $\boldsymbol{p} = \boldsymbol{X}'\boldsymbol{v}$. The loading vector is a representative for the samples (sometimes called objects) of $\boldsymbol{X}$ (rows of $\boldsymbol{X}$). If $\boldsymbol{w} = (0, 0, \dots, 0, 1, 0, \dots, 0)$ we are selecting a variable of $\boldsymbol{X}$ (a column). The different choices of $\boldsymbol{w}$ and $\boldsymbol{v}$ reflect the way we want to carry

* Institute of Applied Construction and Production, Danish Technical University, bldg 358, 2800 Lyngby, Denmark, e-mail: ah@akp.dtu.dk.

out the analysis. If a large portion of $\boldsymbol{X}$ is redundant, e.g., $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ and $\boldsymbol{X}_2'\boldsymbol{Y} = \boldsymbol{0}$, it may be necessary to eliminate the redundant part $\boldsymbol{X}_2$. Although the data are large, the rank is typically small. In a regression context it is rare that more than ten components are selected. The basic magnitude to work with is the prediction variance,

$$\mathrm{Var}(y(\boldsymbol{x}_0)) = s^2\, \boldsymbol{x}_0'\, (\boldsymbol{X}'\boldsymbol{X})^+ \boldsymbol{x}_0 \tag{1}$$

Here $\boldsymbol{x}_0$ is a new sample and $y(\boldsymbol{x}_0)$ is our estimate of the response value for this sample. If there are many variables, but the rank is small, it is important to be careful in the modelling procedure, otherwise we may end up with a model that cannot be used for prediction purposes. The modelling task consists thus of two parts:

(a) Obtain a good fit, i.e., a small value of the residual variance $s^2 = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^+\boldsymbol{X}')\boldsymbol{y}/(N - A)$;

(b) Control the size of the model variation, $\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^+\boldsymbol{x}_0$.

Here N is the number of samples and A the number of components or variables in the regression model. The basic idea of the Heisenberg modelling procedure is to carry out the modelling procedure in steps, where at each step a component $(d, \boldsymbol{t}, \boldsymbol{p})$ (d a scaling constant, $\boldsymbol{t}$ a score vector and $\boldsymbol{p}$ a loading vector) is selected. When the data matrix $\boldsymbol{X}$ is decomposed in this way, the associated generalized inverse $\boldsymbol{X}^+$ is computed at the same time. The component is accepted if the improvement in the fit is relatively better than the increase in the model variation. By estimating the parameters precisely, we get more stable estimates than by traditional methods. Even if other methods are used, it can be advantageous to apply the Heisenberg modelling procedure in order to detect where a possible extreme model variation is. The Heisenberg modelling procedure is a stepwise procedure where one component is chosen at a time. This is in contrast to a traditional modelling procedure that starts with a model for data and carries out a statistical testing procedure to reduce the model. There are several problems with this traditional approach to modelling when applied to data from modern measurement equipment. Even if the practical rank of data is low, it is often possible to obtain an exact numerical solution to the given problem. If the full rank solution is used, we may be testing against a residual variation that has no practical interpretation. Also the classical tests (e.g., t- and F-test) do not check the size of the model variation involved. A test may be significant beyond any limits, but the model variation so large that the associated results are useless. If the data values $(\boldsymbol{X}, \boldsymbol{Y})$ can be described by a multivariate normal distribution, the standard theory states that the residual variation, $\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}$, and the inverse $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ are stochastically independent. This means that the residual variance and the model variation are stochastically independent. The implication of this fact in the data analysis is that the variation we may expect from the residual variance does not tell us anything of the variation we may expect from the model variation. What does this mean? In practice, we always work with estimated models. We tend to select a variation (variables, effects, factors and others) according to a measure of fit. This is the case if we are using 'maximum likelihood estimation and associated likelihood ratio tests,' 'least squares and significance tests' and similar procedures. For example, in stepwise regression we see this often. When searching for a variable, we may find one that gives a good improvement in the fit, but the associated score variable (the marginal variable) is so small that the values may be below the noise level in the data. Including such a variable in the model can be disastrous both because the improvement in fit may be 'accidental' and because by including it the model variation may blow up. Therefore, besides these methods that are only based on measures of fit, we must also use methods or measures that give us the magnitude of the model variation or at least control the size of it.

2. An Approach to Linear Regression

We shall consider an approach to linear regression that is an illustration of the Heisenberg modelling procedure. Suppose that there are given N samples that are organized as an $N \times K$ matrix $\boldsymbol{X}$. Let $\boldsymbol{Y}$ be the associated 'response' values, an $N \times M$ matrix. How should we select a weight vector $\boldsymbol{w}$ such that score vector $\boldsymbol{t}$, $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$, is good to use in predicting the response values? This means that we want to find a representative for the variables to use when new response values are to be estimated. There are two basic aspects of this task:

(a) The fit is to be as good as possible,

(b) The model variation should be as small as possible.

Natural ways to formulate these requirements are as follows:

(a) *The size of projection of* $\boldsymbol{Y}$ *onto* $\mathbf{t}$ *is to be as large as possible:* $\max|\boldsymbol{Y}'\boldsymbol{t}|^2/(\boldsymbol{t}'\boldsymbol{t})$,

(b) *The size of the score vector* $\mathbf{t}$ *is to be as large as possible:* $\max|\boldsymbol{t}|^2$.

A score vector that satisfies (a) will give a good fit. Since the model variation is the inverse of the size of the score vector, we are basically minimizing the model variation when requiring (b). In the experimental design, when we are planning new experiments, the primary objective is (b). We want to be sure that the model applies to the situations we may be facing in future applications. These two criteria, (a) and (b), are not in harmony with each other. If (a) is maximized, (b) may be small and vice versa. One way to reach a compromise when needing to maximize two expressions is to take their products,

$$\max\left(|\boldsymbol{Y}'\boldsymbol{t}|^2/(\boldsymbol{t}'\boldsymbol{t})\right) * \left(|\boldsymbol{t}|^2\right) = \max\ |\boldsymbol{Y}'\boldsymbol{t}|^2 = \max\ \boldsymbol{w}'\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{w} \qquad (2)$$

subject to $|\boldsymbol{w}| = 1$. Using the Lagrange multiplier technique, we get as a solution to the eigenvalue task,

$$\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{w} = \lambda\boldsymbol{w} \qquad (3)$$

As $\boldsymbol{w}$ we select the eigenvector associated with the largest eigenvalue. The resulting score vector $\boldsymbol{t}$, $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$, is our choice of the representative of the variables when doing the regression analysis. This choice is the one used in PLS regression as shown in (Höskuldsson, 1996). In case there is only one response variable, $\boldsymbol{y}$, the weight vector $\boldsymbol{w}$ can be found without iteration methods,

$$\boldsymbol{w} = \frac{\boldsymbol{X}'\boldsymbol{y}}{|\boldsymbol{X}'\boldsymbol{y}|} \tag{4}$$

When regression analysis has been carried out on this score vector $\boldsymbol{t}$, the data matrix $\boldsymbol{X}$ is adjusted for this score vector, and we start over again to find the next score vector.

3. The Heisenberg Modelling Procedure

Historical background. The H-principle, or the Heisenberg principle of mathematical modelling of data, has its conceptual background in the discussions of the 1920s concerning description of physical systems. Heisenberg pointed out that there are certain limitations concerning our description of the physical world. Heisenberg's theory concerns 'conjugate magnitudes' which cannot be observed exactly at the same time. An example of such conjugate magnitudes is the position and momentum (speed) of an elementary particle. He formulated his famous Uncertainty Inequality as

$$(\Delta\text{Position}) * (\Delta\text{Momentum}) > \text{constant}$$

where the constant is related to the Planck constant of light. Although Heisenberg formulated his uncertainty inequality in terms of microphysical systems, it has been transferred to 'macrophysical' systems. The basic measurement questions are of general importance, not only for microphysical systems. In philosophy and methodology of sciences it has often been formulated by stating that the description of a physical system and the associated precision are mutually conjugate magnitudes. The uncertainty inequality takes here the form

$$(\Delta\text{Description}) * (\Delta\text{Associated Precision}) > \text{constant} \tag{5}$$

The H-principle is concerned with a mathematical formulation of (5). The basic idea of the H-principle is to carry out the computations in steps, where at each step we seek to balance the improvement in the description we are looking for with the associated precision we in fact can obtain. The balance is obtained by minimizing the left-hand side of (5). When using the H-principle, we require that the quality of what we get is to be good and that it functions well when it is applied. Thus the H-principle suggests that equal weight is given to the improvement in the description (in terms of the fit or other measure) and the functioning of the improvement in the mathematical model.

Formulation of the H-principle. The starting point is a mathematical formulation of (5). The description corresponds to the mathematical 'fit' and the precision to the inverse of the variance. The H-principle is as follows:

(a) Carry out the modeling in steps such that at each step we determine the improvement in the criterion (fit, maximizing function, etc.) and the associated precision.

(b) Balance the improvement in the criterion and change in performance. If no preferences are present, choose the improvement according to

$$\max(\Delta\text{Criterion}) * (\Delta\text{Precision}) = \max(\Delta\text{Fit})/(\Delta\text{Variance})$$

(c) The present change in the parameter vector is accepted if we get a relatively better improvement in the criterion (fit, etc.) than the increase in the model uncertainties.

(d) Adjust the present data with respect to the choices that were made before starting modeling at a new step.

The H-principle suggests that you should carry out the modeling procedure in steps and at each step you should compute a measure of fit and associated precision. The analysis is to be carried out in steps in order to be able to correctly identify the noise level in the data. It is also necessary to carry out the computations in steps in order to be able to detect deviations, e.g. from linearity, and the presence of groups and gaps in the data. This suggestion of the H-principle seems to be the one that is most criticized. It is opposite to the standard statistical procedure that starts with a model, estimates the unknown parameters in the model and tests the significance of parameters or parts of the model by statistical tests. The lack of success of this standard procedure in industrial applications is due to that we are 'forcing' a model on the data that may not be consistent with the data. The reason for the lack of consistency is often that the rank in the data is smaller than suggested by the model.

The Heisenberg uncertainty inequality and (5) are concerned with minimization tasks. But in (b) it is formulated as a maximization task when we work with fit and variance. The reduction in fit is negative, but we take the absolute value of the improvement in fit. Therefore the task becomes a maximization one. In (d) it is suggested to adjust $\boldsymbol{X}$ with what was selected at this step. This means that $\boldsymbol{X}$ is to be orthogonalized with respect to what was selected. This is suggested because then the reduced $\boldsymbol{X}$ becomes independent of what has previously been selected. Thus at each step we have a similar situation and we wish to determine, according to (b), a new variable/component.

Associated with the H-principle there is a decomposition algorithm that simultaneously decomposes $\boldsymbol{X}$ and its generalized inverse $\boldsymbol{X}^+$ into rank one components (Höskuldsson, 1994; 1996),

$$\boldsymbol{X} = d_1\boldsymbol{t}_1\boldsymbol{p}_1' + d_2\boldsymbol{t}_2\boldsymbol{p}_2' + \cdots \qquad (6)$$

$$\boldsymbol{X}^+ = d_1\boldsymbol{r}_1\boldsymbol{s}_1' + d_2\boldsymbol{r}_2\boldsymbol{s}_2' + \cdots \qquad (7)$$

The generalized inverse $\boldsymbol{X}^+$ satisfies $\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{X} = \boldsymbol{X}$.

The role of the vectors. The algorithm works with three pairs of vectors, $(\boldsymbol{w}_a, \boldsymbol{v}_a), (\boldsymbol{t}_a, \boldsymbol{p}_a)$ and $(\boldsymbol{r}_a, \boldsymbol{s}_a)$. Here $(\boldsymbol{w}_a)$ and $(\boldsymbol{v}_a)$ are determined from the criteria given by the practical problem in question. By appropriate choices of $(\boldsymbol{w}_a)$ and $(\boldsymbol{v}_a)$ we can get most of the statistical procedures that are used today. $(\boldsymbol{t}_a)$ and $(\boldsymbol{p}_a)$

reduce $\boldsymbol{X}$ by rank one, and $(\boldsymbol{r}_a)$ and $(\boldsymbol{s}_a)$ reduce $\boldsymbol{X}^+$ by rank one. When $\boldsymbol{w}_a$ or $\boldsymbol{v}_a$ are selected, they can be scaled to be of length one. This is the choice used in the algorithm. In the algorithm we do not scale $\boldsymbol{t}_a$ or $\boldsymbol{p}_a$. The scaling is performed by the constant d_a. Note that a new set of vectors $(\boldsymbol{w}_a, \boldsymbol{v}_a), (\boldsymbol{t}_a, \boldsymbol{p}_a)$ and $(\boldsymbol{r}_a, \boldsymbol{s}_a)$ is determined at each step of the algorithm. This means that, when these vectors have been computed, the results are judged. If the results are not satisfactory, we can revise these vectors. This is especially important when there is both a row and a column criterion. It often happens that at a certain step one of the criterion is not selecting anything. In this case it can be recommended to drop this criterion and to use only the other one. A new set of vectors is computed using only one criterion. Some of the geometric properties of the vectors $(\boldsymbol{t}_a, \boldsymbol{p}_a)$ and $(\boldsymbol{r}_a, \boldsymbol{s}_a)$ are summarized in the equations

$$\boldsymbol{S}'\boldsymbol{T} = \boldsymbol{D}^{-1}, \qquad \boldsymbol{R}'\boldsymbol{P} = \boldsymbol{D}^{-1} \tag{8}$$

where $\boldsymbol{S} = (\boldsymbol{s}_a)$, $\boldsymbol{T} = (\boldsymbol{t}_a)$, $\boldsymbol{P} = (\boldsymbol{p}_a)$ and $\boldsymbol{R} = (\boldsymbol{r}_a)$. $\boldsymbol{D}$ is the diagonal matrix with $d'_a s$ on the diagonal.

Numerical precision. The constant d_a is used for scaling purposes. This is done in order to secure the numerical stability of the algorithm. All computations in the algorithm are products and inner products except the computation of d_a. This means that the algorithm can be programmed to give very precise numerical results. For large matrices it is important to be careful with the numerical stability. This way to arrange the computations secures the stability of results, even for large matrices containing, say, thousands of rows and/or columns. If there is a large difference in the sizes of the elements in $\boldsymbol{X}$, it can often be recommended to scale the data, e.g., such that each column of $\boldsymbol{X}$ has length or variance one.

Plots. In the applications of the algorithm it is useful to look at plots of the vectors computed at each step. A short guideline for using the plots is as follows:

(a) $\boldsymbol{y}$ *vs* $\boldsymbol{t}_a$. *Added variable plots.* They show how the new component relates to the response variable. The plots should show some degree of linearity. If the scatter plot does not show any kind of linearity, it indicates that the component and the later ones should not be used.

(b) $\boldsymbol{t}_a$ *vs* $\boldsymbol{t}_b$. *Score plots.* They show how the components vary. They are useful in detecting outliers, gaps and groups in data. It may show some special relationship between score vectors that should be taken into consideration.

(c) $\boldsymbol{p}_a$ *vs* $\boldsymbol{p}_b$. *Loading plots.* They show how the variables vary. They are useful in detecting special variables and groups of variables.

(d) $\boldsymbol{r}_a$ *vs* $\boldsymbol{r}_b$. *Transformation plot.* They show how the score values are computed. They are useful in detecting the influence of variables and groups of variables.

Besides the decomposition algorithm there is also associated a collection of formulae that are used to identify the noise level.

4. Only a Column Criterion

We shall consider closer the case where we only have a column criterion. This means that we only choose $\boldsymbol{w}_a$. The weight vector $\boldsymbol{v}_a$ is chosen as $\boldsymbol{v}_a = \boldsymbol{t}_a/|\boldsymbol{t}_a|$. We shall restrict our discussion to the case where we do not compute the score vectors. This may be actual when we are solving the normal equations for $\boldsymbol{B}$,

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B} = \boldsymbol{X}'\boldsymbol{Y} \tag{9}$$

We shall now consider the basic algorithm of the H-principle to carry out the computations of the regression coefficients $\boldsymbol{B}$. The algorithm is formulated in terms of sample values. This is easiest, because it shows how the computations are actually carried out. The algorithm is as follows:

0. *Compute* $\boldsymbol{G} = \boldsymbol{X}'\boldsymbol{X}$. Let $\boldsymbol{G}_0 = \boldsymbol{G}$, $\boldsymbol{U} = \boldsymbol{X}'\boldsymbol{Y}$, $\boldsymbol{U}_0 = \boldsymbol{U}$, $\boldsymbol{F} = \boldsymbol{Y}'\boldsymbol{Y}$, $\boldsymbol{F}_0 = \boldsymbol{F}$, $\boldsymbol{E} = \boldsymbol{I}_K$, $\boldsymbol{B}_0 = 0$.

1. *Determine* $\boldsymbol{w}_a$
   Select $\boldsymbol{w}_a$ by some criterion.

2. *Compute components*
   $\boldsymbol{p}_a = \boldsymbol{G}_{a-1}\boldsymbol{w}_a$, $d_a = 1/(\boldsymbol{w}_a'\boldsymbol{p}_a)$

3. *Rank one reduction of* $\boldsymbol{G}$
   $\boldsymbol{G}_a = \boldsymbol{G}_{a-1} - d_a\boldsymbol{p}_a\boldsymbol{p}_a'$

4. *The transformation vector*
   $\boldsymbol{r}_a = \boldsymbol{E}_{a-1}\boldsymbol{w}_a$
   Update $\boldsymbol{E}$
   $\boldsymbol{E}_a = \boldsymbol{E}_{a-1} - d_a\boldsymbol{r}_a\boldsymbol{p}_a'$

5. *Adjust covariance and residual variance*
   $\boldsymbol{q}_a = \boldsymbol{U}_{a-1}\boldsymbol{r}_a$
   $\boldsymbol{U}_a = \boldsymbol{U}_{a-1} - d_a\boldsymbol{p}_a\boldsymbol{q}_a'$
   $\boldsymbol{F}_a = \boldsymbol{F}_{a-1} - d_a\boldsymbol{q}_a\boldsymbol{q}_a'$

6. *Regression coefficients*
   $\boldsymbol{B}_a = \boldsymbol{B}_{a-1} + d_a\boldsymbol{r}_a\boldsymbol{q}_a'$

7. New iteration? If a new component is to be selected, set a to $a+1$ and go to 1.

The same algorithm is used for the theoretical values. Suppose that x is normally distributed with mean $\boldsymbol{\mu}_x$ and covariance matrix Σ_{xx}. Similarly, suppose that y is normally distributed with mean $\boldsymbol{\mu}_y$ and covariance matrix Σ_{yy}. Let Σ_{xy} be the theoretical covariance matrix between x and y. Then we use this algorithm with $\boldsymbol{S} = \Sigma_{xx}$, $\boldsymbol{U} = \Sigma_{xy}$ and $\boldsymbol{F} = \Sigma_{yy}$. The weight vector $\boldsymbol{w}_a$ can be computed in many ways. Each choice reflects the way we want to carry out the analysis. Here are some examples of some common choices:

(a) $\boldsymbol{w}_a = (0,0,\ldots,0,1,0,\ldots)$. Here we are selecting variables according to some criterion.

(b) $\boldsymbol{w}_a$ is found as an eigenvector to the covariance matrix, $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{w}_a = \lambda\boldsymbol{w}_a$. This choice leads to Principal Component Regression (PCR).

(c) $\boldsymbol{w}_a$ is found as an eigenvector to the eigensystem, $\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{w}_a = \lambda\boldsymbol{w}_a$. This choice leads to PLS regression. If we are working with the theoretical covariance matrix, the eigensystem is $\Sigma_{xy}\Sigma_{xy}'\,\boldsymbol{w}_a = \lambda\,\boldsymbol{w}_a$.

Each type of choice of $\boldsymbol{w}_a$ leads to the specific type of regression analysis. Numerically, we are approximating the linear least squares solution $\boldsymbol{B}$ to eqns. (9). What actually is going on is that data is split up into *elementary units* with the purpose of being able to judge each unit. The decompositions of data and figures into units are as follows:

$$\begin{aligned}
\boldsymbol{G} = \boldsymbol{X}'\boldsymbol{X} &= d_1\boldsymbol{p}_1\boldsymbol{p}_1' + d_2\boldsymbol{p}_2\boldsymbol{p}_2' + \cdots + d_A\boldsymbol{p}_A\boldsymbol{p}_A' + \cdots + d_K\boldsymbol{p}_K\boldsymbol{p}_K'\\
\boldsymbol{U} = \boldsymbol{X}'\boldsymbol{Y} &= d_1\boldsymbol{p}_1\boldsymbol{q}_1' + d_2\boldsymbol{p}_2\boldsymbol{q}_2' + \cdots + d_A\boldsymbol{p}_A\boldsymbol{q}_A' + \cdots + d_K\boldsymbol{p}_K\boldsymbol{q}_K'\\
(\boldsymbol{X}'\boldsymbol{X})^{-1} &= d_1\boldsymbol{r}_1\boldsymbol{r}_1' + d_2\boldsymbol{r}_2\boldsymbol{r}_2' + \cdots + d_A\boldsymbol{r}_A\boldsymbol{r}_A' + \cdots + d_K\boldsymbol{r}_K\boldsymbol{r}_K'\\
\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} &= d_1\boldsymbol{r}_1\boldsymbol{q}_1' + d_2\boldsymbol{r}_2\boldsymbol{q}_2' + \cdots + d_A\boldsymbol{r}_A\boldsymbol{q}_A' + \cdots + d_K\boldsymbol{r}_K\boldsymbol{q}_K'\\
\boldsymbol{X} &= d_1\boldsymbol{t}_1\boldsymbol{p}_1' + d_2\boldsymbol{t}_2\boldsymbol{p}_2' + \cdots + d_A\boldsymbol{t}_A\boldsymbol{p}_A' + \cdots + d_K\boldsymbol{t}_K\boldsymbol{p}_K'\\
\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{B} &= d_1\boldsymbol{t}_1\boldsymbol{q}_1' + d_2\boldsymbol{t}_2\boldsymbol{q}_2' + \cdots + d_A\boldsymbol{t}_A\boldsymbol{q}_A' + \cdots + d_K\boldsymbol{t}_K\boldsymbol{q}_K'\\
\boldsymbol{V} = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} &= d_1\boldsymbol{q}_1\boldsymbol{q}_1' + d_2\boldsymbol{q}_2\boldsymbol{q}_2' + \cdots + d_A\boldsymbol{q}_A\boldsymbol{q}_A' + \cdots + d_K\boldsymbol{q}_K\boldsymbol{q}_K'
\end{aligned}$$

In the analysis we find a component $(d_a, \boldsymbol{t}_a, \boldsymbol{p}_a)$ that we are selecting from $\boldsymbol{X}$ and from it we compute the remaining vectors. Furthermore, we only select A components, because we find that further components may improve the fit but they will not improve the prediction ability of the model. If the score vectors $(\boldsymbol{t}_a)$ are not computed at each step, like in the algorithm above, they can be computed afterwards from $\boldsymbol{t}_a = \boldsymbol{X}\boldsymbol{r}_a$.

In the standard analysis we use $\boldsymbol{X}'\boldsymbol{X}/N$ (or $\boldsymbol{X}'\boldsymbol{X}/(N-1)$ in the case of centered data) as an estimate of the theoretical covariance matrix Σ_{xx}. Sometimes we may expect some theoretical structure of the covariance matrix Σ_{xx}. In this case we may have to iterate to get an optimal solution, where $\boldsymbol{X}$ is replaced by $\boldsymbol{X}\Sigma_{xx}^{-1/2}$ at each iteration. In other situations we may want to specify some theoretical structure of the covariance matrix Σ_{xx} and compute the solution vector $\boldsymbol{B}$ on the basis of this covariance matrix. We can use the formulae to compute some figures of interest, e.g., we may want to compute the score vectors that are associated with this theoretical covariance matrix. We do this from the formula above, $\boldsymbol{t}_a = \boldsymbol{X}\boldsymbol{r}_a$.

Constrained parameters. It is sometimes needed to constrain the solution $\boldsymbol{B}$ that is obtained. Suppose that we want a constrained solution where the constraints are given as $\boldsymbol{CB} = \boldsymbol{F}$. These constraints are included in the algorithm by expanding $\boldsymbol{X}$ by $\boldsymbol{C}$. In fact, we place $\boldsymbol{C}$ on the top of $\boldsymbol{X}$. Then we reduce $(\boldsymbol{C}, \boldsymbol{X})$ with respect to equations containing $\boldsymbol{C}$. Here only the *forward procedure* is needed. The reduced $\boldsymbol{X}_C$ enters the algorithm and provides us with a solution. This solution goes then through the *backward phase* of the equations of $\boldsymbol{C}$. The resulting solution $\boldsymbol{B}$ will then satisfy the constraints $\boldsymbol{CB} = \boldsymbol{F}$.

5. Application to Time Series Analysis

A popular model for the time series (x_t) is the autoregressive model (AR),

$$x_t = b_1 x_{t-1} + b_2 x_{t-2} + \cdots + b_p x_{t-p} + \epsilon_t \tag{10}$$

where the value at time t is a linear combination of previous values (up to lag p) apart from some small random variations, ϵ_t. At time t we know the previous errors, $\epsilon_s, s < t$. It is therefore natural to include in the model the errors of previous periods. This leads to the Autoregressive Moving Average model (ARMA),

$$x_t = b_1 x_{t-1} + b_2 x_{t-2} + \cdots + b_p x_{t-p} + \epsilon_t + a_1 \epsilon_{t-1} + \cdots + a_q \epsilon_{t-q} \tag{11}$$

In the basic algorithm that is associated with the H-principle it is assumed that we have a data matrix $\boldsymbol{X}$ and $\boldsymbol{Y}$, and we want to do regression of $\boldsymbol{Y}$ upon $\boldsymbol{X}$. In the case of (10) it is simple to generate $\boldsymbol{X}$. The first row of $\boldsymbol{X}$ is $(x_1, x_2, \ldots, x_p)$, the second row of $\boldsymbol{X}$ is $(x_2, x_3, \ldots, x_{p+1})$, and so on. In order to generate $\boldsymbol{X}$ in the case of (11) an iterative procedure is needed. We initialize $\boldsymbol{X}$ in the same way as in the case of the AR model. The residuals obtained are then extended to $\boldsymbol{X}$ by adding columns to $\boldsymbol{X}$ corresponding to the residuals found. This new $\boldsymbol{X}$ is now used to find new residuals that are again inserted in $\boldsymbol{X}$. This is continued until convergence. Usually only few iterations (less than, say, 20) are needed. At convergence we have a matrix $\boldsymbol{X}$ that contains the data values that are used in the analysis. For large models the standard linear least squares solution is often unstable. This way of arranging or generating the data makes it possible to use the techniques presented in (Höskuldsson, 1996) in the analysis of time series data. They provide us with a new tool for the analysis. Plots involving score variables are important. If the score values are not stable over the period, it indicates that the process generating the data is not stationary. If we are working with large models, we typically do not select all components. This procedure allows us to 'peel off' the part of variation in the data that only contributes to the model variation. The methods in (Höskuldsson, 1996) for finding dimensions in linear models are useful in the time series context, see (Höskuldsson, 1998b) for different applications of this approach to time series analysis. It also contains a procedure showing how Kalman filtering may be embedded within this framework.

6. Application to Nonlinear Modelling

We shall now show how we can extend this approach to nonlinear modeling of data. In PLS regression we find a linear surface that will give us a stable projection. Similarly, we can ask for a quadratic surface or a surface of a higher order.

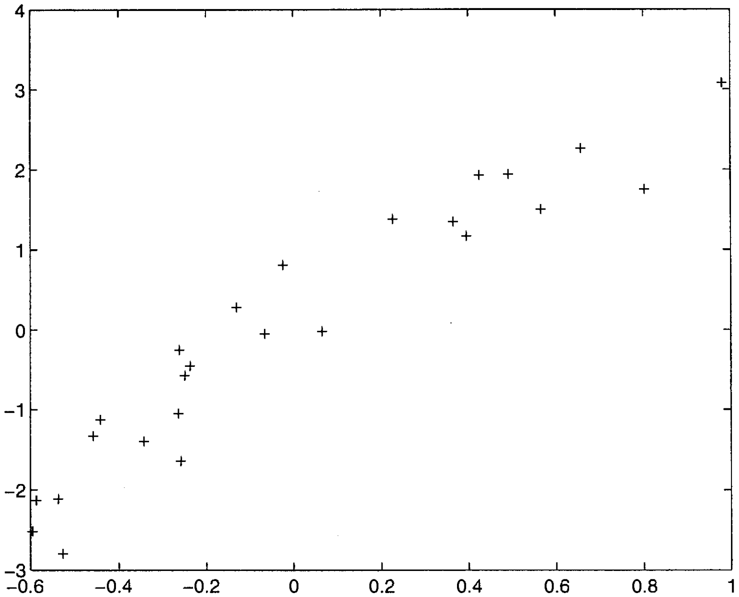

Fig. 1. Fat content vs the first PLS score values (data values centered, 25 samples).

The nonlinear procedure that we develop here will be applied to the NIR (Near-Infra Red) data analysed in Section 11.3 in (Höskuldsson, 1996). In industry there is a great interest in NIR measurements, because they are easy to make and e.g., in food industry the measurement instruments need not touch the food. There are given 19 x-variable of NIR measurements on pieces of meat. Each measurement value is the logarithm of the light intensity reflected. The y-variable is the fat content of the meat. 25 samples were measured. The x-variables are highly correlated. Each x-variable shows a quadratic relationship with the y-variable. In order to see this curvature clearly, we compute the first PLS score vector and plot the y-variable against it. This is shown in Fig. 1. From the figure we can see a clear curvature in data. This is probably due to the scale used that the logarithm of the light intensity is not a linear scale.

It is also of some interest to see the score plots of the two first score vectors. This is shown in Fig. 2. We see a quadratic relationship between the score vectors. This suggests that the relationship in X-space should be included in the modelling procedure. It is therefore natural to seek a surface in lower dimensions to describe the response values. When defining score variables, we are defining new variables from the given ones. We shall be using these variables to define the surfaces. The first task is to find the first score variable, $\boldsymbol{t}_1$, such that a polynomial in this variable would give a good approximation to the response variable. When we consider the data values, the squared variable associated with $\boldsymbol{t}_1$ has the coordinates that are the squared ones of $\boldsymbol{t}_1$. We shall denote by $\boldsymbol{t}_1^2$ the vector of squared coordinates, $\boldsymbol{t}_1^2 = (t_1^2, t_2^2, \dots)$. For two vectors $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ the element-wise product is denoted by $\boldsymbol{t}_1 \otimes \boldsymbol{t}_2$. The first task

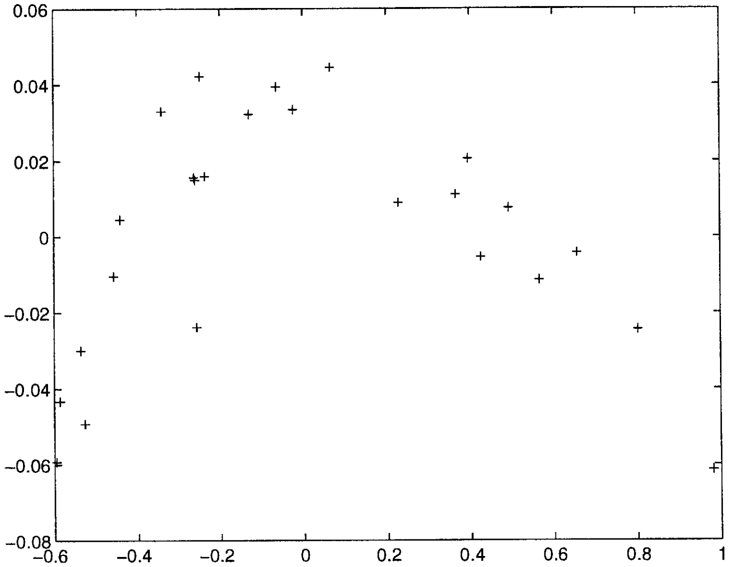

Fig. 2. The second PLS score vector vs the first one (data values centered, 25 samples).

is to answer the question: how to choose $\boldsymbol{w}$ such that $\boldsymbol{t}_1$, $\boldsymbol{t}_1 = \boldsymbol{X}\boldsymbol{w}$, and $\boldsymbol{t}_1^2$ give us a good description of the response variable $\boldsymbol{y}$. The criterion that we shall use is

$$\max\left[(\boldsymbol{y}'\boldsymbol{t}_1)^2 + (\boldsymbol{y}'\boldsymbol{t}_1^2)^2\right] \tag{12}$$

subject to $|\boldsymbol{w}| = 1$. Why is this a good criterion for finding $\boldsymbol{w}$? In (Höskuldsson, 1996, Ch. 7), it is shown that the score vector $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$ determined by maximizing the squared covariance $(\boldsymbol{y}'\boldsymbol{t})^2$ for $|\boldsymbol{w}| = 1$ has some desirable properties. It balances the estimation and prediction aspect of the linear model and it leads to PLS regression. When we have two variables, $\boldsymbol{t}_1$ and $\boldsymbol{t}_1^2$, it is natural to look at the squared covariance that they together contribute to, which is the criterion (18). Another argument for choosing this criterion is given in (Höskuldsson, 1992). In the Appendix it is shown how the criterion can be maximized. When $\boldsymbol{w}$ has been found, the vectors $\boldsymbol{t}_1$ and $\boldsymbol{t}_1^2$ are computed. The estimated response values based on these two score vectors, $\boldsymbol{t}_1$ and $\boldsymbol{t}_1^2$, are plotted against the observed response values in Fig. 3. It shows that we already have obtained a fairly good approximation to the y-values.

The next task is to determine $\boldsymbol{t}_2$. Before this is done, the matrix $\boldsymbol{X}$ is adjusted by the vector $\boldsymbol{t}_1$ that has already been selected. The reason for this is that we want $\boldsymbol{t}_2$ to be orthogonal to $\boldsymbol{t}_1$. This procedure will give us new information that is not contained in the first step when we chose $\boldsymbol{t}_1$. On the other hand, it is possible that we may need product terms involving $\boldsymbol{t}_2$ and its power with the previously chosen variables $\boldsymbol{t}_1$ and $\boldsymbol{t}_1^2$. Thus at the next step we need to find $\boldsymbol{w}$ that maximizes

$$(\boldsymbol{y}'\boldsymbol{t}_2)^2 + \big(\boldsymbol{y}'(\boldsymbol{t}_2 \otimes \boldsymbol{t}_1)\big)^2 + \big(\boldsymbol{y}'(\boldsymbol{t}_2 \otimes \boldsymbol{t}_1^2)\big)^2 + (\boldsymbol{y}'\boldsymbol{t}_2^2)^2 + \big(\boldsymbol{y}'(\boldsymbol{t}_2^2 \otimes \boldsymbol{t}_1)\big)^2 + \big(\boldsymbol{y}'(\boldsymbol{t}_2^2 \otimes \boldsymbol{t}_1^2)\big)^2$$

subject to $|\boldsymbol{w}| = 1$. Among these six terms only the first one is found to have predictive information. The term $\boldsymbol{t}_2^2 \otimes \boldsymbol{t}_1^2$ is statistically significant, if it is measured by

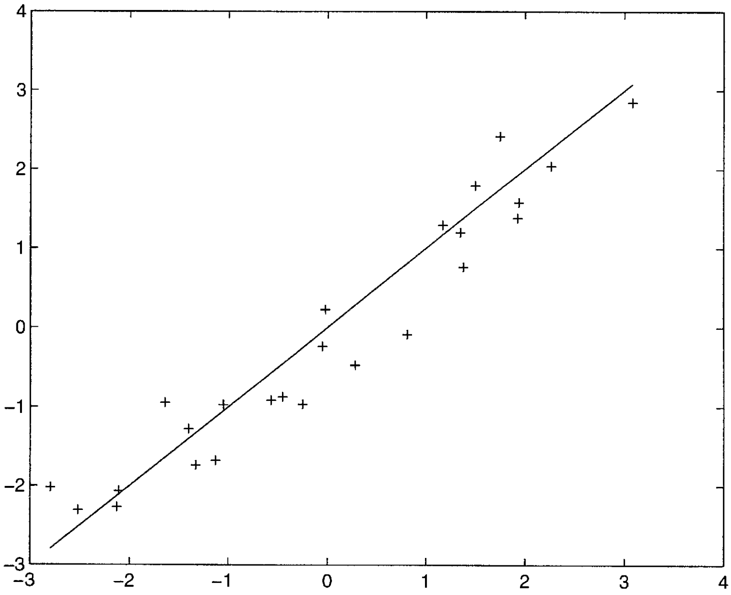

Fig. 3. The estimated response values based on $\boldsymbol{t}_1$ and $\boldsymbol{t}_1^2$ vs the response values.

a t-test. But it is so small that it cannot be recommended to use it, ($|\boldsymbol{t}_2^2 \otimes \boldsymbol{t}_1^2|^2 = 0.023$ or 0.007 percentage of the total size). Thus, only $\boldsymbol{t}^2$ should be used. The estimate response values based on $\boldsymbol{t}_1$, $\boldsymbol{t}_1^2$ and $\boldsymbol{t}_2$ against the response values are shown in Fig. 4.

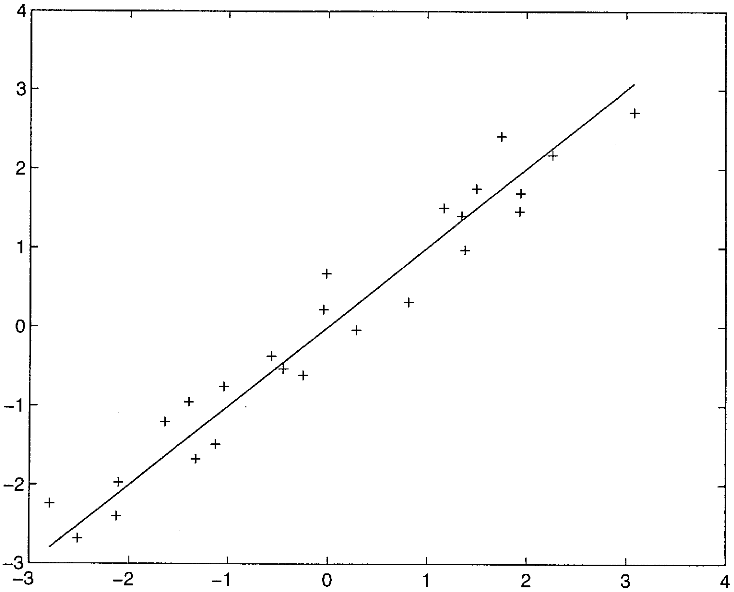

Fig. 4. The estimated response values based on $\boldsymbol{t}_1$, $\boldsymbol{t}_1^2$ and $\boldsymbol{t}_2$ vs the response values.

The following table shows the results of the modeling associated with Fig. 4:

Var t	X	ΣX R_X^2	y	Σy R^2	$(y't)^2$	$(t't)$	t-test
t_1	73.25	73.25	90.626	90.626	313.522	5.258	14.91
t_1^2	26.54	99.78	1.198	91.824	1.579	2.004	1.80
t_2	0.14	99.93	2.945	94.769	0.021	0.011	3.44

The measure of fit, R^2, obtained here, is 94.769%. It is possible to get a higher degree of fit. But this would not improve the prediction abilities of the model. The extra score variables are too small to be used.

Summary of the procedure. We have presented here a procedure to find a surface in a low dimension that adequately describes the values of the response variable. When the given data show a clear sign of curvature, like in the case of the present data, the linear modelling procedure is very sensitive to the given samples. If only two samples at one of the extremes (e.g., at large y-values) are excluded from the analysis, there are large changes in the regression coefficients. The present procedure suggests to use a quadratic surface in t_1, t_1^2 and t_2. The surface is a saddle or a valley, where the samples are geometrically situated. This surface is considerably more robust than the linear plane found by, e.g., stepwise regression. The quadratic surface changes only very little if, say, two samples are excluded from the analysis. The quadratic surface appropriately models the curvature found in data, making the predictions much more reliable than in the linear model.

7. Statistical Process Control

In today's industry there is a great interest in Statistical Process Control (SPC). There are several reasons for the increased interest in this subject. The main reasons are perhaps the following:

(a) *Large amount of data values.* Many companies make a large amount of regular measurements on the processes involved. Although the main purpose of the measurements has been to check that the processes are within certain limits, there is a need for more advanced or elaborate control of the development of the data values.

(b) *Technology.* It is cheap to install new measurement points at different places in the production process. Earlier it did cost a lot of money to arrange the collecting of data such that they could be viewed in one 'control room.' Now this can be arranged without any costs or insignificant ones. The location can even be distant from where the data are collected. The engineer can also in his office 'logon' the console, where an operator is supervising his process.

(c) *Prevention.* Usually there are indicators of a breakdown and similar events before the event actually happens. The situation is often similar to that before earthquakes: at some time before it actually happens, we receive data that indicate what actually will happen. This aspect of prevention has caused increased attention paid to new types of measurements like, e.g., acoustic measurements.

Implementation of SPC in industrial environments is usually not simple. There are many reasons for this. One of the most difficult problem is simply to find variables in the production process that are important for the final quality. Furthermore, the manufacturing processes are driven by inertial elements that give high correlations both over time and among the variables. There may be, say, 50 variables that should be simultaneously supervised, but the rank needed may be ten or less. This means that supervision should be carried out by ten or fewer score variables.

Here we shall summarize the author's experience in SPC.

8. Chart of Means

Choice of centering. When setting up a control chart for the means, there are certain choices to be made concerning the centering value. Examples are

1. target value required from the measurements,
2. average sample values in the data,
3. the median values in the data.

We use 2 if the data values are normally distributed and 3 otherwise. When decomposing the data as $\boldsymbol{X} = \boldsymbol{TP}'$, we may have wanted to subtract one of these three choices from the data before the analysis. The results will in principle depend on our choices of this preprocessing the data, although differences are usually small.

Individual Shewart charts. When carrying out a control chart analysis on the score variables, we require them to be orthogonal. This means that the columns of $\boldsymbol{T}$ are orthogonal. The variance of a column i in $\boldsymbol{T}$ is $s_i^2 = (\boldsymbol{t}_i'\boldsymbol{t}_i)/(N-1)$. If there are only two or three score variables, we can draw the plots in two or three dimensions. MacGregor and his associates (Kourti *et al.*, 1995; Nomikos and MacGregor, 1995) have succeeded in presenting the scores as three-dimensional plots containing the score values and the region of normal variation.

Hotelling $\boldsymbol{T}^2$ control chart. There has been some interest in combining all variables into one plot. One such plot can be based on the T^2 statistic. It is given by $T^2 = n\boldsymbol{x}_j'\boldsymbol{S}^{-1}\boldsymbol{x}_j = n|\boldsymbol{R}'\boldsymbol{x}_j|^2 = n\Sigma t_i^2$. Here (t_i) is the score vector that corresponds to $\boldsymbol{x}_j$. There are a few things that we should be aware of when using the Hotelling T^2 statistic. Normally we should not use all score variables in the equation. If we have many score vectors, we should be careful when using the Hotelling T^2 statistic. The reason is that we are adding the squared deviations for each score vector. There can be significant variation in one score vector, which disappears when working with the sum in Hotelling T^2. Conversely, although less frequently, Hotelling T^2 can be significant but the individual terms (the separate score vectors) are not significant.

We should not work with Hotelling T^2 except the situation where there are only few score vectors, say, four or fewer.

CUSUM technique. CUmulative SUM technique is popular to detect small changes on the level of the process. The change in the level is often measured in relation to the standard deviation. Let σ be the standard deviation and suppose that the change in the process is from μ to $\mu + a \times \sigma$. This means that the measurements (x_i) change from $\mu + \epsilon$ to $\mu + a \times \sigma + \epsilon$. The cumulative sums $S_i = \Sigma(x_j - \mu)$ change as follows:

$$\begin{array}{ll} S_1 = \epsilon_1 \simeq \epsilon, & S_1 \simeq a \times \sigma + \epsilon \\ S_2 = \epsilon_1 + \epsilon_2 \simeq \epsilon, & S_2 \simeq 2a \times \sigma + \epsilon \\ S_3 = \epsilon_1 + \epsilon_2 + \epsilon_3 \simeq \epsilon, & S_3 \simeq 3a \times \sigma + \epsilon \end{array}$$

This shows that, if there is a change in the level of the process, the points in a (i, S_i) plot will follow a straight line with slope equal to $a \times \sigma$. In a visual inspection there should be at least five points falling on a clear straight line in order that we can say that there is a change in level. In books and manuals significance testing is usually based on nomograms or 'V-angle.' We have found it equally efficient to base the testing on the simple measures of the normal distribution. The following table gives the mean value and variance of the cumulative sums involving powers of (x_i):

Sum	Mean value	Variance	Mean value
	$E(x_i) = 0$		$E(x_i) = \mu_i$
$S_1 = \Sigma_1^n x_i$	$E(S_i) = 0$	$\mathrm{Var}(S_1) = n\sigma^2$	$E(S_1) = \Sigma_1^n \mu_i$
$S_2 = \Sigma_1^n x_i^2$	$E(S_2) = n\sigma^2$	$\mathrm{Var}(S_2) = 2n\sigma^4$	$E(S_2) = n\sigma^2 + \Sigma_1^n \mu^2$
$S_3 = \Sigma_1^n x_i^3$	$E(S_3) = 0$	$\mathrm{Var}(S_3) = 15n\sigma^6$	$E(S_3) = 3\sigma^2 \Sigma_1^n \mu_i + \Sigma_1^n \mu_i^3$
$S_4 = \Sigma_1^n x_i^4$	$E(S_4) = 3n\sigma_4$	$\mathrm{Var}(S_4) = 96n\sigma^8$	$E(S_4) = 3n\sigma^4 + 6\sigma^2 \Sigma_1^n \mu_i^2 + \Sigma_1^n \mu_i^4$

For example, a 99 percentage interval for values of S_i would be given by $E(S_i) \pm 2.576 \times \sqrt{(}\mathrm{Var}(S_i))$ for i=1,2,3 and 4. Normally, we use the cumulative sum S_1. In case the data values follow the distribution the use of higher moments may be efficient. If a random variable x follows the normal distribution with zero mean and variance σ^2, the third moment is zero, $E(x^3) = 0$ and the fourth moment is given by $E(x^4) = 3\sigma^4$. This property of the normal distribution, i.e., $E(x) = 0, E(x^2) = \sigma^2, E(x^3) = 0, E(x^4) = 3\sigma^4$, does not characterize the distribution theoretically. But the Edgeworth expansion shows that the distribution is normal within the order of $O(1/N)$, where N is the number of samples. For a sample of data these properties can be tested by some measures of skewness and kurtosis. If these measures are not significant, this usually indicates that the sample can be described by the normal distribution. In these cases the use of higher moments can be efficient.

In practice, we are often interested in detecting as soon as possible a change in the level of the process. In these cases it is useful to apply these measures with the last values first. We say that we take the *backward* cumulative sums. The *forward* cumulative sums detect the change in the level in the beginning of the process and the backward cumulative sums detect it at the end of the process. For an application, see (Höskuldsson, 1996).

9. Process Variability

It is also important to control the variability of the process. It is summarized by the covariance matrix $\boldsymbol{\Sigma}$. In the following, we shall suppose that the scaling of the decomposition of $\boldsymbol{X}$ has been done such that $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}'$ and the columns of $\boldsymbol{T}$ have unit length. Then $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{P}\boldsymbol{P}'$ and $\boldsymbol{R} = (\boldsymbol{P}^{-1})'$. An important measure is the result of the likelihood ratio test that the covariance matrices $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ are equal,

$$W = \text{const} - n\log\left(|\boldsymbol{A}|/|\boldsymbol{\Sigma}|\right) + \text{tr}\,(\boldsymbol{A}\boldsymbol{\Sigma}^{-1}) \tag{13}$$

Here $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X}$. For details, see (Siotani *et al.*, 1985). In practice, this function requires full rank. Instead, it is better to work with the *J-divergence*,

$$J = \frac{1}{2}\text{tr}\,(\boldsymbol{A}\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}\boldsymbol{A}^{-1} - 2\boldsymbol{I})$$

The interpretation of J-divergence is the same. J follows approximately a χ^2 distribution. If the computed value of J is not significant, we cannot distinguish between the normal distributions having $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$ as their covariance matrices. As the formula states, we also need full rank of $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$. But it can be rewritten so that it can be used for the reduced rank case. Let $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{P}'$ be the covariance matrix for the whole process and $\boldsymbol{A} = \boldsymbol{P}_i\boldsymbol{P}_i'$ for a subprocess. Then we estimate the inverse by $\boldsymbol{\Sigma}^{-1} = \boldsymbol{R}\boldsymbol{R}'$ and $\boldsymbol{A}^{-1} = \boldsymbol{R}_i\boldsymbol{R}_i'$. The J-divergence becomes

$$J_i = \frac{1}{2}\text{tr}\left((\boldsymbol{R}'\boldsymbol{P}_i)(\boldsymbol{R}'\boldsymbol{P}_i)' + (\boldsymbol{R}_i'\boldsymbol{P})(\boldsymbol{R}_i'\boldsymbol{P})' - 2\boldsymbol{I}_a\right) \tag{14}$$

Here $\boldsymbol{I}_a$ is the $a \times a$ identity matrix and we are working with a columns in $\boldsymbol{P}, \boldsymbol{P}_i, \boldsymbol{R}$ and $\boldsymbol{R}_i$. The idea is that both $\boldsymbol{R}'\boldsymbol{P}_i$ and $\boldsymbol{R}_i'\boldsymbol{P}$ are approximately the unit matrix if there is a normal variability in the process. In the application we plot J_i against i, where J_i is based on the subprocess we are studying.

10. Out of Control Signal

In the application of multivariate control charts there is a practical problem of interpretation of out of control signals. The cause of this signal can be an outlier in one of the samples, change in the level of one or more variables, change in the correlation between some variables or some other things. The problem can be difficult to solve. But there are several things that can be done to handle the situation. We shall discuss some of them.

Selection of variables. An important issue is to select the variables that should be used in the multivariate control charts. Before the analysis we may decide that only significant variables should be used. This can be achieved by selecting *principal variables.* This is done in the basic algorithm above by choosing $\boldsymbol{w}_a = (0,0,\ldots,0,1,0,\ldots)$ where the index of 1 corresponds to the variable that maximizes $\boldsymbol{x}_i'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{x}_i$ for $i = 1,\ldots,K$. This variable is the one that has the largest predictive abilities among the K variables. Although this procedure is good in selecting variables, it has some disadvantages. When one variable has been selected, $\boldsymbol{X}$ is adjusted for the selected variable. This has the effects that we usually only select few variables. In PLS regression, it can be advantageous to have somewhat more variables than the ones selected as principal variables.

Revision of variables. When the decomposition has been carried out and we have studied the plots, we sometimes see that some variables should be removed. The procedure that can be recommended is to sort the variables according to the size of $\boldsymbol{x}_i'\boldsymbol{T}\boldsymbol{T}'\boldsymbol{x}_i$ for $i = 1,\ldots,K$. Then we cut-off, when a variable and the remaining ones are not contributing significantly to $\boldsymbol{T}$, i.e. when $\boldsymbol{x}_i$ and $\boldsymbol{T}$ can be considered as mutually orthogonal. In a simulation experiment this was the only measure that could correctly identify all the x-variables that were significant for the PLS regression (Höskuldsson, 1996, Sec. 11.7). This procedure gives more variables than the principal variable procedure. This is often favorable for PLS regression, because it stabilizes the regression.

Simplification of the loading matrix. The loading matrix **P** shows how we add the score vectors to give the observed measurements. When performing process control we would like as many coefficients to be zero as possible. It is not a good procedure just to make some coefficients equal to zero, because the different coefficients are correlated. We shall briefly describe a procedure that can be used to identify the coefficients that can be set to zero and estimate the non-zero coefficients. The starting point is the J-divergence, $J = \mathrm{tr}\,(\boldsymbol{A}\Sigma^{-1} + \Sigma\boldsymbol{A}^{-1} - 2\boldsymbol{I})/2$. Suppose that $\boldsymbol{A}$ depends on some parameter, θ. If we differentiate J with respect to θ, we get

$$\frac{\partial J}{\partial \theta} = \frac{1}{2}\mathrm{tr}\left((\Sigma^{-1} - \boldsymbol{A}^{-1}\Sigma\boldsymbol{A}^{-1})\frac{\partial \boldsymbol{A}}{\partial \theta}\right) \tag{15}$$

It is interesting to note that the differential of the log-likelihood function is a similar expression with the first term Σ^{-1} replaced by $\boldsymbol{A}^{-1}$. Since $\boldsymbol{A}$ is approaching Σ and are equal at convergence, we see that J-divergence is an approximation to the log-likelihood function. In the present procedure Σ is the covariance matrix associated with the loading matrix, $\Sigma = \boldsymbol{P}\boldsymbol{P}'$. $\boldsymbol{A}$ is the approximating covariance matrix, $\boldsymbol{A} = \boldsymbol{P}_i\boldsymbol{P}_i'$. When finding the value of the next $p_{i,j}$, we need the derivative of $\boldsymbol{A}$ with respect to $p_{i,j}$. It is given by

$$\frac{\partial \boldsymbol{A}}{\partial p_{i,j}} = \boldsymbol{e}_i\boldsymbol{p}_j' + \boldsymbol{p}_j\boldsymbol{e}_i' \tag{16}$$

with $e_i = (0,\ldots,1,\ldots)$ and p_j the j-th column of P. The procedure for finding the non-zero values in P and estimate their values is as follows:

0. Initialize P_0 by finding the largest values in each column of P. These numerically largest values are inserted in P_0. The values found must be in different rows of P. Other values in P_0 are zeros.
1. Compute the inverse $R_i = P_i(P_i'P_i)^{-1}$ that is a linear combination of columns of $\mathrm{P_i}$ and satisfies $R_i'P_i = I$.
2. Use the formulae (15) and (16) to compute the estimates of non-zero elements in P_i. This is done by equating (15) to zero and inserting (16) for the derivative in (15).
3. Find the next candidate element in P to be non-zero.
4. Use the J-divergence to evaluate if it pays to work with more non-zero elements. If it does, start again at 1.

In this procedure, the J-divergence goes to zero, because when all elements in P have been selected the matrices A and Σ are equal. We stop when the value of J is well within the range given by the χ^2 distribution. Typically, we find that around two thirds of the elements in the loading matrix P can be considered zero. This simplifies considerably the interpretation of out-of-control signals. It gives relatively few variables to study when we get values of a score vector that are not conforming to the model. If the number of columns of P is much smaller than the number of rows, it may be necessary to expand P to full rank by adding columns to P. Note that the derivative of A with respect to $p_{i,j}$ depends only on p_j. Thus it does not depend on an extension of P. Furthermore, there must be given a certain procedure to select the non-zero element of P, e.g., according to the size of the elements.

Appendix

We shall briefly treat the mathematics of finding surfaces of lower dimensions. At the a-th step we want to find w and compute the score vector as $t_a = Xw$ such that the score vector, the powers of the score vector and their product terms with previously selected terms are in a certain sense optimal. In prediction the basic measure is the covariance, see (Höskuldsson, 1996). Therefore we want to maximize the squared covariance involving these t-variables, i.e., to maximize

$$\begin{aligned} f(w) &= (y't_a)^2 + (y'(t_a \otimes t_b))^2 + (y'(t_a \otimes t_c))^2 + \cdots \\ &\quad + (y't_a^2)^2 + (y'(t_a^2 \otimes t_b))^2 + (y'(t_a^2 \otimes t_c))^2 + \cdots \\ &\quad + (y't_a^3)^2 + (y'(t_a^3 \otimes t_b))^2 + (y'(t_a^3 \otimes t_c))^2 + \cdots \end{aligned}$$

There appears to be many terms. But if the previously selected ones are collected in a matrix H, the new terms are

$$t_a \quad H \otimes t_a \quad t_a^2 \quad H \otimes t_a^2 \quad t_a^3 \quad H \otimes t_a^3 \ \ldots \tag{17}$$

We want the squared covariance of the response variables with these terms to be as large as possible. The algorithm works with several y-variables at the same time. If the response variables are selected in $\boldsymbol{Y} = (y_{ij})$, each term in $f(\boldsymbol{w})$ has the form $|\boldsymbol{Y}'(\boldsymbol{t}_b^j \otimes \boldsymbol{t}_a^n)|^2$. At the a-th step we know $\boldsymbol{Y}$ and $\boldsymbol{t}_b^j$. Denote by $\boldsymbol{Z}$ the product $\boldsymbol{Z} = \boldsymbol{Y}'\boldsymbol{t}_b^j$. Then a typical term in $f(\boldsymbol{w})$ is $|\boldsymbol{Z} \otimes \boldsymbol{t}_a^n|^2$. In order to maximize $f(\boldsymbol{w})$ we need the derivative of $f(\boldsymbol{w})$. The derivative of $|\boldsymbol{Z} \otimes \boldsymbol{t}_a^n|^2$ with respect to $\boldsymbol{w}$ is

$$\frac{\partial |\boldsymbol{Z} \otimes \boldsymbol{t}_a^n|^2}{\partial \boldsymbol{w}} = 2n\boldsymbol{X}'\boldsymbol{V}\boldsymbol{Z}'\boldsymbol{t}_a^n \tag{18}$$

where $\boldsymbol{V} = \boldsymbol{Z} \otimes \boldsymbol{t}_a^{n-1}$. We want to maximize $f(\boldsymbol{w})$ subject to $|\boldsymbol{w}| = 1$. Using the Lagrange multiplier technique we want to calculate

$$\max\left[f(\boldsymbol{w}) + \lambda(\boldsymbol{w}'\boldsymbol{w} - 1)\right] \tag{19}$$

The solution is

$$\boldsymbol{w} = \frac{f'(\boldsymbol{w})}{\left(\boldsymbol{w}'f'(\boldsymbol{w})\right)} \tag{20}$$

This equation is solved iteratively. A good initial guess for $\boldsymbol{w}$ is when only $\boldsymbol{t}_a$ is used. In this case $\boldsymbol{w}$ is found by maximizing $|\boldsymbol{Y}'\boldsymbol{t}_a|^2$, which is solved by the eigenvalue problem (Höskuldsson, 1998a)

$$\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{w} = \lambda\boldsymbol{w} \tag{21}$$

The convergence of $\boldsymbol{w}$ is fast, usually within ten to twenty iterations. Although the mathematics may seem a little complicated, they may be so organised that they only fill some half a page of a matrix code. MATLAB code for some of the methods presented here and further illustrations is available at the Internet address http://www.predict.dk.

References

Höskuldsson A. (1992): *QPLS regression methods.* — J. Chemometrics, Vol.6, No.6, pp.307–334.

Höskuldsson A. (1994): *Data analysis, matrix decompositions, and generalized inverse.* — SIAM J. Sci. Comp. 15, pp.239–262.

Höskuldsson A. (1996): *Prediction Methods in Science and Technology. Vol 1. Basic Theory.* — Copenhagen: Thor Publishing.

Höskuldsson A. (1998a): *Prediction Methods in Science and Technology. Vol 2. Non-Linear Methods.* — Copenhagen: Thor Publishing.

Höskuldsson A. (1998b): *Prediction Methods in Science and Technology. Vol 3. Analysis of Time Series and Process Data.* — Copenhagen: Thor Publishing.

Kourti T., Nomikos P. and MacGregor J.F. (1995): *Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS.* — J. Process Contr., Vol.5, No.5, pp.277–284.

Nomikos P. and MacGregor J.F. (1995): *Multivariate SPC charts for monitoring batch processes.* — Technometrics, Vol.37, No.1, pp.41–59.

Siotani M., Hayakawa T. and Fujikoshi Y. (1985): *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook.* — Columbus Ohio: American Science Press.