1 General linear models

In many procedures of applied sciences the starting point is a positive (semi) definite matrix **S**. It might be derived from **X** as $\mathbf{S}=\mathbf{X}^T\mathbf{X}$. It can also be derived in many other ways. Let us consider an example.

1^0 Choose a weight vector **w**
2^0 Compute
  score vector: $\mathbf{t}=\mathbf{Xw}$
  loading vector: $\mathbf{p}=\mathbf{Sw}$
  scaling constant: $d=1/(\mathbf{w}^T\mathbf{p})$
  Y-loading vector: $\mathbf{q}=\mathbf{Y}^T\mathbf{t}$
3^0 Adjust **S**, **X** and **Y**
  $\mathbf{S}\leftarrow\mathbf{S}-d\,\mathbf{p}\,\mathbf{p}^T$
  $\mathbf{X}\leftarrow\mathbf{X}-d\,\mathbf{t}\,\mathbf{p}^T$
  $\mathbf{Y}\leftarrow\mathbf{Y}-d\,\mathbf{t}\,\mathbf{q}^T$
4^0 Evaluate results:
  Check if the present set of vectors improves modeling. In this case go to 1^0. Otherwise stop the algorithm.

Box 1. Numerical steps in general linear regression.

In many engineering applications it is desired to formulate special estimation requirements in the optimisation task. An example is where the task is to minimize the expression $\mathrm{tr}(\mathbf{B}^T\mathbf{F}\mathbf{B}) + |\mathbf{Y} - \mathbf{X}\mathbf{B}|^2$. In this case it is desired that both the fit, $|\mathbf{Y} - \mathbf{X}\mathbf{B}|^2$, is small and also the relative size of the regression coefficients, **B**, measured by $\mathbf{B}^T\mathbf{F}\mathbf{B}$. The matrix **F** is a weight matrix that reflects the weights or importance of the regression coefficients. If the expression is differentiated with respect to **B** and the result is set to zero, the result is $\mathbf{F}\mathbf{B}+\mathbf{X}^T\mathbf{X}\mathbf{B} - \mathbf{X}^T\mathbf{Y}=\mathbf{0}$. The solution with respect to **B** is $\mathbf{B}=(\mathbf{X}^T\mathbf{X}+\mathbf{F})^{-1}\mathbf{X}^T\mathbf{Y}=\mathbf{S}^{-1}\mathbf{X}^T\mathbf{Y}$, with $\mathbf{S}=\mathbf{X}^T\mathbf{X}+\mathbf{F}$. Some methods of Kalman Filtering and some other methods in process control can be formulated in this way.

Other types of models can also result in this set of equations. E.g., in the analysis of variance components in the theory of experimental design, it also appears where **F** refers to the variance components part.

Box 1 shows how this type of equations can be solved. There is no restriction on the weight vector **w** except that the resulting loading vector **p** may not be zero, $|\mathbf{p}|\neq 0$. If $\mathbf{S}\neq\mathbf{X}^T\mathbf{X}$, the score vectors will not be orthogonal. For further properties of this algorithm see Ref 1.

In the analysis there are needed the *loading weight* vectors, **r**'s. They are defined from the requirement that $\mathbf{p}_a=\mathbf{S}\mathbf{v}_a$, where **S** is the original **S**-matrix. In Ref 2 it is shown that they can be computed by the equations

$$(1) \qquad \mathbf{r}_a= \mathbf{w}_a - d_1\mathbf{r}_1(\mathbf{p}_1^T\mathbf{w}_a) - \ldots - d_{a-1}\mathbf{r}_{a-1}(\mathbf{p}_{a-1}^T\mathbf{w}_a), \qquad a=1,...,A. \ (\mathbf{r}_1=\mathbf{w}_1)$$

Furthermore, it is shown that collecting the vectors in a matrix, $\mathbf{R}_A=(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_A)$, and similarly for $\mathbf{P}_A$, $\mathbf{W}_A$ and $\mathbf{D}_A$, the equation (1) can be written as ($(\mathbf{p}_a^T\mathbf{w}_a)=1/d_a$)

$$(2) \qquad \mathbf{R}_A = \mathbf{W}_A\,(\mathbf{D}_A\mathbf{P}_A^T\mathbf{W}_A)^{-1}.$$

From this equation it follows that $\mathbf{P}_A^T\,\mathbf{R}_A = \mathbf{D}_A^{-1}$, or $\mathbf{p}_a^T\mathbf{r}_b=0$ for $b\neq a$, and $\mathbf{p}_a^T\mathbf{r}_a =1/d_a$. It is also shown that the score vectors also satisfy $\mathbf{t}_a=\mathbf{X}\mathbf{r}_a$, where here **X** is also the original **X**-matrix.

2 Decomposition of data

It is instructive to look closer at the decompositions that are derived by the algorithms presented. The results are expansions of the matrices as follows:

$$
\begin{array}{ll}
\mathbf{S} = & d_1 \mathbf{p}_1 \mathbf{p}_1^T + \ldots + d_A \mathbf{p}_A \mathbf{p}_A^T + \ldots + d_K \mathbf{p}_K \mathbf{p}_K^T = \mathbf{PDP}^T \\
\mathbf{S}^{-1} = & d_1 \mathbf{r}_1 \mathbf{r}_1^T + \ldots + d_A \mathbf{r}_A \mathbf{r}_A^T + \ldots + d_K \mathbf{r}_K \mathbf{r}_K^T = \mathbf{RDR}^T \\
\mathbf{X} = & d_1 \mathbf{t}_1 \mathbf{p}_1^T + \ldots + d_A \mathbf{t}_A \mathbf{p}_A^T + \ldots + d_K \mathbf{t}_K \mathbf{p}_K^T = \mathbf{TDP}^T \\
\mathbf{X}^T\mathbf{Y} = & d_1 \mathbf{p}_1 \mathbf{q}_1^T + \ldots + d_A \mathbf{p}_A \mathbf{q}_A^T + \ldots + d_K \mathbf{p}_K \mathbf{q}_K^T = \mathbf{PDQ}^T \\
\mathbf{B}=\mathbf{S}^{-1}\mathbf{X}^T\mathbf{Y} = & d_1 \mathbf{r}_1 \mathbf{q}_1^T + \ldots + d_A \mathbf{r}_A \mathbf{q}_A^T + \ldots + d_K \mathbf{r}_K \mathbf{q}_K^T = \mathbf{RDQ}^T \\
\hat{\mathbf{Y}}=\mathbf{XB} & d_1 \mathbf{t}_1 \mathbf{q}_1^T + \ldots + d_A \mathbf{t}_A \mathbf{q}_A^T + \ldots + d_K \mathbf{t}_K \mathbf{q}_K^T = \mathbf{TDQ}^T
\end{array}
$$

Here the vectors are collected in a matrix, e.g., $\mathbf{T}=(\mathbf{t}_1,\mathbf{t}_2,\ldots,\mathbf{t}_K)$. **D** is a diagonal matrix with d_a's in the diagonal. The decomposition of **S** is a rank one reduction, meaning that the rank of say, $\mathbf{S}_a$ is one less than that of $\mathbf{S}_{a-1}$. (Follows from $\mathbf{S}_a\mathbf{w}_a=\mathbf{0}$). For $\mathbf{S}=\mathbf{X}^T\mathbf{X}+\mathbf{F}$, with **F** positive semi-definite, algorithms can be seen as approximating the exact solution **B**. The decompositions look the same for any choice of the weight vectors, $\mathbf{w}_a$'s. The H-principle suggests that an optimal balance should be found between the improvement in fit, $d_A\mathbf{t}_A\mathbf{q}_A^T$, and worsening of the precision, $d_A\mathbf{v}_A\mathbf{v}_A^T$. The main motivation for this approach is the prediction aspect of the model.

Assuming normal distribution, the precision $(\mathbf{X}^T\mathbf{X})^{-1}$ and the residuals $[\mathbf{Y}^T(\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y}]$ are stochastically independent, hence both need to be modelled. 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. If only A terms are used, the matrix **S** is approximated by $\mathbf{S}_A=d_1 \mathbf{p}_1 \mathbf{p}_1^T + \ldots + d_a \mathbf{p}_a \mathbf{p}_a^T + \ldots + d_A \mathbf{p}_A \mathbf{p}_A^T$. Similarly $\mathbf{S}^{-1}$ is approximated by the first A term, $(\mathbf{S}^{-1})_A$. $(\mathbf{S}^{-1})_A$ is the generalized inverse of $\mathbf{S}_A$, $\mathbf{S}_A\,(\mathbf{S}^{-1})_A\,\mathbf{S}_A=\mathbf{S}_A$.

The way of working with this decomposition is similar to eigen value decomposition of **S** after some rescaling. Assume the decomposition is $\mathbf{S}=\mathbf{UEU}$, where **U** is orthonormal and **E** diagonal. Then **P** corresponds to (**EU**) and **D** to $\mathbf{E}^{-1}$. Thus **D** corresponds to the inverse of the eigen values. The reason for this way of scaling is the numerical precision. New vectors are computed like **p**, $\mathbf{p}=\mathbf{Sw}$, where **w** has length 1, i.e., as range of a unit vector. Thus numerical stability is secured, also for very large systems, although e.g., adjustments, step 3^0 above, may be numerically unstable, if one is not careful.

In the applied analysis the samples are measured, but analysis is based on the score values. We shall look closer at the connection between these two types of values. The sample data matrix is given by $\mathbf{X}=\mathbf{TDP}^T$. The i^{th} sample is the i^{th} row of **X**, $\mathbf{x}^i$. It is given by $\mathbf{x}^i = \mathbf{t}^i\,\mathbf{DP}^T$. Denoting $\mathbf{x}=(\mathbf{x}^i)^T$ and $\mathbf{t}=(\mathbf{t}^i)^T$, the relationship can be written as $\mathbf{x}=(\mathbf{PD})\mathbf{t}$. We also have $\mathbf{XR}=\mathbf{T}$, or $\mathbf{x}^i\mathbf{R}=\mathbf{t}^i$, which is written similarly as $\mathbf{t}=\mathbf{R}^T\mathbf{x}$. Thus there are given two sets of transformations:

Sample space to score space: $\mathbf{t}=\mathbf{R}^T\mathbf{x}$.
Score space to sample space: $\mathbf{x}=(\mathbf{PD})\mathbf{t}$.

If all components have been selected, A=K, these transformations are one-to-one. But typically only A<K components are selected. When a new sample $\mathbf{x}_0$ is available, the associated score values are computed by the transformation, $\mathbf{t}_0=\mathbf{R}^T\mathbf{x}_0$. The values of $\mathbf{t}_0$ can then be compared to the rows of **T** to see how these new score values are relatively to the

present ones. Similarly, if there is given a set of score values, $\mathbf{t}_0$, for a sample, the associated sample, $\mathbf{x}_0$, would be estimated by $\mathbf{x}_0=(\mathbf{PD})\mathbf{t}_0$. A transformation matrix **R** is computed for each sub-group of data and each weighing mode.

3 Graphic analysis of data

In the numerical computations there are computed four sets of vectors, $\mathbf{w}_a$, $\mathbf{p}_a$, $\mathbf{r}_a$, and $\mathbf{t}_a$, at each step. In the following it is described how one can 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 the analysis they are computed as shown above 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 one or more of them get generally 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 $\mathbf{t}_a=\mathbf{X}\mathbf{r}_a$. The score vectors define the latent structure. They show what has been used of **X** and how **Y** can be described. 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 **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 **S** is any positive definite matrix, a similar interpretation is used. Pair wise plots of the loading vectors show the correlation structure in data.
- $\mathbf{r}_a$, **the loading weight vector**. It is given by $\mathbf{p}_a=\mathbf{S}\mathbf{r}_a$. If $\mathbf{S}=\mathbf{X}^T\mathbf{X}$, then $\mathbf{t}_a=\mathbf{X}\mathbf{r}_a$. They show how $\mathbf{p}_a$ is derived from the correlations of the original X-variables. Since $\mathbf{S}_0=\mathbf{S}$, it follows that $\mathbf{r}_1=\mathbf{w}_1$. The loading weight vectors are studied in order to know how the original variables generate the latent structure.

Figure 2 shows that the vectors $\mathbf{w}_a$, $\mathbf{p}_a$, and $\mathbf{r}_a$ are of the same size. It also emphasizes that the score vectors $\mathbf{t}_a$ are used for describing both **X** and **Y**, although the primary purpose with the analysis is to describe **Y**. In the applied work much time is spent on analysing how the score vectors describe **X**. Besides the plots mentioned above it can be recommended to look at further plots to study the results of the analysis:

- **Observed versus computed Y-values**. The columns of **Y** are drawn against the corresponding columns of $\hat{\mathbf{Y}}=\mathbf{X}\mathbf{B}_A$. The graphs are supplied by different measures of how good a fit that 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 $\mathbf{w}_a=(0,0,\ldots,0,1,0,\ldots)$, where 1 corresponds to the variable selected. In this case the graphs are called 'Added variable plots', Ref 3.
- **Y-values against the residuals**. The residuals are given as $\mathbf{E}=\mathbf{Y}-\mathbf{X}\mathbf{B}_A$. If the plots of the columns of **Y** against the corresponding column of **E** show systematic variations, it indicates that the modeling 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.

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 positive definite matrix **S**.

4 Case study. Process data

Process data
The data that are considered here are process data. They are published in Ref 4. These are hourly measurements of 12 x-variables and a quality variable y. The process was measured over a period of 289 hours. Thus **X** is a 289×12 matrix and **y** a 289×1 vector. Before analysis the data are auto-scaled (centred and scaled to unit variance).

Principal Component Analysis, PCA
It is often useful to study data by PCA analysis. The solution from PCA can be obtained by choosing **Y**=**X** in the algorithm above. The weight vectors will be the eigen vectors associated with the eigen value system $\mathbf{X}^T\mathbf{X}\mathbf{w}=\lambda\mathbf{w}$. It can also be computed from the Singular Value Decomposition of **X**. It gives **X**=**AFC**, where **F** is a diagonal matrix, and **A** and **C** are orthonormal. In this case the score matrix can be computed as **T**=**AF**. Both **W**, **P** and **R** are proportional to **C** (i.e. a diagonal matrix times **C**). The first task is to look at the first four score vectors, the first four columns of **T**.

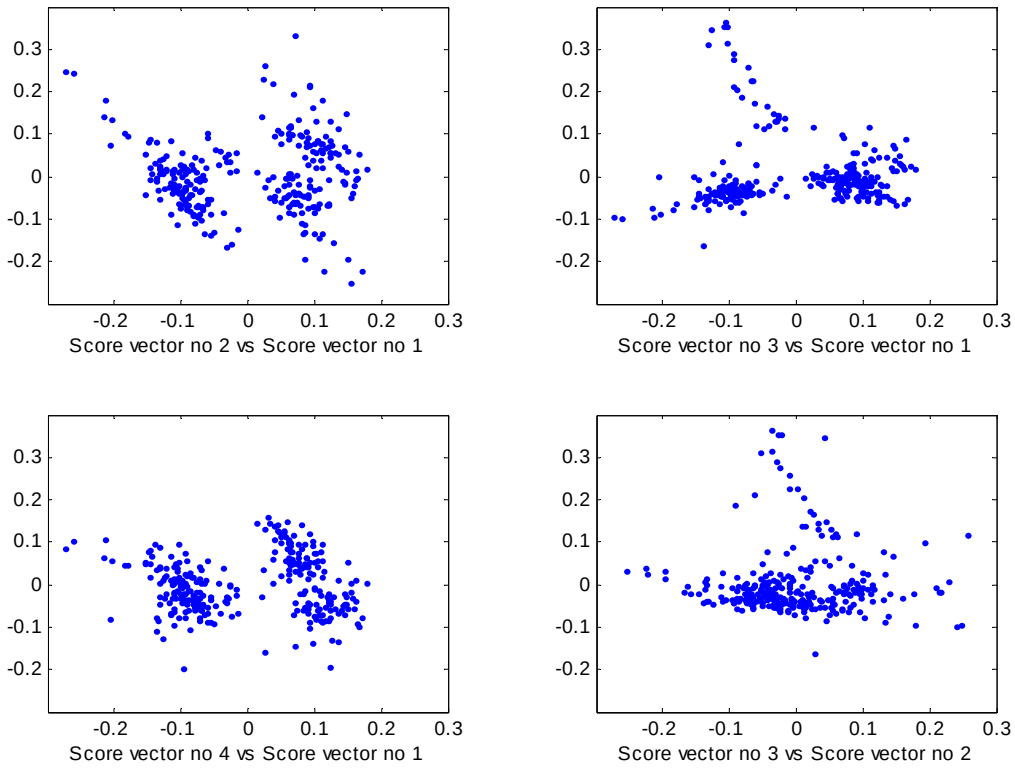

Figure 9. Pair-wise plot of the first four score vectors. Upper left no 2 vs 1, upper right no 3 vs 1, lower left no 4 vs 1 and lower right no 3 vs 2.

Three scatter plots in Figure 9 involving the first score vector show a clear sign of grouping in data. The grouping corresponds the first score vector is negative and positive. The lower right scatter plot shows that score vector no 3 has some special behaviour for values larger than 0.1. Score vector 1 and 3 are plotted in Figure 10 versus time. The plot of the first score vector shows that there is a clear change in the process variables around time 150. The values are positive before that time and negative later. The plot of the third score vector versus time shows that there has happened something at around time 150 until time 170. Furthermore, the score values are on average slightly smaller after time 170 than before time 150.

These changes in the process variables are not studied closer here. But an important issue is, if these features can be detected by the regression analysis. This is studied closer later.

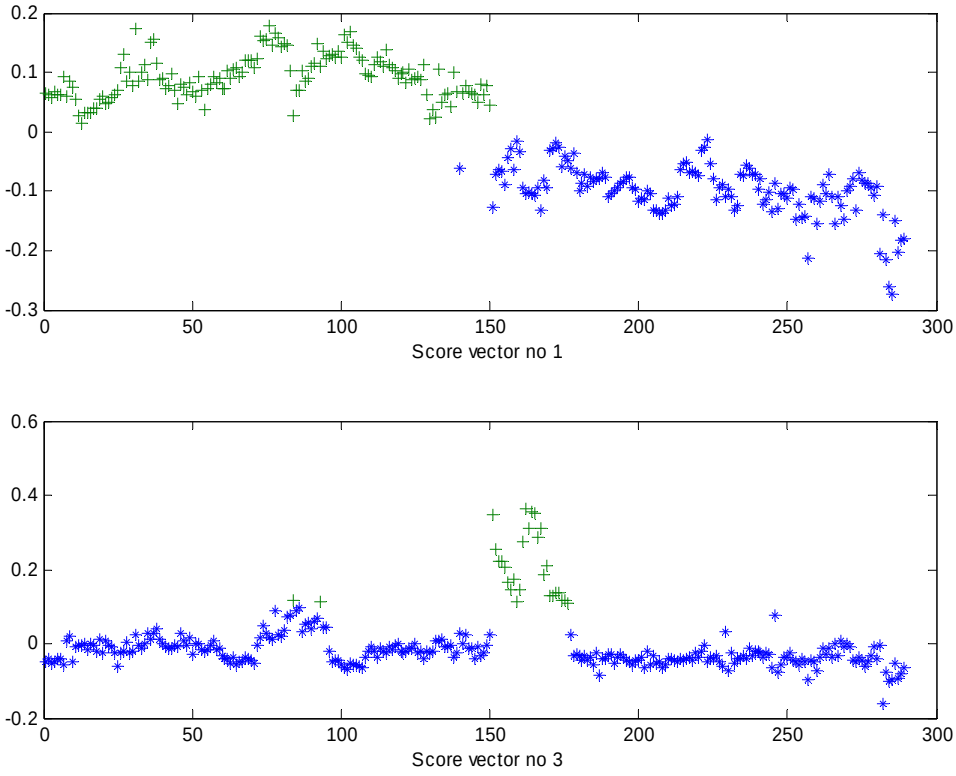

Figure 10. Upper figure plot of the first score vector vs time. Lower one the third score vector vs time, where values larger than 0.1 are marked as +.

Regression analysis

Figure 11 shows the scatter plots of the response variable versus the first four score vectors. A closer study shows that the first four score vectors are significant. The first four score vectors explain R^2=97.8% if the variation of the response variables. The score vectors account for 57.8% of the variation of **X**.

Figure 12 shows the scatter plot of the observed versus computed response variable, when four score vectors are used. A line through (0,0) with slope 45^o is drawn in the figure. It shows that the fit is fairly good, although there are some points that deviate from the line.

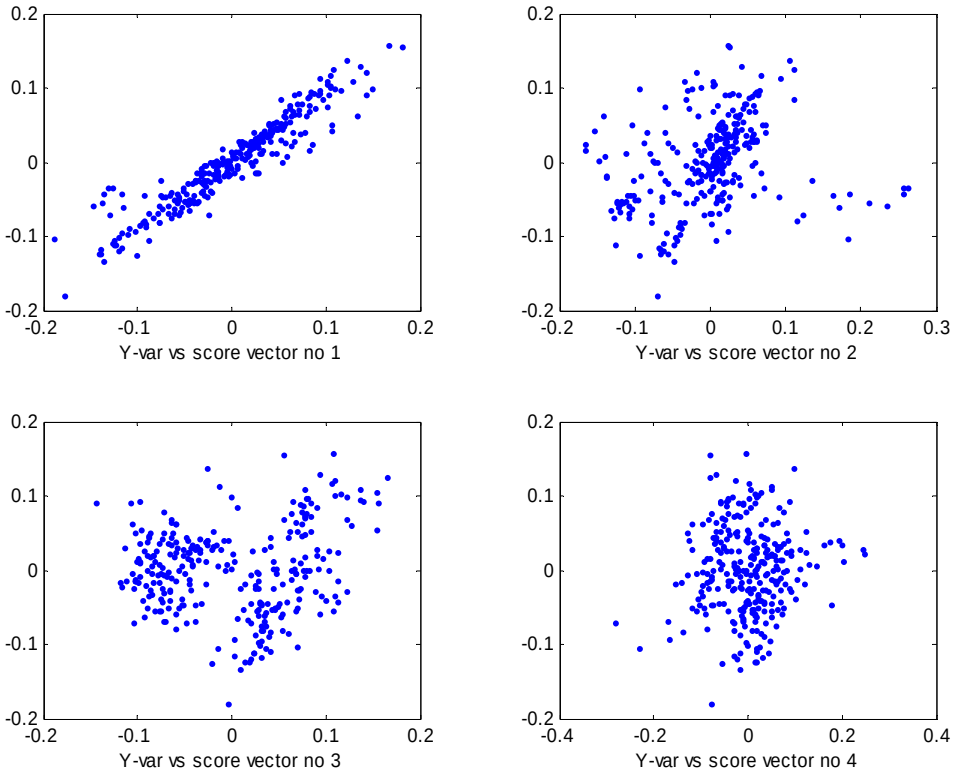

Figure 11. Plot of the values of the response variables against the first four score vectors.

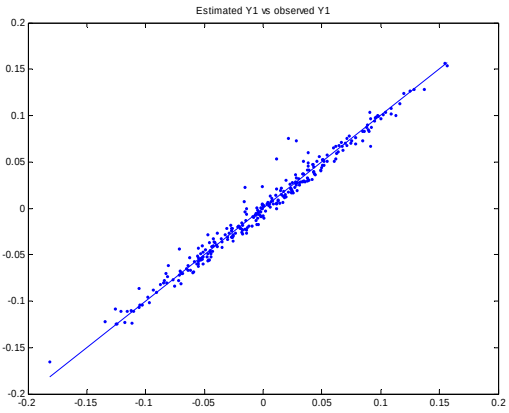

Figure 12. Plot of observed versus computed values of the response variable.

An interesting issue is concerning the scatter plot of score vectors. We know that there are some special features in the process data, and the question is if they appear in the scatter plot.

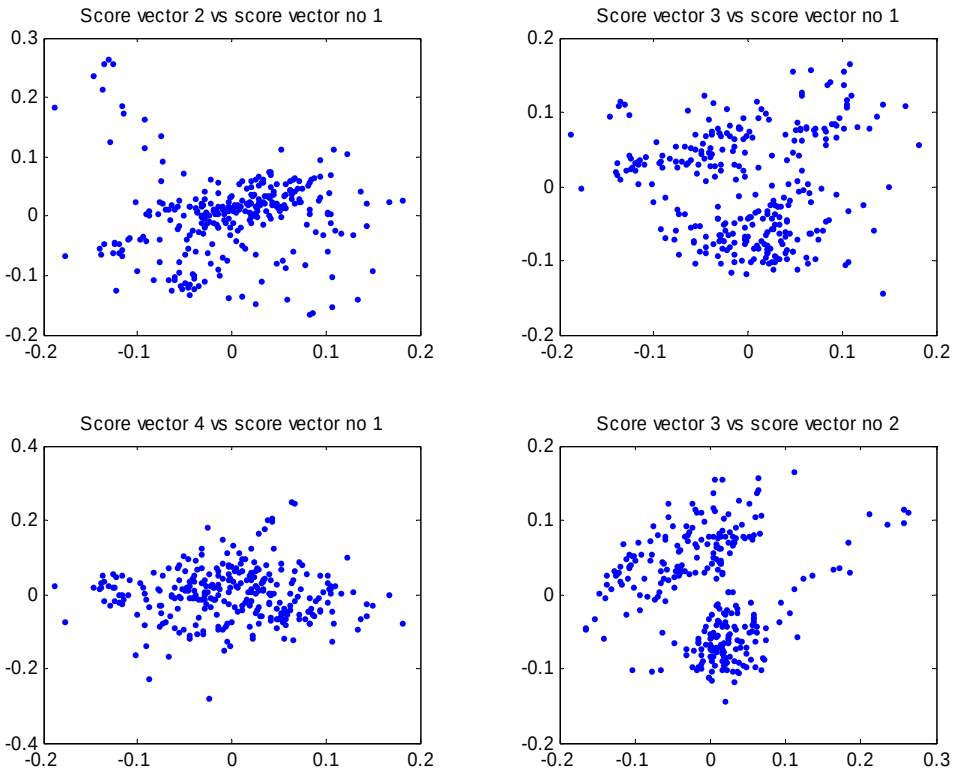

Figure 13. Pair-wise scatter plot of the first four score vectors. Upper left no 2 vs 1, upper right no 3 vs 2, lower left no 4 vs 1 and lower right no 3 vs 2.

A closer study of Figure 13 show that the grouping we find in the scatter plot of score vector no 3 versus no 2 corresponds to the change in the process that happens at time 150 hours. The points in the upper left corner of the scatter plot in the upper left part of the Figure 13 corresponds to the changes in the process between 150 and 170 hours. A plot of the first score vector alone will give a plot similar to the lower one in Figure 10. Thus the conclusion here is that the plots using the score vectors give the same type of information as the plots given by the PCA analysis. But the plots derived from the score vectors of the regression analysis more explicitly show the variation relevant for the response variable.

Ridge regression

It can be shown theoretically that the mean squared error of prediction can be reduced by allowing a small bias in the estimation. In practical terms it means that instead of working with the covariance matrix $\mathbf{S}=\mathbf{X}^T\mathbf{X}$, it can be advantageous to work with $\mathbf{S}=\mathbf{X}^T\mathbf{X}+k\mathbf{I}$, where $\mathbf{I}$ is the K×K identity matrix and k is a (small) positive constant. There exist a number of theoretically motivated formulas for computing k, but in most cases they are not appropriate.

They give generally too large value of k. Instead k is normally found by cross-validation, for instance by a leave-one-out cross-validation. This way of finding k is assumed here.

The application of the H-method can be formulated as follows.

- Find a weight vector **w** such that the resulting score vector **t**=**Xw** is good for describing **Y**. A balanced description is given by the covariance. Thus it is suggested to find **w** such that

  maximize $|\mathbf{Y}^T\mathbf{t}|^2$ = maximize $\mathbf{w}^T(\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X})\mathbf{w}$, subject to $|\mathbf{w}|=1$.

- When **w** has been found, adjust **S**, **X** and **Y** by the results found,

  $\mathbf{S} \leftarrow \mathbf{S} - d\,\mathbf{p}\,\mathbf{p}^T$, where $\mathbf{p}=\mathbf{S}\mathbf{w}$ and $d=1/(\mathbf{w}^T\mathbf{p})$
  $\mathbf{X} \leftarrow \mathbf{X} - d\,\mathbf{t}\,\mathbf{p}^T$, where $\mathbf{t}=\mathbf{X}\mathbf{w}$
  $\mathbf{Y} \leftarrow \mathbf{Y} - d\,\mathbf{t}\,\mathbf{q}^T$, where $\mathbf{q}=\mathbf{Y}^T\mathbf{t}$

  Note that **S** is reduced by rank one by this procedure. Furthermore, the score vectors ($\mathbf{t}_a$) will not be orthogonal. A new weight vector **w** is now found for the reduced matrices **X** and **Y**.

The weight vector **w** can be chosen in many other ways than suggested by the H-method. Here also the only restriction on **w** is that the resulting loading vector **p** may not be zero, $|\mathbf{p}|\neq 0$.

The value of k is found by leaving-one-out cross validation. The value k=0.0019 is the one that give the smallest value of $\sum(y_i-\hat{y}_{-i})^2$, where $\hat{y}_{-i}$ is the estimated response value associated with the i^{th} sample, when the i^{th} sample is not used in the estimation. The value of k is here rather small. We find this often in the case that the effective dimension is small compared to the number of variables. The first four score vectors explain R^2=97.6% of the variation of the response variable and 57.6% of **X**. Further score vectors do not improve the prediction derived from the model. Since the value of k is so small, there will be very little difference between the plots obtained from the regression analysis above. Therefore, the analogous plots are not shown here.

Traditionally, the full rank model is used for making predictions. Using all score vectors correspond to the full rank solution. The explained variation of that solution is R^2=98.3%, but it represents a severe overfitting. A consequence of the overfitting can or be judged by using that the precision term $\mathbf{x}_0{}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0$ can be viewed as having a distribution that is approximately proportional to a χ^2 distribution with degrees of freedom, f, equal to the number of variables/dimension in **X**. If four score vectors are used, the precision term has mean value proportional to around 4, while if 12 are used, it is close to 12, and the constant of proportionality is the same in both cases. Thus, a full rank solution will have *prediction variances that are around three times larger* compared with the ones obtained by the H-method.

One can perhaps say that this comparison is not quite fair, because the H-method builds up a solution by optimising the prediction and determines when to stop. There in no optimisation of prediction, when using the Ridge regression procedure, except for finding the Ridge constant k.