

CovProc methods and variable/subset selection

1. Variable selection

In applied sciences there is considerable interest in variable selection. I.e., to find the variables the ‘generate’ the observed values of the response variables. In theoretical statistics this has been intensively studied⁹⁻¹¹. In the popular program packages like SAS, SPSS, BMDP and others, programs to select variables are the most popular ones. Advanced procedures using ‘Backwards, Forward, Stepwise, Jackknife¹² and other criteria’ have been developed that assist the user in finding the variables.

Traditional approach to select variables is to base the selection on significance testing. That approach has some basic defects. There can be many variables that have effects on the responses, but when looking at them one at the time, they do not show significant contribution on say, 1% level. If on the other hand a latent structure associated with these variables is generated, the latent structure found can provide with an important contribution to explaining the responses. (It is easy to construct data, where the response variable is highly correlated to score variables associated with the smallest eigen values of the correlation matrix, but does not show high correlation to any of the individual variables). Another important shortcoming of traditional significance testing is that it is based on measures that are equivalent to correlation coefficients involving the score vectors (variables). The correlation coefficient is invariant to the size of the score vectors. But small score vectors may cause uncertain models, if they are used. Thus, significance testing may declare variables significant, but if they are used may spoil the modeling task. An example is shown below.

Ranking of variables like e.g., selection of variables according to their importance/significance in the regression analysis is one of the most important studied topics in statistics⁹. Many of the suggested methods are imprecise and data dependent. An example is the stepwise option in the regression analysis in the program packages. A new variable is selected that is the most important one and with this variable it is judged if some of the selected variables can be excluded, because of this new variable. The variables selected have been found significant, and it is not clear what is going on, when the variables are judged again. Another popular method is backward selection of variables, where the least significant variable is eliminated one by one. The risk is here that an important variable is eliminated because of correlations with many other variables. These kinds of search in data for the maximal values of the correlation coefficients are very data dependent. If these methods are used on the first half of the data, one set of result is obtained, while if used on the other half, the results may be very different.

Most of these methods may sound reasonable. Also, the questions that are asked are natural ones. E.g., if there are only a few variables, one can ask if some variables are ‘better’ than some others. On the other hand, when working with many variables, the variables, the \mathbf{X} -data, are replaced by a latent structure \mathbf{T} , and the further analysis is based on \mathbf{T} . There are two fundamental issues, when working with latent structures:

1. **Evaluation of a variable x_i .** We want to include variables, if they ‘improve’ the modeling task.

2. **Size of T.** We want the size of **T** to be as large as possible relative to **X**. The size of **T** is measured as $\text{tr}(\mathbf{T}^T\mathbf{T})=\sum(\mathbf{t}_a^T\mathbf{t}_a)$.

These issues are illustrated in the light of an example. Here a forward selection of variables shall be used. A popular criterion to use is to find the variable that gives largest increase in the explained variation of the response variable. It can be explained as follows. The projection of **y** onto **x_i** is given by $\mathbf{x}_i(\mathbf{y}^T\mathbf{x}_i)/(\mathbf{x}_i^T\mathbf{x}_i)$. The variation of a response variable is $\mathbf{y}^T\mathbf{y}$ and the reduction in the variation due to a variable **x_i** is

$$(1) \quad \mathbf{y}^T\mathbf{y} - (\mathbf{y}^T\mathbf{x}_i)^2/(\mathbf{x}_i^T\mathbf{x}_i) = \mathbf{y}^T\mathbf{y} (1 - (r_{yxi})^2) = \mathbf{y}^T\mathbf{y} - [\mathbf{y}^T\mathbf{y} (r_{yxi})^2],$$

where r_{yxi} is the simple correlation coefficient between **y** and **x_i**, assuming that original data have been centered. (1) shows that to find the variable that gives the largest explanation is equivalent to find the variable that has the largest correlation coefficient with the response variable. The equation (1) is used at each step of the selection of variables. In terms of weighing schemes $\mathbf{w}_a=(0,\dots,0,1,0,\dots)$ and $\mathbf{t}_a=\mathbf{x}_i$ in the reduced **X**-matrix, and **y** in (1) is the reduced one.

No	Var	y before	t _a = x _i	(y ^T x _i)	r _{yxi}	y ^T y (r _{yxi}) ² , %	Cum. Sum, %
1	38	1	1	0.9328	0.9328	87.0031	87.0031
2	152	0.361	0.534	0.1613	0.8381	9.1291	96.1322
3	218	0.197	0.317	0.0371	0.5961	1.3745	97.5066
4	132	0.158	0.184	0.0208	0.7153	1.2756	98.7822
5	31	0.110	0.112	0.0057	0.4649	0.2632	99.0455
6	144	0.098	0.089	0.0067	0.7700	0.5660	99.6115
7	29	0.062	0.037	0.0013	0.5466	0.1161	99.7276
8	16	0.052	0.062	0.0013	0.4042	0.0445	99.7721
9	33	0.048	0.034	0.0008	0.5117	0.0597	99.8318
10	24	0.041	0.034	0.0006	0.4085	0.0281	99.8598
11	614	0.037	0.854	0.0124	0.3890	0.0212	99.8810
12	15	0.034	0.030	0.0005	0.4840	0.0279	99.9089
13	117	0.030	0.077	0.0009	0.3986	0.0145	99.9234
14	543	0.028	0.096	0.0011	0.4334	0.0144	99.9378
15	553	0.025	0.374	0.0052	0.5568	0.0193	99.9571

Table 1. Results from selecting 15 variables from the 1056. FT-IR data. Y=y₃.

The results are to be interpreted as follows. Variable no. 38, $x=x_{38}$, is the one having the largest correlation coefficient with the response variable $y=y_3$, $r_{yx}=0.9328$, among all 1056 x-variables. When this variable has been selected, the **X** matrix is the reduced one. The y-variable is also reduced in similar way. The size of the reduced y-variable is computed as shown in (1). After choosing x_{38} the size is $|y|=(\mathbf{y}^T\mathbf{y})^{1/2}=0.361$. In the reduced matrix **X** the variable no. 152, x_{152} , has the largest correlation coefficient among the 1055 x-variables with the reduced y-variable of 0.8381. In the table there are shown the results from the first 15 x-variables or iterations. The 15 x-variables explain 99.9571% of the variation in the response variable.

This procedure seems intuitively appropriate and available as an option in most program packages in statistics. It is also a procedure that is recommended by many statisticians. This procedure shall be discussed in the light of the two basic issues mentioned above.

When the score vectors are orthogonal, like here, it is customary in the statistical analysis to base significance testing on the procedures that are equivalent to using the correlation coefficient as described above. The 1% significance level of the correlation coefficient with 45 samples is 0.372. Thus all 15 variables above would be judged significant. On the other hand the H-principle suggests that judging the variables should be based on the covariance, $(\mathbf{y}^T \mathbf{x}_i)$. From the table it can be seen that from step 7 and later the covariance is very small. If the situation is studied closer by cross-validation, we also find out that the variables from step 7 and later do not contribute to the prediction concerning new samples. The problem is that the correlation coefficient is invariant to the size of the score vector. Thus, at step no. 7 the correlation coefficient is 0.5466, but the associated score vector has the size $|\mathbf{t}_7|=0.037$. A score vector that is as small as this is very uncertain, because the measurement uncertainty of the scaled x-values is around ± 0.015 . If all variables are included in our model as suggested by the significance testing, an unstable model has been found, because the measurement errors have been allowed to influence our modeling results.

The other issue is concerning the size of the latent structure. If PLS regression is used with three components, we get $\text{tr}(\mathbf{T}^T \mathbf{T}) = \sum_a (\mathbf{t}_a^T \mathbf{t}_a) = (808.6803 + 53.5355 + 15.0727) = 877.2886$. For the case above with 15 x-variables we get $\text{tr}(\mathbf{T}^T \mathbf{T}) = 2.3315$. The total size of \mathbf{X} is $\text{tr}(\mathbf{X}^T \mathbf{X}) = 1056$. Thus the PLS regression accounts for 83.08% of the variation in \mathbf{X} and the procedure above only 0.22% of the variation. In industrial environments the PLS regression is relatively robust concerning small errors in the measurement values, while the procedure above is very sensitive of small errors in the 15 selected x-variables. It can be said that the latent structure generated by the procedure above may be good for describing \mathbf{y} , while it is 'shaky' concerning describing \mathbf{X} .

In conclusion it can be stated that significance testing does not select the variables according to the prediction aspect and it may lead to very 'small' latent structure.

2. CovProc methods

How can these two issues of modeling be combined, when variables are selected? We shall briefly review the new approach to ranking and choosing variables that is called the CovProc (*Covariance Procedures*) methods. It is based on ranking the variables according to their covariance and select ‘an optimal’ number of variables to use in the analysis. The H-principle suggests using $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$ as the measure of strength between \mathbf{X} and \mathbf{Y} . The CovProc suggests sorting the variables according to some weights derived from the matrix $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$ and judge the further analysis based on expanding \mathbf{X} according to the sorting obtained. Thus the steps of the CovProc methods are as follows.

For each step $a=1,2,\dots,A$

1. **Sorting the variables.** Sort a weight vector that has been derived from the matrix $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$, the largest first.
2. **Defining \mathbf{X} .** Let \mathbf{X} consist of the first k columns (variables) corresponding to the k largest values of the weight vector. This is carried out for $k=1,2, \dots, K$.
3. **Judge the results.** Apply a criterion to decide which \mathbf{X} should be used. Compute the score and loading vector based on the ‘optimal’ \mathbf{X} . The full \mathbf{X} is now adjusted for the score and loading vector found and similarly for \mathbf{Y} . a is set to $a+1$, and go to 1.

There are many ways to define the weight vector in 1. A natural choice is the eigen vector of $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$ corresponding to the leading eigen value. Another choice is the diagonal elements of $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$. In case of one response variable these two choices coincide. There are also many criteria that may be interesting to use, when judging the results that are obtained by a given score vector. In a way it is natural to judge the score vector from the point of view of a measure of fit. If the score vector \mathbf{t} is given by $\mathbf{t}_a=\mathbf{X}_{a-1}\mathbf{w}_a$, the size of fit for \mathbf{t} is given by $|\mathbf{Y}^T\mathbf{t}_a|^2/(\mathbf{t}_a^T\mathbf{t}_a)$. The ‘optimal’ of \mathbf{X} is found as the one that, for the associated score vector gives the highest value the measure of fit. It is natural to use a measure of fit, because in 1. the size of the score vector is secured. Thus, CovProc methods can be viewed as a two step methods, where at the first step large score vectors are found and at the second step their performance are judged.

A CovProc method for finding one score vector is illustrated by the MATLAB code in Box 1. To simplify the ideas only one y-variable is chosen in the code. In this case the weight vector is computed as $\mathbf{w}=\mathbf{X}^T\mathbf{y}$, which is sorted according to its numerical values with largest first. With this sorting \mathbf{X} is expanded and the measure of fit for \mathbf{t} is computed. When all possible measures of fit are computed, the index is found that gives the highest value of the fit measure. This is extracted from \mathbf{X} , \mathbf{X}_r , and the

corresponding weight and score vector are computed from the extracted \mathbf{X} .

These ideas are illustrated by an example using the paper mill data.

```
[N,K]=size(X);
w=X'*y;
[w1,I]=sort(-abs(w));
r2=zeros(K,1);
for i=1:K
    w=X(:,I(1:i))'*y;
    t=X(:,I(1:i))*w;
    r2(i)=(y'*t)^2/(t'*t);
end
[rmax,imax]=max(r2);
Xr=X(:,I(1:imax));
w=Xr'*y;
t=Xr*w;
```

Box 1. MATLAB code for one score vector. One y-variable

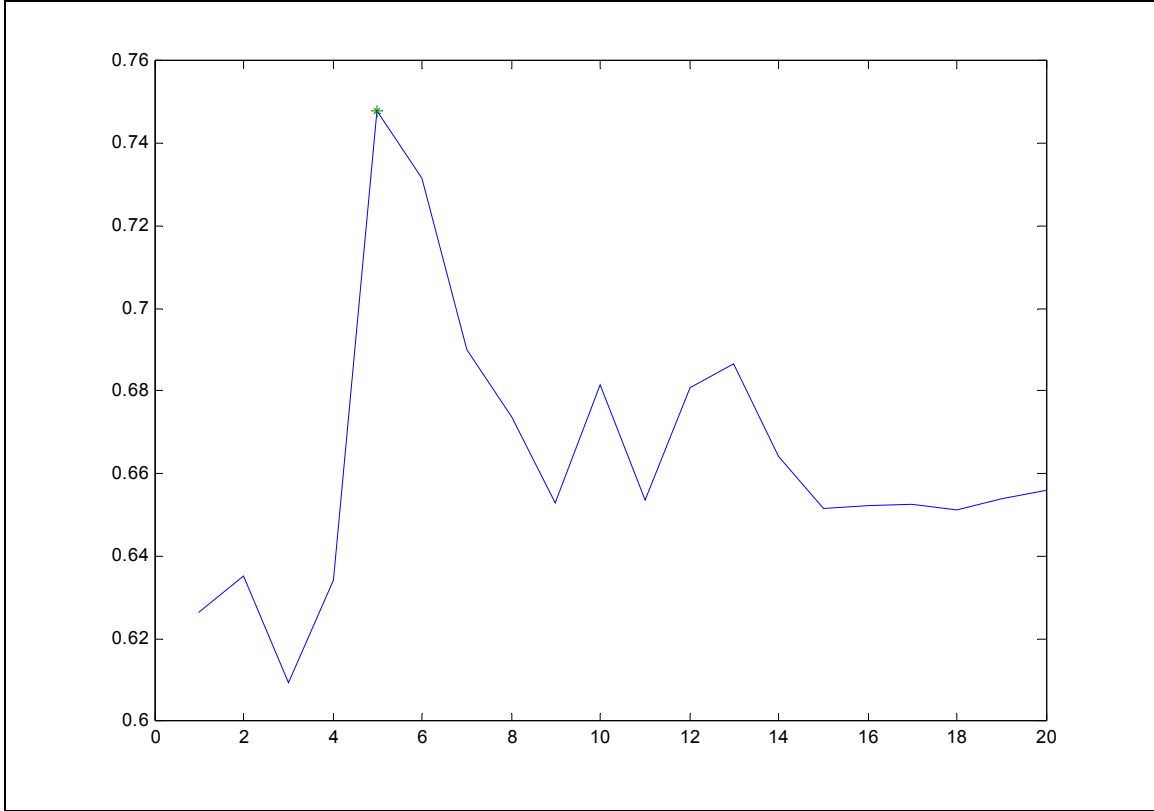


Figure 1. R^2 (r_2) values, explained variation, as function of the number of columns in X . Paper Mill data. $Y=y_4$. X-axis the variable number.

The code shown in Box 1 has been applied to the Paper Mill data, where the response variable is the fourth one. We see that the first score vector, \mathbf{t}_1 , is explaining most variation, given by $(\mathbf{y}^T \mathbf{t}_1)^2 / (\mathbf{t}_1^T \mathbf{t}_1)$, in the response variable, when \mathbf{X} consists of five columns, the ones corresponding to the five largest weights. If these five columns (variables) are used it explains 74.8% of the variation, while only 65.6% if all 20 variables are used.

It is an important issue that the measure of fit, $(\mathbf{y}^T \mathbf{t})^2 / (\mathbf{t}^T \mathbf{t})$, can change much depending on what part of the matrix \mathbf{X} that is being used. Let us look closer at the measure. Suppose that \mathbf{X} consists of two parts, $\mathbf{X}=(\mathbf{X}_1, \mathbf{X}_2)$, and $\mathbf{w}=(\mathbf{w}_1, \mathbf{w}_2)$. Then

$$(\mathbf{y}^T \mathbf{t})^2 = ((\mathbf{y}^T \mathbf{X}_1 \mathbf{w}_1) + (\mathbf{y}^T \mathbf{X}_2 \mathbf{w}_2))^2,$$

$$(\mathbf{t}^T \mathbf{t}) = \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_1 \mathbf{w}_1 + \mathbf{w}_2^T \mathbf{X}_2^T \mathbf{X}_1 \mathbf{w}_1 + \mathbf{w}_2^T \mathbf{X}_2^T \mathbf{X}_2 \mathbf{w}_2.$$

It may be that \mathbf{X}_1 provides with good fit, but \mathbf{X}_2 shows low covariance with \mathbf{y} . It means that $(\mathbf{y}^T \mathbf{X}_1 \mathbf{w}_1)$ is large but $(\mathbf{y}^T \mathbf{X}_2 \mathbf{w}_2)$ is small. But the value of $\mathbf{w}_2^T \mathbf{X}_2^T \mathbf{X}_2 \mathbf{w}_2$ can so large that it spoils the fit. In general, if large percentage of the weight values in \mathbf{w} is small, it is important to reduce \mathbf{X} by removing columns (variables) that have small weights.

If a variable has been left out in each of the A steps of computing the score vectors, the regression coefficient associated with the variable will be zero.

The above procedure can be viewed as analysis of the variables. The \mathbf{X} -matrix is expanded, $X(:,I(1:i))$, and the results judged by some measure. A similar procedure can be used for the samples (objects). In fact \mathbf{X} can be expanded in both ways, both according to variables and samples, $X(J(1:j),I(1:i))$. When judging the results by some measure, we will be looking at a surface instead of a curve like in Fig. 1. The procedure successfully identified both the variables and the past history of data that should be used, giving considerable improvements compared to standard methods.

Above the emphasis has been on the measure of fit. In on-line environments it may be more important to use some prediction measure instead in order to control the prediction ability of the model. Since such procedure is analogous, it is not considered here closer. The general result is that the prediction measures will suggest more variables to use than the fit measures.

Note that above only one score vector has been used to judge the results. The same set of variables can be chosen for all score vectors or a separate set of variables for each score vector. It depends on how we want to use the results.

The CovProc methods have some important advantages:

- They are based on very simple ideas of ranking the variables according to the size of the covariance they show.
- Different measures to evaluate the results can be applied, and we are sure to use variables that show covariance.
- For auto-scaled data we are sure that the score vectors obtained are large.
- In statistics the focus is on the measures of fit, which has the disadvantage that it is invariant to the size of the score vectors used. Here large score vectors are secured before the fit aspect of the model is considered.
- The methods can be used to select or exclude variables or samples according to some criteria that is important for the present situation.
- They are computationally fast, because the basic considerations are simple.
- They can be used for online selection of subsets of data that should be included in the modeling procedure.

3 Case study

In the following industrial data, the Oven data, are used. The y-variable is critical to the quality of the final product. 40 x-variables have been specified that are expected to influence the y-variable or to be of importance for accurate description of the processes. 50 samples were selected of the 40 x-variables and the corresponding y-variable. Data were auto-scaled (centered and scaled to unit variance) before analysis.

Standard PLS regression.

The first task is to carry out a standard PLS regression analysis. Table 2 shows the overall results of the analysis. In order to simplify the presentation, only three measures are shown. The column %X shows the cumulative percentage of how much of X is used for each dimension. The column %y shows the cumulative percentage of the variation of y that is explained. For each dimension a cross-validation was carried out. It consists of dividing the 50 samples into 10 segments. The regression analysis was carried out for the samples in 9 segments, i.e., 45 samples, and the results were used to estimate the response values in the 10th segment, the 5 response values. The cross-validation measure used is

$$Q = [\sum_{\text{segments}} (\sum_{\text{samples}} (y_i - \hat{y}_i)^2/5)/10]^{1/2}.$$

The value of Q is given for the original (un-scaled) data. It will tell us, how well on average we predict in terms of the standard deviation for the 5 samples. From the table it can be seen that cross-validation suggests 5 dimensions. If 5 dimensions are used (score vectors), 49.932% of X is used and 90.394% of y is explained.

No	%X	%y	Q
1	12.801	51.303	2.4326
2	22.815	74.308	2.2410
3	34.842	82.413	2.0499
4	40.984	87.797	2.0113
5	49.932	90.394	1.8916
6	61.404	91.479	1.9631
7	66.254	93.002	2.0976
8	70.343	94.129	2.1071
9	73.463	94.963	2.1827
10	77.634	95.425	2.1716

Table 2. Results of a standard PLS regression

The important issue is: Can the results be improved by the CovProc method? The data are typical in situations with process data. The largest correlation coefficient between the x's and y is 0.526, and they decrease to 0.0015. Only 9 x-variables have a correlation coefficient with y that is larger than 0,275, which is the critical value at the 5% significance level. But people in charge do not want to remove variables. The argument is that the 40 x-variables have been carefully selected, and they as a whole describe the situation in question.

Application of the CovProc method.

The idea is to ‘optimize’ the choice of the weights \mathbf{w} , when computing the score vector, $\mathbf{t}=\mathbf{X}\mathbf{w}$. Therefore, the weights are sorted and the number of non-zeros weight is increased. Thus, first $\mathbf{t}=\mathbf{X}\mathbf{w}$ is computed, with $\mathbf{w}=(0,0,\dots,0,w_i,0,\dots)$ and the size of the corresponding fit, $(\mathbf{y}^T\mathbf{t})^2/(\mathbf{t}^T\mathbf{t})$. The weights are here computed as $w_i=(\mathbf{y}^T\mathbf{x}_i)$, but other choices are possible. Then the score vector \mathbf{t} with $\mathbf{w}=(0,\dots,0,w_i,0,\dots,0,w_j,0,\dots)$ is computed and the size of the fit. Note that w_i is the largest weight value, w_j the second largest and so on. In this way the number of non-zero coefficients is increased. All K values of the size of the fit are computed, and the weights giving the largest value of fit are selected. With the score vector found in this way, the loading vector is computed, $\mathbf{p}=\mathbf{X}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$, the corresponding loading weight vector \mathbf{v} and other measures needed.

Fig. 2 shows the changes in the size of the fit for the first four score vectors.

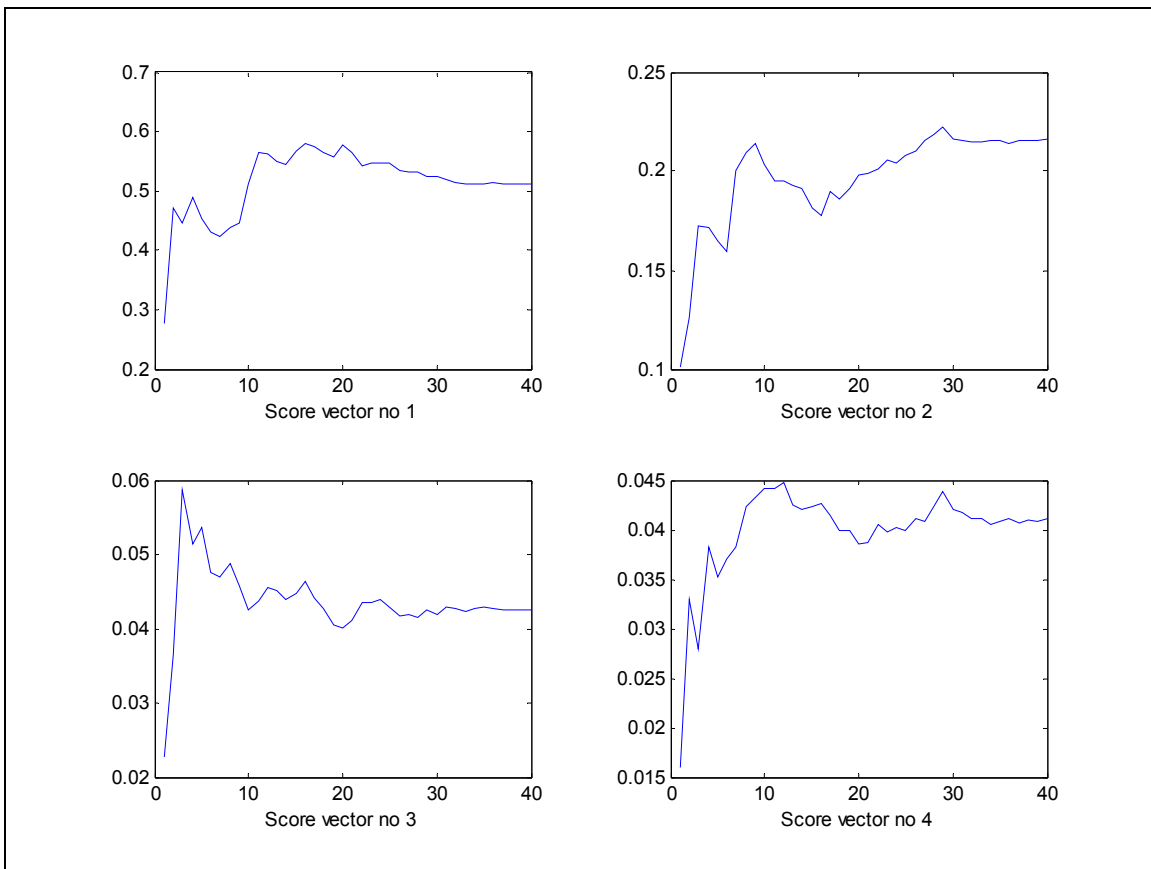


Figure 2. Changes of size in fit, $(\mathbf{y}^T\mathbf{t})^2/(\mathbf{t}^T\mathbf{t})$, for the first four score vectors. The x-axis is the number of variables among the 40 that enter the matrix \mathbf{X} .

From the figures it can be seen that maximum value of the fit, $(\mathbf{y}^T \mathbf{t})^2 / (\mathbf{t}^T \mathbf{t})$, is obtained for 16, 29, 3 and 12 non-zero values of the first four weight vectors respectively. Like shown in Fig. 7, the size of the fit using all of \mathbf{w} can differ a lot from the largest value.

Table 3 gives the numerical results. The results concerning Q are considerably improved. Also, more score vectors are suggested than in Table 2. The cross-validation measure Q is a sample measure that exhibits some sample variations. A further study of Q suggests 8 score vectors. Even if only 5 score vectors are selected, a clear improvement is obtained compared to Table 2. In conclusion it can be stated that a considerable improvement is obtained using the CovProc method.

No	%X	%y	Q
1	11.645	58.091	1.6759
2	19.116	80.327	1.1916
3	29.418	86.184	1.0245
4	35.056	90.658	0.8633
5	46.787	91.837	0.8251
6	55.506	93.186	0.7333
7	58.331	95.024	0.6297
8	62.455	95.627	0.5763
9	66.004	96.026	0.5720
10	70.359	96.281	0.5552

Table 3. Results of COVPROC.

Removing variables. It may be of some interest to see the results, where some of the variables have been removed. The procedure used to find the variables that should be excluded is as follows.

1. Sort the variables. The $K=40$ variables are sorted according to how much they describe the variation of the y -variable. The first x -variable is the one that describes most. Then \mathbf{X} is adjusted for this variable and the next x -variable is found that describes most. This gives a ranking of all 40 x -variables.

2. PLS regression. A number of PLS regression is carried out by expanding the \mathbf{X} matrix by one variable at a time. \mathbf{X} consists first of one variable, the first in the rank. Next the second in the rank is added to \mathbf{X} . This is continued until all variables have been added. For each \mathbf{X} up to 10 score vectors are computed. Thus 10 times 40 PLS regressions are carried out, minus the number, where there are fewer variables than 10. Then the PLS regression is found that gives the best fit. It is registered, which variables are found in that PLS regression.

The result was that the ‘optimal PLS regression’ used 33 x -variables. Then the CovProc method is used to determine optimal weights, where \mathbf{X} now consists of these 33 x -variables. The results are given in Table 4. If the results are compared with Table 3, we see that there is an improvement up to the 6th score vectors, but later picture is unclear. For 8 score vectors suggested in Table 3, the results using all of \mathbf{X} are slightly better.

No	%X	%y	Q
1	11.426	60.912	1.6270
2	19.598	80.689	1.1858
3	34.477	84.879	1.0339
4	40.687	89.611	0.8891
5	45.717	92.334	0.7603
6	51.974	93.848	0.6787
7	61.835	94.649	0.6385
8	68.950	95.391	0.5955
9	73.331	95.950	0.5787
10	76.539	96.356	0.5826

Table 4. Results of COVPROC for \mathbf{X} having 33 x -variables.

Thus, it can be concluded that it is not necessary to remove the least significant variables, when the CovProc method is used. It should be remarked that standard PLS regression

with the 33 x-variables only gives very little change compared to Table 2. Five score vectors give $(\%X, \%y, Q) = (47.547, 91.065, 1.7232)$.

Conclusion. The CovProc method gives considerable improvement compared to standard PLS regression. The measure of fit, R^2 , increases from 90.39% to 95.63%. The cross validation measure Q reduces from 1.892 to 0.576 (for the original un-scaled data). From the point of view of the company this is a clear improvement. It is not necessary to remove the x-variables that have very low correlation with the y-variable. In this case all variables are used at some time, giving a non-zero regression coefficient for all variables. This is an important issue, because the 40 x-variables describe the situation and they are used in further analysis. Even if some of the regression coefficients are small, then it is important that the coefficients are there for further mathematical and economic analysis.