Original Article

STRATEGIC COMPARISON OF RIDGE REGRESSION AND PARTIAL LEAST SQUARES IN MODERN DATA ANALYSIS

Emma Sofia Nielsen and Clara Marie Christensen

Center for Advanced Data Analysis, Denmark

Abstract: There has been a long-standing debate over the choice between Ridge Regression (RR) and Partial Least Squares (PLS) in the field of statistics and chemometrics. Statisticians argue that RR is firmly grounded in a well-established mathematical framework, making it a preferable choice. In contrast, chemometricians tend to favor PLS, which employs projection onto orthogonal vectors, akin to Canonical Correlation (CC). PLS maximizes the covariance between X- and Y-score vectors, while CC focuses on correlation. Both PLS and CC share a closely related theoretical foundation, making PLS an attractive option.

One of the key advantages of PLS is its ability to handle datasets with more variables than samples, with validation techniques like cross-validation and test sets available for result validation. Additionally, graphical tools aid researchers in exploring the data. Previous studies, such as the work by Frank et al. (1993), have attempted to compare RR and PLS, with inconclusive results. The choice between the two methods remains contentious, with papers favoring RR, PLS, or showing mixed findings. This paper aims to contribute to this ongoing discussion and provide insights into the comparative performance of RR and PLS, specifically in the context of chemometrics.

Keywords: Ridge Regression, Partial Least Squares, chemometrics, statistical methods, data analysis, variable selection.

1. Introduction

Over the past several decades, there has been a disagreement about RR and PLS. Statisticians claim that RR is based on a well-developed mathematical foundation. PLS, on the other hand, is based on projection onto orthogonal variables, where the statistical properties are unknown. Therefore, RR is clearly preferable to PLS. Chemometricians favor the use of PLS. PLS uses projection on orthogonal vectors, which are determined in a similar way as at Canonical Correlation (CC). PLS uses maximization of the covariance between X- and Y- score vectors, while CC uses the correlation. The theory of PLS and CC are closely related. PLS has the important advantage that there can be more variables than samples. Validation procedures, like e.g., cross-validation and test sets can be used to validate the results. Graphic procedures assist experimenters in studying the data. Frank

et al. (1993) was one of the first papers to compare RR and PLS. Their conclusion was that RR is slightly better, but the difference is small. The data used was a simulated one, where data are of full rank, but the last singular values are small. The paper was criticized for not using data that are common in chemometrics. Since then, many papers have been published on the comparison, see e.g., search in scholar.google.com. The results obtained in the papers are somewhat mixed. Basak et al. (2002) favor RR, while Irfan et al. (2013) and Wold et al. (1983) favor PLS.

The present paper uses process and spectral data. These data are typical in chemometric work.

In Section 2 we discuss the data that are used here. For the process data, the instrumental data **X** has 25 columns (variables). Singular values of **X** are of order 10^{-6} from the 14^{th} to the 25^{th}. For spectral data, the instrumental data has 40 columns. The last 20 singular values are very small and the last seven of **X** are zero. Thus, both data have reduced rank. However, this does not give any problems, when using RR or PLS. It is common for industrial data they have low or reduced numerical rank. Scaling **X** is important when working with low-rank data. This issue is discussed further.

In Section 3 we present a brief introduction to RR. The use of the Singular Value Decomposition (SVD) of **X** gives precise computation of the RR coefficients and their variances, even for low-rank **X**-matrices.

The low 'practical' rank causes a challenge when determining the Ridge constant. The method used to determine k is by minimizing the size of the residuals in Leave-one-out RR. This procedure is presented in Section 4. It gives Ridge constant $k=2.6\times10^{-5}$ for process data and 6.5×10^{-4} for spectral data. It is common to get a Ridge constant k of order 10^{-5} or smaller. When the Ridge constant is so small, the results of RR are very sensitive to the value of k. The variances of the regression coefficients obtained by OLS show that the modeling task consists of two parts, which are stochastically independent. One part is the fit obtained. The other part is the precision of the estimates. PLS is approaching both parts. This is briefly shown in Section 5.

In Section 6 RR and PLS are applied in the analysis of process and spectral data. The estimation using all data gives similar results for RR and PLS. For test data, PLS and full-rank RR solutions give similar results. However, PLS is slightly better than RR, when low-rank solutions are used. In Section 7 it is argued for using ⌢average cross-validation for judging results. Denote by $_c$ the estimate of **y** obtained from the average of 20 cross-validations It has become stable at the average of 20. R_c^2 is the squared correlation coefficient between **y** and $\widehat{\ }_c$. R_c^2 is used in judging the dimension at PLS and in comparing PLS to RR. It is also used when comparing results from stepwise deletion/selection of variables.

The use of R_c^2 in the evaluation of results is described closer in Section 8.

Backward elimination of variables is important when working with industrial data. The instruments and censors tend to give 'too many' numerical values. An efficient procedure to carry out the backward elimination of variables for RR and PLS is presented in Section 9. There is not a significant difference between RR and PLS when they are applied to this procedure. However, when we study applications to test sets, there is not a clear picture of which method is better. When working with many variables, it is often recommended to use a forward selection of variables. In Section 10 we present a procedure, which has been found efficient when working with many variables. Here we also do not find a significant difference between RR and PLS. When applied to test sets, the picture is also unclear.

In Section 11 we show that for a given Ridge constant k>0, there is a 'noise' matrix **Z** derived from (**X**,**y**), so that the OLS solution based on $\mathbf{X}_1=\mathbf{X}+\mathbf{Z}$ gives the same solution, regression coefficients, as RR. This is used to show that the variances of the regression coefficients in RR are too small. We can use the same algorithm for RR and PLS. Initially, the Ridge constant is estimated. Then, the same algorithm can be used for both. Graphic analysis of data is important in empirical work with data. Section 12 shows some common examples of graphic analysis in chemometrics that is carried out for RR in the same way as for PLS.

In experimental work using PLS, it is known that one should not work with too small score vectors. In RR a large number of score vectors can be small when computing the full-rank solution. Thus, there is an indication that one should work with low-rank RR solutions. Section 13 discusses the results of this paper. It is a challenging issue that RR performs almost equally well as PLS, while the theory of RR is not applicable.

2. Data sets and scaling

The process data are from the production of alcohol, see Höskuldsson et al. (2006). There are 25 process variables. The y-variable is the quality of the product, which is measured at the end of the production process. There are 154 processes. It gives **X** as a 154×25 matrix and **y** as a 154-vector. In the case of test data, **X** is 123×25 and $\mathbf{X}_t$ 31×25, which can be selected in many different ways. The spectral data are FTIR data in the MID-IR range, see Jessen et al. (2014). The FTIR instrument measures the absorbance of infrared light in the liquid, giving 1100 values each time a sample is measured. The initial **X** has 1100 columns (variables). A technician suggests areas, where absorbance may be expected. These areas are studied and those that do not show correlation are deleted. Finally, we end up with 40 wavenumbers (variables) to be used in the analysis.

The response variable is the substance that we want to determine by the FTIR instrument. 200 samples are measured. Thus, **X** is a 200×40 matrix, and **y** 200×1 vector. This is used in the analysis and at cross-validation. However, for test data **X** is 160×40 and $\mathbf{X}_t$ is 40×40. Similarly for **y**. The last 7 singular values of **X** are zero (less than 10^{-20}). The 20th to 33-rd singular values are of order 10^{-4} to 10^{-3}. Thus, **X** has a reduced numerical rank. Experience has shown that it is desirable to reduce the number of variables to 20 to 30. However, this is only a recommendation. Slightly fewer or more variables may be satisfactory.

When working with low-rank data, it is necessary to scale the data. Scaling of columns of **X** and **Y** can be achieved by multiplying the matrices from the right by a diagonal matrix. The linear least squares solution is given by $\boldsymbol{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. If **X** and **Y** are scaled column-wise (by variables), it amounts to the transformations, **X** $\rightarrow$ (**X**$\mathbf{C}_1$) and **Y** $\rightarrow$ (**Y**$\mathbf{C}_2$), where $\mathbf{C}_1$ and $\mathbf{C}_2$ are diagonal matrices. The solution **B** can be obtained from the solution for scaled data, $\mathbf{B}_1$, as follows,

$$\boldsymbol{B} = \boldsymbol{C}_1[\{(\boldsymbol{X}\boldsymbol{C}_1)^T(\boldsymbol{X}\boldsymbol{C}_1)\}^{-1}(\boldsymbol{X}\boldsymbol{C}_1)^T(\boldsymbol{Y}\boldsymbol{C}_2)]\boldsymbol{C}_2^{-1} = \boldsymbol{C}_1\boldsymbol{B}_1\boldsymbol{C}_2^{-1}$$

This equation shows that if we are computing or approximating the linear least squares solution, we can work with scaled data. When we want the solution for the original data, we scale 'back' as shown in the equation. This property is also used when the approximate solution is being computed.

The effect of scaling is better numerical precision. For a small Ridge constant, we are in RR working with a ratio of numbers close to zero. In PLS we are working with projections. Here the adjustment (deflation) is a difference between two matrices, where numbers are close to zero. Scaling secures that the numbers in the computations are approximately of the same size.

Some experimental workers are critical towards scaling of data. The numerical precision at optical instruments (like those of FTIR) is often of the order 10^{-4}. They argue that scaling may cause 'zeros' to be enlarged. But 'zeros' will continue to be small after scaling. Furthermore, scaling may be necessary in order to obtain precise solutions. Variables that have values below detection limits must be analyzed separately.

In the equation below it is supposed that data values are centered. E.g., for a vector $\mathbf{x}_i$ we write $\boldsymbol{x}_i$ instead of $(\boldsymbol{x}_i - \bar{}_i)$. This simplifies the notation. $\mathbf{X}^T$ is the transpose of **X**. The squared length of a vector is $|\mathbf{x}|^2=(\mathbf{x}^T\mathbf{x})=\Sigma_{i=1}^{N} x_i^2$.

3 Ridge Regression

The OLS solution to the linear regression model

(1) $y = b_1 x_1 + b_2 x_2 + \ldots + b_K x_K + \square$ is given by

(2) $\mathbf{b}_O = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

When $(\mathbf{X}^T\mathbf{X})^{-1}$ becomes close to being singular, the solution becomes unstable. In RR it is suggested to stabilize the solution by adding a constant k to the diagonal elements of $(\mathbf{X}^T\mathbf{X})$. The RR solution is now computed as

Original Article

(3) $\boldsymbol{b}_R = (\boldsymbol{X}^T\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

I is the identity matrix and k the Ridge constant. It is often suggested that k should be determined by a Leave-one-out regression. This is considered closer in Section 4.

It is efficient to compute the regression coefficients $\mathbf{b}_R$ by using the Singular Value Decomposition (SVD) of **X**, $\mathbf{X}=\mathbf{USV}^T$. Here **S** is a diagonal matrix with singular values in the diagonal, and **U** and **V** have orthonormal columns. Then the OLS and RR solutions can be written as

(4) $\boldsymbol{b}_O = \boldsymbol{VDU}^T\boldsymbol{y} = \boldsymbol{VD}_0\boldsymbol{f}$

(5) $\boldsymbol{b}_R = \boldsymbol{VD}_1\boldsymbol{U}^T\boldsymbol{y} = \boldsymbol{VD}_1\boldsymbol{f} = \Sigma_{i=1}^{K} d_i/(d_i^2+k)[(\boldsymbol{u}_i^T\boldsymbol{y})\boldsymbol{v}_i]$

where $\mathbf{f}=\mathbf{U}^T\mathbf{y}$, $\mathbf{D}_0$ is a diagonal matrix $(1/d_i^2)$, $\mathbf{D}_1$ is a diagonal matrix $(d_i/(d_i^2+k))$ and d_i is the diagonal element of **S**, $d_i=S(i,i)$.

The standard OLS assumptions are that **y** has the variance $\mathrm{Var}(\mathbf{y})=\sigma^2\mathbf{I}$ and that the expected value of $\mathbf{b}_O$ is $E(\mathbf{b}_O)=\boldsymbol{\beta}$. The estimate $\mathbf{b}_O$ has the variance, $Var(\boldsymbol{b}_O) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. This gives

(6) $tr\big(Var(\boldsymbol{b}_O)\big) = \Sigma_{a=1}^{K} Var(b_a) = \sigma^2\Sigma_{a=1}^{K} 1/d_a^2$

tr() is the trace function. The variance of the RR estimates is

(7) $Var(\boldsymbol{b}_R) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X}+k\boldsymbol{I})^{-1} = \sigma^2\boldsymbol{VD}_2\boldsymbol{V}^T$,

where $\mathbf{D}_2$ is a diagonal matrix, $([d_a/(d_a^2+k)]^2)$. This gives

(8) $tr\big(Var(\boldsymbol{b}_R)\big) = \Sigma_{a=1}^{K} Var\big(b_{R,a}\big) = \sigma^2\Sigma_{a=1}^{K} d_a^2/(d_a^2+k)^2$

The bias can be computed as

(9) $E(\boldsymbol{b}_R)-\boldsymbol{\beta} = [(\boldsymbol{X}^T\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}-\boldsymbol{I})]\boldsymbol{\beta} = -k\boldsymbol{V}(\boldsymbol{D}+k\boldsymbol{I})^{-1}\boldsymbol{V}^T\boldsymbol{\beta}$

Let $\square=\mathbf{V}^T\boldsymbol{\beta}$. Then we get for the squared size of the bias

(10) $\Sigma_{a=1}^{K}[E\big(b_{R,a}\big)-\beta_a]^2 = \mathrm{k}^2\Sigma_{a=1}^{K}\gamma_a^2/(d_a^2+k)^2$

Note, that we cannot compute the bias, when some singular values are exactly zero. This is due to that we cannot compute the estimate of $\boldsymbol{\beta}$, $\mathbf{b}_O$, which is used to compute $\square$.

We now use that for any random variable Z we have $E(Z^2)=\mathrm{Var}(Z)+(E(Z))^2$. This gives

2

(11) $\varphi(k) = {}_{a=1}^{K}\{E(b_{R,a}-\beta_a)\}$

$= {}_{a}^{K}$

$= \sigma^2\,{}_{a=}^{K}$

$_{=1}Var\big(b_{R,a}\big) + {}_{a=1}^{K}[E\left(b_{R,a}\right)-\beta_a]^2$

$_1 d_a^2/(d_a^2+k)^2 + \mathrm{k}^2\ {}_{a=1}^{K}\gamma_a^2/(d_a^2+k$

$)^2$

We get OLS regression, when k=0 and $\varphi(0) = \Sigma_{a=1}^{K} Var(b_a)$. Differentiating (11) with respect to k and let k=0, we get

(12) $\varphi'(0) = -\sigma^2\Sigma_{a=1}^{K} 1/d_a^2$

From the Tailor expansion,

(13) $\varphi(k) \cong \varphi(0) + k\varphi'(0)$,

we see that we can always find a value of k in the neighborhood of zero so that $\varphi(k) < \varphi(0)$. This is the main motivation for RR. By replacing $\boldsymbol{X}^T\boldsymbol{X}$ by $(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})$ for some small value of k, both the variance (8) and the mean squared error (11) can be reduced.

4. The Ridge constant k

The theoretical considerations to determine k do not function well. Instead, it is recommended to determine k by Leave-one-out regression. The procedure is as follows.

One sample among the N samples is left out and $\mathbf{b}_R$ is computed for the (N-1) samples. The estimate (5) is used to compute the y-value of the left-out sample. This is repeated for all samples. Thus, $\hat{}_{est}$ is the result for all samples and the difference ($\mathbf{y}$-$\hat{}_{est}$) shows how well the Leave-one-out regression works. The task is to find k that gives the minimum value of | $\mathbf{y}$-$\hat{}_{est}$|.

There is a unique k, k_{min}, where the value of |$\mathbf{y}$-$\hat{}_{est}$| is at minimum. Furthermore, for the present data, the value of the |$\mathbf{y}$-$\hat{}_{est}$| only increases, when k is smaller or larger than k_{min}. Therefore, it is easy to obtain the minimum value.

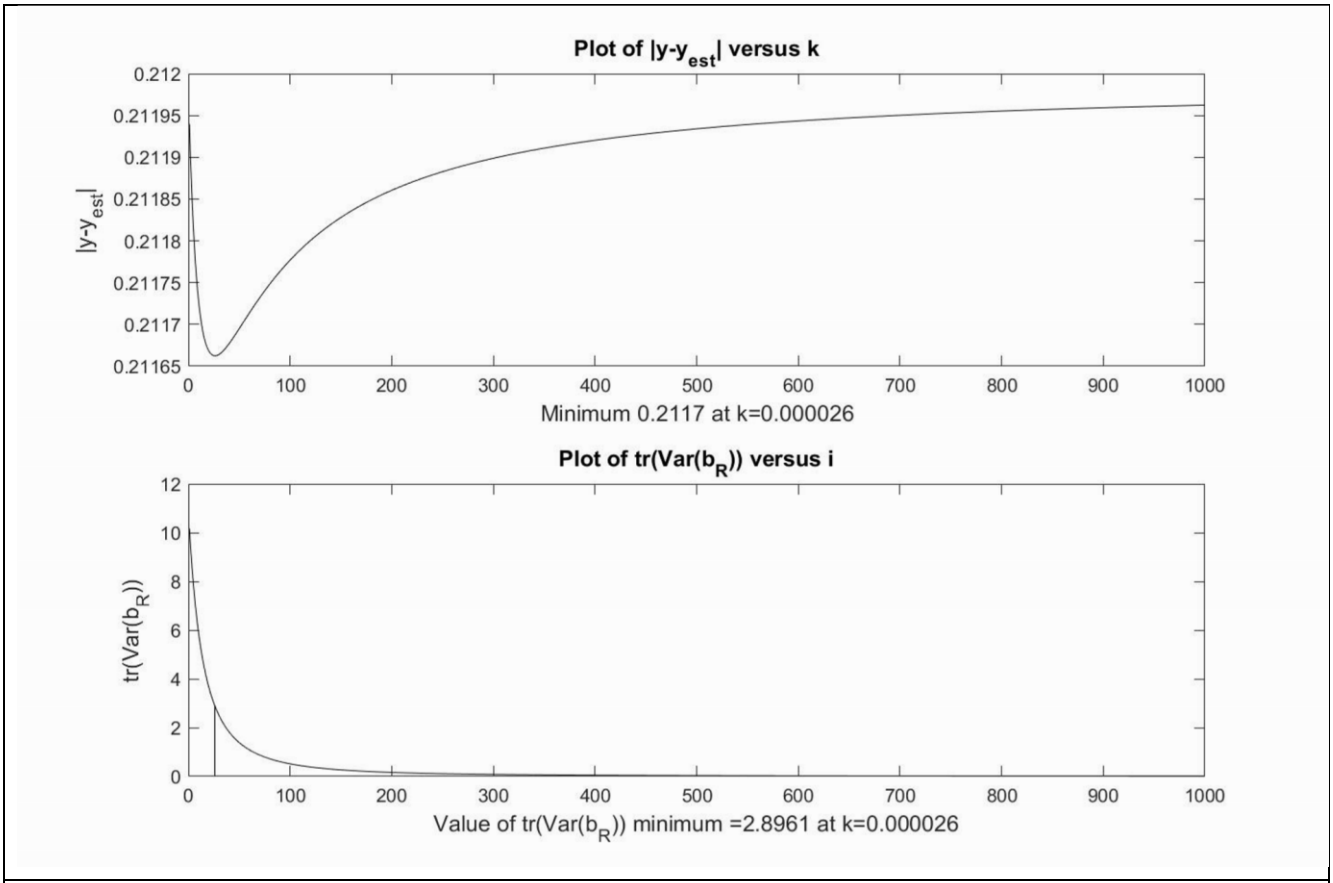


Figure 1. Upper figure plot of |$\mathbf{y}$-$\hat{}_{est}$| versus i. k=i×10^{-6}, i=1,2, … , 10^4. x-axis is the i-values. The lower figure is plot of tr(Var($\mathbf{b}_R$)), (8), versus i. Process data.

We consider first the process data. **X** is here 154×25. Columns of **X** are scaled by their standard deviation. The Ridge constant is generally small for X-matrices that are not of full rank. For the process data we expect k to be small. We compute |$\mathbf{y}$-$\hat{}_{est}$| for k=i×10^{-6}, i=1,2,3, … , 10^4. The results are illustrated in Figure 1. The minimum value of |$\mathbf{y}$-$\hat{}_{est}$| is 0.212 and is obtained for i=26 or k=0.000026. The figure shows increasing values of |$\mathbf{y}$-$\hat{}_{est}$| for decreasing i less than 26 and also for increasing i-values larger than 26. The total variance (8) is

Original Article

tr(Var($\mathbf{b}_R$))=2.8961, which is unrealistically small for the 25 variables. When working with different k-values (e.g., for i=500 to 1000 in Figure 1), we see clear evidence that the value of tr(Var($\mathbf{b}_R$)) is too small. This is studied closer in Section 11.

For the spectral data **X** is 200×40. They are also centered and columns scaled by their standard deviation. Figure 2 shows the results for the spectral data.

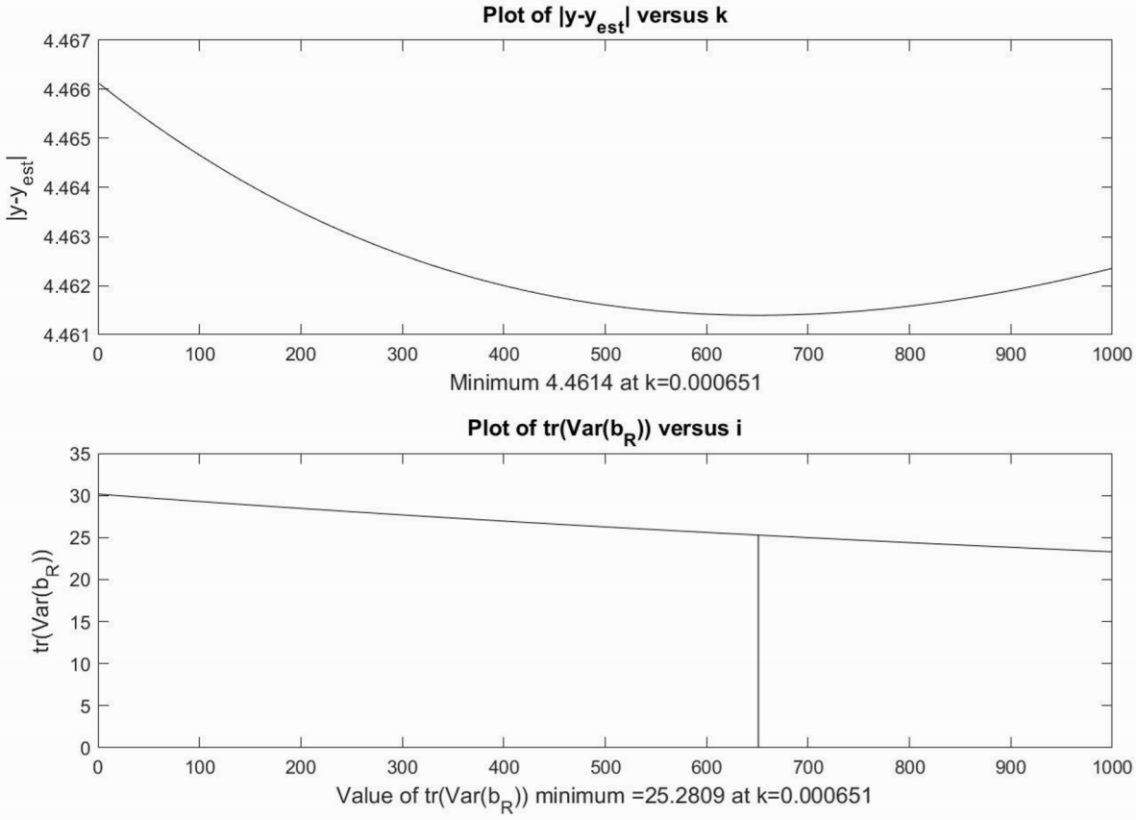


Figure 2. Upper figure plot of |**y**-$\hat{}_{est}$| versus i. k=i×10^{-5}, i=1,2, … , 10^4. x-axis is the i-values. The lower figure is a plot of tr(Var($\mathbf{b}_R$)), (8), versus i. Spectral data.

The smallest value of |**y**-$\hat{}_{est}$|=4.4614 is obtained for k=0.000651=6.51×10^{-4}. The value of the total variance is here tr(Var($\mathbf{b}_R$))=25.2809. From the figure we see that by making k larger, the value of tr(Var($\mathbf{b}_R$)) can be made smaller. For both data we get a unique value for the minimum of |**y**-$\hat{}_{est}$|.

5 PLS Regression

We shall here briefly explain the background for PLS Regression.

Score vectors are used as regression components. An X-score vector **t** is given by

(14) $\mathbf{t} = w_1 \mathbf{x}_1 + w_2 \mathbf{x}_2 + \ldots + w_p \mathbf{x}_p = \mathbf{Xw}$

Similarly, a Y-score vector is given by **u**=**Yq**. **w** and **q** are unknown weight vectors. They are determined by maximizing the covariance between **t** and **u**,

(15) maximize $\mathbf{t}^T\mathbf{u}$ = maximize $\mathbf{w}^T\mathbf{XYq}$, s.t. |**w**|=|**q**|=1

It can be shown that the maximization task leads to the eigensystem

(16) $\mathbf{X}^T \mathbf{Y}\mathbf{Y}^T \mathbf{X}\,\mathbf{w} = \square\,\mathbf{w}$

When the X-score vector **t** has been determined, **Y** is projected onto it,

Original Article

$(\mathbf{Y}^T\mathbf{t})/\mathbf{t}^T\mathbf{t})\ \mathbf{t}$

The estimated $\mathbf{Y}$, $\hat{\mathbf{Y}}$, based on A X-score vectors, is computed as

(17) $\hat{Y} = (\boldsymbol{Y}^T\boldsymbol{t}_1)/(\boldsymbol{t}_1^T\boldsymbol{t}_1)\boldsymbol{t}_1 + \quad (\ ^T\ )/(\ ^T\ ) \ldots + \ \boldsymbol{Y}\ \boldsymbol{t}_A \quad \boldsymbol{t}_A\boldsymbol{t}_A\ \boldsymbol{t}_A$

When a score vector $\mathbf{t}$ has been determined, $\mathbf{X}$ and $\mathbf{Y}$ are adjusted (deflated) by this score vector

(18) $\mathbf{X} \leftarrow \mathbf{X} - d\ \mathbf{t}\ \mathbf{p}^T$, $d=1/(\mathbf{t}^T\mathbf{t})$ and $\mathbf{p}=\mathbf{X}^T\mathbf{t}$

(19) $\mathbf{Y} \leftarrow \mathbf{Y} - d\ \mathbf{t}\ \mathbf{q}^T$, $\mathbf{q}=\mathbf{Y}^T\mathbf{t}$

The adjustment (18) gives both orthogonal score vectors and a reduction of $\mathbf{X}$ by rank 1. These adjustments can be numerically unstable. Therefore, for low-rank data, data should be scaled.

The weight vectors $\mathbf{w}$ and $\mathbf{q}$ can be determined by the Singular Value Decomposition of $\mathbf{X}^T\mathbf{Y}$,

$\mathbf{X}^T\mathbf{Y} = \mathbf{U}\ \mathbf{S}\ \mathbf{V}^T$, $\mathbf{w}=\mathbf{u}_1$ and $\mathbf{q} =\mathbf{v}_1$

This shows that equal importance is given to $\mathbf{X}$ and $\mathbf{Y}$. The significance of this can be seen by looking at the variance of the OLS regression coefficients (written for one y-variable),

(20) $\mathrm{Var}(\mathbf{b}_O) \approx s^2\times(\mathbf{X}^T\mathbf{X})^{-1} = [\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}]\times[(\mathbf{X}^T\mathbf{X})^{-1}]/(N-K)$

Assuming a multivariate normal distribution for data, the two terms in the squared brackets are stochastically independent. Therefore, both need to be addressed, when computing the regression coefficients. For further analysis, see Höskuldsson (2017).

6 Application to test sets

Consider first the process data. Data are divided into calibration data ($\mathbf{X}_1$,$\mathbf{y}_1$) containing 123 samples, and test data ($\mathbf{X}_t$,$\mathbf{y}_t$) containing 31 samples. The test data are selected as every 5th, no. 2,7,12, … . Before selecting test samples, we may randomize the samples and select the samples from those. However, this is not done here. PLS is applied to ($\mathbf{X}_1$,$\mathbf{y}_1$). The test set is ($\mathbf{X}_t$,$\mathbf{y}_t$). The regression coefficients $\mathbf{b}_P$ are computed for each dimension, 1,2, …, 25. The estimated y-values for the test data are $\hat{}_{t,est}=\mathbf{X}_t\mathbf{b}_P$. In Table 1 is shown the standard deviations of ($\mathbf{y}_t - \hat{}_{t,est}$), s_p. We see that the smallest value is found at dimension 11, $s_p=0.0161$.

We carry out the analysis in Section 4 for ($\mathbf{X}_1$,$\mathbf{y}_1$). The revised numbers are $k=1.8\times10^{-5}$, $|\mathbf{y}-\hat{}_{est}|=0.1942$, and $\mathrm{tr}(\mathrm{Var}(\mathbf{b}_R))=5.614$. At the RR analysis, the regression coefficients are computed for each 1 to i in (5). The Ridge constant $k=1.8\times10^{-5}$ is used. The RR coefficients $\mathbf{b}_R$ are used to estimate the $\mathbf{y}_t$-values by $_{r,est}=\mathbf{X}_t\mathbf{b}_R$. The last column in Table 1 shows the standard deviation of ($\mathbf{y}_t - \hat{}_{r,est}$), s_r. At dimension 11 the value of s_r is also 0.0162. A full rank solution also gives $s_r=0.0162$.

In conclusion, we can state that there is not difference between the results of PLS and RR.

The same analysis is carried out for the spectral data. Here $\mathbf{X}_1$ is 160×40 and $\mathbf{X}_t$ is 40×40. The analysis in Section 4 is carried out for ($\mathbf{X}_1$,$\mathbf{y}_1$). The results are $k=0.000061$, $|\mathbf{y}-\hat{}_{r,est}|=4.295$, and $\mathrm{tr}(\mathrm{Var}(\mathbf{b}_R))=45.165$.

The regression coefficients are computed at each dimension. For PLS coefficients $\mathbf{b}_P$ we compute by $\hat{}_{p,est}$ $=\mathbf{X}_t\mathbf{b}_P$ and the standard deviation of the residuals, ($\mathbf{y}_t -$), s_p. The 40 s_p-values are plotted in Figure 3 and drawn as a curve, '.-'. Similarly for RR. The estimated $\mathbf{y}_t$-values are computed at each dimension by $\hat{}_{r,est}=$ $\mathbf{X}_t\mathbf{b}_R$. s_r is the standard deviation of ($\mathbf{y}_t - \hat{}_{r,est}$). It is also plotted as a curve using '--'. The smallest value of s_p is obtained at dimension 25, $s_p=0.2777$. At dimension 25 we get for RR $s_r=0.2869$. The residual standard deviation is slightly smaller for PLS. At dimension 33 the standard deviation s_r is equal to 0.2769.

In conclusion, we can state that PLS is slightly better than RR. When the full-rank solution of RR is used, there is for practical purposes very small difference between PLS and RR.

7. Cross-validation procedures

In n-fold cross-validation, samples are randomly divided into n groups. The analysis is carried out for samples in (n-1) groups and results are applied to the nth one. This is carried out so that each group is excluded once. Leave-one-out regression is an example of cross-validation, where n=N. Usually, n=10 is selected. The result of cross-

validation is an estimate $\mathbf{y}_c$ of $\mathbf{y}$, where each value of $\mathbf{y}_c$ is estimated by 90% of the samples. We compute the squared correlation coefficient, R_c^2, between $\mathbf{y}$ and $\mathbf{y}_c$, and the standard deviation, s_c, of ($\mathbf{y}$ - $\mathbf{y}_c$). A cross-validation procedure can be repeated several times, e.g., 20 times, $\mathbf{y}_{c,1}$, … $\mathbf{y}_{c,20}$. The average $\mathbf{y}_c$=($\mathbf{y}_{c,1}$+ …+ $\mathbf{y}_{c,20}$)/20 will be relatively stable. There can be uncertainties in one cross-validation, which give variation in R_c^2 and s_c. There can be a different reason for this. E.g., there can be relatively many y-values that are small and a few large ones, or groups in data, etc. We may need a stable estimate of $\mathbf{y}_c$ in order to be able to distinguish between the x-variables, because there can be high correlations between variables like in the case of the spectral data. The average of 20 cross-validations is normally fine.

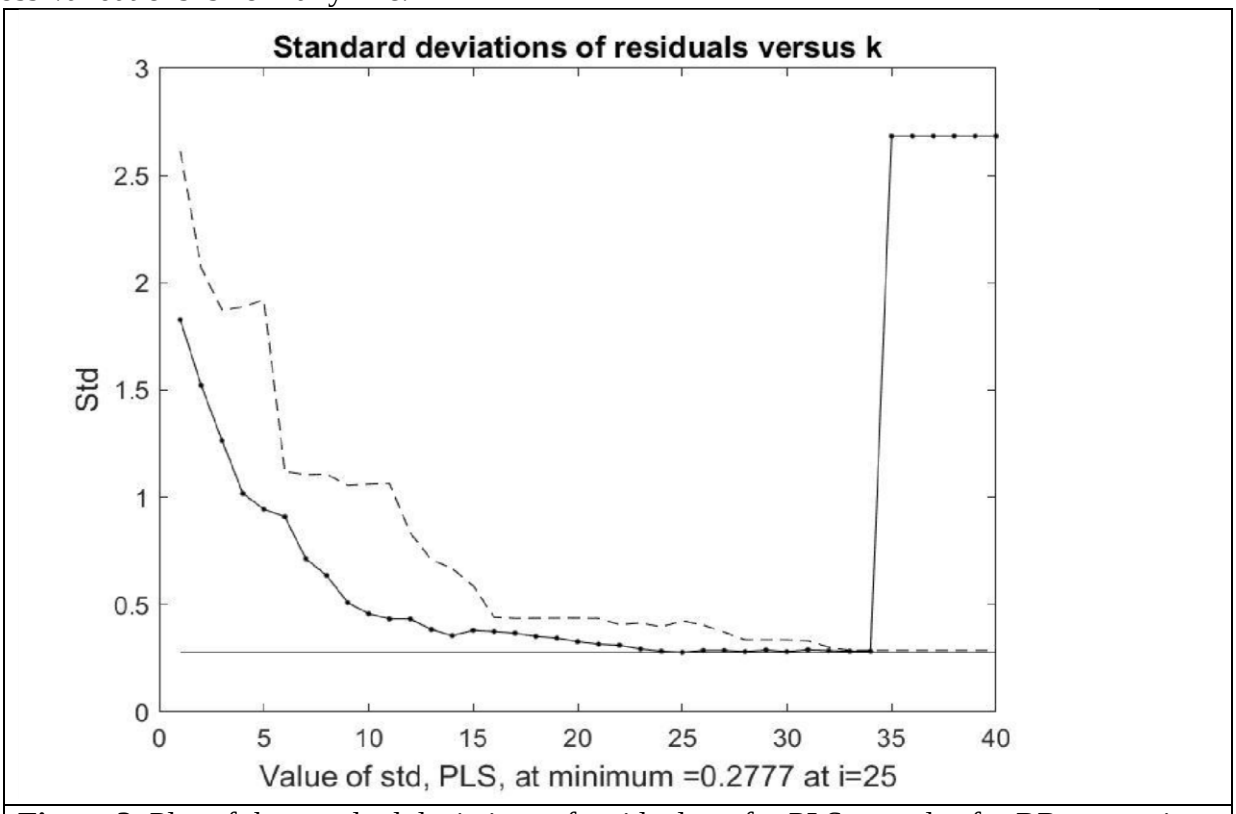


Figure 3. Plot of the standard deviations of residuals, s_t for PLS .-, and s_r for RR --. x-axis is the dimension. Spectral data.

It is important that each group in the cross-validation is representative for all data. This can be achieved in many ways. The samples can be mixed randomly before a cross-validation. Also, the samples can be sorted according to the y-values, or the first PLS score vector, or the first PCA score vector or some other criterion. Random selection is then based on the sorted samples, which also can be randomly mixed before cross-validation.

8. Criterion of comparison of RR and PLS

There are many ways to compare the two methods. The criterion chosen here is the value of R_c^2 for the average $\mathbf{y}_c$ of 20 cross-validation. $R_c^2$is the squared correlation coefficient between $\mathbf{y}$ and $\mathbf{y}_c$. In the variable selection methods below there is given a 'pool' of variables. In backwards deletion we want to determine a variable that should leave the pool. In forward selection we want to find the variable that is to be added to the pool. In RR we start, for a given pool, to determine the Ridge constant k. It is determined, like shown in Figure 1, by finding the smallest value of |$\mathbf{y}$-$\hat{}_{est}$|. This Ridge constant is used in the cross-validations and in the application to a test set.

Original Article

Full rank solution is computed each time we compute the RR regression coefficients in the variable selection methods.

In PLS, also for a given pool, we register the value of $\mathbf{y}_c$, the average cross-validation, at each dimension. For the spectral data the dimension is from 1 to 33. The last 7 score vectors are zero. We compute R_c^2 for each dimension, $R_{c,1}^2$, $R_{c,2}^2$, …, $R_{c,33}^2$. R_c^2 for PLS is the largest value of $R_{c,i}^2$ for i=1,2, …, 33. This dimension is used for each deletion/selection of variables of the pool. When working with test set, PLS is carried out using this dimension and results applied to the test set. The disadvantage of this criterion is that $R_{c,i}^2$-values tend to be almost equal for a range of dimensions. For instance, $R_{c,15}^2$, …, $R_{c,25}^2$, may be all almost equal, while $R_{c,25}^2$ happen to be the largest. This may not be the best dimension for the test set. Therefore, we need to be careful in the interpretation of the results, when we compare results on test data.

9. Backward deletions of variables

We shall here only work with the spectral data. The engineer, who is responsible for the data, has suggested using 40 variables (wavenumbers). The experience is that we should use between 22 and 32 variables in future analyses. We shall study this task by using RR and PLS.

For cross-validation, all data are used. For test data, the data are divided into calibration data, 160 samples, and test data, 40 samples. Variables are eliminated one by one, starting with a pool of 40 variables and continuing until 15.

The following magnitudes are computed.

1) The number of variables in the pool, initially 40 variables

2a) For RR, the RR constant is determined using all variables in the pool. It is used at cross-validations and at test set, for each variable that is deleted from the pool.

2b) For PLS determine the dimension to use in the analysis

3) For all variables in the pool: Delete one variable from the pool, and compute 4) to 9) without this variable by RR

4) The squared correlation coefficient, R^2, between $\mathbf{y}$ and $\hat{}_{r,e}=\mathbf{X}\mathbf{b}_R$, $\mathbf{b}_R$ the RR coefficients

5) The standard deviation, $s=|\mathbf{y}-\hat{}_{r,e}|/(N-1)^{½}$

6) R_c^2 between $\mathbf{y}$ and $\mathbf{y}_c$, where $\mathbf{y}_c$ is the average over 20 cross-validations

7) $s_c=|\mathbf{y}-\mathbf{y}_c|/(N-1)^{½}$

8) R_t^2 between $\mathbf{y}_t$ and $\hat{}_{t,e}$, where $\hat{}_{t,e}=\mathbf{X}_t\mathbf{b}$, $\mathbf{b}$ is $\mathbf{b}_R$

9) $s_t=|\mathbf{y}_t-\hat{}_{t,e}|/(N-1)^{½}$

10) A variable is deleted from the pool that gives the largest values of R_c^2 Steps 2) to 10) are repeated until 15 variables in the pool.

The steps of the computations are

0. Initially, (2a) to (9) are computed for the 40 variables. Initial pool consists of all 40 variables. This gives the first line in Table 2.

1a. In the case of RR, compute the Ridge constant k for variables in the pool.

1. Delete a variable from the pool. 4) to 9) are computed without this variable. This is carried out for all variables in the pool. RR is used in each regression. Same k is used for all deleted variables.

2. Delete the variable from the pool, which gives the largest value of R_c^2, the squared correlation coefficient between $\mathbf{y}$ and the average of 20 values of $\mathbf{y}_c$.

3. If there are variables left in the pool, go to 2a) for RR (and 2b) for PLS).

Note, that the RR constant k is computed before the cross-validations and use of test set. This value of k is used in the computation of $\mathbf{b}_R$ for each cross-validation and each analysis, when a variable is deleted from the pool.

Original Article

Table 2 shows the results for RR. Initially, R_c^2=98.684%. The largest value of R_c^2 is obtained at 31 variables, where R_c^2=98.787% and sc=0.3206. Thus, RR suggests to use 31 variables in the future. We see that there is relatively small variation in the numbers before and after deletion of a variable. The same procedure is applied for PLS. In the analysis 4) to 9) the dimension used in PLS is the one found before deleting a variable. This dimension is used in the computation of 4) to 9), when a variable is deleted. In 4) to 9) the PLS estimates $\mathbf{b}_P$'s are used. We get a table for PLS that is similar to Table 2. It is not shown here. The largest value of R_c^2 is 98.771%, which is found at
29 variables. And s_c=0.3226. The numbers for R_c^2 and s_c are very close to each other for RR and PLS. One cannot state if one is better than the other.

Table 2. Backward deletion of variables, RR

1)	2)	3)	4)	5)	6)	7)	8)	9)
0	0	0,000375	99,254	0,2514	98,684	0,3343	99,236	0,2785
40	15	0,000375	99,240	0,2537	98,716	0,3301	99,257	0,2751
39	8	0,000640	99,237	0,2542	98,730	0,3283	99,261	0,2747
38	30	0,000421	99,239	0,2539	98,734	0,3277	99,260	0,2747
37	23	0,000422	99,236	0,2543	98,738	0,3272	99,248	0,2767
36	40	0,000372	99,227	0,2559	98,756	0,3247	99,157	0,2929
35	16	0,000352	99,227	0,2559	98,753	0,3253	99,159	0,2927
34	20	0,000359	99,202	0,2600	98,779	0,3216	99,027	0,3153
33	3	0,000383	99,202	0,2600	98,786	0,3207	99,028	0,3152
32	13	0,000385	99,196	0,2610	98,774	0,3224	99,041	0,3126
31	**36**	**0,000527**	**99,195**	**0,2611**	**98,787**	**0,3206**	**99,044**	**0,3123**
30	5	0,000513	99,195	0,2611	98,775	0,3223	99,043	0,3124
29	39	0,000500	99,195	0,2611	98,769	0,3230	99,043	0,3124
28	10	0,000500	99,195	0,2611	98,772	0,3226	99,043	0,3124
27	32	0,000501	99,188	0,2623	98,773	0,3225	99,056	0,3108
26	38	0,000829	99,171	0,2650	98,766	0,3234	99,087	0,3051
25	33	0,000823	99,141	0,2697	98,753	0,3251	99,100	0,3039
24	31	0,002614	99,104	0,2755	98,716	0,3298	98,960	0,3275
23	22	0,003089	99,088	0,2780	98,713	0,3301	98,863	0,3422
22	1	0,002228	99,069	0,2808	98,710	0,3306	98,875	0,3385
21	29	0,001071	99,019	0,2883	98,664	0,3363	98,602	0,3790
20	25	0,001067	98,954	0,2977	98,588	0,3458	98,736	0,3593
19	11	0,000675	98,882	0,3077	98,493	0,3573	98,491	0,3924
18	24	0,000149	98,881	0,3079	98,517	0,3544	98,473	0,3949
17	7	0,000422	98,592	0,3453	98,225	0,3877	98,342	0,4139
16	9	0,000353	98,054	0,4060	97,535	0,4570	98,482	0,3998
15	35	0,000195	97,585	0,4523	97,009	0,5034	98,131	0,4414

Original Article

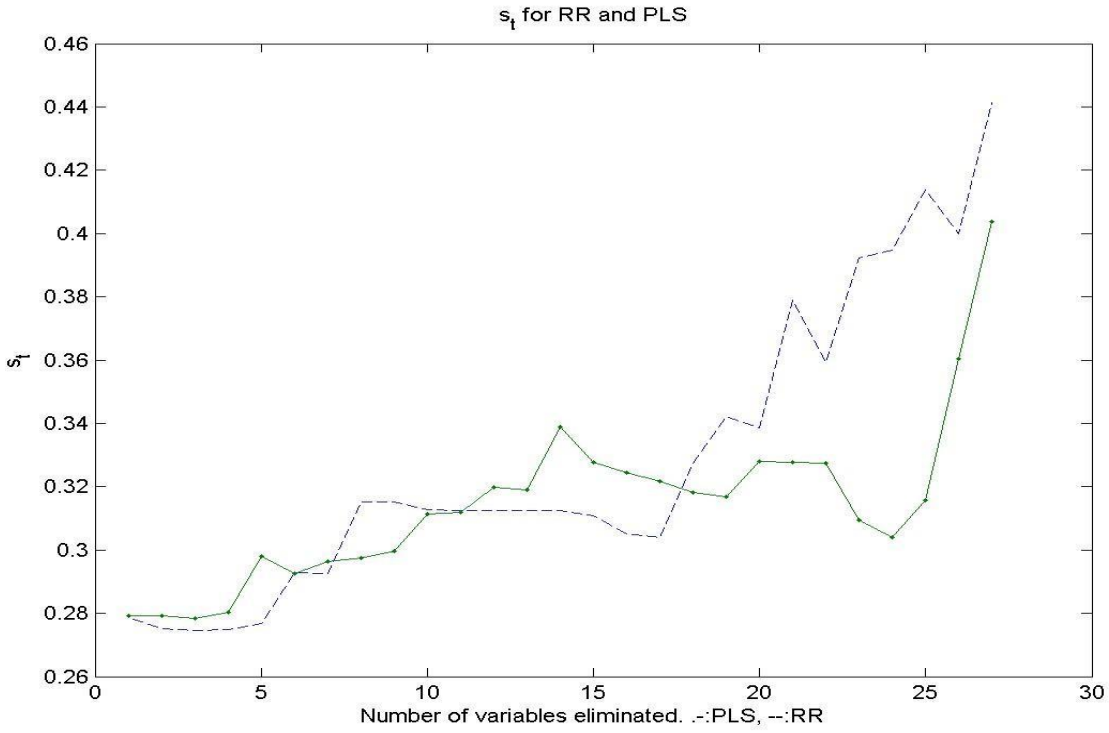


Figure 4. The values of s_t were obtained for RR and PLS at the backwards deletion of variables. .- is those from PLS, -- from RR.

Consider now the residual standard deviation for the test set, s_t, given by 9). Figure 4 shows the values of s_t for both RR and PLS, which is obtained at each deletion, from 1 to 26 (40 down to 15) deleted variables. When cross-validation is used, RR suggests that 9 variables should be deleted, while PLS 11. For these numbers the values of s_t are close to equal. Otherwise, there is some difference between RR and PLS. RR gives smaller values from 1 to 7 and 12 to 18 deleted variables, while PLS gives smaller values for 19 and more are deleted. As mentioned above, the application of PLS to test set is sensitive to the dimension used in PLS. Therefore, further study is needed in order to find out, which is better, RR or PLS, when applied to test set. This is not considered here.

10 Forward selection of variable

Similar analysis like in in previous section can be carried out for forward selection of variables. The steps are: Select the variable having the largest correlation coefficient with **y**. Initially the pool of variable consists of this variable.

1a. In case of RR, compute the Ridge constant k for variables in the pool. 1b.

Determine the dimension to use for PLS for variables in the pool.

For each variable not in the pool, compute 4) to 9) of previous section

0. Add a variable to the pool of variable. Carry out cross-validation using the variables in the pool and this variable. Carry this out for all variables not in the pool, one at a time. PLS/RR is used in each regression.
1. Add the variable permanently to the pool, which gives the largest value of R_c^2, the squared correlation coefficient between **y** and average $\mathbf{y}_c$.
2. If there are variables left not on the pool, go to 1a. for RR and 1b. for PLS.

Original Article

In all cross-validations the average of 20 cross-validations is used. Like at backward deletion of variables, the Ridge constant is computed and the dimension used in PLS are computed for a given pool of variables and before selection of variables. We get as output from the computations that is similar to Table 2, except the they start with smaller values of R_c^2. It is not shown here.
In Figure 5 is shown the values of R_c^2 , the upper figure, and of R_t^2, the lower figure.
When selecting 27th or later variables, the values of R_c^2 are practically the same for PLS and RR. R_c^2 and R_t^2 increase in the beginning faster for PLS than for RR. The largest values of R_c^2 are obtained around 30 variables for both PLS and RR. Thus, almost the same conclusion is obtained for forward selection of variables as for backwards deletion. This holds for both PLS and RR.
In conclusion, the difference between PLS and RR is small. For the test set there are differences until around 26 selected variables. However, as mentioned earlier, there are some uncertainties in the results for test set, when PLS is used due to the criterion used, the maximal R_c^2-value.

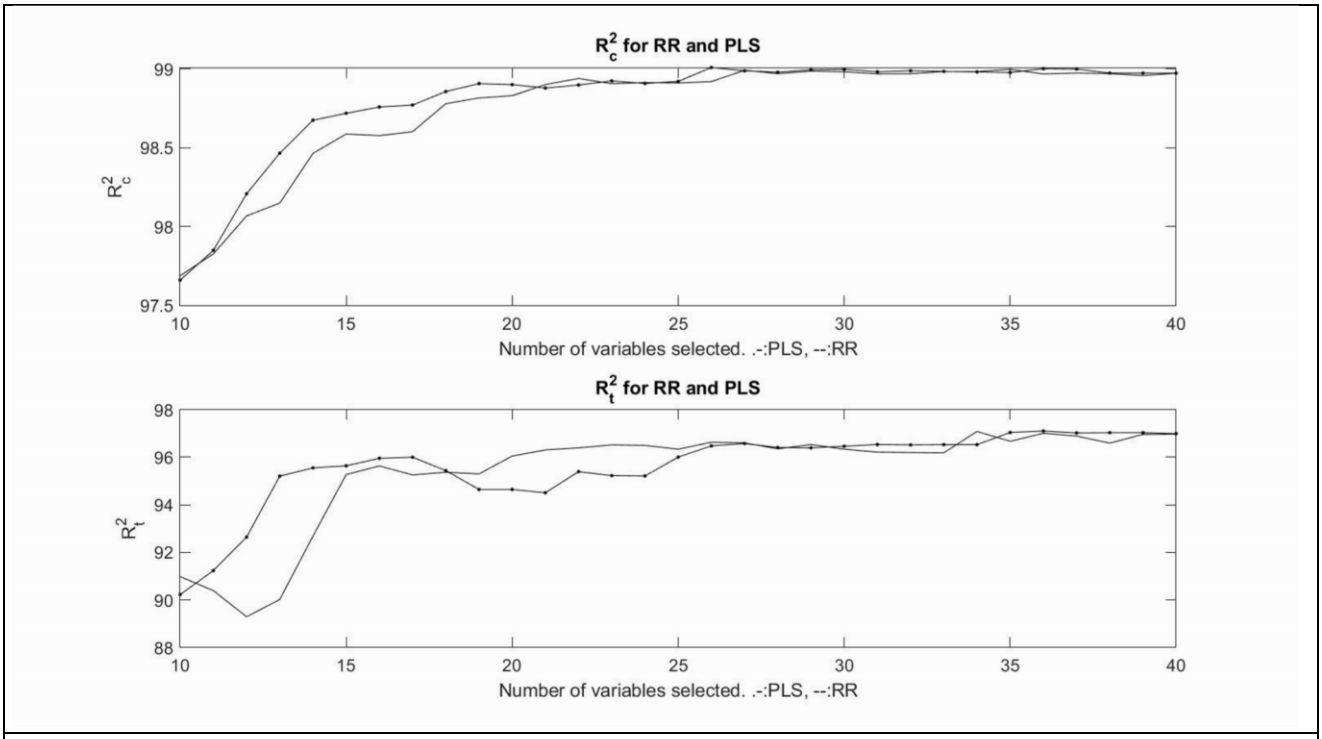


Figure 5. Upper figure R_c^2-values obtained for RR and PLS at forward selection of variables. Lower figure the R_t^2-values for test set. .- is those from PLS, -- from RR.

11 RR estimation as OLS

When working with different analysis of RR, we get a clear impression that the total variance of RR, (8), is too small. We shall consider this closer.
The variance matrix for the OLS regression coefficients is $Var(\boldsymbol{b}_O) \square s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. When there is collinearity in data, the precision matrix, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, tends to be large. Even if the precision matrix is far from being singular, there may be problems in using OLS. This is the case, when the there are many variables and correlation among all or most of the variables, OLS may give wrong or misleading results, like e.g., declare a variable significant, although it is not. For the process data, this happens already at 10 variables. (It is a serious problem in industry that popular

program packages in statistics use OLS as a standard for regression analysis). By replacing the estimate of $\mathbf{b}_O$ by $\boldsymbol{b}_R = (\boldsymbol{X}^T\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, we get both smaller regression coefficients and smaller variance matrix for $\mathbf{b}_R$.

A question is: Is there a matrix **Z** containing small values derived from (**X**,**y**), so that the RR estimate $\mathbf{b}_R$ is equal to the OLS estimate, when using $\mathbf{X}_1=\mathbf{X}+\mathbf{Z}$ instead of **X**? This is an important question, because if affirmative, would allow us to evaluate the RR solution by using the theory of OLS. In RR we search for a good value of k, but treat the results as if the value is fixed and given beforehand.

The answer is in fact affirmative. For a given value of k, we can use (**X**,**y**) to determine **Z** so that for $\mathbf{X}_1=\mathbf{X}+\mathbf{Z}$ we have

(21) $\boldsymbol{X}_1^T\boldsymbol{X}_1 = \boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I},\ \boldsymbol{X}^T\boldsymbol{Z} = d\boldsymbol{I}\ and\ \boldsymbol{Z}^T\boldsymbol{y} = \boldsymbol{0}$, where d is some constant

It follows from (21) that the OLS solution using $\mathbf{X}_1$ is the same as the one of RR. When k is small, the values in **Z** are also small. The values in **Z** can be viewed as 'noise' values derived from (**X**,**y**), which are added to **X**. The derivation of **Z** is somewhat technical. Instead of going through the details, we show in Box 1 a Matlab program that carries out the computations. First, the matrix **Q** contains **y** and **X**. **W** is an orthogonalization of **Q**. **C** is lower triangular. $\mathbf{B}_2$ is determined so that $\mathbf{B}_1{}^T\mathbf{B}_1+\mathbf{B}_2{}^T\mathbf{B}_2=\mathbf{I}$. The matrix $\mathbf{Z}=\mathbf{A}\times\mathbf{B}$ is the desired matrix.

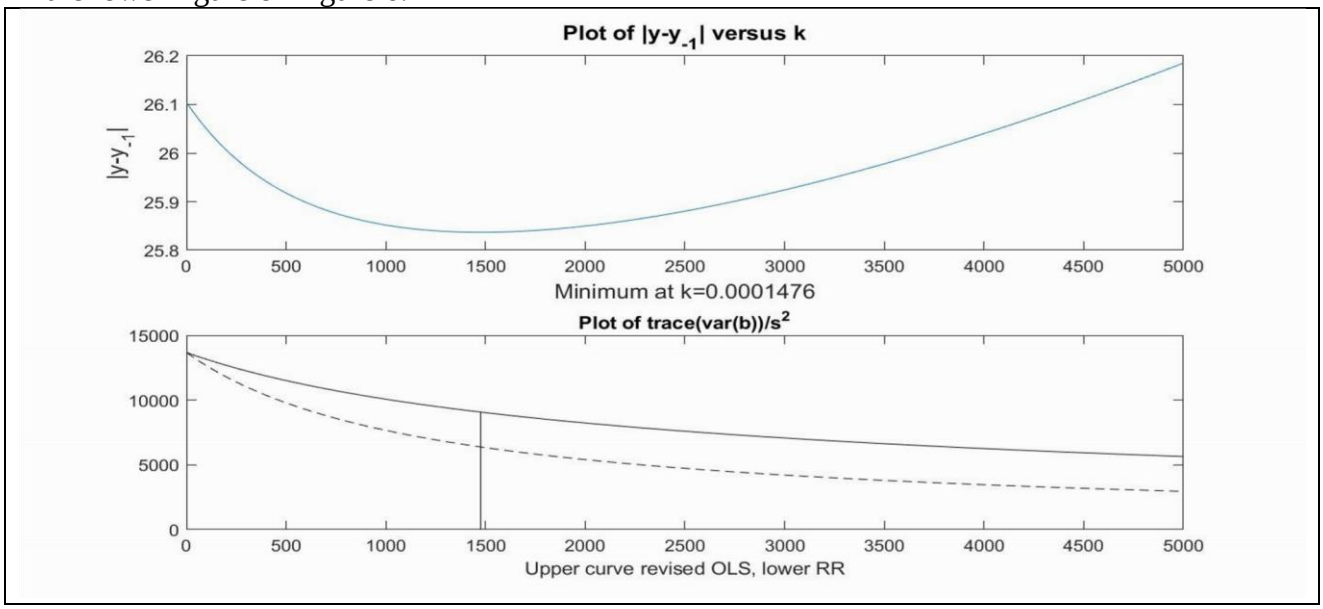
The chol.m subroutine requires that the matrix is non-singular. An error message is given, if the matrix is too close to being singular. The subroutine can be modified to allow zero diagonal elements in **F**.

From the algorithm it can be seen that the number of samples needs to be large enough, (N-1)>2K, where **X** is N×K.

Let us summarize the procedure. For a given value of the Ridge constant, k, a 'noise' matrix **Z** is added to **X** so that the OLS solution using

$\mathbf{X}_1=\mathbf{X}+\mathbf{Z}$ gives the same estimates of the regression coefficients as RR. **Z** has the property that it is orthogonal to **y**, $\mathbf{Z}^T\mathbf{y}=\mathbf{0}$. These results are illustrated by the spectral data, where the first 19 variables are used. (More than 19 variables gives an error message in Matlab, when using chol.m).

The upper figure in Figure 6 shows the values of $|\mathbf{y}-\hat{}_{est}|$ around the minimum value. $\hat{}_{est}$ is computed by Leave-one-out regression as explained earlier. The minimum is found at k=0.0001476. For this value of k, the matrix $\mathbf{X}_1$ is found. The sizes of OLS precision matrix for $\mathbf{X}_1$ and similar for RR (divided by s^2) are plotted in the lower figure of Figure 6.



Original Article

Figure 6. Upper figure a plot of $|\mathbf{y}-\hat{}_{est}|$ versus k. k=i×10^{-7}, i=1,2, … , 5000. x-axis is the i-values. Lower figure are plots of (22) and (23). Upper curve for the revised OLS, (22), lower curve for RR, (23).

We have $\boldsymbol{b}_O^* = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y}$ is the revised OLS solution. The upper curve is computed from

(22) $tr(Var(\boldsymbol{b}_O^*)/\sigma^2) = tr(\boldsymbol{X}_1{}^T\boldsymbol{X}_1)^{-1} = tr(\boldsymbol{X}^T\boldsymbol{X}+k\boldsymbol{I})^{-1} = \Sigma_{a=1}^K 1/(d_a^2+k)$

The lower curve from

(23) $tr(Var(\boldsymbol{b}_R)/\sigma^2) = \Sigma_{a=1}^K d_a^2/(d_a^2+k)^2$

Here (d_a) are the singular values of **X**. The difference between (22) and (23) is

(24) $k\Sigma_{a=1}^K 1/(d_a^2+k)^2$

Thus, (23) is always smaller than (22). The RR approach is somewhat not satisfactory. A linear model is assumed, $\mathbf{y}\sim N(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I})$, which might not be correct, because a low rank **X** places some restrictions on $\boldsymbol{\beta}$. The RR method does not use this model. It uses the same way (regularization) of computing the solution, but uses the model to compute the variances of the solution.

When working with data, it is clear that the OLS variances from the modified **X** are more reliable than those of RR. Thus, there is a clear indication of that the variances of the RR solution are too small. The size of the difference is given by (24).

12 Graphic analysis of data in RR

The same algorithm can be used for RR as for PLS, see Höskuldsson (2015). The only difference is that at entry we use $\mathbf{S}=\mathbf{X}^T\mathbf{X}+k\mathbf{I}$ as covariance matrix instead of $S=\mathbf{X}^T\mathbf{X}$ for PLS. This allows us to carry out graphic analysis of data for RR in much the same way as for PLS.

We shall use spectral data for illustration. The Ridge constant is k=0.0001476. The upper most two figures in Figure 7 show the y-values plotted against the first two score vectors. The explained y-variation for the first X-score vector $\mathbf{t}_1$ is 33.73% and for the second $\mathbf{t}_2$ it is 36.25%. A line through (0,0) is inserted ($b=(\mathbf{y}^T\mathbf{t}_1)/(\mathbf{t}_1{}^T\mathbf{t}_1)$ for the first line and $b=(\mathbf{y}^T\mathbf{t}_2)/(\mathbf{t}_2{}^T\mathbf{t}_2)$ for the second line). We use the scatter plots to study linearity, extreme samples and special features in data like scatters at small y-values (compared to detection limit) or large y-values (sometimes instrumental error). We may get a score vector that gives a better fit than the one $\mathbf{t}_1$ given here. However, the present score vector $\mathbf{t}_1$ is both extracting variation for **X** and explaining variation of **y**. This explains also that $\mathbf{t}_2$ describes more of the variation of **y** than $\mathbf{t}_1$.

Note, that the score vectors are not orthogonal for RR. However, here they are very close to being orthogonal, because the Ridge constant is so small.

In the middle two figures we consider the plots of loading vectors. We study the grouping of variables and the sizes of the loadings. The basis for the interpretation is the case, when $\mathbf{S}=\mathbf{X}^T\mathbf{X}$ is the correlation matrix and the rank is two. In this case $\boldsymbol{S} = \boldsymbol{p}_1\boldsymbol{p}_1^T + \boldsymbol{p}_2\boldsymbol{p}_2^T$. If points are far from zero and next to each other, we say that the associated variables are closely related. The points are shown with variable numbers to ease the interpretation. The reliability of this way of looking at the loading plots depends on the percentage explained. The higher they are the more reliable it is. For low percentages, like here, it is for guidance only.

The lowest two figures show scatter plots of score vectors. The first score vector $\mathbf{t}_1$ explains 56.50% of the variation of **X**, $\mathbf{t}_2$ 25.79% and $\mathbf{t}_3$ 5.54%. We are looking for special features in the X-samples, like for instance, groups, gaps and dependence between the score vectors. For process data, e.g., we sometimes see 'movements' of points. By numbering the points, we can see when and how samples 'develop'. Here again, we use the percentage as guidance for the conclusions. For instance, if we detect groups in data, when we look at e.g., the plot of the 6th score vector versus the 5th, the interpretation may not be reliable, if the percentages are small, say

less than 1%. We validate signals of grouping in data by appropriate cross-validation. It should be emphasized that in chemometrics the study of the score plots is important for learning to know the variation in data.
Same figures can be made by a PLS analysis. The figures are almost alike, because the Ridge constant is so small, k=0.0001476.

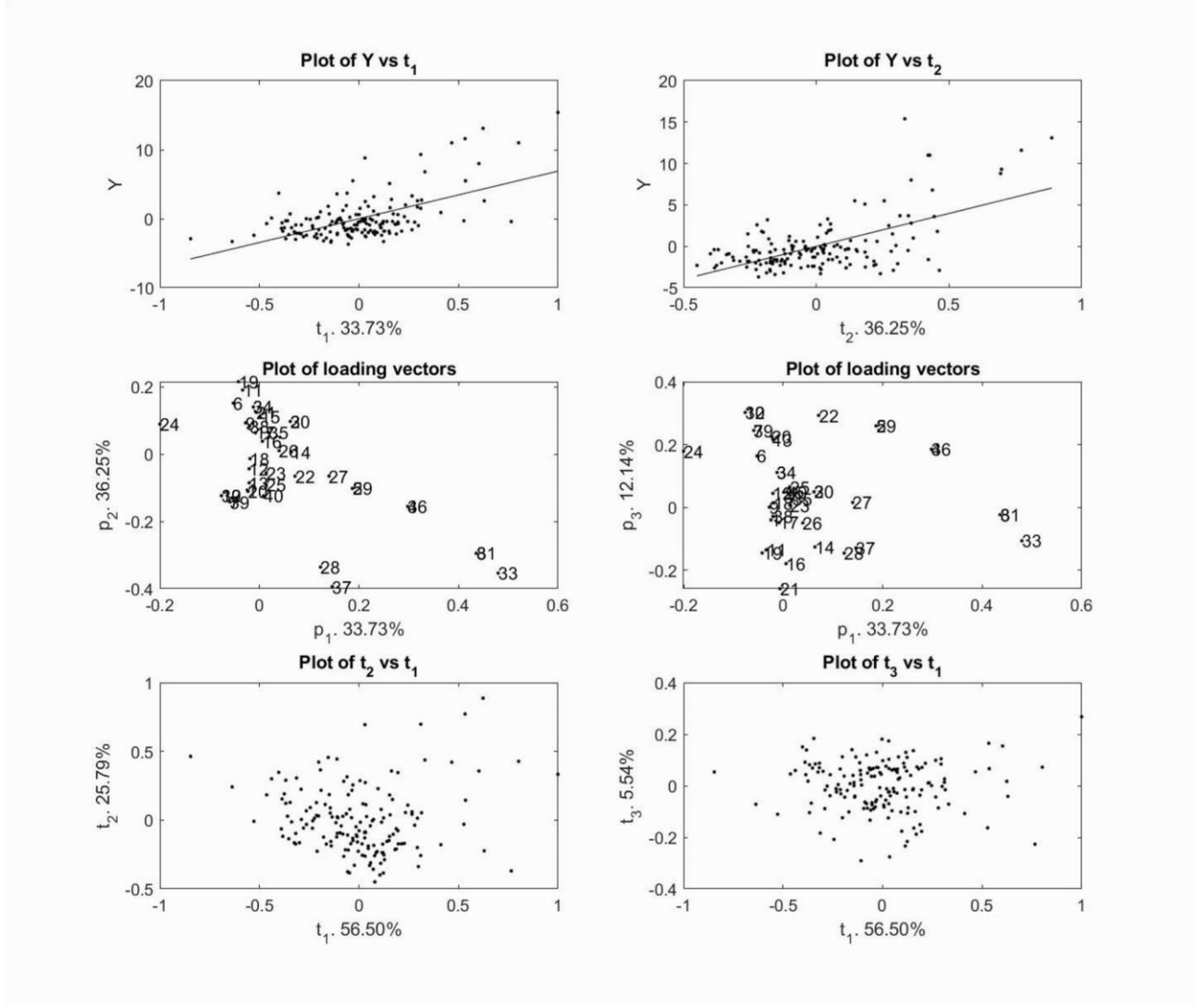


Figure 7. Ridge Regression. Upper most two plots are plot of **y** versus the first two score vector, $\mathbf{t}_1$ and $\mathbf{t}_2$. The middle two plots are scatter plots of the loading vectors, $\mathbf{p}_2$ vs $\mathbf{p}_1$ and $\mathbf{p}_3$ vs $\mathbf{p}_1$. The lowest two figures are the scatter plots of score vectors, $\mathbf{t}_2$ vs $\mathbf{t}_1$ and $\mathbf{t}_3$ vs $\mathbf{t}_1$.

13. Discussion

The data sets that have been used here are typical in chemometric work. The process data have a 25 variables and practical rank 13. The spectral data have 40 variables and rank around 19.
When comparing RR with PLS we use the average of 20 10-fold cross-validation. The average cross-validated **y**, $\mathbf{y}_c$, is a stable measure. When RR and PLS are used for these data, we do not find significant difference. The values of R_c^2 and s_c at average cross-validation are approximately equal. The reason for that the values obtained by RR and PLS are almost equal, is that the values of RR, when SVD of **X** is used, do not change at the dimension

used for PLS, as shown in Table 1. When applied to test set, which is 20% of the data, the values of R_t^2 and s_t are often similar. However, these numbers can be different for these two methods. RR and PLS were applied to stepwise selection/deletion of variables, which is based on average cross-validation. Here we find that RR performs equally well as PLS. When applied to test sets, there can be some difference in the results of RR and PLS. In PLS the largest value of R_c^2 may not be appropriate for test sets.

In industry, selection of variables is an important issue. $\mathbf{y}_c$, which is the average of $\mathbf{y}_{c,i}$'s from 20 cross-validations, is a stable magnitude. By using it in variable selection/deletion, we get efficient methods for finding the variables that should be used.

A test set that is 20% of data is relatively large. It is known that for spectral data there are uncertainties in the X-values. Therefore, we may expect slightly different results for different test sets. RR can be carried out by the same algorithm as PLS. The Ridge constant k is added to the diagonal of the covariance matrix before analysis. It is the experience in chemometrics that one should not use the full rank solution. This indicates that one should study the feasibility of using a solution that is not of full rank. It is a disadvantage of RR to use terms in (5) that have very small or zero singular values, or small score vectors.

We see from Figure 1 that the total variance (8) can be very sensitive to the choice of k. Here, a small increase in k may give (8) close to zero. It is shown that RR amounts to adding small 'noise' values to **X**. The OLS solution of the modified **X** gives the same solution as RR. The theory of OLS confirms the impression from the empirical work that the total variance (8) is too small.

14. Conclusion

We have studied RR and PLS for data that are typical in chemometric work. There is a unique Ridge ⌢constant k, that gives the minimum value of $|\mathbf{y}\text{-}\hat{}_{est}|$, where $_{est}$ is obtained by Leave-one-out RR. The Ridge constant k obtained in this way is typically very small. When maximal value of R_c^2 for the average of 20 cross-validations, $\mathbf{y}_c$, the results obtained by RR and PLS are close to equal. This also holds, when RR and PLS are used for variable selection/deletion, and dimension of PLS are at the maximal values of R_c^2. It is shown that RR amounts to adding small 'noise' values to **X**. OLS applied to the modified **X** gives the same solution as RR. The theory of OLS tells us that the theory of RR cannot be applied to data. We find RR efficient in modelling chemometric data. It also efficient in variable selection/deletion procedures. However, we cannot recommend using the theory of RR in analyzing the parameter estimates.

References

S.C. Basak, D. Mills,D.M. Hawkins &H.A. El-Masri (2002), Prediction of tissue-air partition coefficients: A comparison of structure-based and property-based methods, SAR and QSAR in Environmental Research, Vol 13, Issue 7-8, pp 649-665 https://doi.org/10.1080/1062936021000043409

Frank, I. E. and Friedman, J. H. (1993), A statistical view of some chemometrics regression tools" (with discussion), Technometrics 35, 109–135. doi:10.1080/00401706.1993.10485033

Höskuldsson A., O. Rodionov and A. Pomerantsev (2006). "Path modeling and process control", Chemometr. Intell. Lab. Syst. 76, 257–269.

Höskuldsson A., (2015). "A common framework for linear regression." Chemometr. Intell. Lab. Syst. 146 250–262. DOI: dx.doi.org/10.1016/j.chemolab.2015.05.022

Höskuldsson A., (2017). Application of the H-principle of Mathematical Modelling. Advances in Statistical Methodologies and Their Application to Real Problems, Chapter 3. DOI:10.5772/66153

Original Article

Irfan Muhammad, Maria Javed and Muhammad Ali Raza (2013), Comparison of Shrinkage Regression Methods for Remedy of Multicollinearity Problem, Middle-East Journal of Scientific Research 14 (4): 570-579, DOI: 10.5829/idosi.mejsr.2013.14.4.488

Jessen et al. (2014), "Simultaneous determination of glucose, triglycerides, urea, cholesterol, albumin and total protein in human plasma by Fourier transform infrared spectroscopy: Direct clinical biochemistry without reagents." Clinical Biochemistry, Volume 47, Issues 13–14, pp. 1306–1312.

doi:10.1016/j.clinbiochem.2014.05.064

S. Wold, H. Martens & H. Wold (1983), The multivariate calibration problem in chemistry solved by the PLS method, Conference paper. Matrix Pencils pp 286–293, DOI: 10.1007/BFb0062108