Chapter 3

Applications of the H-Principle of Mathematical Modelling

Agnar Höskuldsson

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/66153

Abstract

Traditional statistical test procedures are briefly reviewed. It is pointed out that significance testing may not always be reliable. The author has formulated a modelling procedure, the H-principle, for how mathematical modelling should be carried out in the case of uncertain data. Here it is applied to linear regression. Using this procedure, the author has developed a common framework for carrying out linear regression. Six regression methods are analysed by this framework, two stepwise methods: principal component regression, ridge regression, PLS regression and an H-method. The same algorithm is used for all methods. It is shown how model validation and graphic analysis, which is popular in chemometrics, apply to all the methods. Validation of the methods is carried out by using numerical measures, cross-validation and test sets. Furthermore, the methods are tested by a blind test, where 40 samples have been excluded. It is shown how procedures in applied statistics and in chemometrics both apply to the present framework.

Keywords: linear regression, common framework for linear regression, stepwise regression, ridge regression, PLS regression, H-methods

1. Introduction

Regression analysis is the most studied subject within theoretical statistics. Numerous books have been published. Advanced program packages have been developed that make it easy for users to develop and study advanced models and methods.

Advanced program packages such as SAS and SPSS have been used by students since the 1970s and the program packages have become very advanced. The user fills out a 'menu' similar to those at restaurants. The program then carries out the analysis as requested providing the users' possibility of carrying out highly advanced analysis of the data. Users rely upon that they can make interpretation of the output in the way they learned at school. For example,

if a variable/factor is significant, its presence in the model improves the predictions derived from the modelling results.

Today measurement instruments are becoming more and more advanced, e.g. optical instruments are becoming popular in industry. They may give thousands of values each time a sample is measured. In applied research, the tendency is to include as much information as possible in order to cover possible alternatives of the situation. When adding interactions or non-linear terms, we also see data having hundreds or thousands of variables.

There are three basic challenges in applied data analysis, which are as follows: (1) Data in applied sciences and industry are typically generated by instruments that produce many variables. In these cases, the data have **latent structure**, which means that the data values are geometrically located in a low-dimensional space. However, mathematical models and methods usually assume that the **X**-data have full rank. In these cases, the models are often incorrect and may give imprecise results.

(2) Typically, scientists develop solutions that are optimal or unbiased. This is a natural approach, because the derived solutions have important properties, when data satisfy the given model. Simulations based on the model confirm the optimality, but often in practice the data often do not satisfy the assumptions of the model. Forcing an optimal solution on the data may give results that have bad or no prediction ability. A better solution may then be obtained **by relaxing** on the optimality or unbiasedness of the solution.

(3) It is a tradition to base results of regression analysis on testing the significance of the variables in the model. However, the results may not always be reliable. The influence of a variable can be so small that it may be of no importance, see Section 3.3. At professional organizations, it is considered to be serious problem for the researchers in the field.

H-principle is a prescription of how a solution to a mathematical model should be generated, when data are uncertain. It is proposed that the determination of the solution should be carried out in steps and compute the improvement in the solution and the associated precision at each step. It is suggested that the solution at each step should be an optimal balance between the improvement and the associated precision. Determination of the solution continues as long as it is supported by the given (uncertain) data.

The author has developed a framework linear regression, which is inspired by the H-principle. Using this framework, the same algorithm is applied for carrying out different types of regression analysis. Most regression methods based on linear algebra can be carried out within this framework. Here, we carry out the analysis of six different methods. Numerical and graphic analyses of the results are the same for all the regression methods.

In Section 2, we specify the used notation and briefly describe the used data. In Section 3, we consider some basic issues in modelling data. The background for the H-principle is treated. In Section 4, we consider the latent variable regression model. In Section 5, we present the H-principle and show some examples of its usage. In Section 6, we present a common framework for linear regression. In Section 7, we discuss model validation. This is an important topic, but we only consider essential aspects that are used in the analysis of the six methods. In Section 8,

http://dx.doi.org/10.5772/66153

we present the results for six different regression methods: (1) stepwise (forward) regression maximising covariance, (2) stepwise (forward) regression maximising R^2,

(3) principal component regression, (4) ridge regression,

(5) PLS regression and (6) an H-method.

Section 9 briefly discusses the results presented. In Section 10, we mention application of the H-principle to multi-block and path modelling, non-linear modelling and extension to multi-linear algebra. Section 11 presents conclusions.

2. Notation, data and scaling

Matrices and vectors are denoted by bold letters; matrices by upper case and vectors by lower case. In order to facilitate the reading of the equations, different types of indices are used for the steps and for the matrices/vectors. The letters *a* and *b* are related to the steps in the algorithm. The letters *i, j* and *k* are used for the indices within a matrix/vector. It is assumed that there is given instrumental data **X**, N × K matrix, and response data **Y**, N × M matrix. The regression data are denoted by (**X, Y**). In some cases, only one *y*-variable, **y**, is used. This is done to simplify the equations. It is assumed that data are centred, which means that average values are subtracted from each column of **X** and **Y**. This also makes it easier to read the equations. $\mathbf{x}_j$ is the j^{th} column of **X** and $\mathbf{x}^i$ is the i^{th} row of **X** = (x_{ij}). A *latent variable* τ is a linear combination of the original measured variables,

$$\tau = a_1 x_1 + a_2 x_2 + \ldots + a_K x_K \tag{1}$$

The data used here are from an optical instrument. The instrument gives 1200 values at each measurement. However, technical knowledge of the instrument suggests that only 40 values should be used for determining the substance in question, the *y*-values. Two hundred samples are measured. Forty samples are put aside for a blind test. Thus, the calibration analysis is based on 160 samples. This gives **X** as 160 × 40 matrix and **y** a 160 vector. These data are challenging and represent a common situation, when working with optical instruments (FTIR, NIR, fluorescence etc.).

It is sometimes important to determine if data should be scaled or not. Scaling can be obtained by multiplying **X** and **Y** by diagonal matrices. If $\mathbf{C}_1$ and $\mathbf{C}_2$ are diagonal matrices, the scaling of **X** is done by the transformation, **X**←(**X**$\mathbf{C}_1$) and of **Y** by **Y**← (**Y**$\mathbf{C}_2$). The linear least squares solution for (**X, Y**) is given by $\boldsymbol{B} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}$. This solution can be obtained from the linear least squares solution for scaled data, $\mathbf{B}_1$, as follows,

$$\boldsymbol{B} = \boldsymbol{C}_1[\{(\boldsymbol{X}\boldsymbol{C}_1)^{\boldsymbol{T}}(\boldsymbol{X}\boldsymbol{C}_1)\}^{-1}(\boldsymbol{X}\boldsymbol{C}_1)^{\boldsymbol{T}}(\boldsymbol{Y}\boldsymbol{C}_2)]\boldsymbol{C}_2^{-1} = \boldsymbol{C}_1\boldsymbol{B}_1\boldsymbol{C}_2^{-1} \tag{2}$$

This shows that when computing the linear least squares solution, we can work with the scaled data. The original solution is obtained by scaling 'back' as shown in the equation. We use this property also, when we compute the approximate solution. The effect of scaling is better

numerical precision. Scaling is necessary for the present data. If data are not scaled (e.g. to unit variance), numerical results beyond dimension of around 15 may not be reliable. The reason is that Eqs. (20) and (23) are sensitive to small numerical values. Scaling is much debated among researchers. When scaling is used, one must secure that all variables have values above the 'noise' level of the instrument. This is a difficult topic, which is not considered closer here. If original data follow a normal distribution, the scaled ones will not. However, for large sample number like given here, the differences will be negligible.

3. Linear regression

3.1. Traditional regression model

It is assumed that there are given instrumental data **X** and response data **y**. A linear regression model is given by

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon \tag{3}$$

The x-variables are called independent variables and the y-variable is the dependent one. When the parameters **β** have been estimated as **b**, the estimated model is now

$$y = b_1 x_1 + b_2 x_2 + \ldots + b_K x_K \tag{4}$$

When there is given a new sample $\mathbf{x}_0 = (x_{10}, x_{20}, \ldots, x_{K0})$, it gives the estimated or predicted value $y_0 = b_1\ x_{10} + \ldots + b_K\ x_{K0}$. There can be many estimates for the regression coefficients. Therefore, Greek letters are used for the theoretical parameters and Roman letters for the estimated values. It is common to use the linear least squares method for estimating the parameters. It is based on minimizing the residuals, $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \rightarrow$ minimum, with respect to **β**. The solution is given by

$$\boldsymbol{b} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}. \tag{5}$$

It is common to assume that the residuals in Eq. (3) are normally distributed. This is often written as $\mathbf{y}\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$. It means that the expected value of **y** is $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and the variance is $Var(\mathbf{y}) = \sigma^2\mathbf{I}$, where **I** is the identity matrix. The linear least squares procedure coincides with the maximum likelihood method in case the data follow a normal distribution. Assuming the normal distribution, the parameter estimate **b** has the variance given by

$$Var(b) = \sigma^2(X^TX)^{-1}. \tag{6}$$

Here σ^2 is estimated by

$$\sigma^2 \cong \Sigma_1^N e_i^2/(N-K) \tag{7}$$

where the residuals, e_is, are computed from $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$.

http://dx.doi.org/10.5772/66153

3.2. Assumptions and properties

It is assumed that the matrix $\mathbf{X}^T\mathbf{X}$ has an inverse. Furthermore, it is assumed that the samples are independent of each other. In case the model (Eq. (3)) is correct and data follow the normal distribution, the estimate (Eq. (5)) has some important properties compared with other possible estimates:

a. It is unbiased, $E(\mathbf{b}) = \boldsymbol{\beta}$.

b. The estimate (Eq. (5)) has among all linear unbiased estimates of $\boldsymbol{\beta}$ the smallest variance. It means that if $\mathbf{b}_1$ is another unbiased estimate of $\boldsymbol{\beta}$, then $Var(\mathbf{b}_1) - Var(\mathbf{b})$ is a semi-positive definite matrix.

There is a general agreement that the variance matrix $Var(\mathbf{b})$ should be as small as possible. The matrix $(\mathbf{X}^T\mathbf{X})^{-1}$, the *precision matrix*, shows how precise the estimates $\mathbf{b}$ are. The properties (a) and (b) state that assuming the linear model (Eq. (3)), the solution (Eq. (5)) is, in fact from a theoretical point of view, the best possible one. It often occurs that the model (Eq. (3)) contains many variables. Interpretation of the estimation results (Eq. (4)) is an important part of the statistical analysis. Therefore, it is common to evaluate the parameters with the aid of a significance test. If a variable is not significant, it may be excluded. A parameter associated with a variable is commonly evaluated by a *t*-test. A *t*-test of a parameter b_i is given by

$$t = \frac{b_i}{\sqrt{Var(b_i)}} \cong \frac{b_i}{\sqrt{(s^2 \times s^{ii})}} \tag{8}$$

Here s^2 is given by Eq. (7) and s^{ii} is the ith diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$.

This is the motivation that the program packages such as SAS and SPSS in statistics compute the linear least squares solution in the case of linear regression analysis as the initial solution. Significance of the parameters in Eq. (4) is shown using Eq. (8).

The problem in using the least squares solution appears when the precision matrix becomes close to singular. If the computer program (SAS or SPSS) detects that the computation of $\mathbf{b}$ may be imprecise due to close to singularity of the precision matrix, the user is informed that the estimates $\mathbf{b}$ may be imprecise. However, the information is related to the judgement of the situation and the numerical precision of the computer. There are practical issues long before the question of the numerical uncertainties of the estimates, $\mathbf{b}$'s.

Consider two examples. Suppose that $\mathbf{X}$ is $N \times 2$ and that $\mathbf{x}_1 = c\,\mathbf{x}_2 + \boldsymbol{\delta}$, for some constant c, where $|\boldsymbol{\delta}| < 10^{-5}$. We may be able to compute $(\mathbf{X}^T\mathbf{X})^{-1}$. However, inference on the model (Eq. (4)) may be uncertain or incorrect. At the other example, suppose that $\mathbf{X}$ is $N \times 40$, but that practical rank is 15. Assume that $\mathbf{X}^T\mathbf{X} = \mathbf{X}_1{}^T\mathbf{X}_1 + \mathbf{X}_2{}^T\mathbf{X}_2$, where $\mathbf{X}_2 = (\mathbf{x}_{2,1}\ \mathbf{x}_{2,2}\ \ldots\ \mathbf{x}_{2,40})$. If all $\mathbf{x}_{2,i}$ $i = 1,\ldots,40$ are small, say $|\mathbf{x}_{2,i}| < 10^{-5}$, inference from Eq. (4) may be uncertain or incorrect. The first example may not be realistic. However, the second one is realistic. It is common that data in applied sciences and industry are of reduced practical rank. In these cases, the model (Eq. (3)) is incorrect and leads to uncertain or incorrect results.

3.3. Stepwise linear regression

We shall consider closer the procedure of stepwise linear regression. We do that by using the Cholesky factorization $\mathbf{X}^T\mathbf{X} = \mathbf{F}\mathbf{F}^T$, where $\mathbf{F}$ is lower triangular. Then we can write the columns of the data matrix $\mathbf{X}$ as

$$\begin{aligned} \boldsymbol{x}_1 &= F_{11}\boldsymbol{t}_1 \\ \boldsymbol{x}_2 &= F_{21}\boldsymbol{t}_1 + F_{22}\boldsymbol{t}_2 \\ &\cdots \\ \boldsymbol{x}_K &= F_{K1}\boldsymbol{t}_1 + F_{K2}\boldsymbol{t}_2 + \ldots + F_{KK}\boldsymbol{t}_K \end{aligned} \tag{9}$$

The vectors ($\mathbf{t}_i$) are mutually orthogonal and of length 1, $\mathbf{t}_i{}^T\mathbf{t}_j = \delta_{ij}$. The significance of $\mathbf{x}_K$, when $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{K-1}$ are given, is computed from Eq. (8) with

$$b_K = \frac{\boldsymbol{y}^T(F_{KK}\boldsymbol{t}_K)}{F_{KK}^2} = \frac{(\boldsymbol{y}^T\boldsymbol{t}_K)}{F_{KK}} \text{ and } s^{KK} = 1/F_{KK}^2 \tag{10}$$

This gives

$$\boldsymbol{t} = \frac{(\boldsymbol{y}^T\boldsymbol{t}_K)}{s} \tag{11}$$

This shows that the *t*-test for the significance of $\mathbf{x}_K$ is independent of the size F_{KK}. When the selection of a new variable is carried out among many variables (e.g. 500), there is a considerable risk that F_{KK} is so small that the marginal effect of $\mathbf{x}_K$ is of no importance although it is being declared as statistically significant. The issue is that the user of program packages is not informed of, if a significant variable improves the prediction of the response variable or not.

3.4. Variance of regression coefficients

It is instructive to study closer the variance of the regression coefficient in the linear least squares model, Eq. (6). It can be written as

$$Var(\boldsymbol{b}) \cong [\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] \times [(\boldsymbol{X}^T\boldsymbol{X})^{-1}]/(N-K) \tag{12}$$

An important objective of the modelling task is to keep the value of Eq. (12) as small as possible. Equation (12) is a product of two parts. For these two parts it can be shown that, when a new variable is added to the model

i. the residual error, $[\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}]$, always decreases

ii. the precision matrix, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, always increases

(In theory these measures can be unchanged, but in practice these changes always occur). Note that the two terms, (i) and (ii), are equally important and appear in a symmetric way in Eq. (12). If the data are normally distributed, it can be shown that

a. the residual error, $[\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}]$,

b. the precision matrix, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$,

http://dx.doi.org/10.5772/66153

are stochastically independent. This means that the knowledge of one of them does not give any information on the other one. If, for instance, a certain fit has been obtained, then there is no information on (b). In order to know the quality of predictions, we must be compute (b) and use Eq. (12). Knowing (a), the results concerning the precision matrix can be good or bad. Program packages only use the *t*-test or some equivalent measure to evaluate to test significance of the variables/components. There is no information on the precision matrix. We must compute the precision matrix together with the significance testing in order to find out how well the model is performing. **However, a modelling procedure *must* include both (a) and (b) in order to secure small values of Eq. (12)**. The procedure must handle both the decrease of (i) and the increase of (ii) during the model estimation.

4. Latent variable model

The measurement data in applied sciences and industry often represent a 'system', where variables are dependent on each other from chemical equilibrium, action-reaction in physics and physical and technical balance. Geometrically, the samples are located in a low-dimensional space. In these cases, the model (Eq. (3)) is **incorrect** (apart from simple, designed laboratory experiments) and the nice theory above may not applicable. Why is this the case? When a new sample is available, we cannot automatically insert it into Eq. (4) and compute the *y*-value. We must secure that the new sample is located geometrically along the samples used for analysis, the rows of **X**. A correct model for this kind of data is

$$y = \alpha_1\tau_1 + \alpha_2\tau_2 + \ldots + \alpha_A\tau_A + \varepsilon. \tag{13}$$

Here (τ_a) are *latent variables* and (α_a) are the regression coefficients on the latent variables. The regression task is both to determine the latent variables, τ's, and to compute the associated regression coefficients. The resulting Eq. (13) is then converted into Eq. (4) using Eq. (26). The model (Eq. (13)) is called **latent structure** linear regression. It is sometimes difficult to make 'good' interpretation of the variables, when using latent variables. However, by using appropriate modelling procedure, it is possible to obtain a fairly good interpretation of the variables (see Section 5.4).

5. The H-principle of mathematical modelling

5.1. Background

In the 1920s, there were large discussions on the measurement aspects of quantum mechanics. W. Heisenberg pointed out that there are certain magnitudes that cannot be determined exactly at the same time. His famous uncertainty inequality states that there is a lower limit to how well conjugate magnitudes can be determined at the same time. The position and momentum of an elementary particle is an example. For that example, the inequality is

$$\Delta(\text{position}) \times \Delta(\text{momentum}) \geq \text{constant} \tag{14}$$

The lower limit of the inequality is related to the Planck's constant of light. These considerations are based on physical theory. In practice, it means that there are some restrictions on the outcome of an experiment. The results may depend on the instrument and the phenomenon being studied. The restrictions are detected by the application of the measurement instrument. The uncertainty inequality and the associated theory are a kind of guidance, when carrying out experiments.

5.1.1. The H-principle

When modelling data, there is an analogous situation. Instead of a measurement instrument we have a mathematical method. The conjugate magnitudes are ΔFit and ΔPrecision and they cannot be controlled at the same time. It is recommended to carry out the modelling in steps and at each step evaluate the situation as prescribed by the uncertainty inequality. At each step, it is assumed that there is given a weight vector. The recommendations of the H-principle of mathematical modelling are the following (expressed in the case of regression analysis):

1. Carry out determining the solution in steps. You specify how you want to look at the data at this step by formulating how the weights are computed.
2. At each step compute
   a. expression for improvement in fit, Δ(Fit)
   b. the associated prediction, Δ(Precision)
3. Compute the solution that minimises the product Δ(Fit) × Δ(Precision)
4. In case the computed solution improves the prediction abilities of the model, the solution is accepted. If the solution does not provide this improvement, the modelling stops.
5. The data are adjusted for what has been selected; restart at 1.

Note that the H-principle is a recommendation of how to proceed in determining the solution of a mathematical problem, when data are uncertain. It is not like maximum likelihood, where a certain function is to be maximized. However, it is necessary to compute the improvement in the solution at each step and the associated precision. The optimal balance is described in Section 3 and the situation is evaluated to find out if data are following along.

5.1.2. Application to linear regression

Let us consider closer how this applies to linear regression. The task is to determine a weight vector **w** according to this principle. For the score vector $\mathbf{t} = \mathbf{Xw}$ we have

a. Improvement in fit: $|\boldsymbol{Y}^T\boldsymbol{t}|^2/(\boldsymbol{t}^T\boldsymbol{t})$
b. Associated variance: $\Sigma/(\boldsymbol{t}^T\boldsymbol{t})$

If $\mathbf{S} = \mathbf{X}^T\mathbf{X}$ the adjustment (Eq. (20)) gives orthogonal score vectors, **t**'s. Therefore, the variance, (*b*), is used for the precision. Treating **Σ** as a constant the task is to maximize

$$\left[\frac{|\boldsymbol{Y^T t}|^2}{\boldsymbol{t^T t}}\right] \times \frac{\boldsymbol{1}}{\left[\frac{1}{\boldsymbol{t^T t}}\right]} = \boldsymbol{w^T X^T Y Y^T X w} \tag{15}$$

This is a maximization task because improvement in fit is negative. The maximization is carried out under the restriction that **w** is of length 1, |**w**|=1. By using the Lagrange multiplier method, it can be shown that the task is an eigenvalue task,

$$\boldsymbol{X^T Y Y^T X w} = \lambda \boldsymbol{w} \tag{16}$$

In case there is only one *y*-variable, the eigenvalue task has a direct solution

$$w = \frac{X^T y}{|X^T y|} = \frac{C}{|C|} \tag{17}$$

These are the solutions used in PLS regression. Thus, we can state that the PLS regression is consistent with the H-principle. H-methods take as a starting point the PLS solution and determine how the prediction aspect or the solution can be improved.

5.2. Interpretation of the modelling results

The fit obtained by the score vector **t** is

$$\frac{(\boldsymbol{y^T t})^2}{(t^T t)} = \frac{(\boldsymbol{C^T C})^2}{\boldsymbol{C^T S C}} = \left[\frac{c_1^2}{f} + \frac{c_2^2}{f} + \ldots + \frac{c_K^2}{f}\right]^2 \tag{18}$$

where $f = \sqrt{\boldsymbol{C^T S C}}$

The regression coefficient can be written similarly as

$$\frac{(\boldsymbol{y^T t})}{(\boldsymbol{t^T t})} = \frac{(\boldsymbol{C^T C})^{1.5}}{\boldsymbol{C^T S C}} = \left[\frac{c_1^2}{g} + \frac{c_2^2}{g} + \ldots + \frac{c_K^2}{g}\right]^{1.5} \tag{19}$$

where $= (\boldsymbol{C^T S C})^{\frac{1}{1.5}}$.

At each step, we can see how much each variable contributes to the fit and the regression coefficient..

6. A common framework for linear regression

6.1. Background: views on the regression analysis task

There are many ways to carry out a regression analysis. Here we present a general framework for carrying out linear regression that includes most methods based on linear algebra. The basic idea is to separate the computations into two parts: the first part is concerned on how it is preferable to look at data. In practice, there can be different emphasis on the regression

analysis. Sometimes it is important to determine important variables. In other cases, it may be the predictions derived that is important. The other part is the numerical algorithm to compute the solution vector and associated measures. At the first part, it is assumed that a weight matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$ is given. Each weight vector should be of length one, although this is not necessary. The weight vectors ($\mathbf{w}_a$) specify how one wants to look at the data. They may be determined by some optimization or significance criteria. They may also be determined by a standard regression method such as PLS regression. In Section 6.4, we show the choices in the case of the six regression analysis. The role of the weight vectors is only to compute the loading vectors. The other part of the computations is a numerical algorithm, which is the same for all choices of **W**. There is no restriction on the weight vectors except that they may not give zero-loading vector, $\mathbf{p}_a \neq \mathbf{0}$.

6.2. Decomposition of data

The starting point is a variance matrix **S** and a covariance matrix **C. S** can be any positive semi-definite matrix, but **C** is assumed to be the covariance, $\mathbf{C} = \mathbf{X}^T\mathbf{Y}$. The algorithm is formulated for multiple *y*'s.

Initially $\mathbf{S}_0 = \mathbf{S}$ and $\mathbf{C}_0 = \mathbf{C}$ and $\mathbf{B} = \mathbf{0}$. For $a = 1, \ldots, K$:

$$\text{Loading vector} : \boldsymbol{p}_a = \boldsymbol{S}_{a-1}\boldsymbol{w}_a \tag{20}$$

$$\text{Scaling constant} : d_a = \frac{1}{\boldsymbol{w}_a^T\boldsymbol{p}_a} \tag{21}$$

$$\text{Loading vector} : \boldsymbol{q}_a = \boldsymbol{C}_{a-1}^T\boldsymbol{w}_a \tag{22}$$

$$\text{Loading weight vector} : \boldsymbol{v}_a \tag{23}$$

The loading weight vectors are computed by, see reference [1],

$$\boldsymbol{v}_1 = \boldsymbol{w}_1, \boldsymbol{v}_a = \boldsymbol{w}_a - d_1(\boldsymbol{p}_1^T\boldsymbol{w}_a)\boldsymbol{v}_1 - \ldots - d_{a-1}(\boldsymbol{p}_{a-1}^T\boldsymbol{w}_a)\boldsymbol{v}_{a-1}, a = 2, 3, \ldots, K \tag{24}$$

S is adjusted by the loading vector $\mathbf{p}_a$ and similarly for **C**,

$$\boldsymbol{S}_a = \boldsymbol{S}_{a-1} - d_a\boldsymbol{p}_a\boldsymbol{p}_a^T \tag{25}$$

$$\boldsymbol{C}_a = \boldsymbol{C}_{a-1} - d_a\boldsymbol{p}_a\boldsymbol{q}_a^T \tag{26}$$

The adjustment of **S** is of rank one reduction. **S** also reduces in size by

$$tr(d_a\boldsymbol{p}_a\boldsymbol{p}_a^T) = d_a\boldsymbol{p}_a^T\boldsymbol{p}_a = \frac{\boldsymbol{w}_a^T\boldsymbol{S}^2\boldsymbol{w}_a}{\boldsymbol{w}_a^T\boldsymbol{S}\boldsymbol{w}_a} > 0. \tag{27}$$

The loading weight matrix **V** satisfies $\mathbf{V}^T\mathbf{P} = \mathbf{D}^{-1}$, see reference [1].

6.3. Expansion of matrices

At $a = K$ we get $\mathbf{S}_K = \mathbf{0}$. Expanding **S** and **C** we get

$$\boldsymbol{S} = d_1\boldsymbol{p}_1\boldsymbol{p}_1^T + \ldots + d_A\boldsymbol{p}_A\boldsymbol{p}_A^T + \ldots + d_K\boldsymbol{p}_K\boldsymbol{p}_K^T = P\boldsymbol{D}P^T \tag{28}$$

$$C = \boldsymbol{X}^T\boldsymbol{Y} = d_1\boldsymbol{p}_1\boldsymbol{q}_1^T + \ldots + d_A\boldsymbol{p}_A\boldsymbol{q}_A^T + \ldots + d_K\boldsymbol{p}_K\boldsymbol{q}_K^T = \boldsymbol{PDQ}^T \tag{29}$$

Inserting appropriate matrices we get

$$\boldsymbol{S}^{-1} = d_1\boldsymbol{v}_1\boldsymbol{v}_1^T + \ldots + d_A\boldsymbol{v}_A\boldsymbol{v}_A^T + \ldots + d_K\boldsymbol{v}_K\boldsymbol{v}_K^T = \boldsymbol{VDV}^T \tag{30}$$

$$\boldsymbol{B} = \boldsymbol{S}^{-1}X^TY = d_1\boldsymbol{v}_1\boldsymbol{q}_1^T + \ldots + d_A\boldsymbol{v}_A\boldsymbol{q}_A^T + \ldots + d_K\boldsymbol{v}_K\boldsymbol{q}_K^T = \boldsymbol{VDQ}^T \tag{31}$$

$$\widehat{\mathrm{Y}}^T\widehat{\mathrm{Y}} = d_1\boldsymbol{q}_1\boldsymbol{q}_1^T + \ldots + d_A\boldsymbol{q}_A\boldsymbol{q}_A^T + \ldots + d_K\boldsymbol{q}_K\boldsymbol{q}_K^T = \boldsymbol{QDQ}^T \tag{32}$$

If $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$, we can expand **X** and **Y** in a similar way. Compute a score matrix **T** by $\mathbf{T} = \mathbf{XV}$. This gives

$$\boldsymbol{X} = d_1\boldsymbol{t}_1\boldsymbol{p}_1^T + \ldots + d_A\boldsymbol{t}_A\boldsymbol{p}_A^T + \ldots + d_K\boldsymbol{t}_K\boldsymbol{p}_K^T = \boldsymbol{TDP}^T \tag{33}$$

$$\widehat{\mathrm{Y}} = d_1\boldsymbol{t}_1\boldsymbol{q}_1^T + \ldots + d_A\boldsymbol{t}_A\boldsymbol{q}_A^T + \ldots + d_K\boldsymbol{t}_Kq_K^T = \boldsymbol{TDQ}^T \tag{34}$$

The score vectors are mutually orthogonal. This follows from $\mathbf{T}^T\mathbf{T} = \mathbf{D}^{-1}$. Generally, only A terms of the expansions are used, because it is verified that the modelling task cannot be improved beyond A terms. For the proof of geometric properties of the vectors in these expansions, see reference [1].

It can be recommended to use a test set, ($\mathbf{X}_t$, $\mathbf{Y}_t$), when carrying out a regression analysis. Centring (and scaling if used) is done on the test set by using the corresponding values from (**X, Y**). The estimated y-values are computed as $\hat{\mathbf{Y}}_t = \mathbf{X}_t\,\mathbf{B}$ and the score vectors for the test set as $\mathbf{T}_t = \mathbf{X}_t\,\mathbf{V}$. It is often useful to study the plots, showing columns of **Y** against columns of **T** (in stepwise regression, the plots are called added variable plots). Similarly, we can plot the score vectors of the test set, the columns of $\mathbf{T}_t$, against columns of $\mathbf{Y}_t$.

6.4. Choices of weight vectors

The weight vectors for the six regression methods are as follows:

(i) Stepwise regression, maximize covariance:

$$w_a(ii) = 1 \text{ for } C_{ii} = \max_{i=1}^{K}|C_i| \text{ and } = 0 \text{ otherwise} \tag{35}$$

(ii) Stepwise regression, maximize R^2:

$$w_a(ii) = 1 \text{ for} \frac{(\boldsymbol{y}^T\boldsymbol{x}_{ii})^2}{(\boldsymbol{x}_{ii}^T\boldsymbol{x}_{ii})} = \max_{i=1}^{K}\frac{(\boldsymbol{y}^T\boldsymbol{x}_i)^2}{(\boldsymbol{x}_i^T\boldsymbol{x}_i)}, \text{ and } = 0 \text{ otherwise} \tag{36}$$

(iii) and (iv) Eigenvector of S:

$$\text{For } \boldsymbol{S} = \boldsymbol{UEU}^T,\ \boldsymbol{w}_a = \boldsymbol{u}_a \tag{37}$$

(v) PLS regression:

$$\boldsymbol{w}_a = \boldsymbol{C}_a / |\boldsymbol{C}_a| \tag{38}$$

where $\mathbf{C}_a$ is the reduced covariance.

(vi) H-method:

The weight vector of PLS regression is sorted with largest first, $\left(w_1^{(s)}, w_2^{(s)}, \ldots, w_K^{(s)}\right)$. Columns of $\mathbf{X}$ are rearranged to match this sorting, $\mathbf{X}^{(s)}$.

$$\boldsymbol{w}_{a,m} = \left(w_1^{(s)}, w_2^{(s)}, \ldots, w_m^{(s)}, 0, \ldots, 0\right) \tag{39}$$

The index m is chosen, which gives the best explained variation,

$$\boldsymbol{t}_i = \boldsymbol{X}^{(s)} \boldsymbol{w}_{a,i},\ \frac{(\boldsymbol{y}^T \boldsymbol{t}_m)^2}{(\boldsymbol{t}_m^T \boldsymbol{t}_m)} = \max_{i=1}^{K} \frac{(\boldsymbol{y}^T \boldsymbol{t}_i)^2}{(\boldsymbol{t}_i^T \boldsymbol{t}_i)} \tag{40}$$

For further details of this method, see reference [2]. It has been applied to different bio-assay studies, where there can be many variables (3000 or more). Comparisons with several other methods have been shown that this method is preferable to work within the bio-assay studies, see references [3, 4]. Several other methods to improve the PLS solution have been developed.

7. Model validation

7.1. Numerical measures

The Mallow's Cp value and Akaike's information measure are commonly presented as results in a regression analysis. Mallow's Cp value is given by

$$C_p = \frac{|\boldsymbol{y} - \hat{\boldsymbol{y}}_A|^2}{\sigma^2} + 2A - N \tag{41}$$

As σ^2 we use the residual variance, s_A^2, at the maximal number of steps in the algorithm. C_p is an estimation of '(total means squared error)/σ^2'. The interpretation of C_p is

i. it should be as small as possible.

ii. its value should be as close to A as possible

iii. deviations of C_p from A suggests bias

Akaike's information measure is given by

$$AIC_A = N(\log(s_A^2) + 1) + s(A + 1) \tag{42}$$

It is an information measure that states the discrepancy between the correct model and the one obtained at step A. The number of components, A, is chosen that gives the smallest value of Eq. (32).

Both C_p and AIC have the property that they are not dependent on the given linear model, [5, 6]. Therefore, they can be used for all six methods considered here.

A *t*-value for the significance of a regression coefficient is given by

$$t\text{–value} = \frac{(\boldsymbol{y}^T \boldsymbol{t}_A)}{s_A} \tag{43}$$

Here $\mathbf{t}_A$ is the Ath score vector of unit length. The significance, *p*-value, of the *t*-value can be computed using the *t*-distribution. Although the assumptions of a *t*-test are not valid for any of the six methods, it is useful to look at the significance. One can show that theoretically this value should be larger than 2 in order to be significant.

When comparing methods, it is useful to look at the estimate for the variance, $Var(\mathbf{b})$, of the regression coefficients. We compute

$$(\text{trace}(Var(\boldsymbol{b}_A)))^{1/2} = (s_A^2 \sum_{a=1}^{A} d_a (\boldsymbol{v}_a^T \boldsymbol{v}_a))^{1/2} \tag{44}$$

We cannot use this measure to determine the dimension. However, it is useful in comparing different methods.

7.2. The covariance

Eq. (15) is equivalent to the singular value decomposition of the covariance **C**. Therefore, it is suggested that the modelling of data should continue as long as the covariance is not zero. The dimension of a model can be determined by finding, when $\mathbf{C} = \mathbf{X}^T\mathbf{y} \cong \mathbf{0}$ for reduced matrices. One procedure is to study the individual terms, $(\mathbf{x}_i{}^T\mathbf{y})$. Assume that data can be described by a multivariate normal distribution with a covariance matrix $\boldsymbol{\Sigma}_{xy}$. Then, it is shown in reference [7] that the sample covariances $(\mathbf{x}_i{}^T\mathbf{y})/(N-3)$ are approximately normally distributed. If $\sigma_{xi,y} = 0$, it is shown in reference [7] that approximate 95% limits for the residual covariance, $(\mathbf{x}_i{}^T\mathbf{y})/(N-3)$, are given by

$$\pm 1.96\sqrt{N}\sigma_{xi}\sigma_y/(N-3) \cong \pm\, 1.96\sqrt{N}s_{xi}s_y/(N-3) \tag{45}$$

Thus, when modelling stops, it is required that all residual covariances should be within these limits. If $\sigma_{xi,y} = 0$, the distribution of the residual covariance approaches quickly the normal distribution by the central limit theorem. Therefore, it is a reliable measure to judge, if the covariances have become zero or close to zero.

Another approach is to study the total value, $\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y} = \Sigma_i(\mathbf{x}_i{}^T\mathbf{y})^2$. If the covariance matrix $\mathbf{\Sigma}_{xy}$ is zero, $\mathbf{\Sigma}_{xy} = \mathbf{0}$, the mean, $\mu = E\{(\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y})\}$, and variance, $\sigma^2 = Var\{(\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y})\}$, can be computed [8]. If the covariance is not zero, $\mathbf{\Sigma}_{xy} \neq \mathbf{0}$, it can be shown that $E\{(\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y})\} > \mu$. The upper 95% limit of a normal distribution $N(\mu,\sigma^2)$ is $\mu + 1.65\sigma$. This is used for mean and variance. In the analysis, it is checked if $\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y}$ is below the upper 95% limit (a one-sided test)

$$\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y} < \text{trace}(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{y}^T\boldsymbol{y})/N + 1.65\sqrt{2\text{trace}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{y}^T\boldsymbol{y})^2/N^2} \tag{46}$$

When this inequality is satisfied, there is an indication that modelling should stop.

In the analysis in Section 8, we use the *p*-value,

$$p\text{–value} = P\{\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y} > \mu | N(\mu, \sigma^2)\} \tag{47}$$

This analysis has been found useful for different types of regression analysis.

7.3. Cross-validation and test sets

In stepwise regression, a search is carried out to find the next best variable. In PLS regression, a weight vector $\mathbf{w}$ is determined that gives the maximal size of the *y*-loading vector $\mathbf{q}$. When search or optimization procedures are carried out there is a considerable risk of overfitting. Evaluation of the results by standard statistical significance tests may show high significance, while other measures may show that the results are not significant. The uncertainties of predictions of new samples depend on the location of the samples in question. The further away from the centre (average values), the larger the uncertainties are. On the other hand, it is important to have large values, both large *x*-samples and *y*-values, in order to obtain stable estimates. In chemometrics, these special features of prediction are well known. A 10-fold cross-validation is often used. The samples are divided randomly into 10 groups. Nine groups are used for calibration (modelling) and the results applied to the 10th group. This is repeated for all groups so that all *y*-values are computed by a model, $\mathbf{y}_c$, that is using 90% of the samples. Experience has shown that this procedure may not always function well. As an example, one can mention the case, where the *y*-values have very skew distribution. There can be big difference in the results from cross-validation depending on how well large *y*-values are represented in the groups. It may be necessary that each group has a similar 'profile' as the total set of samples. For the present data, the *y*-values show very skew distribution (log(log(*y*-values)) show a normal distribution). It may be necessary that each group is representative for the whole data. This can be achieved by *ordered cross-validation*. Here, the samples are ordered in some way, which reflects the variation or sizes in data. In a 10-fold ordered cross-validation, the first group of samples is number 1, 11, 21, etc. of the ordered samples. The second group of samples is number 2, 12, 22, etc. In this way, each of the 10 groups of samples is representative with respect to the chosen ordering. The cross-validation is now carried out in the usual way.

As a result of cross-validation, we compute

$$D_a = 1 - |\boldsymbol{y} - \boldsymbol{y}_c|^2/|\boldsymbol{y}|^2 \; for \; a = 1, 2, \ldots, A \tag{48}$$

D_a is computed for each dimension. In the analysis below sorted *y*-values are used in the ordered cross-validation. A test set for the analysis is also selected in a similar way. Samples

are ordered according to the first PLS score vector. 15% of the samples or 160 × 0.15 = 24 samples are used in the test set ($\mathbf{X}_t$, $\mathbf{y}_t$). Thus, the analysis in Section 8 is based on 136 samples. Equation (38) is also used, when applying results to a test set ($\mathbf{y} = \mathbf{y}_t$ and $\mathbf{y}_c = \mathbf{X}_t \times \mathbf{b}$).

8. Results for six different methods

8.1. Preliminary remarks

We have chosen here data that are challenging to work with. They are representative for many situations that industry and research projects find themselves in, when working with optical instruments. It is expected that large dimension is needed to model the data. The aim of the present work is to show how commonly used measures in applied statistics and popular chemometric procedures fit within the present framework. Therefore, we have chosen data, where there is not big difference between the six methods chosen here. We show that the different measures/procedures do not agree on what dimension should be used. Thus, the data analyst is choosing one set of measures that will typically be based on experience. It is a part of industry standard, see reference [9], to develop a calibration method. When the development work is finished, a blind test is carried out for 40 new samples. The data analyst should not know the new 40 reference values, y-values. The estimate for the new 40 samples should be evaluated by another person. This procedure is used here: the last 40 samples are put aside and the calibration analysis will be based on the other samples. When analysis has been completed for the six methods, we apply the results to the 40 samples. Optical instruments may give thousands of values for each measurement that is carried out (the one used here gives 1200 values). Experts often point out which part of the data should be used. Here it has been proposed to work with 40 variables, which all show significant correlation with the y-values. An upper limit for the dimension is chosen here to be 20. There is not a numerical problem in computing the precision matrix. The ratio between the largest and the smallest eigenvalue of $\mathbf{S}$ is $\lambda_1/\lambda_{40} = 9.0 \times 10^6$. However, experience suggests that the dimension should be less than 20.

8.2. Modelling results for six methods

In order to make readings of the following tables easier, we shall use the following headings for the tables.

i. C_p, Eq. (31), where σ^2 is the residual variance at dimension 20. For dimension 20, we get $C_p = 19$

ii. AIC, Eq. (32), where $s_A{}^2$ is the residual variance at dimension A

iii. the t-value Eq. (33), where s_A is the residual standard deviation at dimension A

iv. two-sided p-value for the t-test

v. the size of covariance $|\mathbf{X}^T\mathbf{y}|$ at dimension A

vi. one-sided p-value for the significance of (v), see Eq. (37)

vii. the estimate (Eq. (34)) for $\sqrt{\text{trace}(Var(\mathbf{b}_A))}$

viii. ordered cross-validation, samples ordered by *y*-values, Eq. (38)

ix. test set results, samples ordered by values of the first PLS score vector (Eq. (38)).

The improvement in fit, size of score vectors, etc. are not shown in the tables below. Only dimensions from 10 to 20 are shown. Dimensions from 1 to 9 are all significant for all methods.

8.2.1. Forward stepwise regression, maximal covariance

Here a variable is selected at each step that has maximal covariance $|(\mathbf{x}_i^T\mathbf{y})|$ for the reduced data (**Table 1**).

The 15th variable selection has the smallest AIC value and highly significant *t*-value. Therefore, it seems reasonable to choose 15 variables. Cross-validation and test set have reached the high level at dimension 10 or 11. This indicates some overfitting by choosing 15 variables.

8.2.2. Forward stepwise regression, maximal R^2-value

Here a variable is selected at each step that has maximal size of $(\mathbf{x}_i^T\mathbf{y})^2/(\mathbf{x}_i^T\mathbf{x}_i)$ for the reduced data (**Table 2**).

It seems appropriate to choose 16 variables. Seventeenth is at the boundary of being also significant. C_p for 16 variables indicates some bias present. Cross-validation and test set have reached high values at step 10, which indicates some overfitting.

8.2.3. Principal component regression

Here score vectors are selected that correspond to the associated eigenvalues (**Table 3**).

Score vector no 14 is not significant, and perhaps should be excluded. It seems appropriate to choose dimension 17 here. It is common to remove score vectors that are not significant.

No.	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
10	90.5	–1571.9	12.65	0.0000	0.0518	0.000	0.23	0.9577	0.9761
11	88.0	–1572.2	–1.78	0.0765	0.0186	0.033	0.24	0.9609	0.9747
12	81.4	–1575.7	2.52	0.0127	0.0121	0.000	0.27	0.9640	0.9744
13	62.1	–1590.0	4.10	0.0001	0.0111	0.000	0.31	0.9631	0.9760
14	29.9	–1618.3	–5.61	0.0000	0.0070	0.000	0.43	0.9626	0.9804
15	**18.3**	**–1629.1**	**3.64**	**0.0003**	**0.0053**	**0.002**	**0.46**	**0.9658**	**0.9786**
16	19.1	–1627.3	–1.09	0.2772	0.0022	0.617	0.50	0.9662	0.9797
17	18.4	–1627.1	1.63	0.1045	0.0017	0.664	0.61	0.9663	0.9804
18	20.1	–1624.4	–0.60	0.5525	0.0012	0.744	0.64	0.9657	0.9804
19	17.4	–1626.5	2.18	0.0307	0.0013	0.462	0.78	0.9652	0.9813
20	19.0	–1623.8	–0.59	0.5539	0.0006	0.806	0.97	0.9656	0.9811

Table 1. Nine measures at stepwise selection of variables having maximal covariance.

http://dx.doi.org/10.5772/66153

No	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
10	87.4	–1591.7	3.64	0.0004	0.0411	0.058	0.34	0.9668	0.9815
11	75.4	–1599.5	–3.24	0.0014	0.0419	0.033	0.33	0.9662	0.9814
12	60.2	–1610.5	–3.69	0.0003	0.0301	0.000	0.33	0.9645	0.9789
13	49.7	–1618.4	–3.24	0.0014	0.0052	0.622	0.36	0.9643	0.9798
14	33.8	–1632.0	4.01	0.0001	0.0088	0.028	0.86	0.9645	0.9807
15	30.3	–1634.3	–2.26	0.0249	0.0073	0.117	2.25	0.9663	0.9794
16	26.2	–1637.4	2.40	0.0175	0.0056	0.365	2.25	0.9677	0.9780
17	**23.8**	**–1638.9**	**–2.05**	**0.0422**	**0.0076**	**0.023**	**2.28**	**0.9701**	**0.9778**
18	22.7	–1639.1	1.76	0.0807	0.0026	0.541	2.38	0.9698	0.9789
19	20.2	–1640.9	2.10	0.0370	0.0011	0.821	3.33	0.9698	0.9823
20	19.0	–1641.3	1.79	0.0752	0.0021	0.541	4.72	0.9684	0.9836

Table 2. Nine measures at stepwise selection of variables having maximal increase in R^2.

No	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
10	854.5	–1250.7	–2.40	0.0175	0.0586	0.000	0.15	0.7984	0.9176
11	452.2	–1348.6	–11.08	0.0000	0.0555	0.000	0.16	0.8722	0.9521
12	421.2	–1356.5	3.24	0.0014	0.0322	0.000	0.20	0.8783	0.9498
13	171.2	–1464.5	–11.72	0.0000	0.0302	0.000	0.19	0.9334	0.9682
14	172.8	–1461.7	–0.48	0.6294	0.0108	0.000	0.23	0.9292	0.9674
15	97.6	–1509.7	–7.32	0.0000	0.0108	0.000	0.27	0.9421	0.9718
16	66.2	–1533.4	5.12	0.0000	0.0071	0.000	0.32	0.9508	0.9729
17	**27.3**	**–1568.8**	**–6.21**	**0.0000**	**0.0053**	**0.000**	**0.35**	**0.9566**	**0.9729**
18	29.2	–1565.8	0.36	0.7170	0.0023	0.000	0.43	0.9563	0.9721
19	31.1	–1562.8	0.25	0.7991	0.0023	0.000	0.54	0.9553	0.9722
20	19.0	–1574.8	3.75	0.0002	0.0023	0.000	0.67	0.9579	0.9743

Table 3. Nine measures at principal component regression.

However, it may not always be a good practice. The *t*-test is not a valid test and the score vectors beyond dimension 20 are so small that they have no practical importance.

8.2.4. Ridge regression

Here score vectors are selected that correspond to the associated eigenvalues of $\mathbf{S} = \mathbf{X}^T\mathbf{X} + k\mathbf{I}$. The value of k is estimated by leave-one-out cross-validation, see reference [1]. The optimal value of k is $k = 0.0002$ (**Table 4**).

It seems also appropriate here to choose dimension 17. The value of k is based on the full model. However, there is no practical difference between dimension 17 and a full model.

8.2.5. PLS regression

When working with PLS regression, each step is analysed closer. The score vectors associated with the test set are computed as $\mathbf{T}_t = \mathbf{X}_t \times \mathbf{V}$. The correlation coefficient between response values of the test set, $\mathbf{Y}_t$, and the 12th score vector is –0.058 and the associated *p*-value is 0.197. Therefore, we can conclude that the 12th score vector does not contribute to the modelling task. We can thus conclude that the dimension should be 11 (**Table 5**).

8.2.6. H-method, maximal R^2 *value*

Here, it can be recommended to use 13 components. When 13 have been selected, there is no covariance left (see steps (v) and (vi)) (**Table 6**).

No	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
10	782.7	–1250.6	–2.39	0.0176	0.0586	0.000	0.15	0.7984	0.9176
11	410.6	–1347.7	–1103	0.0000	0.0555	0.000	0.16	0.8722	0.9520
12	382.1	–1355.4	3.22	0.0015	0.0322	0.000	0.20	0.8783	0.9497
13	151.7	–1461.5	–11.59	0.0000	0.0302	0.000	0.19	0.9334	0.9681
14	153.3	–1458.6	–0.48	0.6345	0.0108	0.000	0.23	0.9292	0.9673
15	85.8	–1504.0	–7.10	0.0000	0.0108	0.000	0.28	0.9421	0.9717
16	57.9	–1525.8	4.93	0.0000	0.0071	0.000	0.32	0.9508	0.9731
17	**23.7**	**–1557.7**	**–5.90**	**0.0000**	**0.0053**	**0.000**	**0.35**	**0.9566**	**0.9734**
18	25.6	–1554.8	0.34	0.7345	0.0023	0.000	0.43	0.9563	0.9727
19	27.5	–1551.7	0.23	0.8168	0.0023	0.000	0.53	0.9553	0.9728
20	19.0	–1560.0	3.24	0.0014	0.0023	0.000	0.65	0.9579	0.9744

Table 4. Nine measures at ridge regression.

No	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
10	207.0	–1514.5	7.14	0.000	0.0296	0.000	0.11	0.9426	0.9738
11	**137.0**	**–1553.0**	**6.59**	**0.000**	**0.0116**	**0.000**	**0.14**	**0.9498**	**0.9735**
12	107.8	–1570.9	4.55	0.000	0.0100	0.000	0.17	0.9568	0.9725
13	67.3	–1600.5	5.75	0.000	0.0101	0.000	0.20	0.9587	0.9744
14	58.0	–1607.1	3.02	0.003	0.0024	0.563	0.26	0.9628	0.9754
15	39.7	–1622.5	4.22	0.000	0.0040	0.000	0.35	0.9656	0.9762
16	26.6	–1634.4	3.78	0.000	0.0024	0.088	0.43	0.9654	0.9775
17	22.4	–1637.7	2.43	0.016	0.0023	0.013	0.47	0.9671	0.9804
18	23.5	–1635.6	0.96	0.340	0.0005	0.849	0.52	0.9673	0.9804
19	21.1	–1637.3	2.08	0.039	0.0004	0.822	0.85	0.9679	0.9826
20	19.0	–1638.7	2.02	0.045	0.0006	0.641	1.12	0.9674	0.9816

Table 5. Nine measures at PLS regression.

No	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
10	243.3	–1488.5	8.86	0.000	0.0334	0.000	0.13	0.9499	0.9731
11	116.1	–1559.0	9.13	0.000	0.0168	0.000	0.18	0.9556	0.9759
12	75.4	–1587.5	5.65	0.000	0.0096	0.005	0.21	0.9576	0.9757
13	**68.3**	**–1591.9**	**2.65**	**0.009**	**0.0086**	**0.001**	**0.22**	**0.9592**	**0.9767**
14	38.1	–1617.4	5.33	0.000	0.0027	0.546	0.50	0.9615	0.9753
15	30.9	–1623.2	2.90	0.004	0.0040	0.020	0.56	0.9637	0.9764
16	26.5	–1626.6	2.45	0.015	0.0022	0.362	0.59	0.9654	0.9798
17	22.5	–1629.8	2.42	0.016	0.0018	0.275	0.63	0.9671	0.9823
18	20.8	–1630.7	1.91	0.058	0.0011	0.633	0.70	0.9674	0.9825
19	19.0	–1631.7	1.94	0.054	0.0008	0.691	0.81	09674	0.9815
20	19.0	–1630.8	1.41	0.161	0.0006	0.744	0.97	0.9672	0.9816

Table 6. Nine measures at H-method determining maximal R^2 value along the covariance.

	Stepw. max cov	Stepw. max R^2	PCR	Ridge regression	PLS regression	H-method Max R^2
$(1-\lvert y_n-\hat{y}_n\rvert^2/(y_n^T y_n))$	0.9346	0.9346	0.9330	0.9333	**0.9472**	0.9391
$\sqrt{Var(\mathbf{b})}$	0.46	2.28	0.35	0.35	**0.14**	0.22

Table 7. Comparison of results from six methods.

8.3. Evaluation of modelling results

When modelling has been carried out, industry standards recommend to apply the model to 40 new samples, ($\mathbf{X}_n$, $\mathbf{y}_n$). The estimated values are $\hat{\mathbf{y}}_n = \mathbf{X}_n\mathbf{b}$. $\mathbf{X}_n$ are the new X-values and $\mathbf{b}$ is the regression coefficients from the method. Results are shown in **Table 7**.

From **Table 7**, it can be seen that PLS regression is slightly better than the other methods. The table also shows the estimate of the standard deviations of the regression coefficients. It shows that the PLS regression has much smaller value than the other methods. In conclusion, the PLS regression can be recommended for determining y-values for future samples.

Note that the values in the first row in **Table 7** are smaller than those were obtained for cross-validation and test sets during the calibration analysis. It indicates that the new 40 samples deviate in some way from the 160 samples that were used in the analysis. This is not explored further here.

8.4. Confidence interval for regression coefficients

Approximate confidence intervals for regression coefficients can be obtained by the procedure presented in reference [10]. Let e_i be the ith residual obtained by a regression method. Define

$$h_i = \boldsymbol{x}^i \boldsymbol{S}_A^{-1} \boldsymbol{x}^{iT} \text{ and } r_i = \frac{e_i}{1-h_i},\ i = 1, 2, \ldots, N. \tag{49}$$

A new set of residuals e_i^* are defined by randomly sampling from the modified residuals $r_1, \ldots, r_N$. By repeating the generation of (e_i^*), a new set of regression coefficients are obtained. This can be repeated say 200 times to get a confidence interval for the regression coefficients.

9. Discussion

Interpretation of **Tables 1–6** reflects personal experience. Others may select the dimension in a different way. The main issue is that the dimension of all the selected methods is less than 20. If a full model including all 40 variables is estimated, we are overfitting by a dimension over 20. Regression coefficients become large and inference from the estimation not reliable.

The advantage of working with latent variable is that the first ones collect a large amount of variation. This is well known from principal component analysis. Interpretation of variables cannot be done directly in latent variable models. However, using Eqs. (17) and (18) we can set up equations that show how individual variables contribute to the fit and regression coefficients.

10. Applications of the H-principle

The H-principle has been extended to many areas of applied mathematics. It has in general been successful due to the presence of latent structure in data. The implementation of the H-principle is different from area to area. Furthermore, 'household' administration may be needed in order to keep the number of variables low (cf. Mallows theory). In many cases, it has opened up for new mathematics.

In reference [11], it has been extended to multi-block and path modelling, giving new methods to carry out modelling of organized data blocks. The importance of these methods is due to the fact that the methods of regression analysis are extended so that regression models are computed between data blocks. Methods of regression analysis can be used to evaluate relationships between data blocks. Complicated assumptions like we see in structural equations modelling are not needed.

Linear latent structure regression can be viewed as determining a low-dimensional hyperplane in a high-dimensional space. In reference [12], it is extended to finding low-dimensional second, third and higher order surfaces in latent variables. Deviations from linearity often appear as curvature for low and high sample values, which can be handled by these surfaces. In reference [13], it has been applied to non-linear estimation that may give good low-rank solutions, where full-rank regularized solutions do not give convergence.

In reference [14], it is extended to multi-linear algebra, where there are many indices in data, e.g. $\boldsymbol{X} = (x_{ijk})$ and $\mathbf{y} = (y_{ij})$. The basic issue in multi-linear algebra is defining the inverse. This is solved by defining directional inverses for each dimension. It makes it possible to extend

http://dx.doi.org/10.5772/66153

methods of ordinary matrix analysis to multi-linear algebra. These methods have been successfully applied to multi-linear data and to growth models.

The H-principle has been applied to several areas of applied statistics. Here we briefly mention a few.

In time series analysis and dynamic systems, the objective of modelling is both to describe the data and also to obtain good forecasts. Thus, the requirement to a latent variable is both that it describes **X** and also gives good forecasts. Traditional models only focus on the description of **X**. By requiring that the latent variables also should give good forecast, better models are obtained.

In pattern recognition and classification, the objective of modelling is both to obtain good description of each group of data (that are detected or given a-priori) and to get low-error rate of classification. There are different ways to implement the H-principle in these areas, which have been developed. Applications show that these methods are superior to those based on statistical theory (e.g. linear and quadratic discriminate analysis based on the normal distribution, principal component analysis of each group of data).

11. Conclusion

We have presented a short review of standard regression analysis. Theoretically, it has important properties, which makes it a standard approach in popular program packages. However, we show that the results obtained may not always be reliable, when there is a latent structure in data. This is a serious problem in industry and applied sciences, because data typically have latent structure.

The H-principle is formulated in close analogy to the Heisenberg uncertainty inequality. It suggests that in the case of uncertain data, the computation of the mathematical solution should be carried out in steps, where at each step an optimal balance between the fit and associated precision should be obtained. A general framework is presented for linear regression. Any set of weight vectors can be used that do not give loading vectors of zero size. Using the framework, six different regression methods are carried out. It is shown that the methods give low rank solutions. The same type of numerical and graphic analysis can be carried out for any type of regression analysis within this framework. Traditional analysis in applied statistics and graphic analysis, which is popular in chemometrics, can be carried out for each method that uses the framework.

The algorithm can be viewed as an approximation to the full rank solution. Modelling stops, if further steps are not supported by data. Dimension measures, cross-validation and test sets are used to identify, when the steps are supported by data.

Acknowledgements

The cooperation with Clinical Biochemistry, Holbæk Hospital, Denmark, in the Sime project is highly appreciated. The author appreciates the use of the data from the project in the present article.

Author details

Agnar Höskuldsson

Address all correspondence to: ah@agnarh.dk

Centre for Advanced Data Analysis, Denmark

References

[1] Höskuldsson A, A common framework for linear regression, *Chemometrics and Intelligent Laboratory Systems* 146 (2015) 250–262. DOI: 10.1016/j.chemolab.2015.05.022

[2] Reinikainen SP, Höskuldsson A, COVPROC method: strategy in modeling dynamic systems, *Journal of Chemometrics* 17 (2003) 130–139. DOI: 10.1002/cem.770

[3] McLeod G, et al. A comparison of variate pre-selection methods for use in partial least squares regression: a case study on NIR spectroscopy applied to monitoring beer fermentation, *Journal of Food Engineering* 90 (2009) 300–307. DOI: 10.1016/j.jfoodeng.2008.06.037

[4] Tapp HS, et al., Evaluation of multiple variate methods from a biological perspective: a nutrigenomics case study, *Genes Nutrition* 7 (2012) 387–397. DOI: 10.1007/s12263-012-0288-4

[5] *https://en.wikipedia.org/wiki/**Mallows's_Cp***

[6] *https://en.wikipedia.org/wiki/**Akaike_information_criterion***

[7] Siotani M, Hayakawa T, Fujikoshi Y, *Modern Multivariate Analysis: A Graduate Course and Handbook*. American Science Press: Columbus, Ohio, 1985.

[8] Höskuldsson A: *Prediction Methods in Science and Technology*, Vol. 1, Thor Publishing: Copenhagen, 1996, ISBN 87-985941-0-9

[9] Clinical and Laboratory Standards Institute, *http://shop.clsi.org/chemistry-documents/*

[10] Davison AC, Hinkley DV: *Bootstrap Methods and their Application*, Cambridge University Press: Cambridge, New York, 1997.

[11] Höskuldsson A, Modelling procedures for directed network of data blocks, *Chemometrics and Intelligent Laboratory Systems* 97 (2009) 3–10. DOI: 10.1016/j.chemolab.2008.09.002

[12] Höskuldsson, A. The Heisenberg modelling procedure and applications to nonlinear modelling. *Chemometrics and Intelligent Laboratory System* 44 (1998) 15–30. DOI: 10.1016/S0169-7439(98)00111-7

[13] Höskuldsson A, H-methods in applied sciences, *Journal of Chemometrics* 22 (2008) 150–177. DOI: 10.1002/cem.1131

[14] Höskuldsson A, Data analysis, matrix decompositions and generalized inverse, *SIAM Journal on Scientific Computing*, 15 (1994) 239–262. DOI: 10.1137/0915018