Basic issues in modelling industrial data

In recent years there have been intensive developments in data acquisition instruments. Many of these instruments are within the optical and radio areas. In the optical area there are many new and advanced types of instruments. An example of a popular instrument is a NIR (Near Infra-Red) instrument. One sample obtained by the instrument gives typically 1056 values (absorbance at specific wavelengths), corresponding to 1056 variables, but there are NIR instruments that can give say, 8000 values as a result of one sample. Other examples of optical instruments are IR, NMR, RAMAN. Typically for these instruments are that they give thousands of values for each sample they measure. This situation is appearing more and more clearly in applied sciences. This is creating ***a shift in the paradigm in applied sciences***. The traditional methods within statistical and numerical sciences to analyse data characterised by thousand of variables are not very efficient. The weaknesses of these sciences in analysing these types of data are of different kinds, some of which are considered closer here.

In order to meet the challenges of these new developments in the data acquisition area there has been developed a new methodology to handle these types of situations. It has been called the H-methods. The name has been chosen because of the close analogy with the Heisenberg uncertainty principle in quantum mechanics. When modelling data with many variables the mathematical model is the tool we are working with. In the modelling steps the mathematical model starts to interfere with the data in a similar way as prescribed in the Heisenberg uncertainty principle.

The H-methods are recommendations of how we should solve mathematical models, when data are uncertain. The basic idea is to carry out the modeling task in steps, where at each step we seek find an optimal balance between the fit (or an improvement of the solution) and the associated prediction. Associated with the H-methods there have been developed very general algorithms that build up solutions in terms of rank one parts. Each of the parts is a result of optimization task involving fit and precision, such that all parts are in certain sense optimal at the respective step of the analysis. The H-methods provide with a conceptual basis for some of the chemometric methods. The procedure can also be applied to most numerical methods for data analysis, which are based on multivariate analysis, to judge the performance of the solutions that these methods give.

The H-methods have been developed by the author since 1992 and applied to different types of industrial data. The methodology has been used to develop new algorithms and analysis methods to handle industrial data containing many variables. Some of the basic issues of modelling data are discussed in the light of these developments.

When modeling industrial data, there are certain modeling issues that are fundamental for successful analysis. The important ones are the following five.

1. Prediction variance. The primary objective in industry is the prediction associated with new samples. If we assume standard regression analysis the prediction variance of a response value $y(\mathbf{x}_0)$ associated with a new sample $\mathbf{x}_0$ is given by

$$\mathrm{Var}(y(\mathbf{x}_0))=\sigma^2(1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)$$

$$\cong |\mathbf{y}-\hat{\mathbf{y}}|^2\,(1+\mathbf{x}_0^T\,(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{x}_0)\,/(\mathrm{N-K})$$

This equation shows that there are two objectives of modelling, both the residual variance $s^2=|\mathbf{y}-\hat{\mathbf{y}}|^2\,/(\mathrm{N-K})$ and the model variation that is given by $(1+\mathbf{x}_0^T\,(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{x}_0)$. A successful modelling task must have analysed both parts of the prediction variance.

2. Fit and precision. It is a fundamental result of multivariate statistics that the least squares fit, $|\mathbf{y} - \hat{\mathbf{y}}|^2=\mathbf{y}^T (\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$, is stochastically independent of the precision, $(\mathbf{X}^T\mathbf{X})^{-1}$ (assuming normally distributed data). It means that a knowledge of the residual variation, $|\mathbf{y} - \hat{\mathbf{y}}|^2$, does not provide with any information on the size or variation of $(\mathbf{X}^T\mathbf{X})^{-1}$. Therefore it is necessary to involve the precision, $(\mathbf{X}^T\mathbf{X})^{-1}$, in one way or another in the modeling task.

3. Forward analysis. When there are many variables, the emphasis is to obtain a stable solution. The exact (or unbiased) solutions often have little interest, because they are typically unstable in the sense that it may change drastically by small and immaterial changes in the data. Therefore it is appropriate to find the solution by a 'forward' procedure similar to the one used at the Gauss method of solving linear equations. This approach of finding mathematical solutions allows us to judge the solution in an analogous way as is possible at the Gauss method.

4. Decrease of error of fit and increase in model variation. When the model is enlarged (e.g. more variables (dimensions) in linear regression), we obtain decrease in the error of fit, $|\mathbf{y} - \hat{\mathbf{y}}|^2$, but we pay the price of increased model variation, $(\mathbf{X}^T\mathbf{X})^{-1}$. In the modeling task it is important to analyze the price. The analogy with the Heisenberg uncertainty principle appears here. At a certain stage of computing the solution the price becomes too high, even though the decrease in the error of fit is in itself significant.

5. Mean squared error. When a mathematical model is used repeatedly, like e.g. at on-line modeling of data, the mean squared error of deviations, ($y_i - \hat{y}_i$=observed−estimated), is of central importance. The mean squared error is the sum of the variance and squared bias. The Mallow's C_p measure is a way to look at the situation. The formulae show that the critical measures for the success are the dimension and the squared bias. The general results are that the dimension should be as low as possible and that there should be an appropriate balance between dimension and squared bias.

These five issues are fundamental, when modeling industrial data due to the low rank we find in data. Even if there are 1000 variables, the variation that is important for instance for regression analysis is located in say, six dimensional subspace that may be identified by six score (or latent) variables. It is necessary to use the above considerations, the five issues, in properly identifying the six score variables.

The algorithms based on the H-methods have proven their superiority in handling industrial data. They compute simultaneously the decomposition of the data matrix $\mathbf{X}$ and the associated generalized inverse, $\mathbf{X}^+$. At some methods the starting point is a covariance matrix $\mathbf{S}$. In these methods the algorithms generate a simultaneous decomposition of $\mathbf{S}$ and $\mathbf{S}^+$. The reason for the success is that using the H-methods is that they find 'the largest' possible part of $\mathbf{X}$ or $\mathbf{S}$, and 'the smallest' possible dimension that are appropriate, i.e., the fewest terms in $\mathbf{X}^+$ or $\mathbf{S}^+$.

The program packages SAS, SPSS, BMDP and other popular statistical software compute the exact solutions as prescribed by the method. The underlying methods are typically based on least squares method or maximum likelihood. In case there is numerical singularity, when computing the solution, the software tells about the problem and informs that inference from the results may not be reliable. The solution is then reduced by significance testing until a satisfactory result has been found. This standard procedure requires a certain degree of overfitting in order to be able to arrive at a significant model. There are some problems or difficulties, when applying this procedure to industrial data. The significance testing is based on the residual variation, which can be difficult to interpret. E.g., in testing in a too detailed model, the residual variation can be due to the numerical accuracy of the measurement instrument or the number of digits used. But the basic problem is that the least squares solution does not contribute with any information on the precision, $(\mathbf{X}^T\mathbf{X})^{-1}$. When we have arrived at an appropriate solution by significance testing, we do not know of the prediction ability of the

solution. For methods based on the maximum likelihood principle we also have some practical problems that make it difficult or impossible to apply to industrial data. In order to illustrate the problems let us look at the likelihood function assuming the multivariate normal distribution. The log-likelihood function is given by

$$l(\boldsymbol{\theta},\mathbf{S}) = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \mathrm{tr}\,\mathbf{S}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}).$$

For industrial data the covariance matrix **S** tends to be large (a 100 times 100 is a small covariance matrix) and statistically of low rank. In these cases finding the solutions to the specified parameters can be a very unstable numerical process. It is popular procedure by many statisticians to estimate the full model and test the significance of each parameter by using the χ^2-distribution. A significance of one parameter can be compared to the significance value of the χ^2-distribution with one degree of freedom. But the likelihood function is so large and unstable that this procedure typically does not give satisfactory results, when working with industrial data.

The typical situation, when working with industrial data, is that we have many variables, say 1000, but as far as the regression task is concerned the data is located in a low-dimensional subspace, say 6. The standard methods in the statistical program packages do not appropriately identify the subspace that we should work with. Standard approaches in the program packages, when there are many variables are Ridge Regression (RR), Stepwise Regression (SR) and Principal Component Regression (PCR). PCR is usually not satisfactory because it decomposes **X** independently of the response values **Y**. The odds that we get the appropriate 6-dimensional subspace are very small. RR regularizes the numerical task of finding the optimal linear least squares solution. The RR solution is typically not satisfactory when working with industrial data, because we typically have fewer samples than variables. There are other reasons for that RR does not give satisfactory results. If the H-methods are applied to RR, it can be shown that the full-rank solution that is suggested by RR is very inappropriate, when the actual rank is much smaller than the number of variables. SR is the most used method, when we look at the program packages. If the subspace is six dimensional, we typically need more variables than six, say 10. The 10 variables may give the same results as an application of the H-methods would give as far as the fit is concerned. But typically the precision is worse, which leads to too large prediction variances. There are also other problems with SR, e.g., that the situation may not be robust in the sense that small inaccuracies in the measurement values may have large influence on the results. Small changes in data (e.g. removing 10% of the data) may select a totally different set of significant variables.

In conclusion we can say that the program packages do not provide with satisfactory software to analyze industrial and scientific data. Also, the proposed methods to analyze the data are not satisfactory. We have the same story, when we look at the libraries of numerical subroutines. They typically compute the exact solutions that are unstable and may have very low precision.

Statistical methods are usually concerned with methods based on resampling procedures. In statistical terms it is assumed that the samples are drawn from specific distributions. When working with industrial data it is often important to look symmetrically at variables and samples. The algorithms associated with the H-methods are based on weighing procedures for both variables and samples. This can be useful, when some parts of data are more *important* than other parts. It may be useful to include this experience in the modeling task. The people in charge of data may have some *knowledge of data*. It may be important to model the data such that the results match the knowledge. In some situations, e.g., in process control, we want some variables to have some given *target values*. In the modeling procedure we want as stable estimates as possible for the solution in the light of these prescribed values.

Many optical instruments produce the

measurement result by using an estimated model, e.g. a regression model, to produce the results. The similar situation holds when using models for on-line control. In these situations there are three aspects of the modeling task that are important. The first one is to work with the appropriate part of the data. The second is to provide with stable and reliable predictions. The third is to be able to check the incoming samples for failures. Algorithms based on the H-methods satisfy these requirements. The part of data is selected that shows covariance. The solution is optimized with respect to predictions. And for new samples we can check them for failures by studying how they are located in score space.

Many statisticians work in such a way that they specify a mean value structure for the given data and the expected random variation around the mean values. The use of the H-methods can supplement this approach well. Using it, a stable solution is obtained. Furthermore, a study of the score, loading and causal plots gives useful insight into the special features of the data in relation to the specified model. Also, significance testing is more reliable, when the parameter estimates are not based on severe overfitting.

The methods associated with the H-method generate a latent structure that identifies the stable solution. In the literature and data analysis practice there are some misunderstandings concerning the use of latent structures in mathematical modelling. They can be traced to the education of people in statistical analysis. In e.g., regression analysis people are trained to look at the exact least squares solution of parameters and the estimated standard deviation associated with each parameter in the model. A variable can be removed, if it is not found significant. When working with latent structure we do not have this approach to evaluate parameters in the model. There are given a collection of variables that generate the latent structure, which again produce the parameter estimates. The effects of each variable on the latent structure and on the parameter estimate can of course be studied. And a variable can be removed, if it has no or small effects on the latent structure. In sciences it is natural to ask: Which variables are most significant? In science and industry there is a similar situation like at the psychiatrist that is measuring the mental faculties. All variables are important. The issue is the latent structure they generate. If we have say, 20 variables, but the latent structure is say 5, it can be misleading to show the standard deviations derived from the linear least squares method.

The approaches of the H-methods have been used to generate new methods and ideas for extensive data analysis. An example is path modeling, where the standard regression situation is extended to a network of data blocks. There can be many starting blocks, where new samples are initiated. There can also be several ending blocks, where final results are obtained. Using this approach we can study how estimated samples of the input blocks propagate in the network. Other examples of new theories are non-linear modeling, multi-way data analysis, causal modeling, weighing procedures and others.

The H-methods can be used to provide with stable solutions to specialized statistical models like e.g. variance components (generalized mixed linear models), dynamic models, maximum likelihood solutions and others. Experience has shown that it is advantageous to relax on the exactness of the solution and instead seek a stable solution in the light of the given data. This is especially important for scientific and industrial data due to the typical low rank in data.