1 Classification, ideas and concepts

Classification methods are one of the most studied methods in statistics and computer science. In statistics different methodologies like e.g., maximum likelihood, have been applied to classification. The results are different book on classification, chapters in text books and journals focussing on the topic. In computer science the challenge has been to develop algorithms that are efficient in finding patterns in data, which can be generated in different ways like images, signals, sensory information, radio signals and others. In this area important algorithms have been developed that show ingenious ways to, as one could say, 'find corners in data'. When working with industrial data, these procedures have not been found efficient. There are many reasons for the need of new algorithms and methods. Statistical procedures typically assume that the data have full rank. If data show reduced rank, stepwise selection of variables is recommended or to base the analysis on PCA decomposition of data. Why have the algorithms in computer science not been accepted in industry? The reason is that for industry the keywords are: *Simple, Visual, Stable, Robust and Reliable*. The methods used must be simple and applicable by different people that may not have the experience of a computer scientist. The classical statistical procedures have the possibilities of visualising the results, but they are not stable and robust as required by industry. By insisting on that the methods are reliable in the sense that the procedures must detect outliers, trends and other developments in data, the industry accepts that some methods may be good in 'a laboratory', but may not satisfy the quality and security requirements needed.

The methods presented here are based on the H-principle of mathematical modelling. The basic idea is to build up the model in steps, where at each step both the required task and the associated precision are evaluated. Optimal solution is selected at each step, which secures reliable and stable results. The final solution has thus been found by adding parts, where each part has been optimized with respect to the purpose of the model.

The developed methods have business success both at institutions and industry. The success is partly due to that the new methods secure better classification and predictions than traditional methods, when applied to industrial data, and partly that they provide with graphical analysis of data that effectively shows the inherent variation in data. The graphical tools presented here are easily implemented on 'the production floor' and can be used for on-line control of quality and other tasks.

1.1 Industrial data and statistical methods

In industry the data are typically large. It is also characteristic for industrial data that there is a high degree of redundancy in data in the sense that many variables are expressing similar things. Traditional methods for classification are based on the multivariate normal distribution. These methods can be divided into three parts. The first ones are 'full rank' methods that utilize the inverse of the covariance matrix. Even though the inverse can be computed the resulting methods are unstable when applied to industrial data. The second type of approach is the Principal Component Analysis, PCA. Here the covariance matrix is decomposed in its eigen value decomposition and the resulting components are used. Although this method can often be used, there are other decompositions of data that are more natural. The third approach that is often used is stepwise selection of variables according to their discriminating power. When a variable has been selected, the data is adjusted for the selected variable. The disadvantage of this procedure is that it normally is very data dependent. The sequence of variables selected for the first part of data may not be the same as for the last part. From industrial point of view it is not satisfactory to select say, four variables out of 1000 and not making use of the possibility of working with more stable part of the data.

When the H-principle or the H-method is used, we work with certain tools that will be explained closer.

1.2 Some tools of the H-method

Let us assume that there is given an N time K matrix $\mathbf{X}$. Furthermore, there is an external matrix $\mathbf{X}_e=\mathbf{X}_{external}$, that is an N_1 times K. The task is to both characterize the data matrix $\mathbf{X}$ and to show where it is different from the data matrix $\mathbf{X}_e$. For that purpose the columns of $\mathbf{X}$, the variables, are weighed by a weight vector $\mathbf{w}_1$. The resulting vector $\mathbf{t}$, $\mathbf{t}=w_{1,1}\mathbf{x}_1+\ldots+w_{1,K}\mathbf{x}_K=\mathbf{X}\mathbf{w}_1$, is called the *score vector*. If $\mathbf{w}_1=(0,\ldots,0,1,0,\ldots)$, the score vector is a variable. Similarly, there is a weight vector $\mathbf{w}_2$ for the rows of $\mathbf{X}$, the samples. The resulting vector $\mathbf{p}$, $\mathbf{p}=w_{2,1}\mathbf{x}^1+\ldots+w_{2,N}\mathbf{x}^N=\mathbf{X}^T\mathbf{w}_2$, is called the *loading vector*. If $\mathbf{w}_2=(0,\ldots,0,1,0,\ldots)$, the loading vector is a sample. The weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ reflect the way the analysis is carried out. The scaling constant d is computed as $d=1/(\mathbf{w}_2^T\mathbf{X}\mathbf{w}_1)$. When the weight vectors and d have been computed, the data matrix $\mathbf{X}$ is adjusted for what has been selected,

$$\mathbf{X} \leftarrow \mathbf{X} - d\,\mathbf{t}\,\mathbf{p}^T.$$

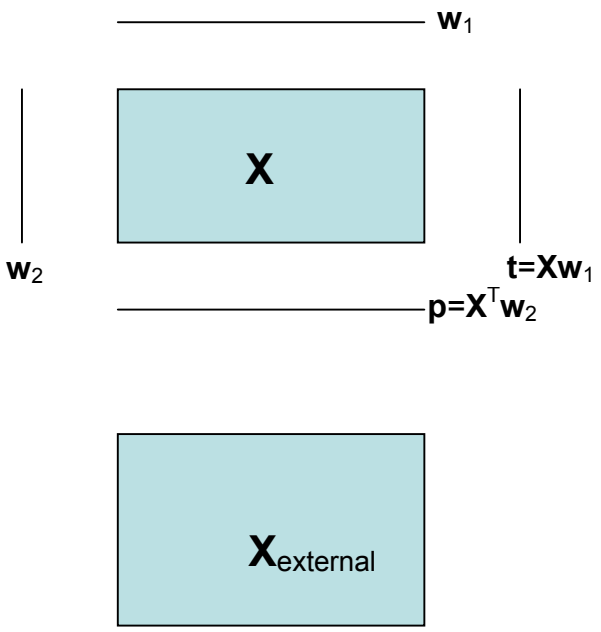

Figure 1.1. Schematic illustration of data and vectors

The data matrices and the vectors are schematically illustrated in Figure 1.1. It shows the sizes of the vectors in question. From a numerical point of view there is large flexibility in the choice of the weight vectors. The only requirement to the weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ is that the $1/d=(\mathbf{w}_2^T\mathbf{X}\mathbf{w}_1)$ may not be zero. Usually, when a new set of samples arrive, $\mathbf{X}_0$, the task is to find out, if some or all belong to the same class as those of $\mathbf{X}$. In that case it is normally desirable to have a collection of score vectors $\mathbf{T}=(\mathbf{t}_1,\mathbf{t}_2,\ldots,\mathbf{t}_A)$ that describe well the special features in $\mathbf{X}$. The basic algorithms always compute a matrix $\mathbf{V}_1$ such that $\mathbf{X}\mathbf{V}_1=\mathbf{T}$. For a new set of samples $\mathbf{X}_0$ the associated score values are computed, $\mathbf{T}_0=\mathbf{X}_0\mathbf{V}_1$. The new score values are then compared to the values of $\mathbf{T}$ to see how the new scores are located in the score space.

The weight vector $\mathbf{w}_2$ can be chosen as the score vector $\mathbf{t}$, which has been found by some method. In that case, and only in the case that $\mathbf{w}_2$ is proportional to $\mathbf{t}$, the score vectors will be orthogonal. It simplifies the interpretation of loading vectors to have orthogonal score vectors. On the other hand it may give more efficient results to weigh samples according to how they differ from those of $\mathbf{X}_e$. The algorithms also always compute a matrix $\mathbf{V}_2$ such that $\mathbf{X}^T\mathbf{V}_2=\mathbf{P}$, where $\mathbf{P}=(\mathbf{p}_1,\mathbf{p}_2,\ldots,\mathbf{p}_A)$ are the loading vectors that have been found in the analysis. For the new samples $\mathbf{X}_0$ the associated loading values are computed in the same way, $\mathbf{X}_0^T\mathbf{V}_2=\mathbf{P}_0$. The new samples are then identified according to the score and loading values found, $\mathbf{T}_0$ and $\mathbf{P}_0$.

1.3 Classification and regression

In Figure 1.2 the situation is schematically illustrated. As indicated in the picture we are both interested in finding score and loading vectors such that they are good to use in linear regression and also identify the special features of **X**. Thus the task is to find a weight vector $\mathbf{w}_1$ such that the resulting score vector $\mathbf{t}=\mathbf{X}\mathbf{w}_1$ is good for regression but also reflects the differences in **X** and $\mathbf{X}_e$. The weight vector $\mathbf{w}_2$ can be used to emphasize the difference in the samples of **X** and $\mathbf{X}_e$. The corresponding weight vectors for $\mathbf{X}_e$ can be applied in a similar way to secure a good regression model and a clear discrimination from **X**.

When a new set of samples $\mathbf{X}_0$ is available, the task is to identify the group that the samples should belong to and the associated prediction of the response values $\mathbf{Y}_0$. Sometimes it is given that the samples belong to the same class as **X**. In that case it is evaluated if the new samples are consistent with those of **X**.

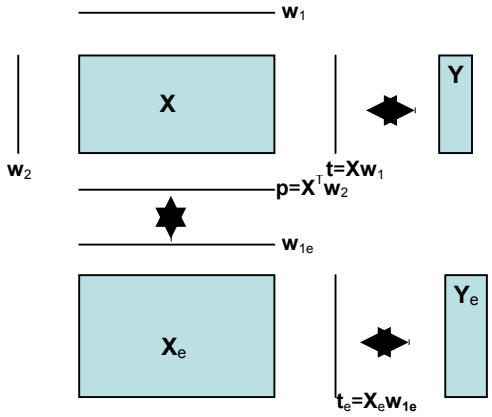

Figure 1.2. Schematic illustration of 'two class' classification and regression

1.4 Weight vectors for discrimination

In linear regression the weight vector $\mathbf{w}_1$ is found such that the score vector is good to use for describing **Y**. Similarly there are needed measures that show how well groups of data can be discriminated. The task that needs to be handled is discussed in a light of an example. Suppose that there are given two sequences of data values:

Treated: 107 109 112 114 119 121 128 139
Not treated: 98 102 103 104 106 109 110 112 113

To what degree ate these two sets of data overlapping? The two sets of data are drawn in Figure 1.3, where the treated ones are marked with an upward line and un-treated with a downward line.

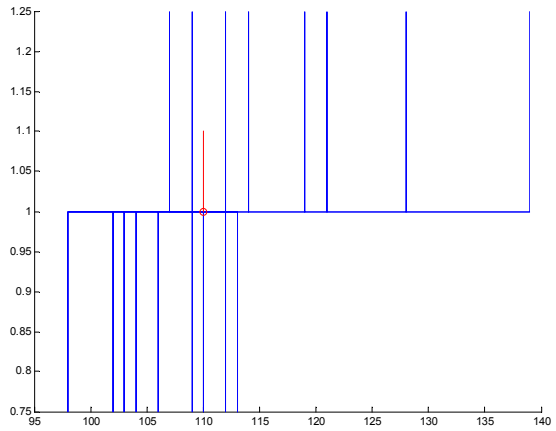

Figure 1.3. Plot of Treated and Not treated data

The figure shows that the values in the Treated group tend to be larger. But there is some overlapping. The field of non-parametric statistical tests contains efficient methods to test for the statistical significance of the difference of the distribution of these two data. A similar approach can be used concerning the degree of overlapping. If the Treated set of data is appended the Not treated ones, we get

98 102 103 104 106 109 110 112 113 107 109 112 114 119 121 128 139

The sorted data are:

98 102 103 104 106 107 109 109 110 112 112 114 113 119 121 128 139

If the simple correlation coefficient between these two data sets, the appended and sorted, is 1, the two sets will be non-overlapping. Here the coefficient is 0.981, which indicates high degree of overlapping.

When there are many variables the weight vector can be computed as the correlation coefficient between the appended and sorted data for each variable. There are of course many other measures that can be used. But this has the advantage that if the correlation coefficient is zero, there is complete overlapping, while if equal 1 there is a complete non-overlapping.

The individual measurements can be replaced by their ranks like is done at the non-parametric tests. It may have some advantage to work with the ranks instead of the original observations. It places the same 'importance' on each measurement value and thus may scale down outlier values. Measures like the correlation coefficient are sensitive to outlier values, and if there are ones, ranks often gives improvements.

1.5 Transformations

The results of the analysis can be viewed as a decomposition of $\mathbf{X}$ as follows:

$$\mathbf{X}=\mathbf{TDP}^T=d_1\,\mathbf{t}_1\,\mathbf{p}_1^T+\ldots+d_A\,\mathbf{t}_A\,\mathbf{p}_A^T+\ldots+d_K\,\mathbf{t}_K\,\mathbf{p}_K^T.$$

Even though it is possible to compute all K components, only A ones are computed or used. The weight vectors $\mathbf{W}_1=(\mathbf{w}_{1,1}, \mathbf{w}_{1,2}, \ldots, \mathbf{w}_{1,A}, \ldots, \mathbf{w}_{1,K})$ tell us how the variables have been weighted. Associated with $\mathbf{W}_1$ there is a transformation matrix $\mathbf{V}_1$ that shows how the samples map into score values. ($\mathbf{V}_1$ is sometimes called the *causal matrix* because it shows how the samples transform into the score values in the latent space). Figure 1.4 shows schematically the transformation from score space to the sample space. It is easy to show that if $\mathbf{x}^i$ is the i^{th} row of $\mathbf{X}$ and $\mathbf{t}^i$ the i^{th} row of $\mathbf{T}$, it follows that $\mathbf{x}^i=(\mathbf{PD})\mathbf{t}^i$. The matrix $\mathbf{V}_1$ is generated such that $\mathbf{XV}_1=\mathbf{T}$. From this follows $\mathbf{t}^i=\mathbf{V}_1^T\mathbf{x}^i$.

The weight vectors in $\mathbf{W}_2$ for the samples similarly generate a transformation matrix $\mathbf{V}_2$ such that $\mathbf{X}^T\mathbf{V}_2=\mathbf{P}$. If $\mathbf{W}_2$ has not entered the analysis, it has been chosen as $\mathbf{T}$, $\mathbf{W}_2=\mathbf{T}$. In that case $\mathbf{V}_2=\mathbf{T}$. For any set of weight vectors $\mathbf{W}_2$ the matrix $\mathbf{V}_2$ satisfies $\mathbf{V}_2\mathbf{T}=\mathbf{D}^{-1}$. Associated with $\mathbf{X}$ there is a generalized inverse computed as

$$\mathbf{X}^+=\mathbf{V}_1\mathbf{D}\mathbf{V}_2^T=d_1\mathbf{v}_{1,1}\mathbf{v}_{2,1}^T+\ldots d_A\mathbf{v}_{1,A}\mathbf{v}_{2,A}^T+\ldots+d_K\mathbf{v}_{1,K}\mathbf{v}_{2,K}^T$$

The generalized inverse $\mathbf{X}^+$ satisfies $\mathbf{XX}^+\mathbf{X}=\mathbf{X}$. The truncated versions of $\mathbf{X}$ and $\mathbf{X}^+$ using A terms also satisfy this equation.

1.6 Transformation and views relative to one group of data

An important and common analysis is to look at the data from the point of view of one class of data. This will be explained closer. The transformation matrix V1 is generated, when analysing the $\mathbf{X}_1$-data. The remaining data can be transformed using the same transformation $\mathbf{V}_1$. This generates two sets of score vectors, one set of score vectors $\mathbf{T}_1$ that are associated with the data and another set of score vectors $\mathbf{T}_{10}$ that are projections of other samples into the score space.

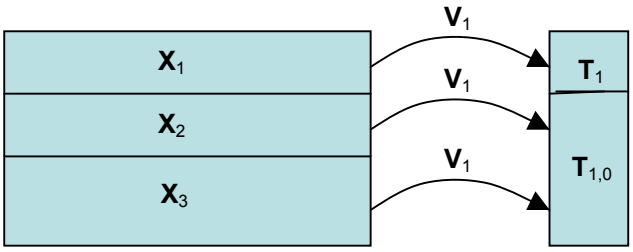

Figure 1.5. Transformation relative to one group

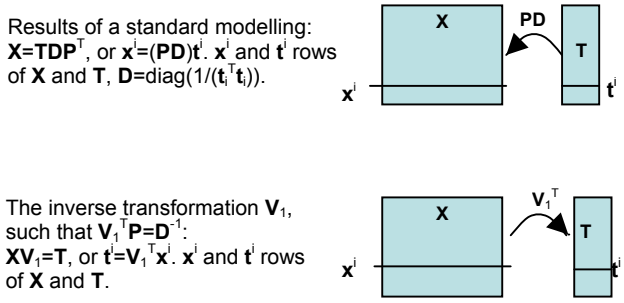

Figure 1.4. Transformations between sample and score spaces

The next task is to find out if it possible to simplify the way the two sets of score vectors are shown. One approach to such a task is to find a rotation matrix $\mathbf{O}$ such that the vectors of $\mathbf{T}_1\mathbf{O}$ are as small as possible, while the vectors of $\mathbf{T}_{10}\mathbf{O}$ are as large as possible. The rotation matrix $\mathbf{O}$ can be found by maximizing the ratio $|\mathbf{T}_{10}\mathbf{O}|^2/|\mathbf{T}_1\mathbf{O}|^2$. The task of finding the rotation matrix is illustrated in Figure 1.6.

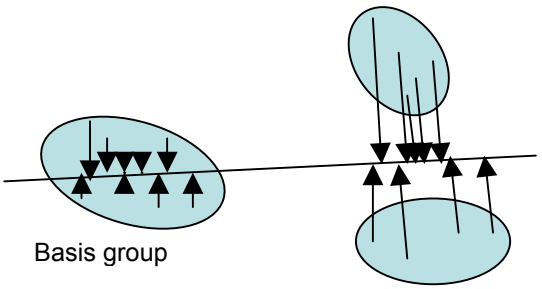

Figure 1.6. Schematic illustration of the projection

It will not be shown here how the rotation matrix is found, but the procedure is illustrated by example.

1.7 Graphic analysis of classes

We shall use here the 'Mushroom data'. The mushrooms have been measured on 16 variables. There are three groups of mushrooms. There are given 7, 8 and 8 samples of the mushrooms. The third group is clearly different from the first two. There are practical problems in separating group one and two. The question is if the graphic procedure above can help in visualising the difference in the groups. In order to show the overall variation in data, a PCA analysis is carried out for all the data. The results are shown in Figure 1.7.

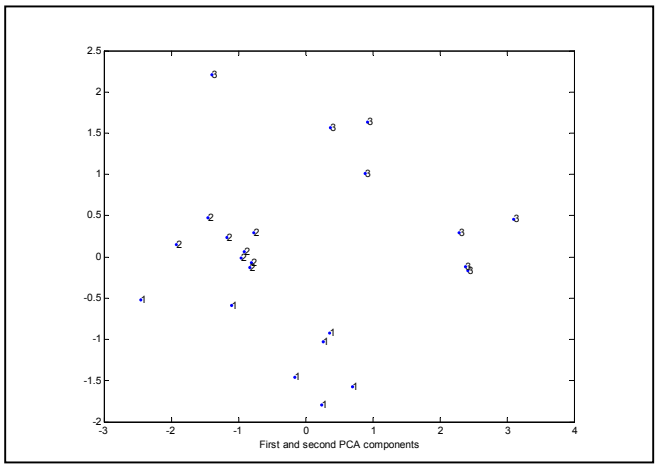

Figure 1.7. First two PCA score vectors of all the data. Groups are marked by numbers.

We can confirm that group number 3 is well separated from the others, while group 1 and 2 are close to each other. The next task is to look at data from the point of view of the first group.

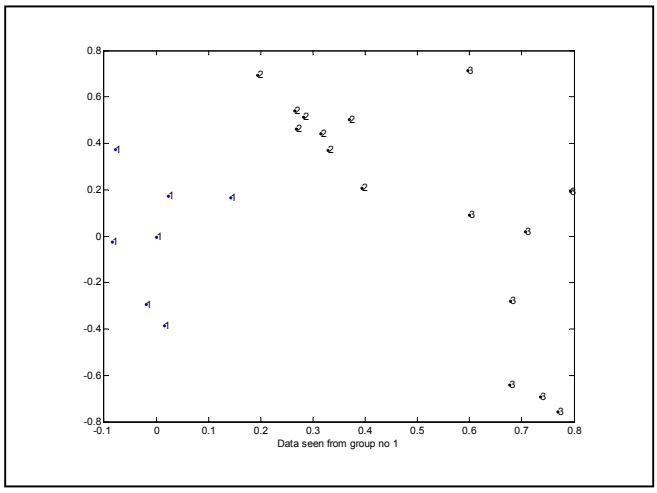

Figure 1.8. Data viewed from the score space of the first group.

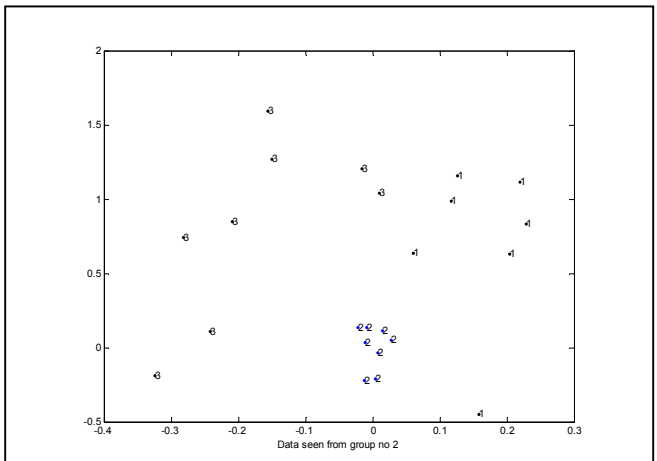

Figure 1.9. Data seen from group no 2.

In Figure 1.8 the centre of the data is the mean values of the class number 1. The score vectors are found with respect to this first class. Other samples are projected onto that score space. It means that the samples are centred by the mean values of the first group and then projected onto the score space. Finally, the score vectors are rotated such that the scatter of points around the centre is small while the score values of group 2 and 3 are far away from the centre. Figure 1.8 shows clearly that the procedure has been successful. Similarly, we see that class no 2 is located around centre showing relatively small variation, while the projected samples give score values that are scattered around the centre, but clearly at a good distance from the centre.

In summary, the three groups can be clearly separated. But we shall not consider here closer how the differences in the groups can be described.

1.8 Summary

Here we have presented some basic ideas and methods that are a part of the H-method. It has been shown how the weighing schemes can be used to 'optimize' the task at each step. When the final results have been found, score values are rotated to give insight into the latent structure that is inherent in the data. This approach is especially important when analysing industrial data. This is due to the relatively low rank that we typically find in industrial data. Using these methods we can appropriately identify variables and samples that are important in describing the differences in data across the groups. By using rank one approximation to the data it is possible to evaluate the data at each step, check for outliers, nonlinearity and other features that may be important for a successful modelling task.

Further reading

1. Höskuldsson, A.: *Prediction Methods in Science and Technology. Vol 1. Basic Teory*. 1996.
Thor Publishing. København. ISBN 87-985941-0-9
2. Martens, H., M. Martens: Multivariate analysis of quality – an introduction. J. Wiley, 2001
3. Munck L., L. Nørgaard, S. B. Engelsen, R. Bro, C. A. Andersson: Chemometrics in food science – a demonstration of a highly exploratory, inductive evaluation strategy of fundamental scientific importance. *Chemometrics and Intelligent Laboratory Systems*, 44 (1998) 31-60.