1 H-methods in Discriminant Analysis

Discriminat analysis is concerned with describing and indentifying groups of data. Here it is assumed that the data X is partitioned into 2 data blocks, $X=(X_1, X_2)$. The H-principle is briefly reviewed. It suggests that modelling should be carried out in steps, where there are two basic aspects of the modelling task. The first is the results obtained at this step and the other the means used to obtain this results. At each step the H-methods balance these two aspects of the modelling task. In section 1.3 some approaches are mentioned that are studied closer in later chapters. The Fisher Iris data is used an example to illustrate an approach and associated graphic illustrations. H-methods are data based. In section 1.5 we discuss some issues related to procedures, where data to large extent determine the final results.

Contents

1.1. Introduction

Pattern recognition methods have been intensively studied since R. Fisher presented maximum likelihood theory for making inference from data. These methods were studied earlier, but Fisher's theory marks a starting of a new era in applied sciences. We can see the degree of studies by searching for 'pattern recognition methods', 'statistical discrimination', 'statistical classification', 'numerical classification' and others in search machines on the Internet like at www.google.com, www.scholar.google.com, www.books.google.com (16.300 books appear when using Pattern Recognition Methods) and others.

These methods have been extended to many different fields of applied sciences. One could mention computer learning, computer vision, intelligence, images and others.

Here is presented a new approach to discrimination. It is based on the H-principle of mathematical modelling. The principle prescribes that the modelling should be carried out in steps, where each step is optimised with respect to the task in question. The use of the H-principle has certain advantages. We get tools to evaluate and validate the results in the light of given data. We get confidence that the mathematical modelling is providing us with results that we can understand and communicate further to others.

It should be emphasised that there are many excellent methods available. The procedures in the program packages SAS and SPSS are very good ones. If one knows these procedures (and others in the literature) well, people are able to do a good job in analysing data. But present methods provide you with some extra tools to work with the data. Furthermore, these new methods optimise the results found.

1.2 A review of the H-principle

The H-principle suggests that the modelling task should be carried out in steps. At each step the terms are setup for what is achieved by the extended model and how the extension performs. It is simplest to illustrate the principle in the case of linear regression, $\mathbf{X}\rightarrow\mathbf{Y}$. Here it is suggested to find a *weight vector* $\mathbf{w}$

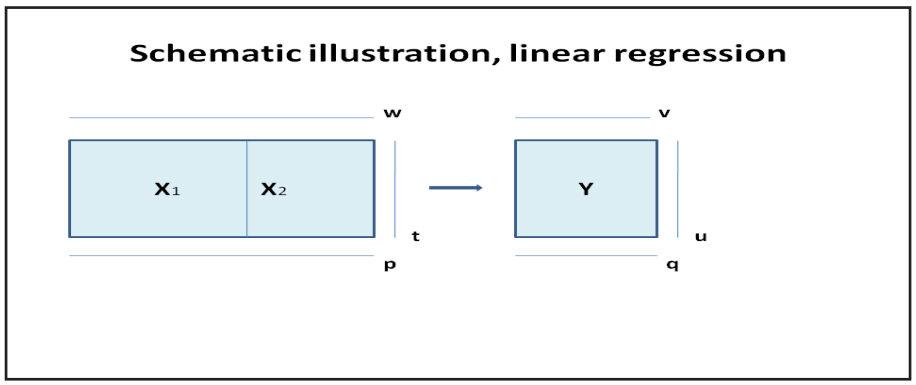

Figure 1.1. Schematic illustration of vectors in linear regression

of the variables so that the resulting *score vector* $\mathbf{t}=\mathbf{Xw}$ has some desirable properties. In linear regression there are two basic aspects of the modeling task. One is that the fit obtained, $[\mathbf{Y}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})]\ \mathbf{t}$. It is required that it should be as large as possible. Another objective of the modeling task is concerning the precision. It is possible to show that the score vector $\mathbf{t}$ should be as large as possible. It is also possible to show that these two requirements are basically independent of each other. One way to balance these two is to follow the Heisenberg uncertainty inequality and maximize

$$(1)\quad [|\mathbf{Y}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})\ \mathbf{t}|^2]\times[1/(\mathbf{t}^T\mathbf{t})] = |\mathbf{Y}^T\mathbf{t}|^2$$

$$= \mathbf{w}^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\ \mathbf{w}, \quad \text{subject to } |\mathbf{w}|=1.$$

The Heisenberg uncertainty inequality suggests minimizing corresponding terms. But the size of improvement in fit is negative. Hence the task turns out as a maximization task. When $\mathbf{w}$ has been found, $\mathbf{X}$ is reduced by rank one by the term $\mathbf{tp}^T/(\mathbf{t}^T\mathbf{t})$, where $\mathbf{p}=\mathbf{X}^T\mathbf{t}$. $\mathbf{Y}$ can also be reduced by the estimated part, $\mathbf{tq}^T/(\mathbf{t}^T\mathbf{t})$, where $\mathbf{q}=\mathbf{Y}^T\mathbf{t}$. But numerically this is not necessary because the score vectors become orthogonal by this adjustment of $\mathbf{X}$. A new step is carried out along the same lines now with reduced $\mathbf{X}$.

It is possible to show that the maximization task (1) is equivalent to the task of finding two weight vectors, $\mathbf{w}$ and $\mathbf{v}$, that give score vectors, $\mathbf{t}=\mathbf{Xw}$, $\mathbf{u}=\mathbf{Yv}$, such that the covariance is maximised

$$(2)\quad \text{maximize } (\mathbf{t}^T\mathbf{u}) = \text{maximize } \mathbf{w}^T\mathbf{X}^T\mathbf{Y}\ \mathbf{v}, \quad \text{subject to } |\mathbf{w}|=|\mathbf{v}|=1.$$

It can be shown that this procedure optimizes the prediction aspect of the model.

In Figure 1.1 the situation is schematically illustrated. Matrices are drawn as rectangles and vectors as lines. $\mathbf{X}$ is partitioned in two parts, $\mathbf{X}=(\mathbf{X}_1\ \mathbf{X}_2)$. The procedure of the H-methods finds $\mathbf{X}_2$ that does not contribute to the modeling task. The weight vector $\mathbf{w}$ gets zeros for indices corresponding to the columns of $\mathbf{X}_2$. The optimization task (1) is to find $\mathbf{w}$ so that the resulting Y-loading vector $\mathbf{q}$, $\mathbf{q}=\mathbf{Y}^T\mathbf{t}=\mathbf{Y}^T\mathbf{Xw}$, is as large as possible. Task (2) is to find $\mathbf{w}$ and $\mathbf{v}$ such that the covariance (inner product) $(\mathbf{t}^T\mathbf{u})$ is as large as possible for $\mathbf{t}=\mathbf{Xw}$ and $\mathbf{u}=\mathbf{Yv}$.

1.3 Application to discrimination analysis

Figure 1.2 illustrates schematically the situation with two sets of data, $\mathbf{X}_1$ and $\mathbf{X}_2$. $\mathbf{X}_1$ is $N_1\times K$ matrix and $\mathbf{X}_2$ $N_2\times K$. The first task is to identify the parts of $\mathbf{X}_1$ and $\mathbf{X}_2$ that do not contribute to the discrimination analysis. This is done by carrying out analysis of each variable. If the distribution for a variable is the same in $\mathbf{X}_1$ and $\mathbf{X}_2$, it is removed from the analysis. This is checked by Wilcoxon test for two independent samples. The result of this analysis is partitioning of $\mathbf{X}_1$ and $\mathbf{X}_2$, $\mathbf{X}_1=(\mathbf{X}_{11}, \mathbf{X}_{22})$ and $\mathbf{X}_2=(\mathbf{X}_{21}, \mathbf{X}_{22})$ such that the number of columns (variables) of $\mathbf{X}_{11}$ and $\mathbf{X}_{21}$ is the same.

There are many different types of analysis that can be carried out. In Figure 1.2 is shown schematically the vectors that are used. The types of analysis that are done are:

1. Find a weight vector $\mathbf{w}_1$ such that the score vector $\mathbf{t}_1$ is good in representing $\mathbf{X}_1$ and show best possible discriminatory properties against $\mathbf{X}_2$.
2. Find a weight vector $\mathbf{v}_1$ such that the loading

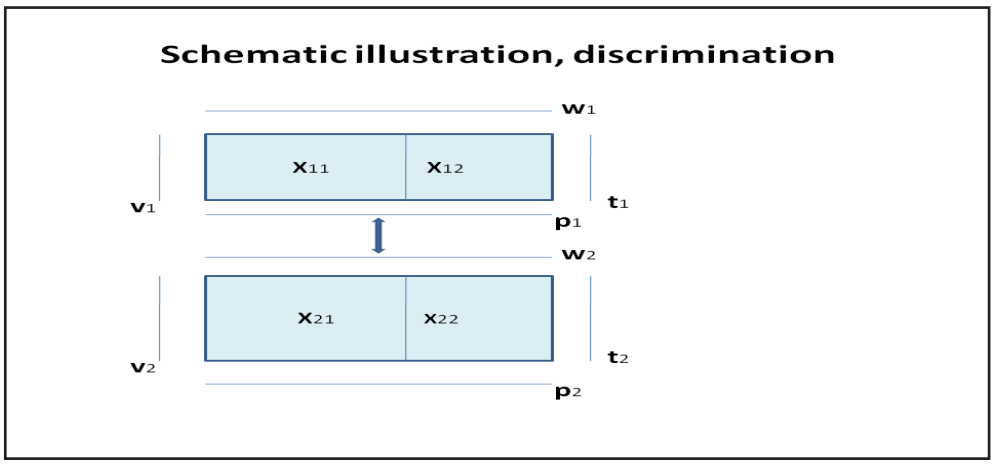

Figure 1.2. Schematic illustration of vectors in discriminant analysis

vector $\mathbf{p}_1$ shows features that express differences to samples of $\mathbf{X}_2$.

3. Combine the weight vectors for $\mathbf{X}_1$, $\mathbf{w}_1$ and $\mathbf{v}_1$, and those of $\mathbf{X}_2$, $\mathbf{w}_2$ and $\mathbf{v}_2$, so that some specific objective is optimised.

In many types of industrial and scientific applications the data have some specific structure. When there are normal conditions, the data are located in an ellipsoid in a high-dimensional space. When special conditions occur, the data values tend to be located outside the ellipsoid. This is often due to extreme low or extreme large values of some of the variables. In these cases it may be important to find the normal region for $\mathbf{X}_{11}$, and display that values of $\mathbf{X}_{21}$ tend to be outside the ellipsoid. This can be done by finding score vectors such that the ratio $|\mathbf{t}_2|^2/|\mathbf{t}_1|^2$ is as large as possible. It is often important to find a subspace in data that exhibit the differences in data.

1.4. Example. Fisher Iris data.

The Iris data were published by R. Fisher in 1936. They have since been widely used for examples in testing methods in discriminant analysis. The data contains 50 iris specimens for each of three species, Setosa, Versicolor and Virginica. For each the sepal length, sepal width, petal length and petal width were measured. The data are thus 150 samples, 50 from each and each sample contains 4 measurement values. The species Setosa differs from the other two. But there is some overlap between the species Versicolor and Virginica.

In Figure 1.3 is shown a plot of petalwidth against petallength. It shows that the group of values for

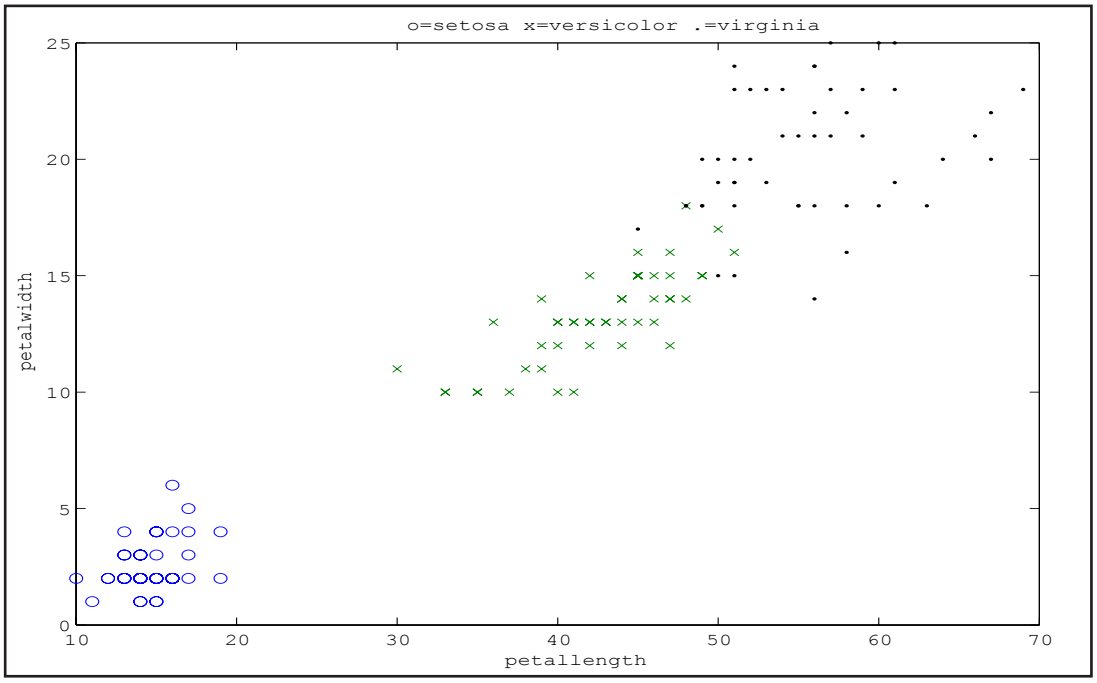

Figure 1.3. Plot of petalwidth against petallength. Groups in data are marked. o=Setosa, x Versicolor and .=Virginica.

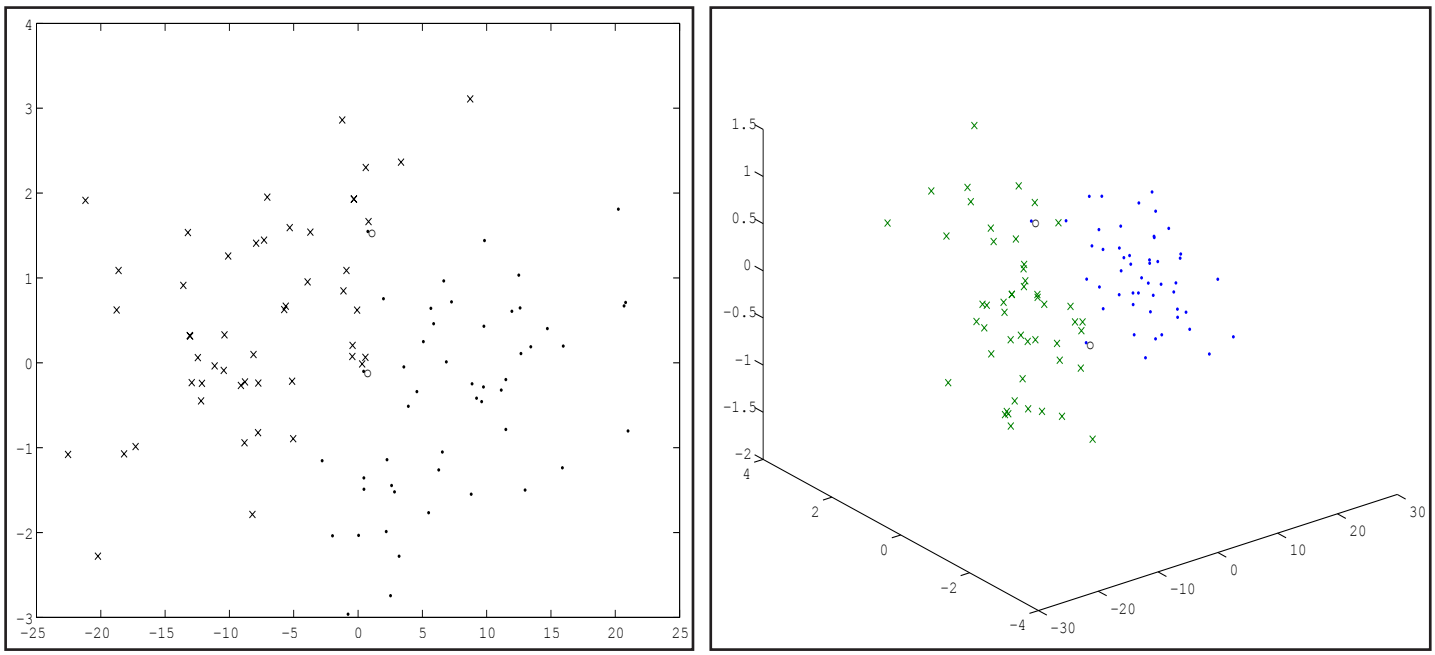

Figure 1.4. On the left plot of t_1, x-axis, against t_2, y-axis, circles are placed at misclassified samples. Versicolor samples are marked by '.' and Virginica samples by 'x'. On the right the corresponding three-dimenional plot.

Setosa are located in the lower left corner and clearly different from the other two groups. The samples for Virginia are clearly smaller than those for Versicolor. But we see clearly some overlap. The challenging task for the Iris data is to find out, how well can Virginica samples be discriminated from those of Versicolor.

Here is presented one analysis to display the difference in the samples from Virsicolor and Virginia. At start the average of all samples is taken and subtracted from all samples of Virsicolor and Virginia. We are interested in score vectors from this centre, which exhibit differences in the samples. The columns of $\mathbf{X}_2$ and $\mathbf{X}_3$ are weighted by the Wilcoxon weights, see section 2.4. The results are score vectors for $\mathbf{X}_2$ and $\mathbf{X}_3$. $\mathbf{X}_2$ and $\mathbf{X}_3$ are adjusted for the respective score vector. A new set of Wicoxon weights is determined. Only significant ones are used. Again the data is adjusted for what has been found and a new set of score vectors found. See chapter 4 for further details. When these three sets of score vectors have been determined, the reduced $\mathbf{X}_2$ and $\mathbf{X}_3$ show no differences. The first two score vectors are plotted in left plot of Figure 1.4. Those of $\mathbf{X}_2$, Versicolor, are plotted by '.', and those of $\mathbf{X}_3$, Virginica, by 'x'. We see that the scatter of points is located within two ellipses. They touch each other at the center, here at (0,0), which is at the average of means. There are two Versicolor samples that seem to be within the ellipse of Virginica samples. They are drawn by a circle.The estimated probabilities of membership of the two samples are shown in Table 1.1. Sample no. 3 of Versicolor has 96.4% probability in the Versicolor group and 61.6% in the Virginica group. Therefore, this sample would be judged belonging to the Virginica group, because it is closer to the center of Virginica samples than to the center of Versicolor samples. It is not an outlier in the Versicolor group, but the values are more normal as Virginica values. Sample no. 5 is also more closely related to Virginica that Versicolor.

If we study closer the scatter of points, we see that there is some overlapping when only the first two score vectors are used, the plot to the left. If classification were only based on these two, 4 samples would be misclassified. In the plot to the right we see that the ellipsoids are well separated apart from the two samples. In MATLAB the scatter of points can be rotated, which more clearly can show the location of the points and how well the ellipsoids are separated.

In conclusion, three sets of score vectors adequately describe the difference between the Versicolor and Virginica samples. The error rate is two samples; Two Versicolor samples are more like Virginica ones. Most algorithms in the literature result in an error rate of 4 or more, where samples of both types would be misclassified.

From	To	Sample	Probability Versicolor	Probability Virginica
Versicolor	Virginica	3	0.964	0.616
Versicolor	Virginica	5	0.955	0.580

Table 1.1. Misclassified samples and associated probabilities for these samples.

1.5 Discussion

It is popular to base discriminat analysis on the multivariate normal distribution. At first it is investigated if the data follow a mulitvariate normal distribution. This is often done by checking the variables and the score vectors found by PCA analysis. Next the discriminating functions are computed. It can be assumed that the covariance matrices are equal across the groups. In this case linear functions are used. If the covariance matrices can not be considered equal, separate covariance matrices are used and we get quadratic discriminating functions. The results are based on these functions and probabilities computed on the basis of the normal distribution. Theoretically, this is a well defined and much studied procedure. But in practice it does not work very well. This is partly due to that data in practice often show reduced rank. In this case the discrimination functions can not be computed. Stepwise selection of variables is then used in order to find variables that should enter the analysis. This is also not good for different reasons, which we shall not consider closer.

The many books available show that this area has been intensively studied. But most of these methods do not fall into the taste of the experimenter. What is the reason? People want to see plots that show how the samples are located, graphic illustrations of the differences and some kind of validation of the results obtained. People used to chemometric methods want to see score vectors that exhibit the variation in data. They want to judge the variation from these score plots.

One scientist explained the situation clearly. "I want methods that hold in court." I want to be able to say that graphically in two or three dimensions this is the region of normal variation for 99.5% of individuals in this group. The other group should be clearly separated such that the conclusion is clear to the audience.

H-methods build up a solution in steps. Each step has a certain objective. It can be to find a score or loading vector that is supposed to do a certain task. There are always two aspects of a modelling task, 1) what is contributing to the task, and 2) how well does the result perform. In the Fisher Iris data example at each step a score vector $\mathbf{t}_1$ for $\mathbf{X}_1$ is needed and similarly $\mathbf{t}_2$ for $\mathbf{X}_2$. Performance for the two vectors is measured by the Wilcoxon test for two samples. Variables are also weighted by the Wilcoxon measure. At the first step all variables contribute to the score vectors. At later steps, two and three, fewer variables are used. Variables are ranked according to the Wilcoxon measure. Variables are added according to this ordering as long as they improve the Wilcoxon measure for the score vectors. The score vectors can be evaluated by another measure than the Wilcoxon one. Then the variables used at each step would be evaluated by the same measure.

The various methods use different measures, but the principles are the same. Necessary and sufficient information (data, variables and/or samples) is found such that the task is optimal at each step. It is optimal with respect to the measures and targets used.

Sometimes some pre-processing it is needed before the analysis can be carried out. Consider the example mentioned in section 1.3. We would like to get a score vector $\mathbf{t}_1=\mathbf{X}_1\mathbf{w}_1$ and $\mathbf{t}_2=\mathbf{X}_2\mathbf{w}_1$ such that the ratio $|\mathbf{t}_2|^2/|\mathbf{t}_1|^2$ is as large as possible. First both $\mathbf{X}_1$ and $\mathbf{X}_2$ are centered with respect to the means of $\mathbf{X}_1$. This task is not well-defined, because we can get the value of $|\mathbf{t}_2|$ large but $|\mathbf{t}_1|=0$. The first task here is to determine the score space that should be used for $\mathbf{X}_1$ and the one for $\mathbf{X}_2$. Using these score spaces we can see how the $\mathbf{X}_2$-samples are located relative to the centre of $\mathbf{X}_1$-data.

H-methods have been criticised for being too data dependent. Outliers or groups in data may receive relatively large influence on the results. Also it may be difficult to assess the importance of the variables in the modelling results.

When we let the data determine the final model, which is done by H-methods, it is important to validate the results. Typically this is done by cross-validation. This way of testing the data will normally reveal outliers and groups in data. H-methods allow proper evaluation of the importance of variables. The variation is described by the score vectors. We can determine how well the variables describe the score vectors. This is normally the best way to describe and order the variables. This important topic is discussed in later chapters.

The basic issue concerning H-methods is the approach that is chosen to establish the final model. The model is tailored to fit the data, where both aspects of the modelling task are considered. At each step the best possible solution is obtained in the view of the criteria used. Different criteria may lead to different solutions.

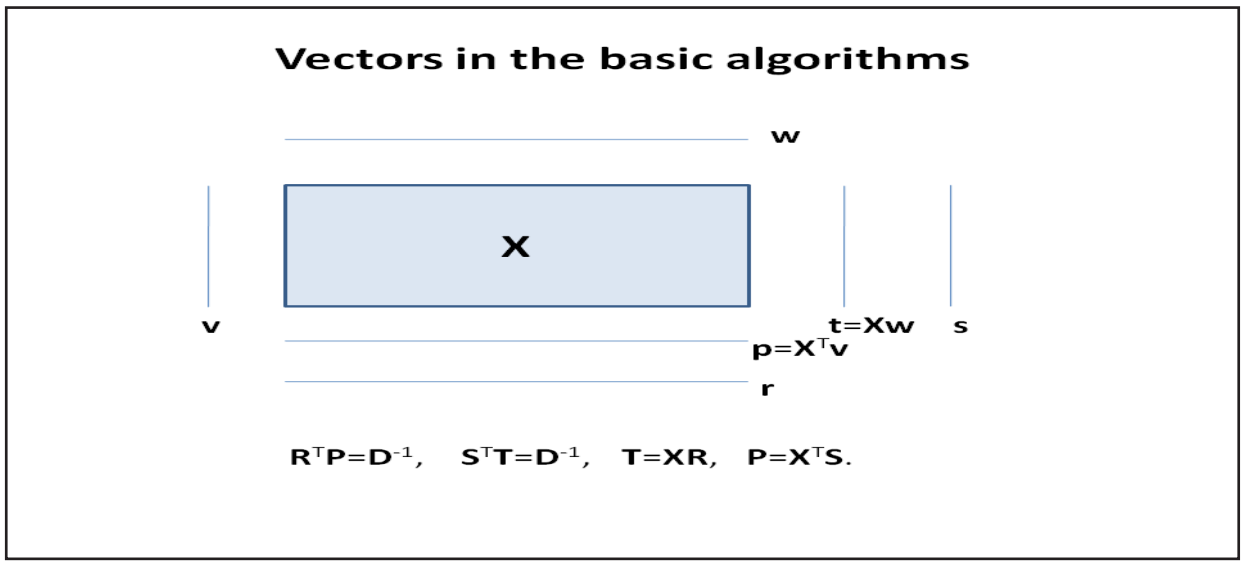

Figure 1.5. Schematic illustrations of vectors in the algorithms.

1.6 Notation

A summary of the notation is as follows:

X	Instrumental data, N×K matrix
W	Weight matrix, K×A
V	Weight matrix, N×A
T	Score matrix, N×A
P	Loading matrix, K×A
D	Scaling constants, diagonal, A×A
R	Loading weight matix, K×A
S	Loading weight matix, N×A
$\mathbf{X}^+$	The generalised inverse of **X**
A	Number of components selected
$\mathbf{X}_a$	Adjusted **X**
$\mathbf{X}_0$	Residual **X**

The indices a,b and c refer to the steps in analysis. The indices i,j and k refer to the elements in the matrices, for instance $\mathbf{X}=(x_{ij})$.

The vectors selected at each step are referred to as a component. The columns of a matrix are written by small letters, e.g., $\mathbf{T}=(\mathbf{t}_1, ...,\mathbf{t}_A)$. Transpose of a matrix is represented by upper case T, the transpose of **X** is $\mathbf{X}^T$.

The instrumental data **X** often represent repeated measurements on different 'objects'. Each column represent specific measurement values of an 'instrument'. Therefore, the notation *variable* is often used. A certain variable is measured repeatedly and the results are given in the column of **X**. Similarly the word *latent variable* is used for a column of **T**. The word latent refers to that the values are derived from the measured ones.

Following relationships hold for the matrices

$$\mathbf{X} = \mathbf{TDP}^T + \mathbf{X}_0$$

$$\mathbf{X}^+ = \mathbf{RDS}^T + \mathbf{X}_0^+$$

$$\mathbf{T}=\mathbf{XW}, \quad \mathbf{P}=\mathbf{X}^T\mathbf{V}$$

$$\mathbf{R}^T\mathbf{P}=\mathbf{D}^{-1}, \quad \mathbf{T}=\mathbf{XR}$$

$$\mathbf{S}^T\mathbf{T}=\mathbf{D}^{-1}, \quad \mathbf{P}=\mathbf{X}^T\mathbf{S}$$

In later chapters the numerical properties of these matrices are analysed closer. If **V**=**T**, the columns of **T** are orthogonal and **S**=**T**. If **W**=**P** the columns of **P** are orthogonal and **R**=**P**. In general neither **T** nor **P** are orthogonal except in the case they are derived from Principal Component Analysis.