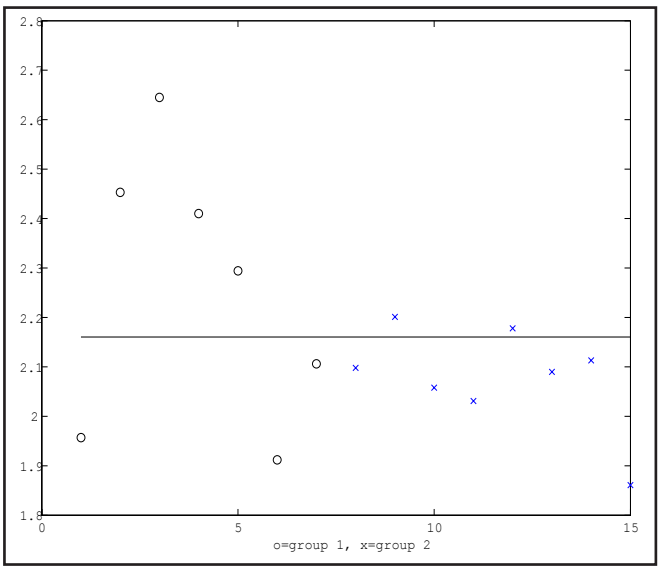

Figure 4.5. Plot of variable 1 for group 1 and 2. 'o' is group 1, 'x' is group 2. Horizontal line the average.

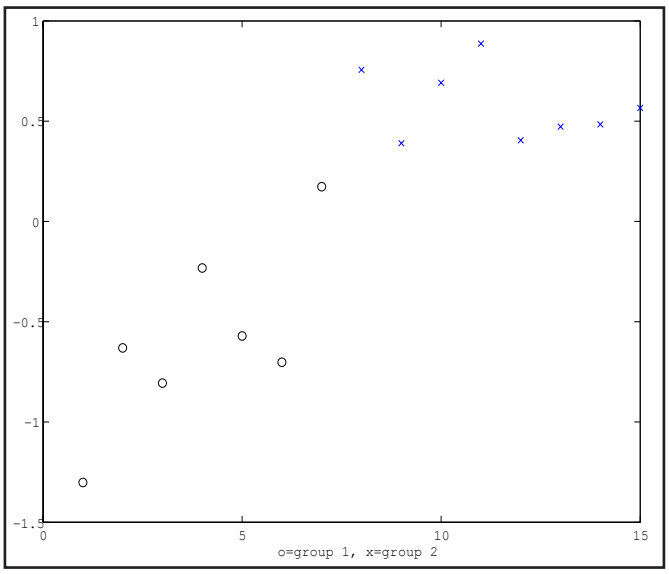

Figure 4.6. Plot of the first pair of score vectors. 'o' is group 1, $\mathbf{t}_1$, and 'x' is group 2, $\mathbf{t}_2$. x-axis sample number.

4.3 Case study. The Mushroom data.

There are given three groups of data. There are 16 variables measured for each group. The number of samples is 7, 8 and 8 resp. The number of samples is relatively small. It means that the results must be 'clear' in order to be reliable.

We shall compare group 1 and 2. The first task is to study the variables pairwise. In Figure 4.5 is shown the data for the first variable. A Wilcoxon test for two samples gives a signficance probability of 11.2%. Thus, there are not a significant difference in the distribution of these two samples for variable 1. Therefore, variable 1 is excluded from the analysis. Similarly, we find that also variables 2, 3, 9 and 12 do not show significance across groups. In conclusion, we shall only work with 11 variables, the variables 1, 2, 3, 9 and 12 being excluded.

The first task is to determine the score vectors $\mathbf{t}_1=\mathbf{X}_1\mathbf{w}$ and $\mathbf{t}_2=\mathbf{X}_2\mathbf{w}$. The variable 4 has the highest Wilcoxon u-value, 3.07. The variables are sorted according to the numerical size of the Wilcoxon u-value. The score vectors based on 5 variables, 4, 6, 8, 13 and 14 give the highest Wilcoxon value, 3.18. The results are shown in Figure 4.6.

The matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are adjusted for the score vectors found and the procedure starts over again. This time the highest value of the Wilcoxon test is found at 11 variables, 1.79. It has a significance probability

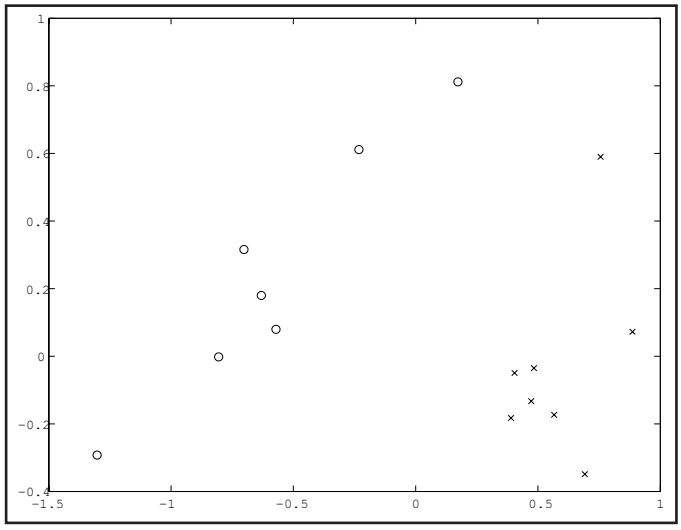

Figure 4.7. Scatter plot of the first two score vectors. Points marked by 'o,' is group 1, 'x' group 2.

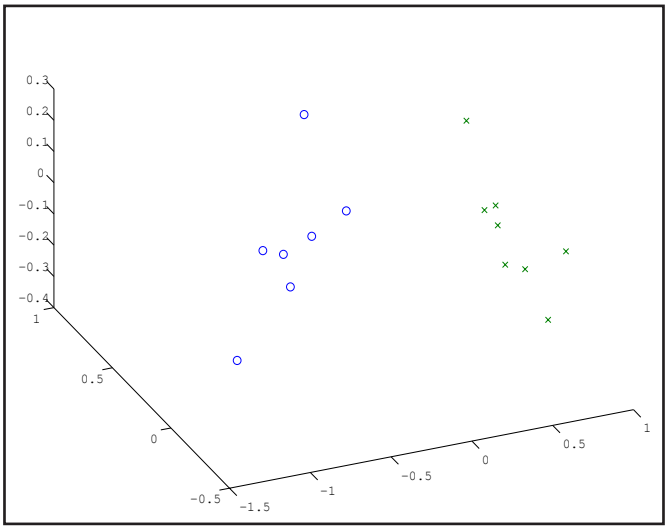

Figure 4.8. Scatter plot of the first three score vectors. Points marked by 'o' is group 1, 'x' group 2.

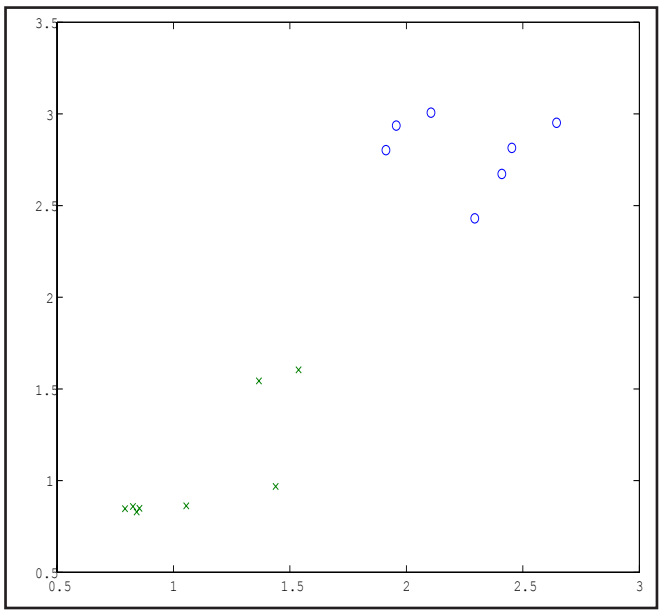

Figure 4.9. Plot of variable $\mathbf{x}_4$ versus $\mathbf{x}_1$. 'o' is group 1 and 'x' group 3.

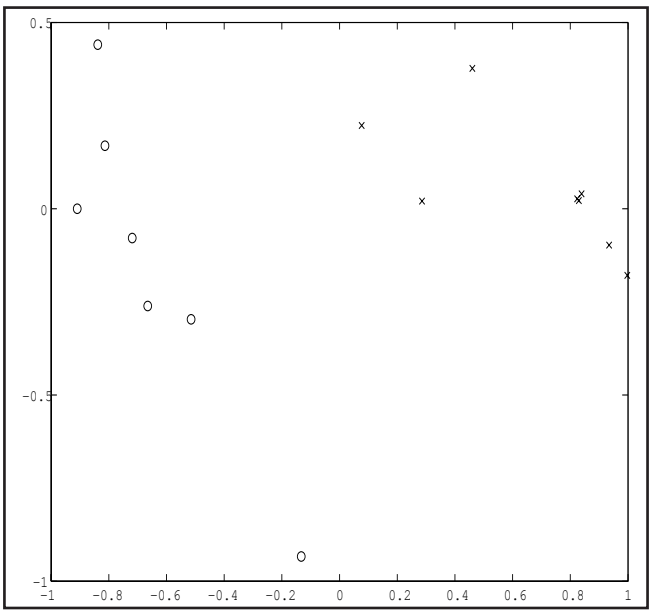

Figure 4.10. Plot of the first two score vectors. 'o' is group 1 and 'x' is group 3.

of 3.6%. In Figure 4.7 is shown the plot of the first two score vectors. It shows that the two groups are well separated.

When determining the third pair of score vectors, the highest Wilcoxon value for the new pair of score vectors is found, when using only variable 16. The value is 0.876 and associated significance probability is 19.2%. The results are not significant and therefore the third pair of score vectors can be excluded. But they may be included here partly because there are so few samples and partly because we get clear separation in three dimensions.

When comparing group 1 and 3, we find the all values of the first six variables in group 1 are larger than the values in froup 3. In Figure 4.9 is variable $\mathbf{x}_4$ (on y-axis) drawn against $\mathbf{x}_1$. The figure shows a clear separation between the groups. In Figure 4.10 is shown the first two score vectors. Only one score vector is significant. There is not a clear advantage here to work with the score vectors, although we get a good geometric description of the differences of the two groups.

We get simialr results, when group 2 is compared to group 3. The first six variables, $\mathbf{x}_1$-$\mathbf{x}_6$, have larger values in group 2 than in group 3. In Figure 4.11 is shown the plot of $\mathbf{x}_4$ (y-axis) versus $\mathbf{x}_1$ for group 2 and 3. It shows that there is a clear difference between these two groups. In Figure 4.12 is shown the plot of the first two score vectors. The first score vector is highly significant, the probability is 0.0005. The

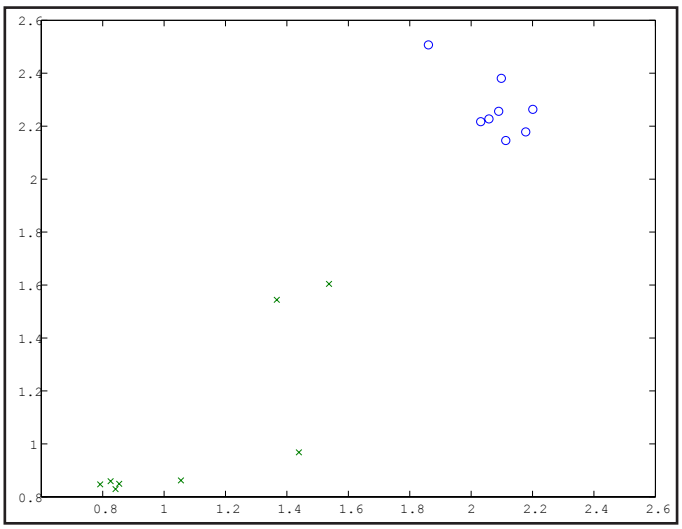

Figure 4.11. Plot of variable $\mathbf{x}_4$ versus $\mathbf{x}_1$. 'o' is group 2 and 'x' group 3.

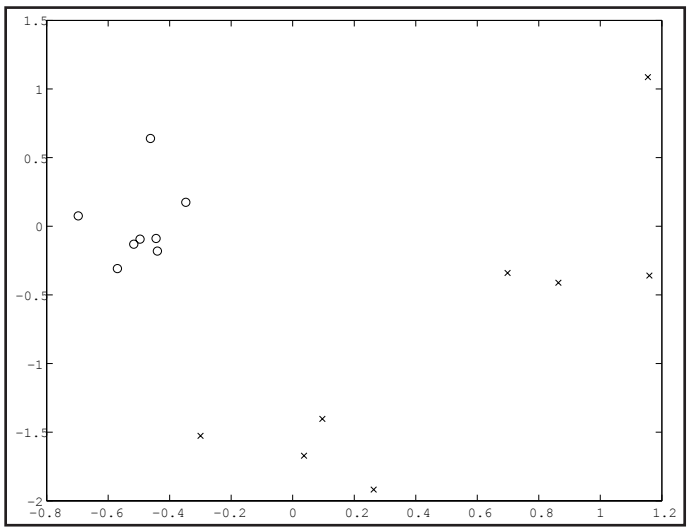

Figure 4.12. Plot of the first two score vectors. 'o' is group 2 and 'x' is group 3.

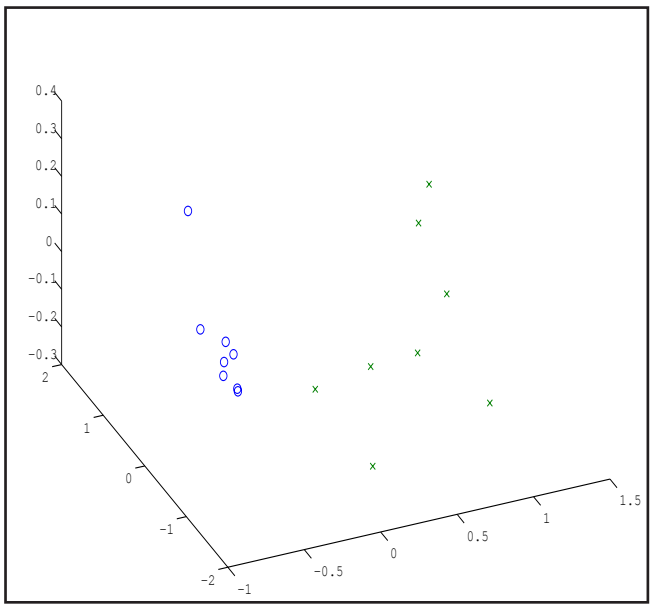

Figure 4.13. Plot of the first three score vectors. 'o' is group 2 and 'x' is group 3.

second score vector is also significant. It is based on eight variables. The Wilcoxon u-values is 2.47 and the probability is 0.007.

The third pair of score vectors are based on two variables and have Wilcoxon u-value of 1.10 and probability 13.5%. Although the pair is not significant, it can be recomended to keep it, because there are so few samples.

When comparing group 2 and 3 it can be recommended to use the score vectors. They identify the subspace, where the variation of data is located.

Conclusion

The H-methods identify the subspace that contains the variation of each group. H-methods secure that we only use the significant part of data for this task. The best possible score vectors, for a given set of criteria, are found. The results are evaluated to secure that they are a significant improvement of the modeling task. For the Mushroom data there are clearly defined subspaces that identify the groups. There are no errors in the classification task. Although the number of present samples is very small, we have arrived at such clear separation that we are confident that new samples will be classified correctly.

Discussion

H-methods have been found superior to other methods for multivariate discrimnat analysis that are based on measurement data. It is easy to construct data, where H-methods may not function well. For instance, if all data are located in a spiral in a multidimensional space, it may be difficult geometrically to find the groups of data on the spiral. But measurement data are typically located in an ellipsoid in a highdimensional space. In these cases the H-methods are efficient in locating the ellipsoids.

In this section we have used Wilcoxon test for two samples. But other tests of significance across groups can be used. But the Wilcoxon procedure is efficient both when data follow normal distribution and when there are some deviations from normality.

What are the disadvantages of the H-methods? The main one is that it is based on searches in data. At each step we search among variables and find those that show significant contribution to the task. Among these we use as many as is needed to get an optimal solution. In chapter 1 it was shown that one sample in group 2 was more similar to group 1 for most of the variables. In present analysis it is assumed that it belongs to group 1. The methods find the variables that supports this. In the figures the sample appears only as being relatively extremely located compared to the other samples (points) in group 2.

Therefore, it is important to use also other methods than H-methods, like the ones described in chapters 2 and 3, in order to detect special features in data, which do not appear in the discriminant analysis.

Scientists are concerned of finding the variables that show difference across groups. This is a different problem, which is studied in chapter 6.

In conclusion, the H-methods provide with efficient and robust ways carry out discriminant analysis on data. The error rate is lower than we see at other methods, when the data are measurement data. The H-methods have no limitations on the number of samples or variables.