Discrimination and biomarker discovery in proteomics

Suzanne Smit[a], Huub C.J. Hoefsloot[a], Age K. Smilde[a],
Marielle J. van Breemen[b], Johannes M. F. G. Aerts[b]

[a]Biosystems Data Analysis, Swammerdam Institute for life Sciences, Universiteit van Amsterdam
[b]Department of Biochemistry, Internal Medicine and Radiology, Academic Medical Center, Universiteit van Amsterdam
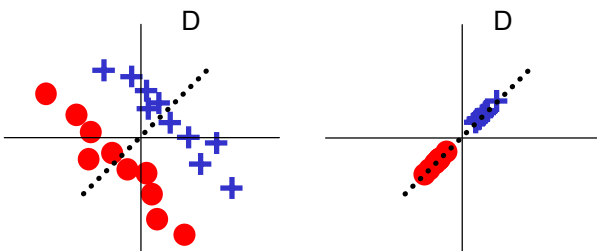
Introduction

With modern techniques it is possible to take a look at many proteins in human tissue or blood at the same time. Since different diseases cause different protein concentrations in blood, measuring all these proteins could enable the early detection of many diseases. To do this, we need to know what the protein pattern for a disease is. We are especially interested in biomarkers: one protein or a combination of a few proteins which presence or absence is indicative for a certain disease.

Method

To find if and how different groups (healthy vs. diseased) can be discriminated we use linear discriminant analysis (LDA). LDA finds directions in variable space for which the variance within the groups is small while the variance between the groups is large. The m/z values (proteins) that are large in these directions are of interest: they have large discriminative power and could be markers for the disease.
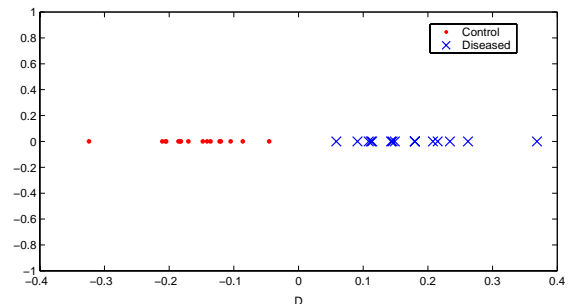
Because of the size of the data (there are many more variables than objects) principal component analysis is used to reduce the data to dimensions that LDA can handle. This is called principal component discriminant analysis, PCDA.

The discriminant vector, D, is the direction in which the groups are best separated
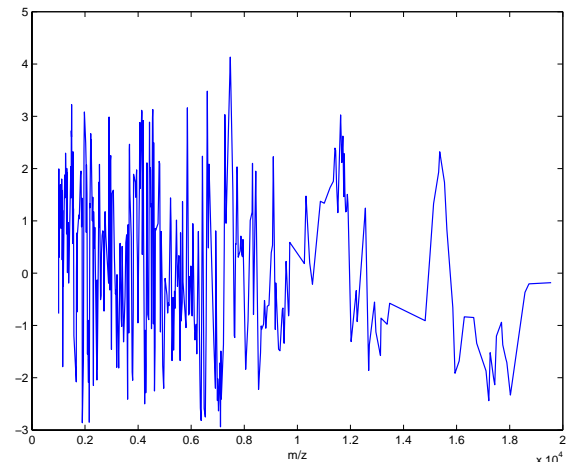
Results

PCDA was used to discriminate between SELDI spectra of 16 control and 15 diseased samples. Before modelling, the data were autoscaled, so that every m/z value, whether its intensity is large or small, can contribute equally. As can be seen below, PCDA can discriminate perfectly between the two groups of samples.

PCDA scores on discriminant vector

Question

The discriminant vector tells what m/z values are important in discrimination and thus possible biomarkers. But how to find biomarkers when there are many important m/z values in the discriminative direction?

Discriminant vector