

# The Application of Multivariate Projection Methods to the Analysis of Large-Scale Proteomic Data Sets

James Campbell Proteome Sciences plc, Institute of Psychiatry, Kings College, London, UK

## INTRODUCTION

Proteome Sciences is a company whose primary goal is the discovery and commercialisation of biomarkers indicating the initiation or progression of disease, efficacy of treatment or underlying causes of drug toxicity. Many of our proteomic data sets consist of quantitative data collected from surface enhanced laser desorption/ionisation (SELDI) mass spectrometry instruments or two-dimensional (2-D) gels. These methods are applied to samples collected from patients with, for example, cancers, neurological or cardiovascular diseases as well as carefully selected controls.

Proteomics data sets tend to have high-dimensionality and thus analysis by univariate methods is hazardous because of the risk of missing interesting correlation structure in the data as well as the increased risk of type I errors. We have used multivariate projection methods, namely principal components analysis (PCA) and partial least squares – discriminant analysis (PLS-DA) to explore our proteomic data sets and overcome many of the problems faced by the more classical statistical methods.

In this poster, a description of an analysis of the ovarian cancer data set (Petricoin *et al.*, 2002) using PCA and PLS-DA is given and the results of this analysis are compared to those of others who have analysed this same data. Additionally, brief descriptions are given of how we have applied these methods to other SELDI and 2-D gel data sets.

## ORIGIN OF THE DATA SET

The ovarian cancer SELDI data set was downloaded from [clinicalproteomics.steem.com](http://clinicalproteomics.steem.com) as plain text files. Each file contained the data for a single spectrum and consisted of 15,154 m/z values and the corresponding intensities. There were spectra available for 162 women with ovarian cancer and 91 women without cancer (controls). Of these available data, the first 25 spectra of the ovarian cancer and control samples were selected as a model training data set. The next 25 spectra of the ovarian cancer and control samples were selected for use as a prediction set, to test the predictive ability of models constructed using the training set. Thus, both training and prediction sets contained 50 samples. The following analyses were all performed in SIMCA-P (Umetrics).

## DATA PRE-PROCESSING AND OVERVIEW

In many kinds of multivariate data analysis methods, it is common practice to perform pre-processing steps such as scaling, centering and where appropriate, transformation. Centering consists of subtracting the mean value from each variable and merely aids in the interpretation of results. Scaling can have a more dramatic effect on the data. Scaling to unit variance or auto-scaling is performed by multiplying each variable by its standard deviation. This can aid methods such as PCA that are sensitive to scale by giving each variable an equal chance of influencing the models parameters. Transformations such as  $\log_{10}$  are commonly used to bring skewed data closer to a normal distribution.

With the spectral data described here, only centering was applied. This approach was taken because the scaling was shown to reduce the usefulness of the information extracted by PCA. With scaling, the variation producing class separation was not apparent until the third PC had been calculated (data not shown) whilst without scaling, the first two PCs showed good separation of the classes (Figure 1).

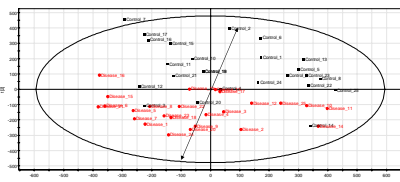


Figure 1. PCA scores plot showing first two PCs. The main direction of separation of the controls (black points) and ovarian cancer samples (red points) is indicated by the arrow.

## TRAINING A PLS-DA CLASSIFIER

The method of modeling used to formally predict the class of new observations was PLS-DA. This method attempts to find variance in the set of predictor variables (X-data) that correlates with variance in the response variables (Y-data). The Y-data set was created by indicating the classes of observations in the training set and a PLS-DA model fitted to the training set X- and Y-data matrices. The model had six significant PLS-components (determined by cross-validation) and had a cumulative fit to the Y-data (R2Y) of 0.97.

The separation of the training observations in the first two PLS-components are shown in Figure 2. The control observations are coloured black and the ovarian cancer observations are coloured red. The green arrow indicates the direction of separation of the ovarian cancer observations from the controls.

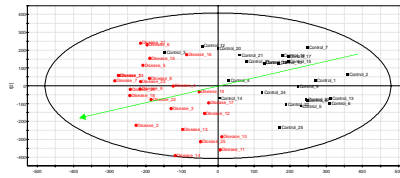


Figure 2. PLS-DA scores plot showing first two components. The main direction of separation of the controls (black points) and ovarian cancer samples (red points) is indicated by the arrow.

Figure 3 shows PLS weights plot corresponding to the first two PLS-components shown in Figure 2. The m/z values of the variables have been omitted for clarity but it can be seen on the plot that several loops of variables extend in the same direction as the separation of the two classes and thus can be considered as contributing strongly to this separation.

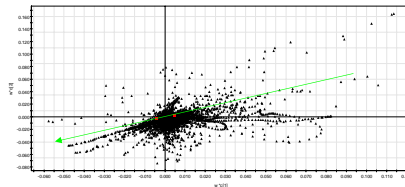


Figure 3. PLS-DA weights plot showing first two components. Each point represents a single m/z value. The direction of separation shown in Figure 2 is indicated by the arrow.

In order to determine which regions of the SELDI spectra contain the information that is most important in driving the separation between the classes, the variable influence on projection (VIP) and PLS regression coefficients can be inspected (Figure 4). The VIP parameters indicate which variables are important in explaining both the X- and the Y-data. The coefficients indicate more concisely than the PLS weights which X-variables are contributing to the modeling of the structure in the Y-variables.

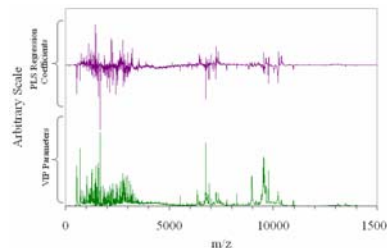


Figure 4. PLS regression coefficients (upper plot in purple) and VIP parameters (lower plot in green) of the PLS-DA model. The plots indicate which regions of the SELDI spectra carry information that drives the separation of the observations shown in Figure 2.

By considering the VIP and coefficients plots shown in Figure 4, it is clear that much of the most important X-data is found in the noisiest part of the spectrum, below 3,500 m/z in the VIP plot and below 1,200 m/z in the coefficients plot. However, there are regions of higher m/z in the spectra which also convey useful information. The VIP plot indicates that many X-variables in the region 5,500 to 13,500 are important in explaining the X- and Y- matrices whilst the coefficients plot indicates that the regions 3,700-5,000 m/z and 8,000-10,000 m/z contain X-variables which explain much of the variation in the Y-data. These regions are thus likely to be of interest for further, more detailed studies.

## PREDICTING NEW SAMPLES

In order to assess the usefulness of any classification model, it is essential to test its predictive ability using known data that was not used during the fitting of the model to the training data. This is sometimes known as a prediction set. Often, when the number of observations available are limited, cross-validation is performed where a portion of the available data is excluded from the training of a model and subsequently predicted using the model. This process is repeated until all the data has been left out once. This approach was not used here and instead, a new prediction set was used to assess the models predictive ability.

The 50 observations in the prediction set were all correctly predicted using the model fitted to the training set, giving sensitivity and specificity of the model to the prediction set of 100%.

## COMPARISON WITH PREVIOUS RESULTS

The sensitivity and specificity of the classification achieved using the PLS-DA model described here compares favorably with the results of other types of classifiers applied to this data set. Petricoin *et al.* (2002) reported a test with sensitivity of 100% and specificity of 95% for the prediction set. Sorace and Zhan (2003) and Alexe *et al.* (2004) both reported tests with maximum sensitivity and specificity using a prediction set and a cross-validation procedure, respectively. Figure 5 shows the PLS weights plot for the first two components with the m/z values found to be discriminatory by other researchers superimposed.

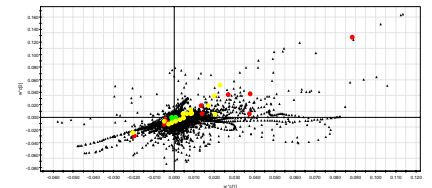


Figure 5. PLS-DA weights plot showing first two components with discriminatory peaks found by other researchers superimposed. The coloured dots show the positions of the m/z values described by Petricoin *et al.* (●), Sorace and Zhan (▲) and Alexe *et al.* (○).

With the exception of the variables reported by Alexe *et al.*, many of the variables highlighted by other researchers lie near the center of the weights plot and thus contribute little to the model developed here.

## OTHER APPLICATIONS

The following are two brief examples of how we have applied PCA to other proteomics experiments.

Standard samples of serum from a single aliquot were spotted on to chips used in a SELDI experiment. Each chip had at least one spot of the standard sample. The data collected from the spots containing the standard samples were analysed using PCA (Figure 6). This approach allowed us to monitor within-chip and between-chip variation and assess whether any systematic problems with chip processing might have occurred.

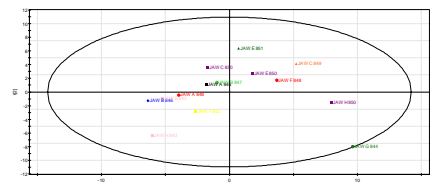


Figure 6. PCA scores of standard samples in a SELDI experiment. Each point represents a spectrum acquired from a standard sample. Samples taken from different spots on the same chip are colored the same.

Bacterial protein expression data derived from 2-D gels were obtained from cultures grown in triplicate on six carbohydrate sources. Figure 7 shows plots produced following the application of PCA.

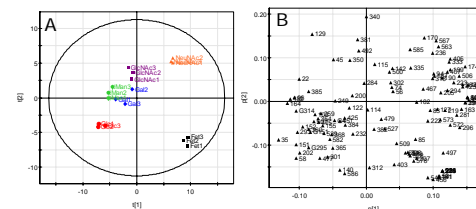


Figure 7. PCA scores and loadings in a bacterial proteomics experiment. A) Relationships between the six treatments are displayed in the PCA scores. B) The loadings show which variables describe the relationships between the treatments.

## References

- Petricoin *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* 2002; **359**:572-577.
- Sorace and Zhan. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 2003; **4**:24.
- Alexe *et al.* Ovarian cancer detection by logical analysis of proteomic data. *Proteomics* 2004; **4**:766-783.

