Short Communication

Parul V. Purohit
David M. Rocke

Center for Image Processing and Integrated Computing, University of California, Davis, USA

Discriminant models for high-throughput proteomics mass spectrometer data

We use several different multivariate analysis methods to discriminate between diseased and healthy patients using protein mass spectrometer data provided by Duke University. Two problems were presented by the university; one in which the responses (diseased or healthy) of the patients were not known and second, when the responses were known. In the latter case, the data can be used as a 'training' set. We attempted both problems. In particular, we use principle component analysis along with clustering methods to discriminate for the first problem set and partial least squares coupled with logistic and discriminant methods when the responses were known. In addition, we were able to detect regions of interest in the spectrum where there were differences in the protein patterns between healthy and diseased patients. There was considerable effort involved in the preprocessing of the data. We used a binning approach to reduce the number of variables rather than peak heights or peak areas. We performed a square root transformation on the data to help stabilize the variance; this in turn made a significant improvement in clustering results.

Keywords: Discriminant / Mass spectrometry / Multivariate PRO 0518

Mass spectrometry of proteins promises to be a very valuable tool in diagnostic applications. There are several challenges to the use of such proteomics data in classification and clustering of samples from diseased and normal patients. In particular, the number of measurements taken *per* sample is very large and even if condensed into peak areas or peak heights, the number of potential predictors greatly exceeds the number of samples in almost all cases. Registration of peaks is difficult. Although the stated accuracy of the *m/z* numbers is within 0.1%, this can encompass a large range of molecular weights on an absolute scale. The variance of the measurements is not constant, with larger peaks having a larger variance.

We propose to tackle the above stated challenges in order to perform patient discrimination and pinpoint certain regions of the spectra that may be of interest in determining the discrimination. As mentioned above, two problems were posted by Duke University. The first data set was unsupervised, where the responses of the patients were unknown. For the second data set which was supervised, the responses of the patients were known. The data in both the data sets were the same but the approaches for the two problems remain different. Both the data sets were given in the raw and processed formats (only the *m/z* and corresponding intensities of the peaks). We worked with the raw form of the data which allowed us to have the same values for the *m/z*'s for all the patients. There were a total of 41 patients with scans *per* patient conducted over 20 different fractions.

The analysis for either the unsupervised or supervised problem involves a two-pronged approach. First, the data have to be preprocessed with the proper background subtracted, normalized and the different fractions combined to obtain one integrated spectrum for each patient. The integrated spectrum is then binned and the data for all patients formatted in a $(n \times p)$ matrix format with n representing the number of patients (41) and p the number of variables (discrete *m/z*'s corresponding to the mean of the *m/z* of each bin). The range of *m/z* in all the raw spectra was restricted only to the region where the processed files indicated peaks which reduced the number of raw data points to $\sim$53 000+ from the original 65 000+ for each spectrum. Since the raw *m/z* interval values for the spectrum increased as a function of *m/z*, we chose bins by a fixed number of *m/z* points (100 for the analysis presented here) rather than using a fixed *m/z* interval. This allowed for a binned spectrum that was

Correspondence: Dr. Parul V. Purohit, Center for Image Processing and Integrated Computing, University of California, One Shields Avenue, Davis, CA 95616, USA
E-mail: pvpurohit@ucdavis.edu
Fax: +1-530-752-8894

Abbreviations: PLS, partial least squares; **Sqrt.**, square root

more in character with the original data as seen in Figs. 1–2 for Patient 1. The number of variables after the binning process drops to 531 for each raw spectrum. Background for each individual spectrum was estimated from the region of the spectrum with no peaks (*m/z* of 170 000 and higher). A flat background was used which was expected to suffice given the multivariate approach to the problems. Each fraction of the spectrum was normalized to an integrated intensity of 1.0 to assure that each spectrum has the same quantity of analyte and that the sum of all the fractions would give a true and accurate spectrum for the patient. There was no normalization conducted between patients to preserve the biological intensity relevance between patients.

The second prong of our approach constitutes multivariate statistics (using SAS). Many of these methods are based on the assumptions that the data are multivariate normal. We conduct data transformation of the spectra prior to statistical analysis in order to stabilize the variance of the data. Ideally, a variable transformation of the type used by Durbin *et al.* [1] would be well suited. However, the lack of replicates in this study precluded us from using this transformation since the parameter cannot be estimated. For this study, we performed a simple log transformation and a square root (Sqrt.) transformation. A constant value dependent on the transformation was added to the entire data set to assure that all the intensities were above 1.0 and 0.0 for the respective transformations so that the transformed values remain above zero. Success of the transformations was measured by performing partial least squares (PLS) regression analysis coupled with logistic and discriminant analysis on the supervised data set. The Sqrt. transformation seemed to be the most suitable for these data sets (Table 1). Figure 3 shows the Sqrt. transformed fraction integrated data set for Patient 1.

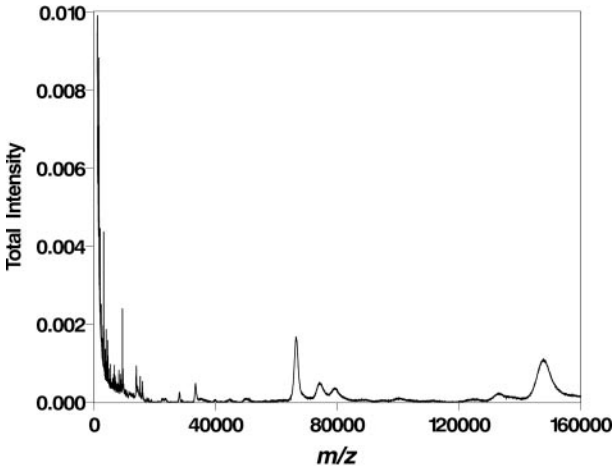

Figure 1. An integrated spectrum (all fractions combined) after background subtraction and normalization in the region of interest for Patient 1.

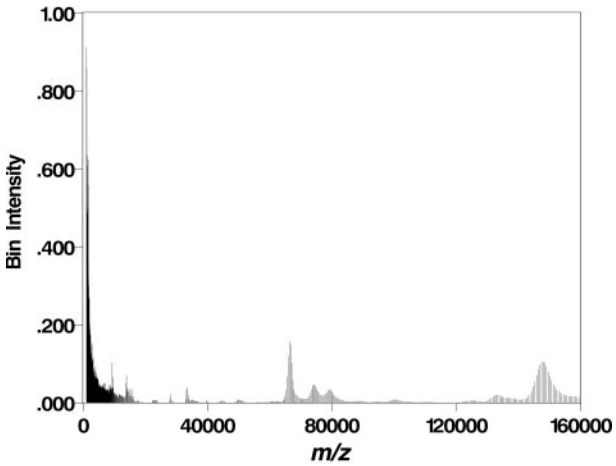

Figure 2. A binned spectrum for Patient 1. The intensity is the integrated intensity for the bin.

The preprocessing of the data as mentioned above renders the data in a 41×531 matrix that is in the right format for multivariate statistical analysis. Even after the binning that reduced the number of variables significantly, the number of variables remains significantly higher than the number of observations (patients). For the unsupervised data set, we perform further dimension reduction by principal component analysis (PCA) selecting the number of factors or principal components by using an eigenvalue plot and observing the 'jog' in the plot [2, 3]. Four factors that accounted for 65% of the variance were used for dimension reduction and further analysis. Eigenvalues and eigenvectors generated from PCA were obtained using a covariance matrix rather than the correlation matrix. This choice associates the large variances with the large eigenvalues and *vice versa* as is usually the

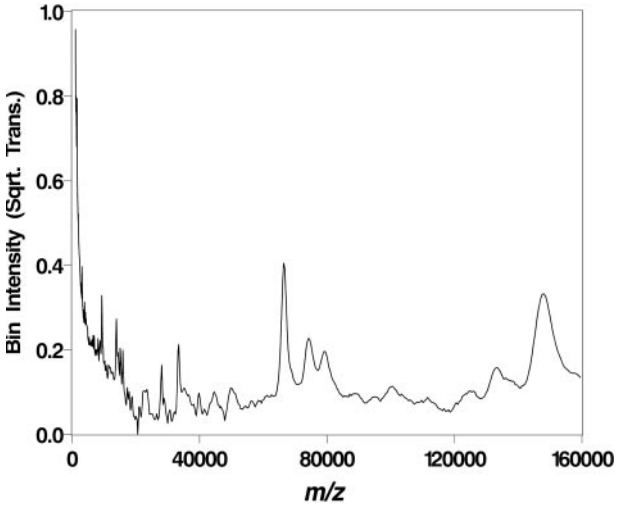

Figure 3. A Sqrt. transformed spectrum for Patient 1.

Table 1. Number of patients correctly identified by combinations of PLS and PCR regression and discriminant methods using various transformations. The second number is the success rate as a percentage. The last row indicates specific patients that were not correctly identified for the Sqrt. transformation. All discriminations were conducted using leave one out cross-validation

	PLS/LOG	PLS/DIS	PCR/LOG	PCR/DIS
No transformation	37/90.2%	39/95.1%	28/68.3%	28/68.3%
Log. transformation	37/90.2%	39/95.1%	28/68.3%	28/68.3%
Sqrt. transformation	41/100%	41/100%	30/75.6%	32/78%
Sqrt. transformation	NA	NA	4, 6, 9, 15, 23, 27, 33, 36, 39, 40	4, 10, 31, 35, 36, 38, 39, 40, 41

case in spectral data. Figure 4 shows a scores plot of Principal Component 1 *vs.* Principal Component 2 that shows the relationships between the patients with respect to each other. Different symbols have been used to indicate the expected responses (obtained from the supervised data) between the sick and healthy; however, this information was not used for the analysis. It is very clear that the patients do not cluster neatly into two groups indicating that results using cluster analysis would be tenuous. There is an outlier (Patient 15) which we removed from the data set to enhance our chances for better discrimination.

Hierarchical cluster analysis was performed on the reduced data matrix using a two-cluster option. Several standard clustering methods were attempted and we obtained the best separation of the two true groups for a correct discrimination of 28 patients (68%) using the flexible-beta hierarchical method [3]. Using this result as a seed for the k-means method did not improve discrimination. Figure 5 shows the results of the clustering where the symbols indicate the expected responses. It is interesting to note that Cluster 1 has only diseased patients whereas Cluster 2 contains all the healthy patients and some of the patients categorized as diseased. It is possible that the diseased patients in this latter cluster may have a different form or different stage of the disease. Further analysis using PLS also suggests this scenario.

For the supervised data set, we reduced dimensions using PLS and then perform logistic (LOG) and discriminant analysis (DIS) for discrimination. We compare these results using principal components regression and

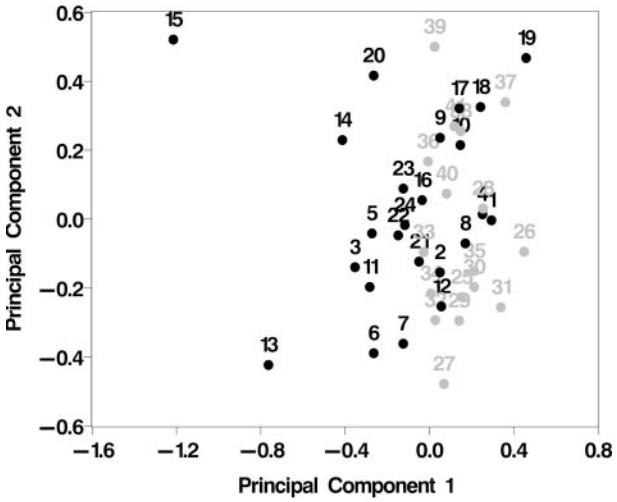

Figure 4. A plot of Principal Component 1 *vs.* Principal Component 2 obtained from PCA analysis of the binned data matrix. The black dots refer to the patients that are expected to be diseased, gray dots to the ones that are expected to be healthy.

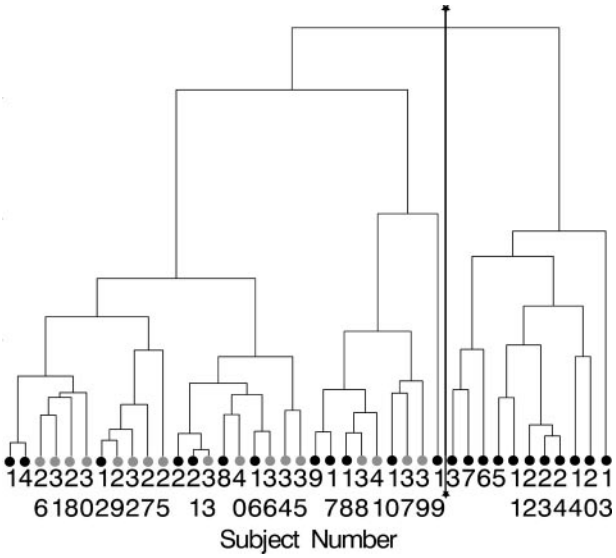

Figure 5. Graphic demonstrating the results of cluster analysis performed on PCA reduced data matrix. The black dots refer to the patients that are expected to be diseased, gray dots to the ones that are expected to be healthy.

then discrimination using LOG and DIS [4–7]. All analysis was conducted using a leave one out cross-validation approach using the rest of the matrix to predict the response of the patient of interest. The number of factors retained was chosen by conducting split-sample validation PLS and calculating the minimum root mean of the predicted residual sum of squares. The correctly identified discriminations achieved using these methods are tabulated in summary form in Table 1 for the various transformations. The last row in the table gives the numbers of the specific patients that were not correctly identified for the Sqrt. transformation. In general, the prediction rates achieved using PLS are higher than that using principal component regression [7]. We have seen similar results for other problems with discrimination issues using these methods. Thus prediction for a new patient would be greatly enhanced if a training set with known patient conditions is used. This training set can be used to calculate PLS factors and subsequently applied to the new patient for prediction calculations.

The weights plot obtained from the PLS analysis provides a powerful tool to recognize the bins that are relevant in terms of the differences between the healthy and diseased patients. The weight factors can be considered to be the eigenvectors of the matrix X^TYY^TX where X represents the data matrix associated with the variables and Y the column matrix associated with the responses. A plot of Weight Factor 1 *vs.* Weight Factor 2 that demonstrates the maximum variance in the predictor-response combination is shown in Fig. 6. Those bins with large weights in absolute value (ones lying at the extremities of the plot) are more important in predicting the variations than the ones lying close to the origin. Several bins with their corresponding bin numbers lying at the extremities of the plot have been labeled along with one bin (Bin 432) that lies fairly close to the center. We expect to find significant differences in the signal of the spectra for the bins lying at the extremities. Figure 7 shows the differences in the raw integrated spectra of six patients for the bin region 118. The patients chosen are Patients 1, 3, 7, 12, 25 and 27 based on their expected responses and the responses as indicated by cluster analysis shown in Fig. 5. Patients 1 and 12 are diseased but have been categorized as healthy in the cluster plot whereas patients 3 and 7 are diseased and categorized as such in the figure. Patients 25 and 27 are healthy and categorized as such in Fig. 5. It can be clearly seen in Fig. 7 that for patients 3 and 7, there is a prominent peak at *m/z* = ~12 900 which is faintly recognized for patient 1 and patient 12 and not at all for patients 25 and 27. This may be a diagnostic peak and is consistent with the results of the cluster analysis. Patients 3 and 7 have an advanced stage of the disease as indicated by the peak and the results shown in Fig. 5, whereas patients 1 and 12 may have an onset of the disease that has not progressed to latter stages. Patients 25 and 27 do not have the diseased as categorized by the absence of the peak in Fig. 7 and through cluster analysis as shown in Fig. 5. The six patients picked may not be entirely representative, but Fig. 4 indicates that distinguishing between healthy and diseased patients remains a difficult task through cluster analysis and different clustering techniques may yield slightly different results from those shown in Fig. 5. However, one can still attach some significance to the peak shown in Fig. 7 that illustrates a peak (or protein) that seems to be a marker of disease. Figure 8 illustrates the lack of any significant difference in spectral patterns in the bin region 432 which is located fairly centrally in the weights plot of Fig. 6. Bins 3–8 which

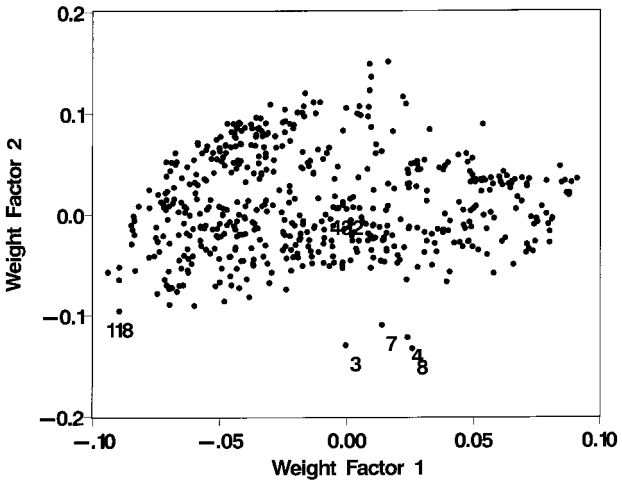

Figure 6. A weights factor plot obtained from PLS analysis of the binned data matrix. A few of the bins lying on the extremities and a bin lying near the origin have been marked and discussed in the text.

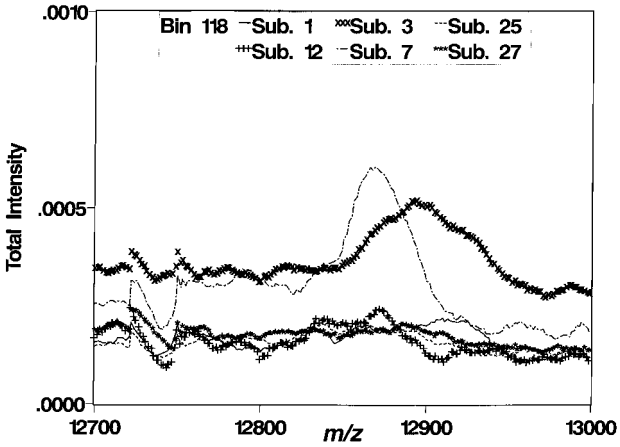

Figure 7. Region of the integrated spectra for six patients (Patients 1, 3, 7, 12, 25 and 27) for bin 118. This bin lies at the extremity in the weights plot as shown in Fig. 6.

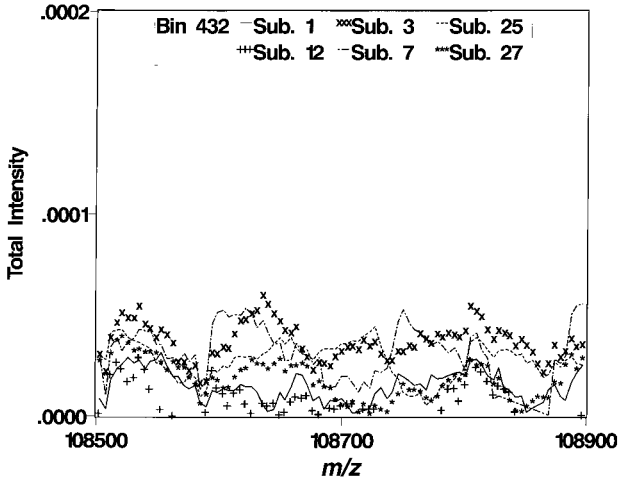

Figure 8. Region of the integrated spectra for six patients (Patients 1, 3, 7, 12, 25 and 27) for bin 432. This bin lies fairly closeness to the origin in the weights plot as shown in Fig. 6.

also lie at the extremities of the plot were not considered for meaningful comparison due to their low *m/z* value and closeness to large fluctuations in the intensity. Study of 'relevant' bins in this fashion may be a key in detecting the minor spectra changes expected between healthy and diseased patients.

Summarizing our results, we have shown that multivariate analysis is a powerful tool in analyzing high-throughput mass spectrometer data for proteomics studies. In particular, for discrimination studies, for an unsupervised data set, a combination of PCA and cluster analysis along with appropriate preprocessing of the data, can lead to meaningful and promising results. Better discrimination can however be achieved if one has a credible training data set that can be used for future predictions as can be seen with our analysis on supervised data using PLS and discrimination methods. We are expecting even further improvements on our discrimination abilities if many of the preprocessing and data steps can be optimized. We have demonstrated that the weights plot obtained by using PLS can be a powerful tool in obtaining diagnostic protein profiles in diseased patients.

Received October 16, 2002

References

[1] Durbin, B., Hardin, J., Hawkins, D., Rocke, D. M., *Bioinformatics* 2002, *18*, s105–s110.

[2] Johnson, R. A., Wichern, D. W., *Multivariate Statistical Analysis*, Fifth Edition, Prentice Hall, NJ 2002.

[3] Khattree, R., Naik, D. N., *Multivariate Data Reduction and Discrimination*, SAS Institute, Cary, NC 2000.

[4] Nguyen, D., Rocke, D. M., *Bioinformatics* 2002, *18*, 1216–1226.

[5] Nguyen, D., Rocke, D. M., *Bioinformatics* in press.

[6] Nguyen, D., Rocke, D. M., in: Lin, S. M., Johnson, K. F. (Eds.), *Methods of Microarray Data Analysis*, Kluwer Academic, Boston 2002, pp. 109–124.

[7] Nguyen, D., Rocke, D. M., *Bioinformatics* 2002, *18*, 39–50.