

Multivariate Processing of Binned HPLC Profiles

L. Scott Ramos & Brian G. Rohrback
Infometrix, Inc., Woodinville, WA

Paper P-1313

HPLC - 2000

24th International Symposium on High Performance
Liquid Phase Separations and Related Techniques

Seattle, Washington, USA
June 24 - 30, 2000

We do chromatography not simply to achieve separation of components but to answer questions about the nature of the sample.

For example, comparisons of HPLC profiles can tell us whether samples are alike or different, provide quality control of a process, or identify fraudulent materials. Such comparisons are best accomplished via pattern recognition, the tools for which are now easily accessible.

HPLC data can be supplied to the pattern recognition software as peak heights or areas, or as the entire chromatographic profile. Each data form has advantages and disadvantages.

Another data form uses the whole profiles which are sub-sampled at a lower frequency than that of the original profile. This report will show that this modulation of the profile data, often called “binning”, can retain the robustness of the pattern recognition ability while reducing the size of the data set by nearly two orders of magnitude.

In order to prepare chromatographic data for multivariate data analysis such as pattern recognition, there are several issues with which we need to be concerned.

Peak Tables

In its simplest mode of operation, the chromatographic software can be directed to export a table of peak information (either heights or areas). To be sure that comparisons among samples look at the same information, this usually means that a peak ID table is prepared ahead of analysis.

The peak ID table contains an entry for every compound expected to occur in analysis of any sample, and includes expected retention times for each of these compounds. The peaks for each sample are compared to those in the peak ID table and, if a corresponding entry exists, the peak's value (height or area) is reported. If an entry in the peak table is not present in the sample, it must still be reported, usually as a zero value.

The drawback to using peak tables is that they are rigid and cannot account for diagnostic peaks in a sample that were not

expected when the peak ID table was created.

In addition, although each entry in the peak ID table has a window of acceptance around the expected retention time, enough variation in absolute retention time can occur within a set of samples such that peaks may elute outside of an expected retention time window. This will cause the peak to be placed in a different "bucket", and the peak table for the sample would be considered different from an otherwise identical sample.

Whole Profiles

Issues with peak assignment errors can be avoided by analyzing the entire chromatographic profile as if it were a spectrum. In fact, analysis of whole profiles has an additional advantage in that peak shoulders, which would not be integrated and therefore would be absent from peak tables, become features in themselves.

Using the whole HPLC profile avoids the issues in peak assignments. However, whole profiles are not without problems. So many data points are acquired on most chromatography systems that the resulting data files become enormous and require

significant processing time. Subtle changes in retention can have significant effects on the multivariate analysis because key peak features do not perfectly align in different chromatograms, of even if those chromatograms are of precisely the same material.

Chromatographic alignment minimizes this problem, but to facilitate such alignment, it is necessary that marker peaks be present, and these must be common to all samples. As few as two markers can correct for a large part of the chromatographic drift, but more markers increase the reliability of alignment.

Markers can be inserted into the chromatographic profiles in a routine way by using prescribed internal standards. These can be co-injected or spiked into the sample solution as a final sample preparation step.

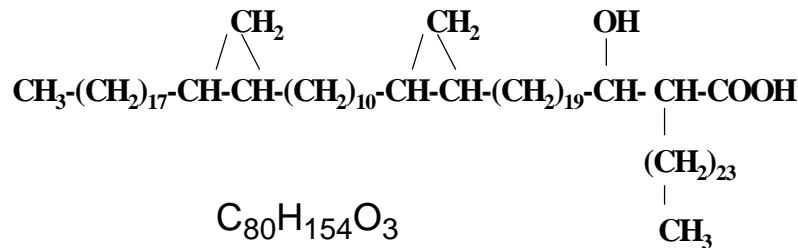
Binning

If using the whole chromatographic profile proves costly, we can modulate the data by subsampling. For example, we could retain 1 out of every 4 data points. Or, we could retain as new data points the mean of every group of 4 points.

In either case, the goal is to reproduce the magnitude of the data vector without, hopefully, discarding diagnostic information.

Because the binned profiles retain the overall shape of the original profile, major features are still present. It is this aspect of the binned data that provides diagnostic utility in a multivariate analysis.

An application from the public health field was used to demonstrate these concepts. Identification of TB in an individual carries with it the burden of both the illness and its treatment. The organism responsible for TB, *Mycobacterium tuberculosis*, is in a genus known for the production of mycolic acids: alpha-branched, beta-hydroxy fatty acids. In fact, these high molecular weight compounds—from 60 to 90 carbons—are species specific. An example is shown below.



Traditional biochemical methods of analysis can be time-consuming and subjective. HPLC has found a role in the identification of Mycobacteria and is now in use in dozens of labs worldwide.

Bacteriology and HPLC

Bacteria specimens are grown on Lowenstein-Jensen slants for 21 days at 35 °C. Titrates are saponified with methanolic

KOH, then derivatized to p-bromophenacyl esters

From an extract of the ester mix, 20 μl are injected into a 4.6 mm x 7.5 cm C₁₈ ultrasphere XL column. The column is initially washed with 80% MeOH / 20% MeCl, then the mobile phase is stepped up to 35% MeOH / 65% MeCl over a 9 minute linear gradient.

The HPLC protocols are outlined in the Standardized Method for HPLC Identification of Mycobacteria, available on the Internet (see the standardized procedures manual for users at the CDC web site: http://www.cdc.gov/ncidod/dastlr/TB/TB_HPLC.htm).

Specimens and Data

Only validated, authentic cultures were used in the study, and the identity of all cultures were reconfirmed by genetic probes, biochemical tests, and/or 16S rRNA sequencing.

From an initial study of ~350 strains, representing 23 species, strains representing 8 very similar species were set aside to evaluate the binning procedures.

The HPLC data were handled differently for the three methods of analysis. All data processing was done on AIA files exported from the chromatographic system.

Peak Data Pretreatment

A peak ID table was generated by observation of common peaks appearing in profiles from the various species; 36 peaks were tabulated. Relative retention time markers were co-added to each sample as internal standards.

The chromatographic system was instructed to report only the calibrated peaks, thus assuring that the same number of peaks appear for each sample, even if absent. Compounds not found were reported as 0. The following chart includes a composite chromatogram from several species with an overlay of a profile of just the heights of the selected peaks.

Each of the species in the subset chosen for this study is represented by the compounds in the late eluting cluster of peaks. These became the focus of the data analysis.

To account for variation in the concentrations of samples, each data vector is normalized by its vector length.

Whole Profile Pretreatment

Because elution times are not reproducible in HPLC, it is imperative that the chromatographic profiles be aligned before attempting multivariate analysis. Each analysis includes the two internal standard peaks, and these were used as markers for the alignment algorithm.

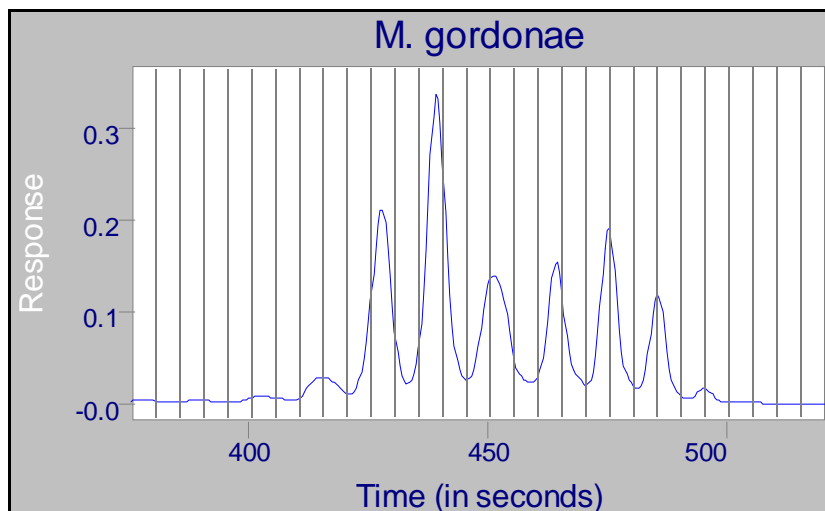
As with the peaks data, the profiles were also normalized as a pretreatment.

Binning Procedures

When binning is performed to reduce the data density, the modulation can be done in various ways. Methods tested included using the first peak in a bin, the tallest peak, the average peak height, and the median peak height.

Although the differences among these procedures were minimal, the tallest peak approach was consistently near the top so was chosen for the remainder of the study.

The following figure shows a Mycobacteria profile of in which the bin size incorporates 5 original data points.



Data Processing

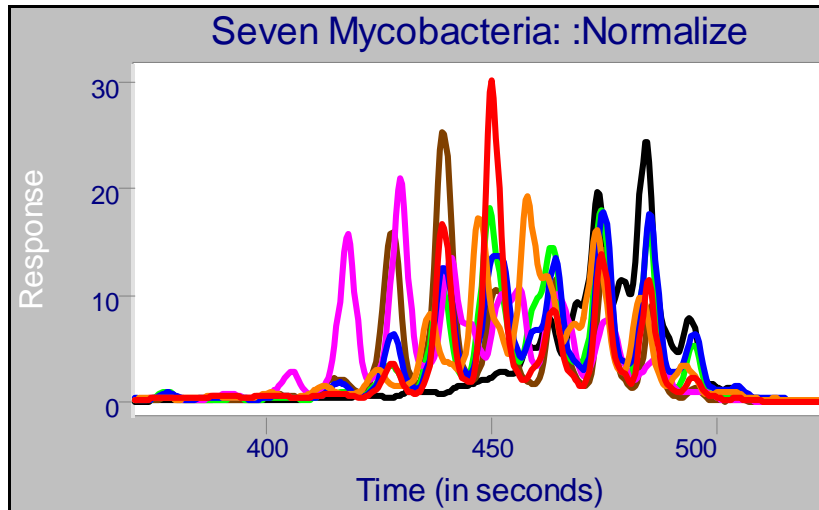
Peak integration and reporting, including export of the AIA files, was done with an Agilent GC ChemStation software package.

Multivariate analysis, including PCA, SIMCA, and the alignment steps were performed by Pirouette software, from Infometrix, Inc. Preprocessing for both algorithms included normalization (to vector length) and mean centering.

Training and evaluation sets were created for each experiment. From the 123 total samples, 87 were selected as training samples and 36 as evaluation samples. Partition into training and evaluation sets was repeated 10 times, with samples from each species divided into the two partitions by random selection.

Included in the evaluation set were 3 samples of one species. Because 3 samples were insufficient to make a quality SIMCA model, these samples were only placed in the evaluation subset and served to indicate false positives.

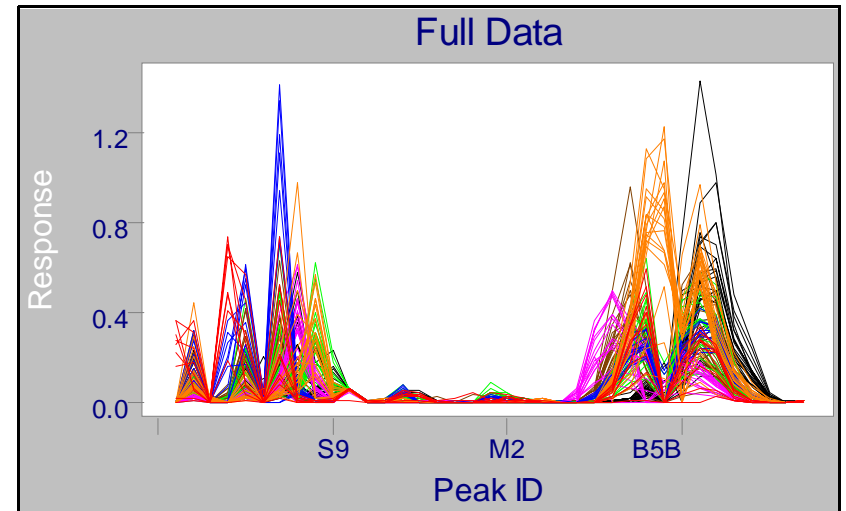
Below is a composite chromatogram, from the diagnostic late-eluting peak cluster, of profiles from each of the 7 similar Mycobacteria species.



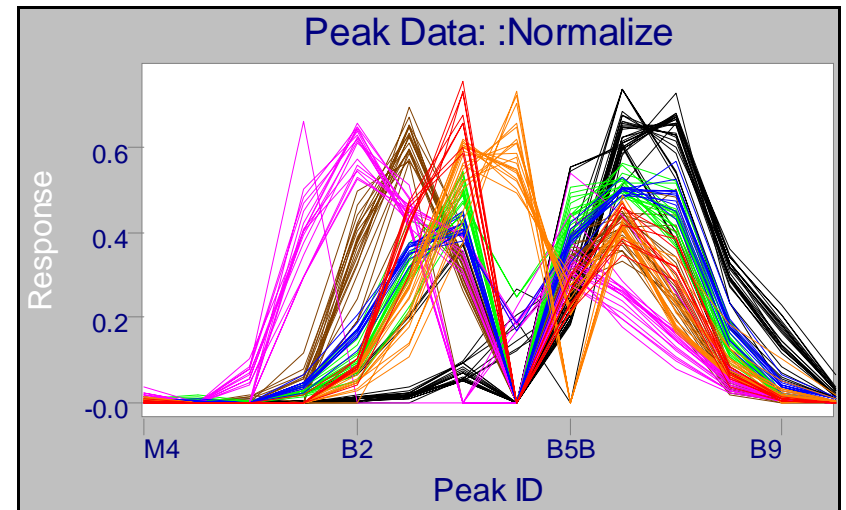
From these data, the profiles were pre-treated according to the approaches below, then processed by the multivariate algorithms.

Peak Data

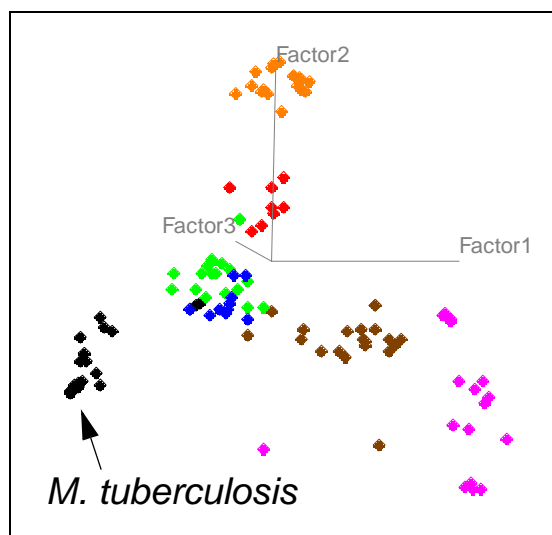
Each chromatogram was processed to integrate peaks, then tabulate their heights based on a peak ID table of 36 expected components. The following chart was made by superimposing a line plot of the peak heights for all of the samples in the study.



The diagnostic peaks that distinguish among the 7 species fall into the late-eluting cluster. The figure below emphasizes this region after each peak height vector was normalized.



Principal Components Analysis was run on these data, and a scores plot is shown below. From the plot, we see that the data fall into clusters, a visual verification that the species can be distinguished by their HPLC peak data.



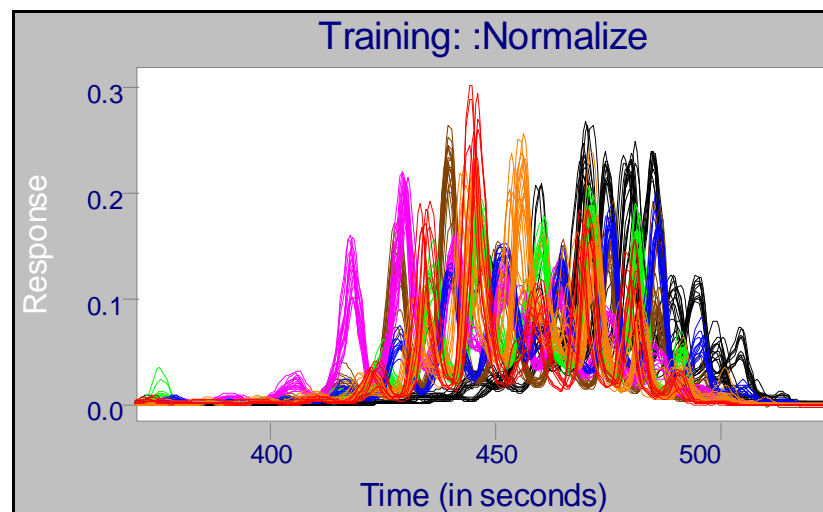
SIMCA models were made for each of the 10 randomly selected training sets, then the mycobacterial species was predicted for the evaluation set. These results are summarized in the first line of Table 1.

Table 1: Species identification success rate

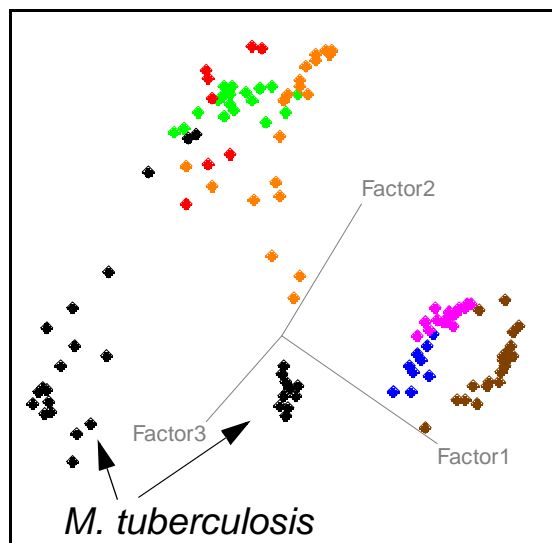
	Average	Std. Dev.	Range
Peak Data	86%	7.72%	72-97%
Profiles	88%	4.55%	81-94%
60 Bins	88%	4.19%	81-94%
30 Bins	87%	4.73%	78-94%

Whole Profile Data

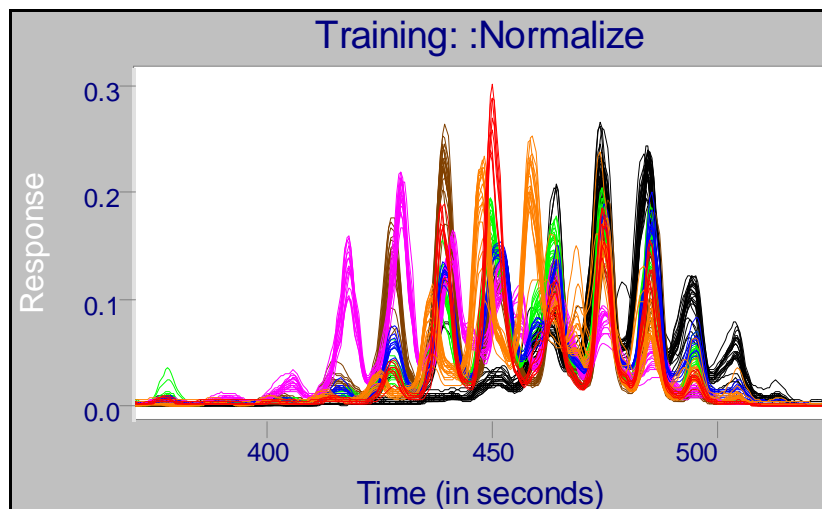
Chromatograms of the whole profile data are shown in the following plot, which emphasizes the diagnostic late-eluting peak cluster. To minimize the effects of different injection amounts, each profile has been normalized. Observe that profiles of the same color tend to band together, with different peaks eluting at different times or at different relative concentrations.



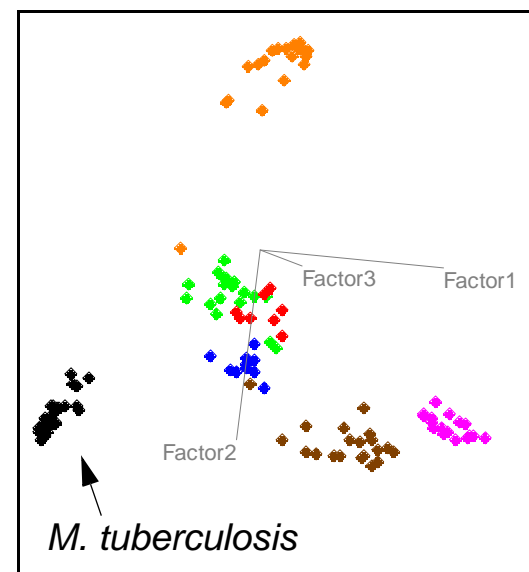
A PCA analysis of these data resulted in the following scores plot. Although the species still group in a distinguishable fashion, there are two clusters of samples of *M. tuberculosis*. This occurs because the samples were analyzed several months apart, and the retention characteristics of the LC column had changed.



An alignment algorithm was applied to these profiles, resulting in a more uniform set of profiles for each species, as shown below.



Now, the PCA scores show that the *M. tuberculosis* samples all cluster into a single group of points.



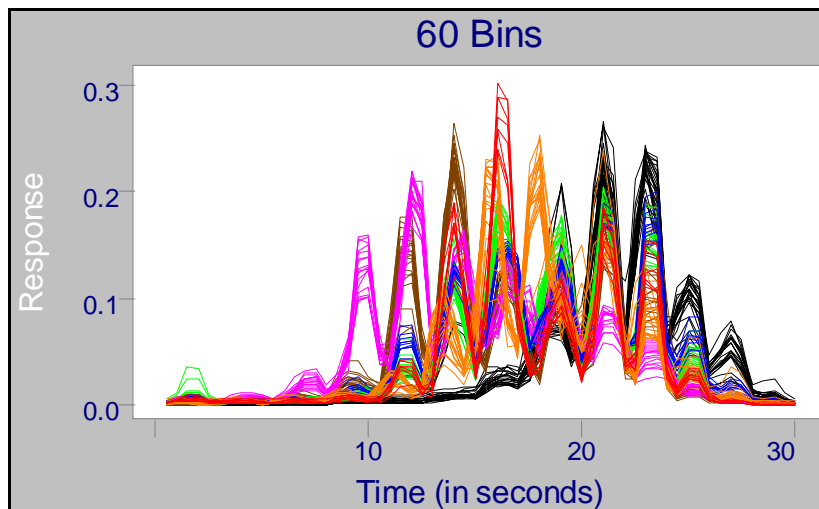
The aligned profile data was next modeled with the SIMCA algorithm, multiple times, and predictions of the mycobacteria species were made of the evaluation subsets. The results of the identification success on the evaluation set are shown in the second line of Table 1.

Note that the success rates for SIMCA on the peaks data and the profiles data are not significantly different. And while neither approach yields perfect predictions, there are several samples in the data set which are clearly outliers. When these outlier samples were withheld from the training set

and placed in the evaluation set, they were not properly classified, thus reducing the classification accuracy.

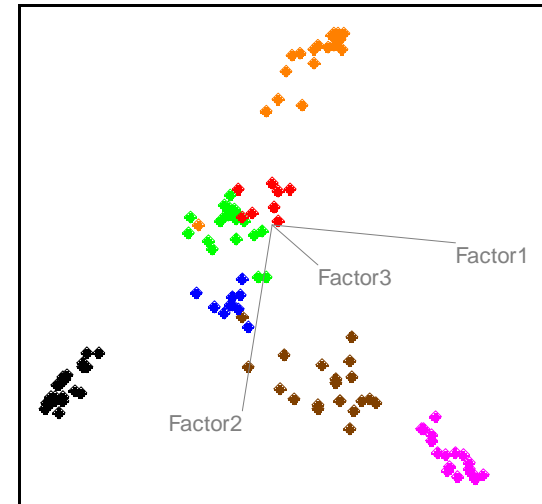
Binned Data

The aligned whole profile data were the starting point for the binning method. Initially, the binning was done such that the data were reduced by a factor of 5, resulting in data points from 60 bins. An overlay of the binned profiles is shown in the following figure.



When these data were analyzed by PCA, the resulting scores were not much

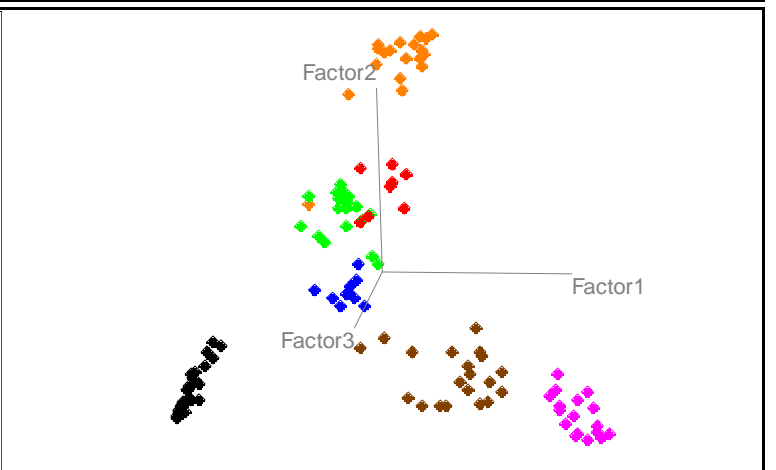
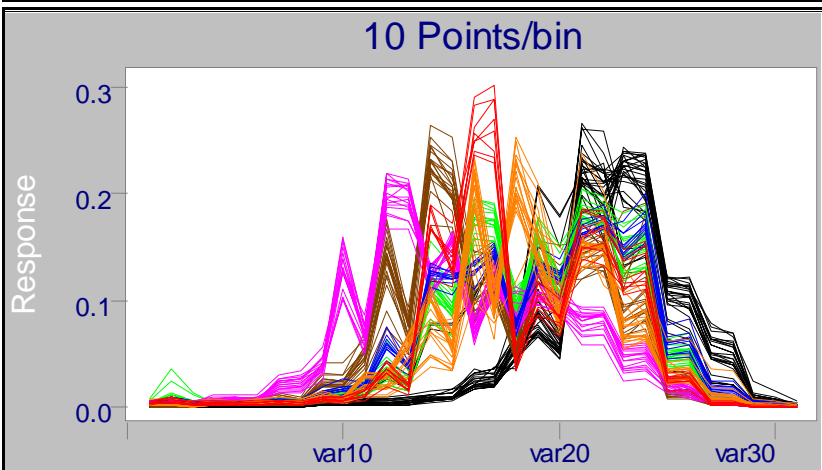
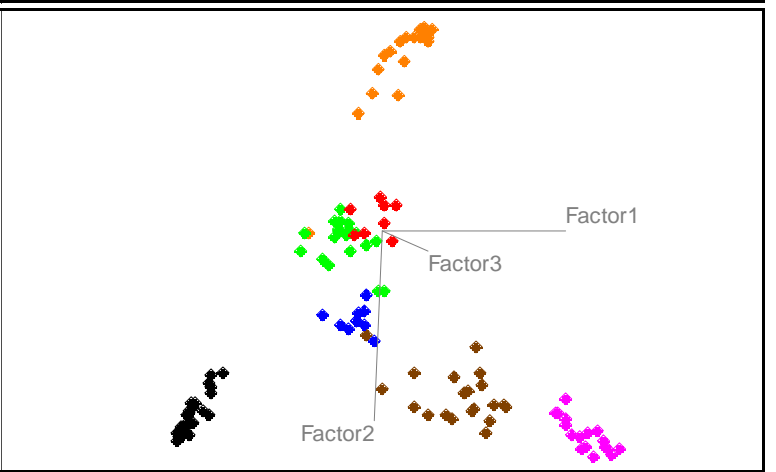
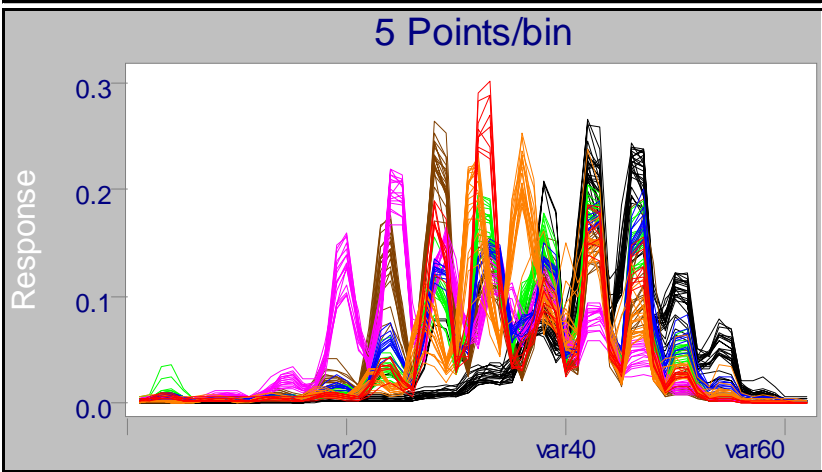
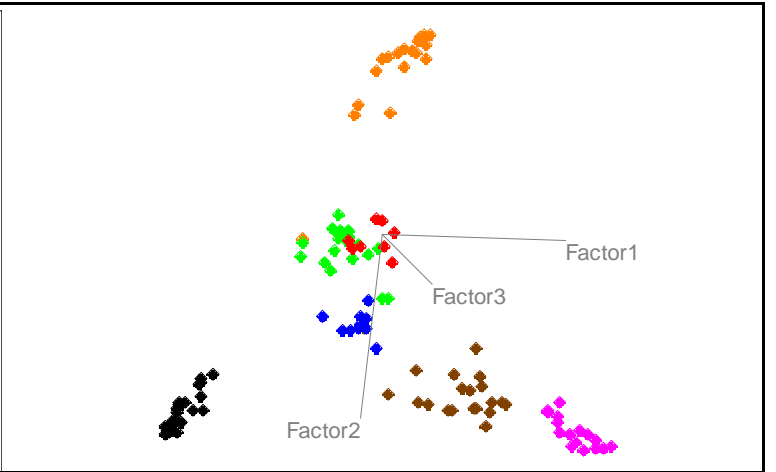
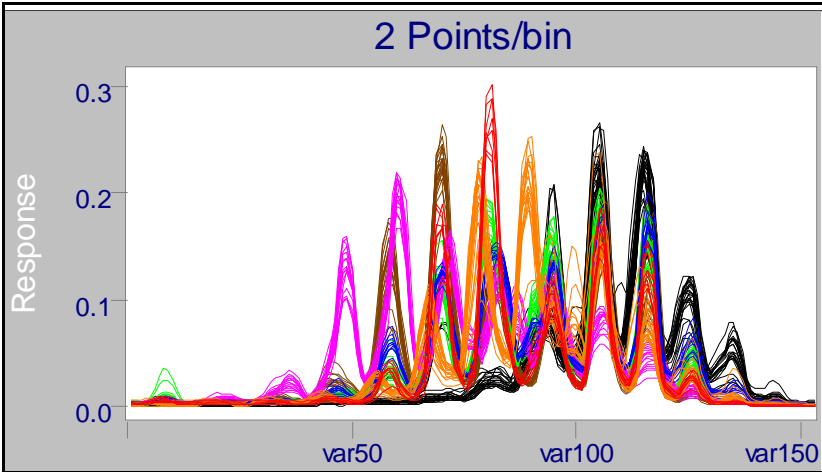
different than those shown earlier for the peaks and profiles data.

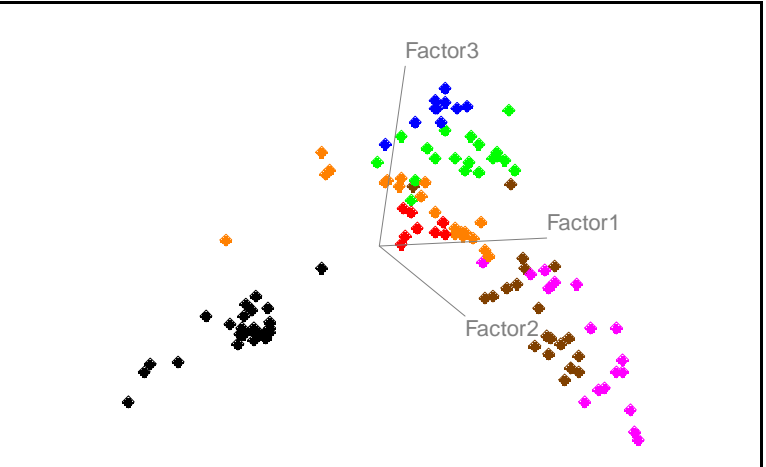
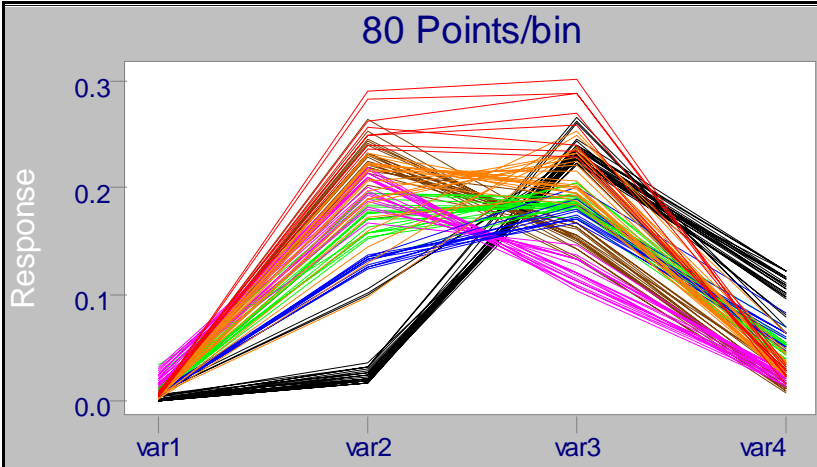
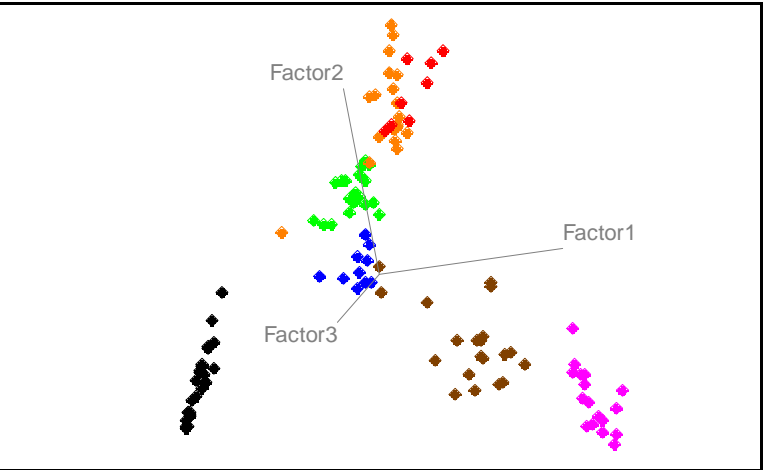
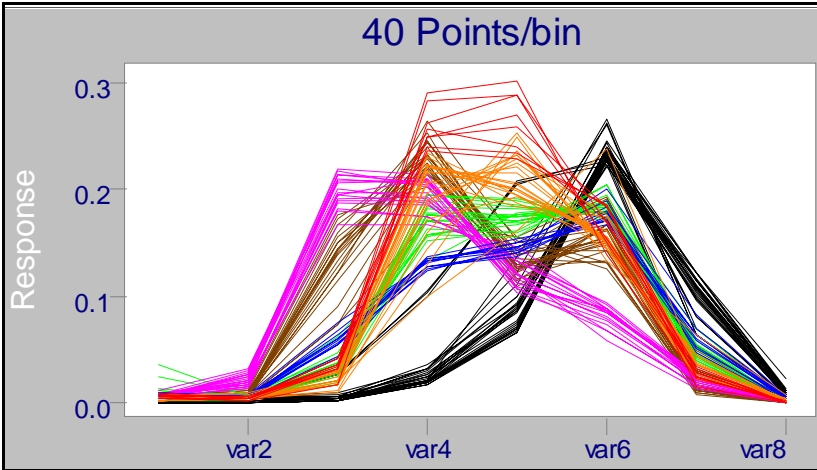
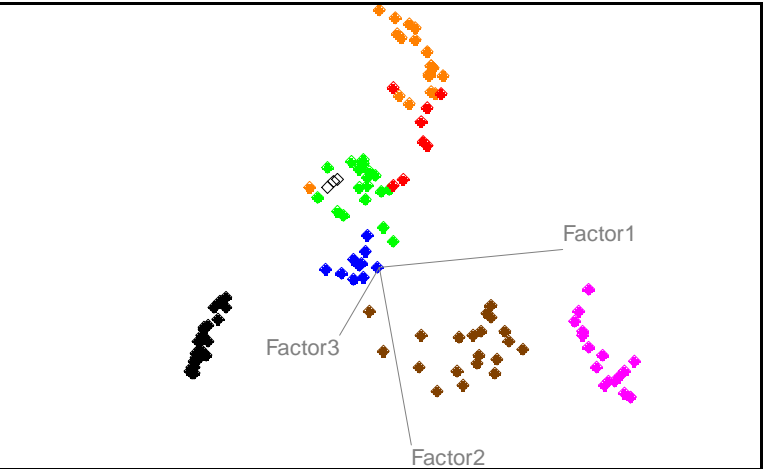
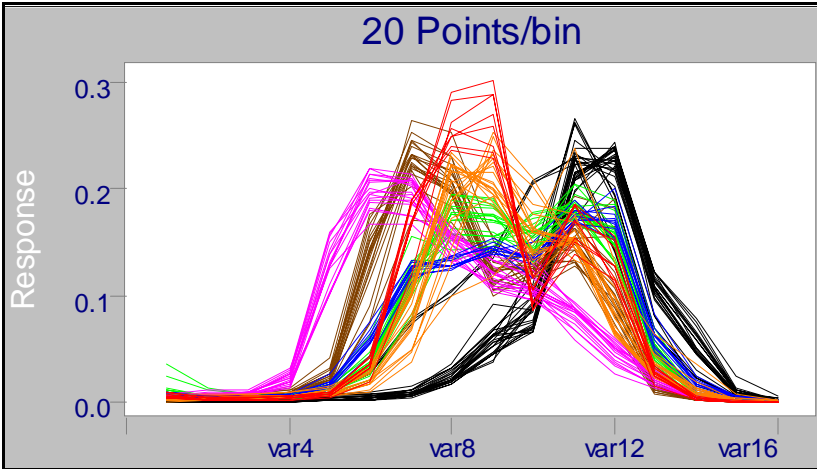


The SIMCA results on these data, shown in the third line of Table 1, are also similar. The binning experiment was repeated, generating only 30 bins, but these results yielded similar results (see Table 1, fourth line).

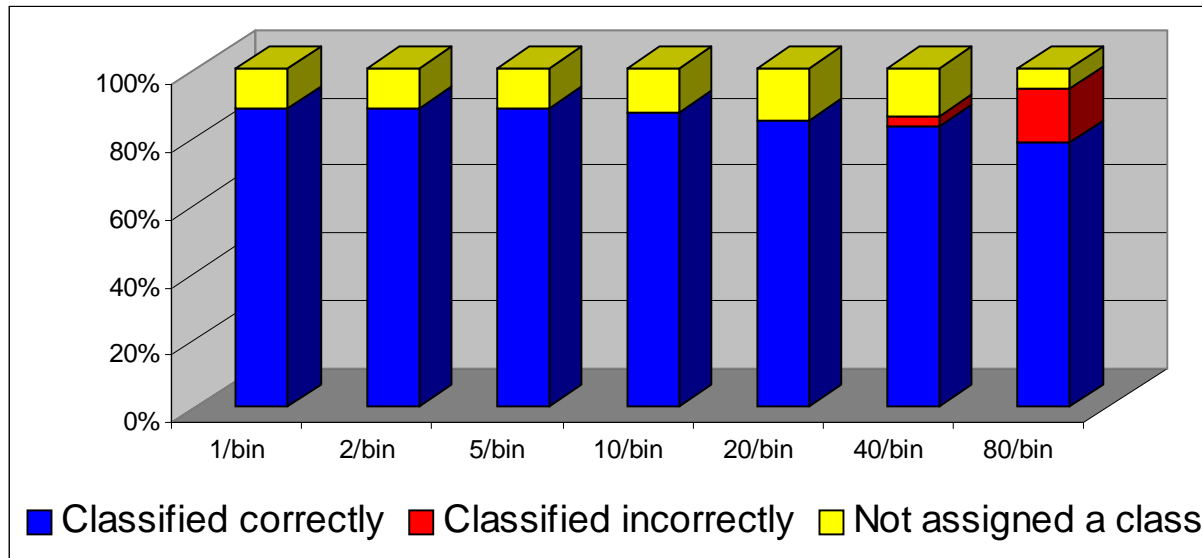
How far could we let the binning algorithm go in reducing the magnitude of the profiles before any degradation of classification success would be noticeable?

The following sets of figures show results from different degrees of binning.





Each of the binning experiments shown above was evaluated by SIMCA: a model was made from each of the randomly generated subsets, then a prediction of species was produced on the corresponding evaluation subset. A summary of these results is illustrated by the following graphic.



SIMCA was able to achieve a successful prediction on over 85% of the evaluation samples even when the number of data points had been reduced by 95% (20 points per bin). This level of classification success was not significantly different than when the whole chromatographic profile was used.

Misclassifications occurred when the sampling rate exceeded 20 points per bin. There is some evidence that this might be happening in the scores plots shown above. At 20 points per bin, the clusters shown at the top of the plot begin to overlap; this overlap becomes more severe at 40 points per bin.

Also, the cluster in brown begins to spread out more, losing its homogeneity. These types of perturbations in the distribution of points within a category will lead to inaccurate classifications.

An average of nearly 10% missed classifications emerge when the number of points per bin is increased to 80. As seen in the line plot above, this degree of sampling resulted in only 4 channels of data. When the goal is a distinction of at least 7 different categories, 4 variables will clearly be insufficient. At 40 points per bin, 8 variables remain; to achieve success at this data density would require that a substantial amount of species specificity be retained in each variable.

- Peak mis-assignments are difficult to eliminate from within the chromatography software.
- Pattern recognition using whole profiles is at least as successful as when using peak data.
- Pattern recognition using whole profiles requires that the profiles be aligned beforehand.
- Binning of whole profiles to reduce the data density does not affect the reliability of pattern recognition. For the data in this study, a 95% reduction in the amount of data collected yielded acceptable results.
- Where liquid chromatography is used as a tool to assist the characterization of samples, it may be possible to reduce the data rate and/or the analysis time and retain the diagnostic ability.
- Interpretation of misclassifications is made easier with profiles because we can look at specific features of the chromatogram.

Acknowledgements

Ken Jost, Texas Department of Health, provided the Mycobacteria data and offered considerable assistance in the understanding of bacterial differentiation.