Improving Investigative Lead Information and Evidential Significance Assessment for Automotive Paint by Development of Pattern Recognition Based Library Search Techniques

Barry K. Lavine
Department of Chemistry
Oklahoma State University
Stillwater, OK 74078

PDQ Database

- Automotive paint systems consist of multiple layers of paint: a clear coat over a color coat which in turn is over one or more undercoats. With the exception of the clear coat, each layer contains pigments and fillers and all layers contain binders.

- Automotive manufacturers tend to use unique combinations of pigments and binders in each paint layer. It is this unique combination that allows forensic scientists to determine the manufacturer, model, and year of a vehicle from a paint chip left at a crime scene.

- PDQ is a database of the physical attributes, the chemical composition and the infrared spectrum of each layer of the original manufacturer’s paint system.

Current Status of PDQ Searches

- PDQ was designed as a general text-based search and retrieval system.

- This text-based search of both physical and chemical characteristics serves as a potent pre-screen to a general infrared spectral search of materials that tend to be chemically very similar to one another.

- The concept is to narrow the list of possible vehicles to a reasonable number of suspects, not to identify an individual vehicle.

Clear Coats

- All too often, only a clear coat paint smear is left at the crime scene of a hit and run collision where damage to vehicles and injury or death has occurred.

- In these cases, the text based portion of the PDQ database will not be able to identify the make and model of the motor vehicle because all modern clear coats applied to any painted metal parts have only one of two possible formulations: acrylic melamine styrene or acrylic melamine styrene polyurethane.

- Search prefilters utilizing pattern recognition methods hold the potential for more specific searches relying less on somewhat subjective text-based characteristics.

Dark Secrets of IR Library Searching Algorithms for PDQ

- Most infrared search algorithms involve some type of point by point numerical comparison between the full spectrum of an unknown and each member of the library.
- These algorithms lack interpretive ability because they treat the spectrum as a set of points rather than as a collection of specific bands.
- Band shifting is not handled well and bands of low intensity, which may be highly informative, are often ignored.
- As the size of the IR library increases, the likelihood of a match will increase even though the spectrum of the unknown may not be present in the library

Search Prefilters

- A search prefilter is a quick test to spot dissimilar spectra, thereby avoiding a complete spectral comparison.
- Utilizing search prefilters, many of aforementioned problems encountered in library searching can be successfully addressed.
- From an interpretive standpoint, the information contained in the prefilter should be based on identifying the factory that has produced the paint.

Development of Search Prefilters for the PDQ Database

- Using the wavelet packet transform, PDQ library spectra will be passed through two scaling filters: a high pass filter and a low pass filter
- The decomposition process which utilizes wavelet coefficients that represent the high and low frequency components of the signal will then be iterated using successive wavelet packets until the required level of signal decomposition is achieved.
- Wavelet coefficients characteristic of plant are identified by a genetic algorithm for pattern recognition analysis.

PDQ Data Set

- IR spectra from six **Chrysler plants (1999)** obtained from the PDQ database were selected for analysis.

- Each of the six plants (BRA, STL, JFN, STH, SAL, and NEW) was represented by at least 10 paint samples.

- For this study, only IR spectra of clear coats from metal substrates obtained using BioRad 40A or BioRad 60 spectrometer were used.

- The data set was divided into a training set of 88 paint samples and a validation set of 3 paint samples.

- The initial focus of the study was the training set samples.

Paint Dataset (91 samples)

- All samples are clear coats
- All samples are from metal parts

Plant	Number of samples
BRA (1)	25
STL (2)	21
JFN (3)	13
STH (4)	9
SAL (5)	12
NEW (6)	11

Part	Number of samples
Roof	68
Hood	9
Fender	10
Door	2
Hatchback	1
Trunk	1

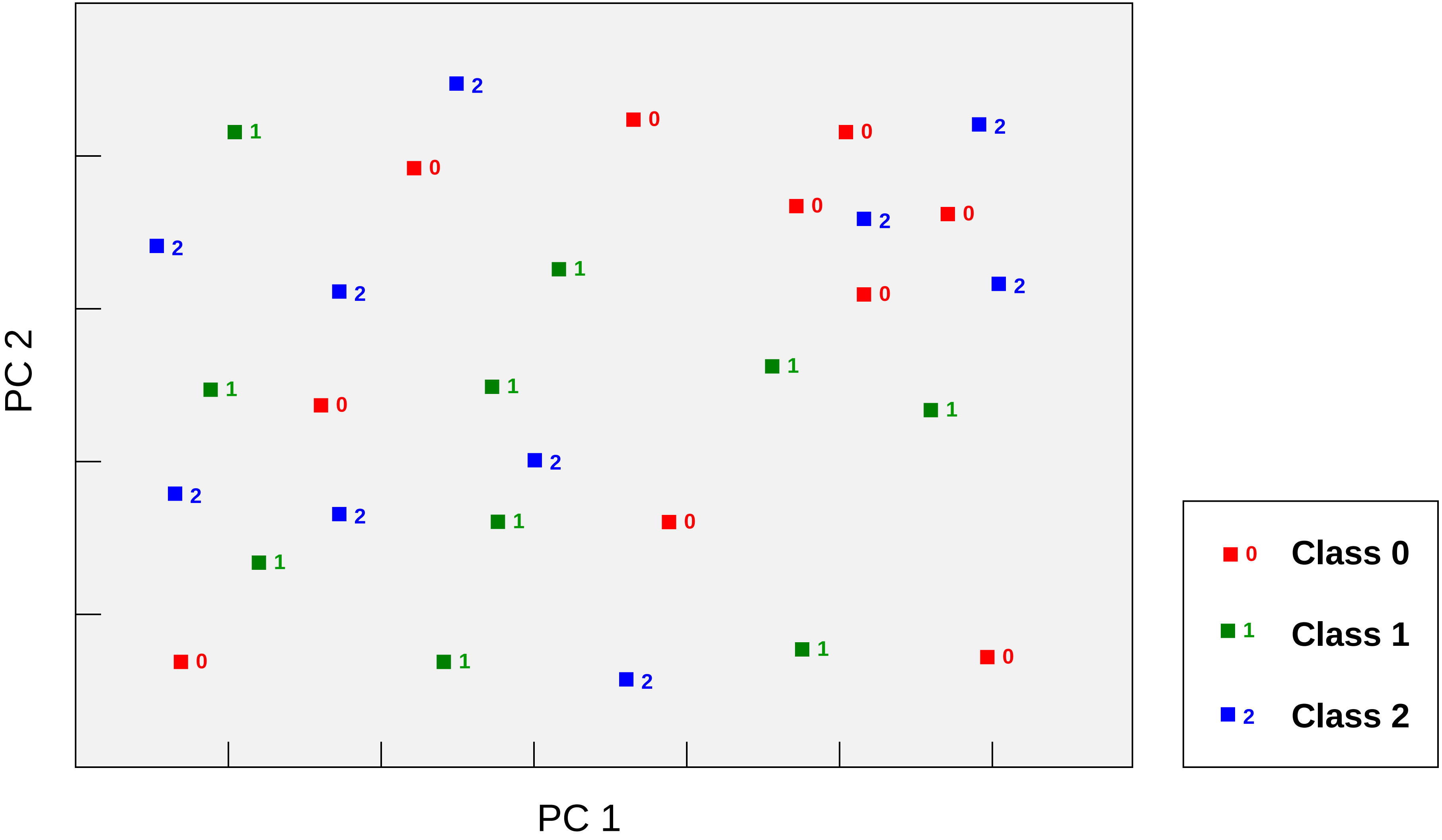
Pattern Recognition Analysis Using the Original Spectral Data

- Each IR spectrum (data vector) was **normalized** to unit length.
- All spectral features were **autoscaled** such that each measurement has a mean of zero and a standard deviation of one.
- **Autoscaling** removed any inadvertent weighing of the data that otherwise would occur due to differences in the magnitude among the measurement variables comprising the data set.

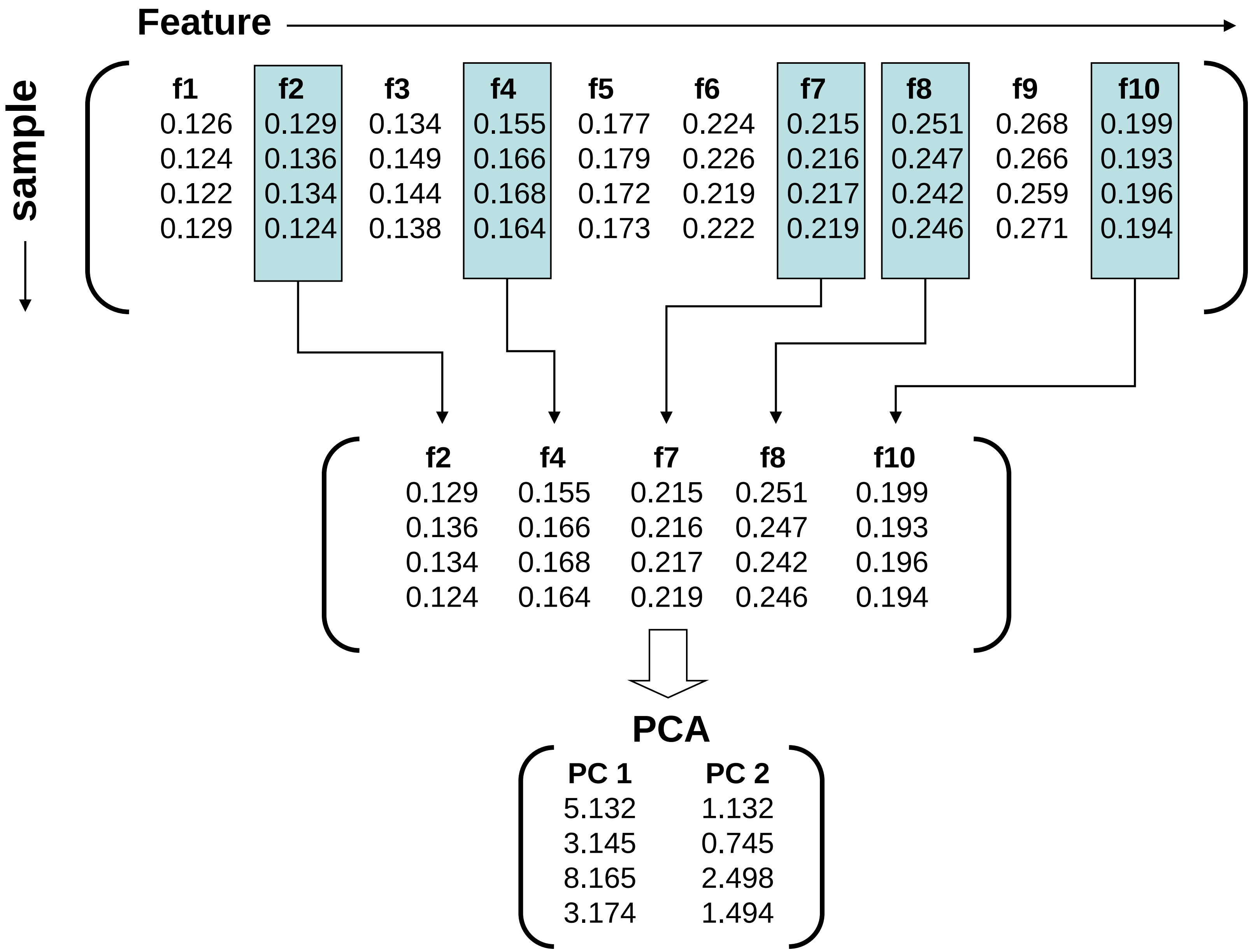
Feature Selection Using the Pattern Recognition GA

- A set of features that optimize the separation of the classes in a plot of the two or three largest principal component of the training set data is identified using a genetic algorithm.
- Because principal components maximize variance, the bulk of the information encoded by these feature sets will be about differences between classes in the data set.

PC plot before **Wavelet Coefficient Selection**



Wavelet Coefficient Selection



Feature

sample

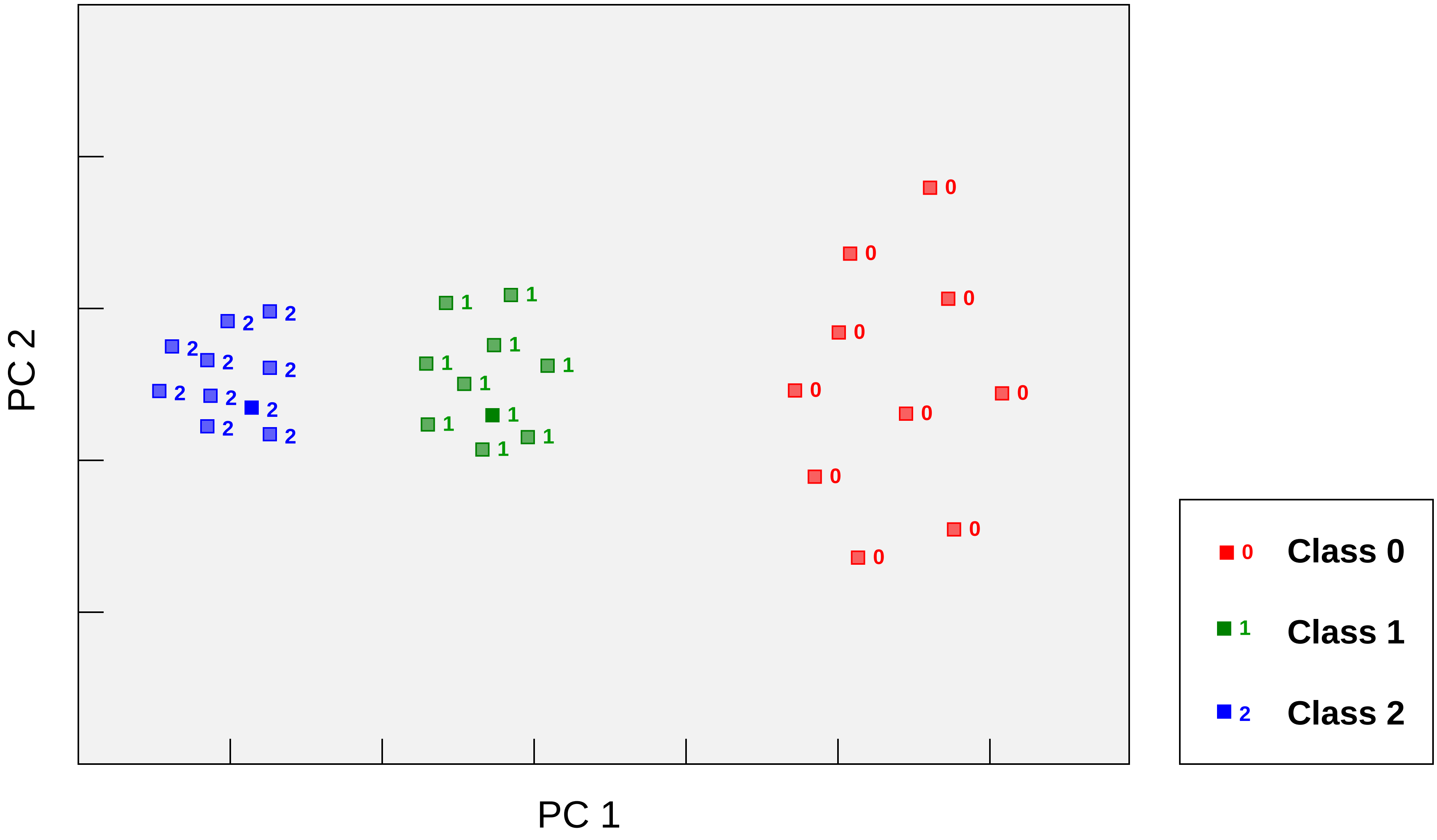
f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
0.126	0.129	0.134	0.155	0.177	0.224	0.215	0.251	0.268	0.199
0.124	0.136	0.149	0.166	0.179	0.226	0.216	0.247	0.266	0.193
0.122	0.134	0.144	0.168	0.172	0.219	0.217	0.242	0.259	0.196
0.129	0.124	0.138	0.164	0.173	0.222	0.219	0.246	0.271	0.194

f2	f4	f7	f8	f10
0.129	0.155	0.215	0.251	0.199
0.136	0.166	0.216	0.247	0.193
0.134	0.168	0.217	0.242	0.196
0.124	0.164	0.219	0.246	0.194

PCA

PC 1	PC 2
5.132	1.132
3.145	0.745
8.165	2.498
3.174	1.494

PC plot after **Feature Selection**



Advantages

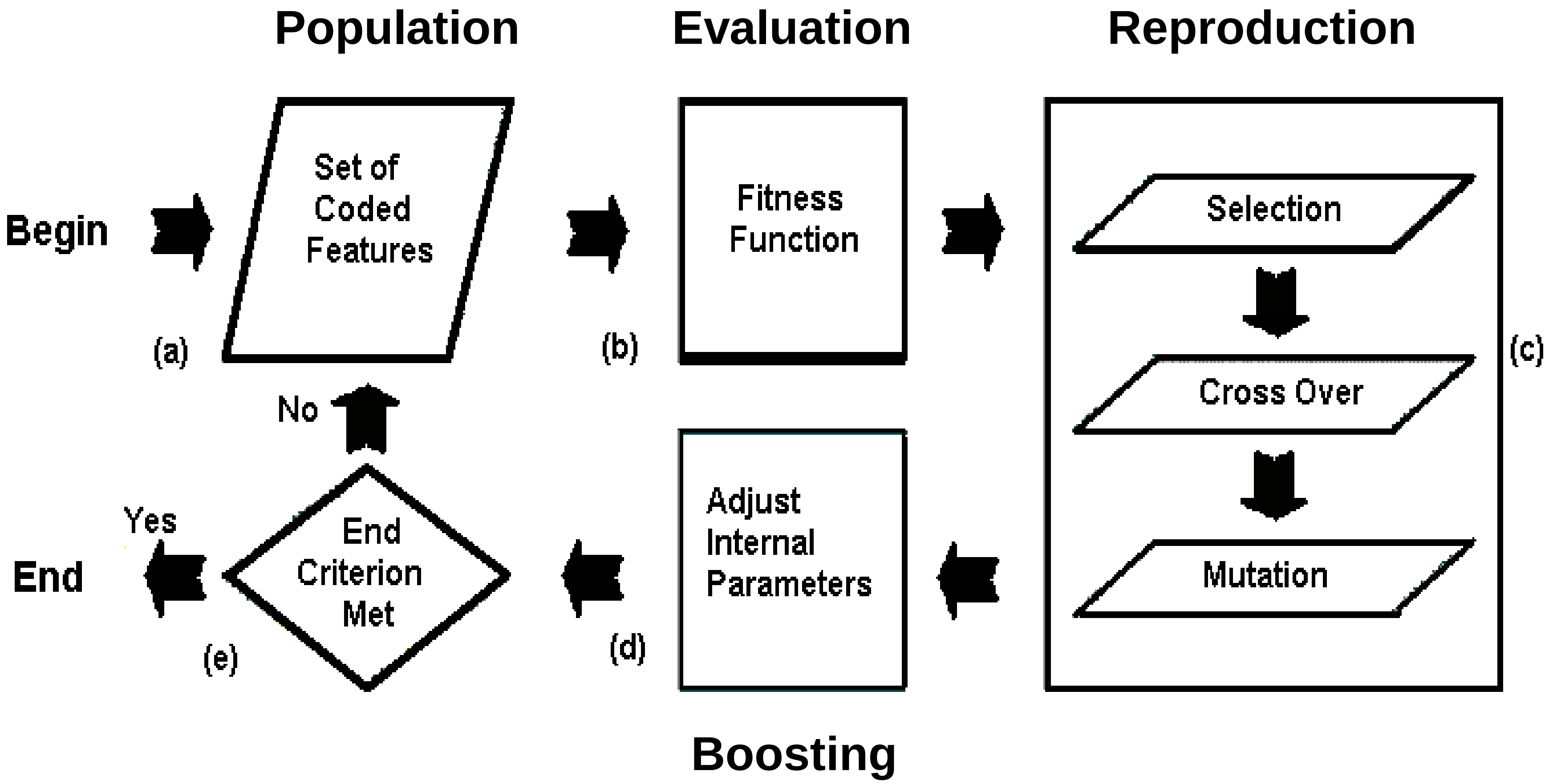
- Chance classification will not be a serious problem since the bulk of the variance or information content of the feature subset selected is about the pattern recognition problem of interest.
- Features that contain discriminatory information about a particular class membership problem are usually correlated, which is why feature selection methods based on PCA are preferred.

Information Filter

- The principal component plot functions as an embedded information filter.
- Feature sets are selected based on their PC plots.
- A good PC plot can only be generated using features whose variance or information is primarily about class differences.
- Hence, PCA limits our search to these types of feature sets, thereby significantly reducing the size of the search space.

Genetic Algorithm

for Pattern Recognition



GA Publications

- B. K. Lavine and Anthony Moores, "Genetic Algorithms for Pattern Recognition Analysis of High Speed Gas Chromatograms of Weathered and Unweathered Jet Fuels," **Buletin Kimia**, 1997, 12(2), 73-86.
- B. K. Lavine, A. J. Moores, H. T. Mayfield, and A. Faruque, "Fuel Spill Identification using Gas Chromatography/Genetic Algorithms-Pattern Recognition Techniques," **Analytical Letters**, 1998, 31(15), 2805-2822.
- B. K. Lavine, A. J. Moores, and L. K. Helfend, "A Genetic Algorithm for Pattern Recognition Analysis of Pyrolysis Gas Chromatographic Data," **J. Anal. Appl. Pyrolysis**, 1999, 50, 47-62.
- B. K. Lavine, A. J. Moores, H. T. Mayfield, and A. Faruque, “Genetic Algorithms Applied to Pattern Recognition Analysis of High Speed Gas Chromatograms of Aviation Turbine Fuels Using an Integrated Jet-A/JP-8 Data Base,” **Microchemical Journal**, 1999, 61, 69-78.
- B. K. Lavine and A. J. Moores, “Genetic Algorithms for Pattern Recognition Analysis and Fusion of Sensor Data,” in Pattern Recognition, Chemometrics, and Imaging for Optical Environmental Monitoring, K. Siddiqui and D. Eastwood (Eds.), **Proceedings of SPIES**, 1999, pp. 103-112.

GA Publications

- B. K. Lavine, J. Ritter, A. J. Moores, M. Wilson, A. Faruque, and H. T. Mayfield, “Source Identification of Underground Fuel Spills by Solid Phase Micro-extraction/High-Resolution Gas Chromatography/Genetic Algorithms,” **Anal. Chem**., 2000, 72(2), 423-431.
- B. K. Lavine, A. J. Moores, and J. P. Ritter, “Underground Fuel Spills, Source Identification,” in the Encyclopedia of Analytical Chemistry: Instrumentation and Applications,” Environment – Water and Waste, 2000, Volume 4,” Edited by R. A. Meyer, pp. 3495-3515.
- B. K. Lavine, D. Brzozowski, J. Ritter, A. J. Moores, and H. T. Mayfield, “Fuel Spill Identification by Selective Fractionation Prior to Gas Chromatography I. Water Soluble Components,” **J. Chromat. Sci.,** 2001, 39(12), 501-506
- B. K. Lavine, D. Brzozowski, A .J. Moores, C. E. Davidson, and H.T. Mayfield, “Genetic Algorithm for Fuel Spill Identification,” **Anal. Chim. Acta**, 2001, 437(2), 233-246
- B. K. Lavine, C. E. Davidson, A. J. Moores, and P. R. Griffiths, “Raman Spectroscopy and Genetic Algorithms for the Classification of Wood Types,” **Applied Spectroscopy**, 2001, 55(8), 960-966.
- B. K. Lavine, A. Vesanen, D. M. Brzozowski, and H. T. Mayfield “Authentication of Fuel Standards using Gas Chromatography/Pattern Recognition Techniques,” **Anal Letters**, 2001, 34(2), 281- 294

GA Publications

- B. K. Lavine, C. E. Davidson, and A. J. Moores, “Innovative Genetic Algorithms for Chemoinformatics, “**Chemometrics & Intelligent Laboratory Instrumentation**, 2002, 60(1), 161-171.
- B. K. Lavine, C. E. Davidson, and A. J. Moores, “Genetic Algorithms for Spectral Pattern Recognition,” **Vibrational Spectroscopy**, 2002, 28(1), 83-95.
- B. K. Lavine, C. E. Davidson, Robert K. Vander Meer, S. Lahav, V. Soroker, and A. Hefetz, “Genetic Algorithms for Deciphering the Complex Chemosensory Code of Social Insects,” **Chemometrics & Intelligent Laboratory Instrumentation**, 2003, 66(1), 51-62.
- B. K. Lavine, C. E. Davidson, C. Breneman, and W. Katt, “Electronic Van der Waals Surface Property Descriptors and Genetic Algorithms for Developing Structure-Activity Correlations in Olfactory Databases,” **J. Chem. Inf. Science,** 2003, 43, 1890-1905.
- B. K. Lavine, C. E. Davidson, and W. T. Rayens, “Machine Learning Based Pattern Recognition Applied to Microarray Data, **Combinatorial Chemistry & High Throughput Screening**,” 2004, 7, 115-131.
- B. K. Lavine, C. E. Davidson, C. Breneman, and W. Katt, “Genetic Algorithms for Clustering and Classification of Olfactory Stimulants,’’ (in Chemoinformatics: Methods and Protocols) J. Bajorath (Ed.), **Methods Mol Biol.,** Humana Press, 2004, 275, 399-426.

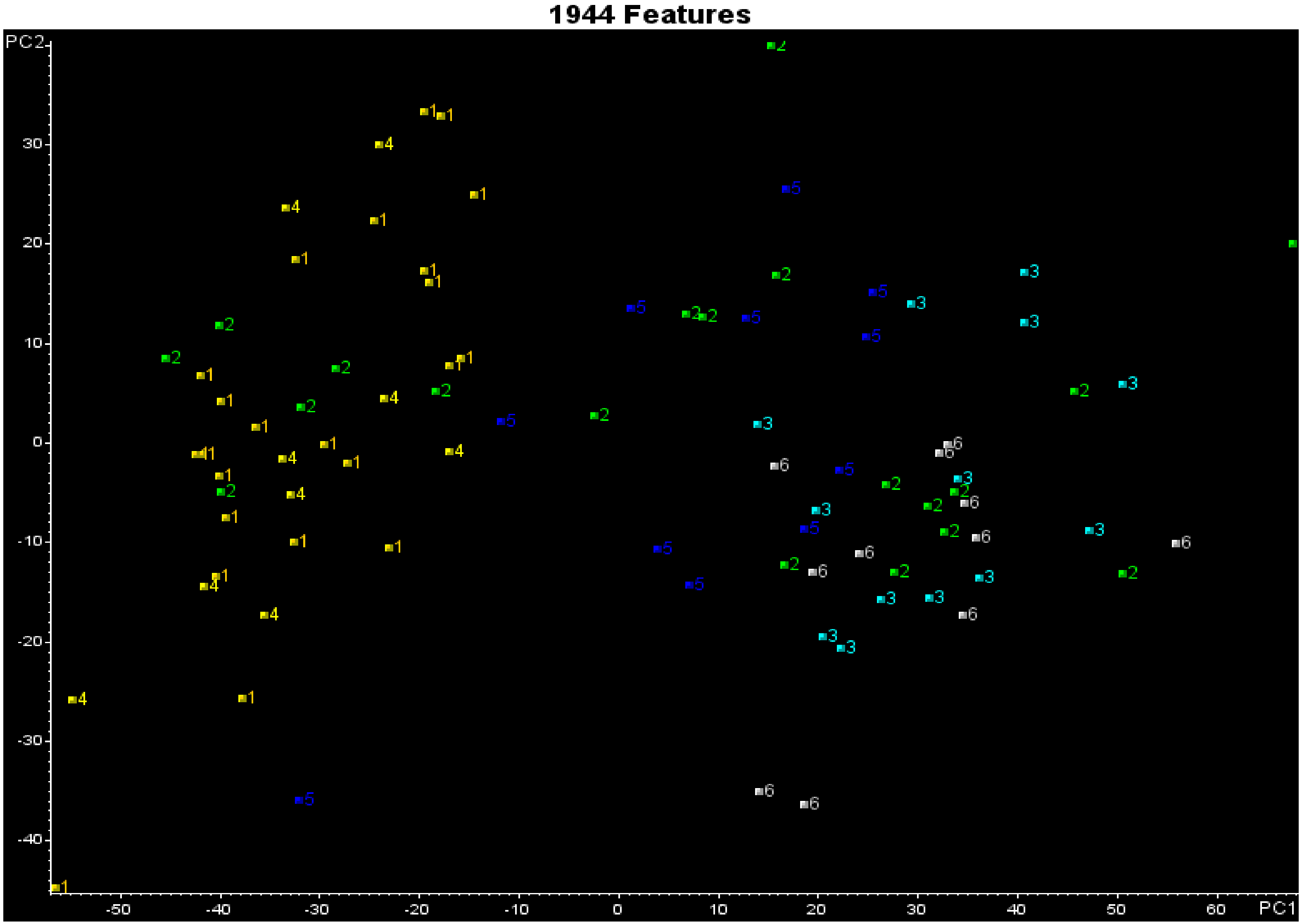
GA Publications

- B. K. Lavine, C. E. Davidson, and D. J. Westover, “Spectral Pattern Recognition Using Self Organizing Maps,” **J. Chem. Inf. Comp. Science,** 2004, 44(3), 1056-1064
- B. K. Lavine, C. E. Davidson, C. Breneman, and W. Katt, “Development of Structure-Activity Olfactory Correlations using Electronic Van der Waals Surface Property Descriptors and Genetic Algorithms,” in Chemometrics and Chemoinformatics, B. K. Lavine (Ed.), **ACS Symposium Series**, 894, 2005, pp. 127-143.
- B. K. Lavine and Mehul Vora, “Identification of Africanized Honeybees,” **Journal Chromatography A**, 2005, 1096, 69-75.
- J. Karasinski, S. Andreescu, O. A Sadik, B. Lavine, and M. N. Vora, “Multiarray Sensors with Pattern Recognition for the Detection, Classification, and Differentiation of Bacteria at Subspecies and Strain Levels,” **Anal. Chem**., 2005, 77(24), 7941-7949.
- G. A. Eiceman, M. Wang, S. Pradad, H. Schmidt, F. K. Tadjimukhamedov, Barry K. Lavine, and Nikhil Mirjankar, “Pattern Recognition Analysis of Differential Mobility Spectra with Classification by Chemical Family,” **Anal. Chem.. Acta**, 2006, 579(1), 1-10
- J. Karasinski, L. White, Y. Zhang, E. Wang, S. Andreescu, O. A Sadik, B. Lavine, and M. N. Vora, “Detection and identification of bacteria using antibiotic susceptibility and a multi-array electrochemical sensor with pattern recognition,” Biosensors & Bioelectronics (2007), 22(11), 2643-2649

Principal Component Analysis

- The first step in this study was to apply principal component analysis (PCA) to the data.
- Using this procedure is analogous to finding a new coordinate system that is better at displaying the information content of the data than axes defined by the original measurement variables.
- This new coordinate system is linked to variation in the data.
- Often, only two or three principal components are necessary to explain most of the information present in spectral data due to the large number of interrelated measurements

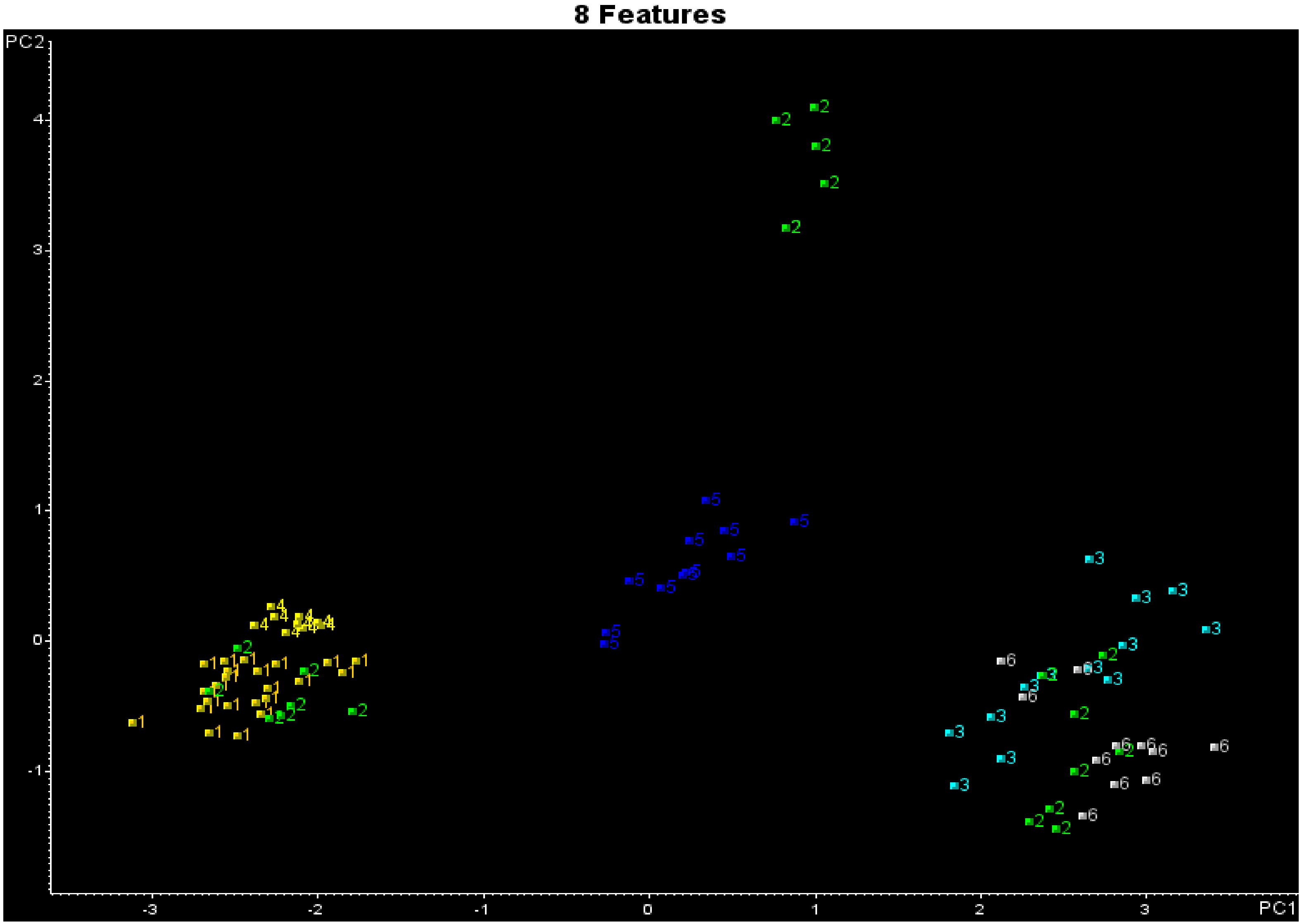
Plot of two largest Principal Components developed from all 1944 wavelengths for the training set



Searching the Solution Space

- The GA identified features by sampling key feature subsets, scoring their PC plots, and tracking those classes and samples that were difficult to classify.
- The boosting routine used this information to steer the population to an optimal solution.
- After 300 generations, the GA identified 8 wavelengths which contain discriminatory information about the manufacturing plant.

Score Plot developed from 8 wavelengths selected by Pattern Recognition GA for the training set



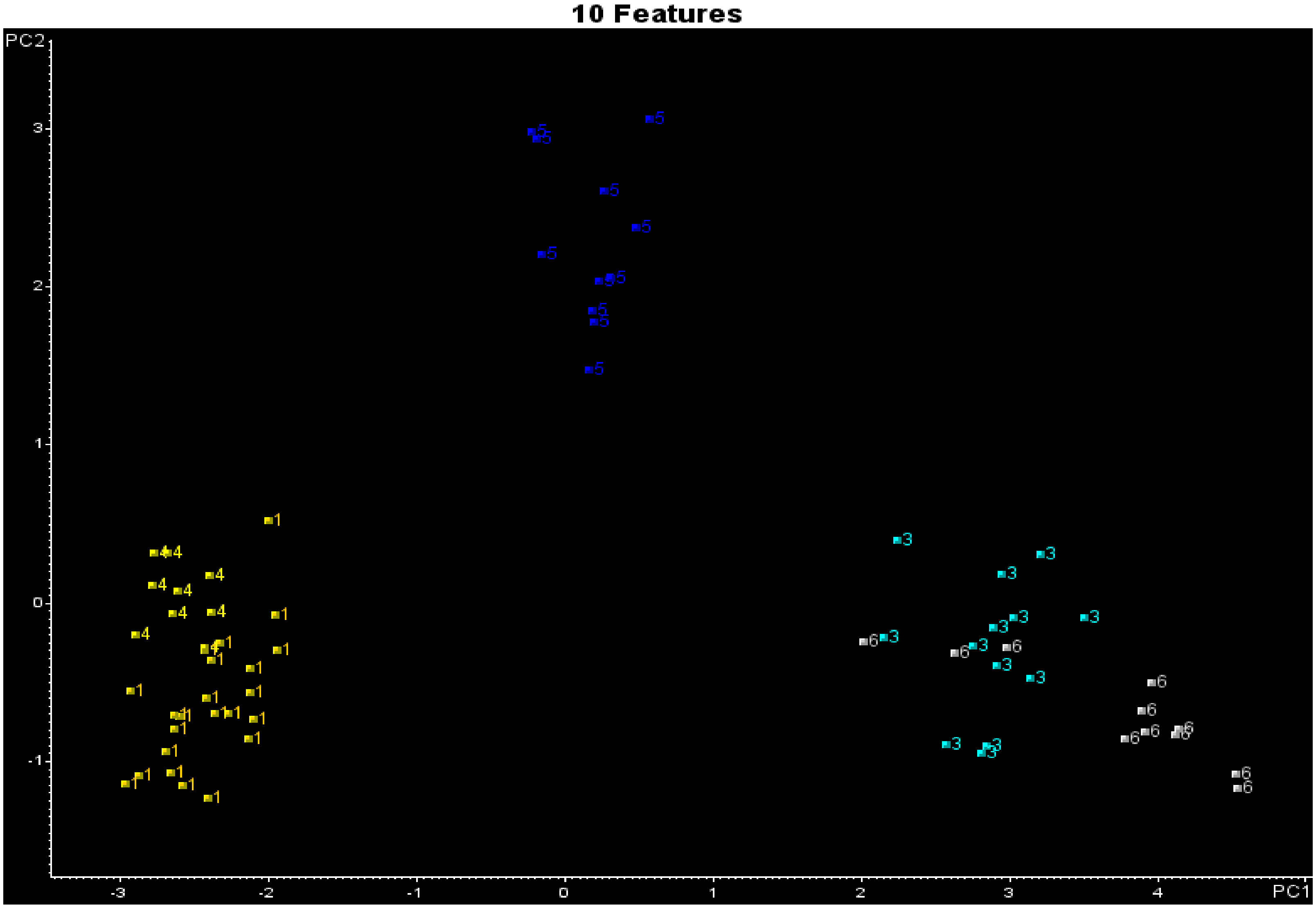
Observations

- Class 2 (STL) appears to be composed of 3 different types of samples.
- Class 5 (SAL), which forms a well defined cluster of points, is well separated from the other manufacturing plants.
- Classes 1 (BRA) and 4 (STH) are adjacent and separated from each other in the PC space.
- Classes 3 (JFN) and 6 (NEW) are adjacent but not fully separated from each other in the PC space.

Removal of STL Samples

- STL (Class 2) samples were removed from the data set and the pattern recognition analysis was repeated again using our genetic algorithm for feature selection.
- 10 wavelengths were identified using the pattern recognition GA.
- The score plot developed from these 10 wavelengths is shown in the following slide.
- Again, the trends that were reported in the previous study are again observed.

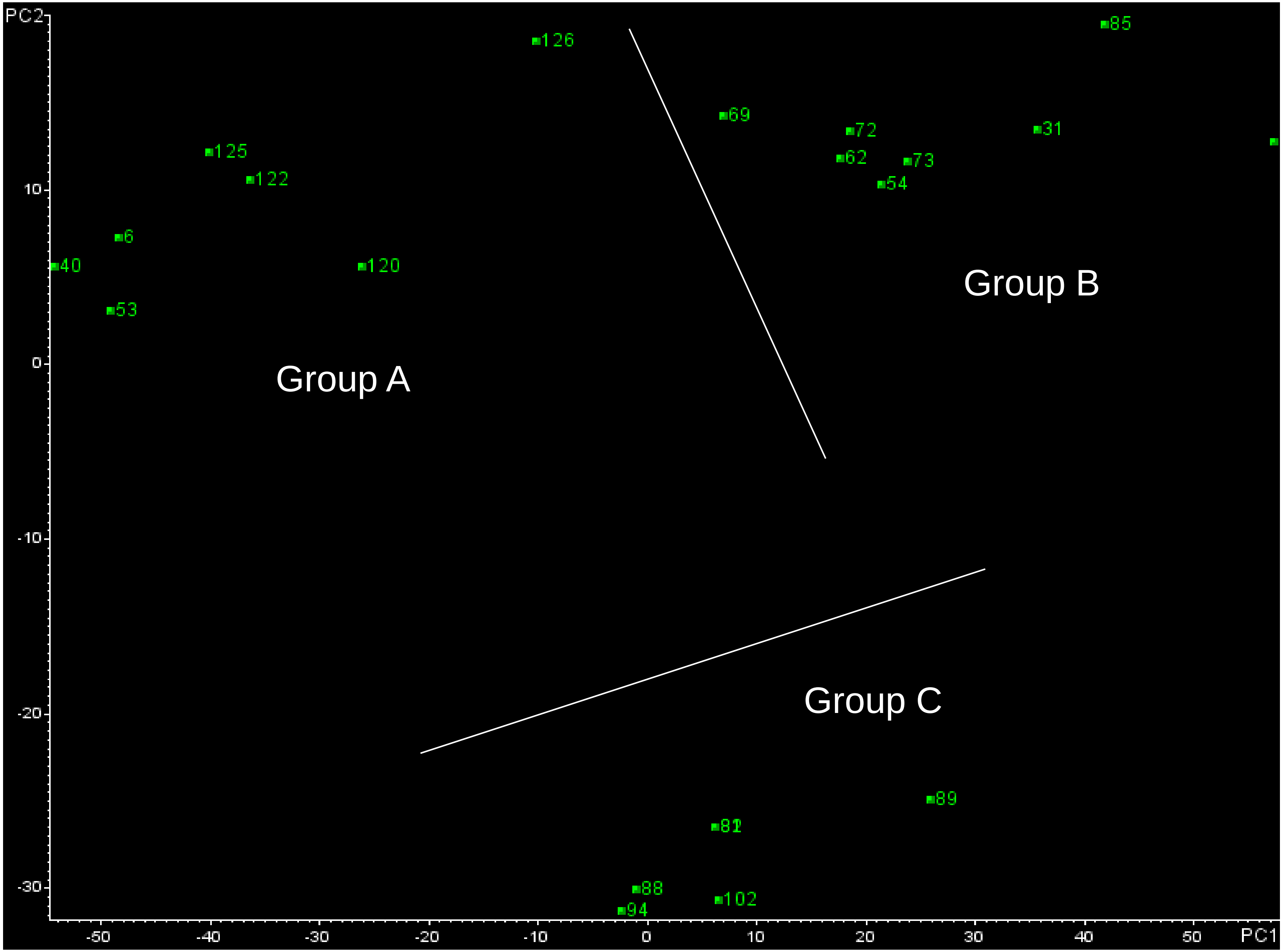
Score Plot developed from 10 Wavelengths Selected by Pattern Recognition GA for the training set



Analysis of IR Spectra

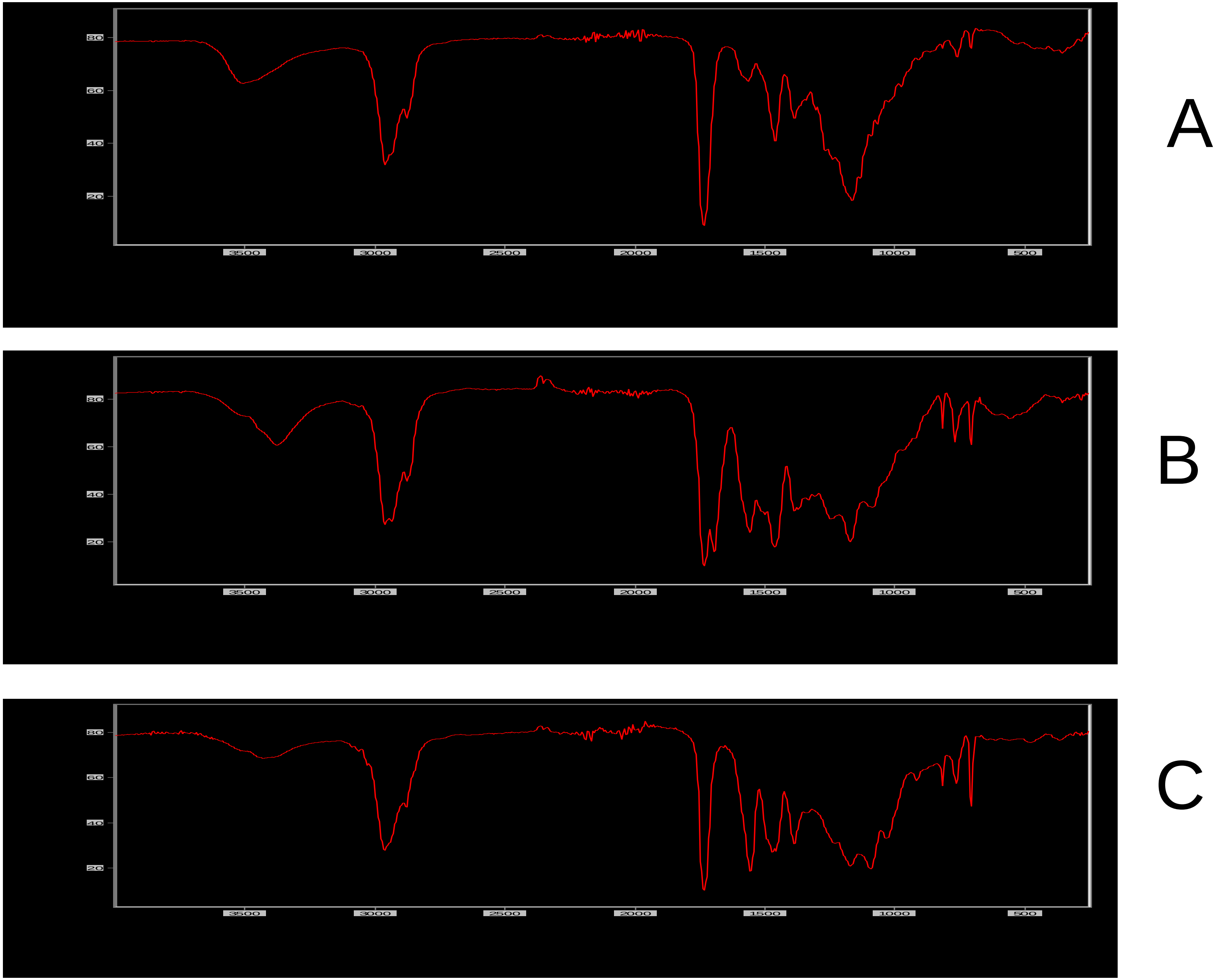
- To better understand the situation involving the STL samples, a PC plot of all 1944 features was generated for the STL samples.
- Clustering in 3 distinct groups is evident, which corresponds to the clustering shown by STL samples in the original 6-class study.
- Each group has a distinctive IR spectrum as shown in the accompanying slides.

1944 Features



Analysis of IR Spectra

- IR spectra from Groups A, B, and C are compared.
- Within each group, the spectra are similar.
- However, spectral differences between groups are evident.


80
60
40
20
3500
3000
2500
2000
1500
1000
500
A
80
60
40
20
3500
3000
2500
2000
1500
1000
500
B
80
60
40
20
3500
3000
2500
2000
1500
1000
500
C

Mystery in St. Louis

- Group A corresponds to SUV’s (Plymouth Voyager, Dodge Grand Caravan, and Chrysler Town and Country).

- Groups B and C correspond to Dodge RAM trucks

- During the production year Chrysler made a change in the clear coat formulation used at the St. Louis North Plant
  - Group B samples fall under the BASF supplied Duraclear II clear coat and has a chemistry of acrylic, melamine, styrene, and polyurethane
  - Group C falls under the DuPont supplied Gen IV AW clear coat and has the chemistry acyclic, melamine, and styrene

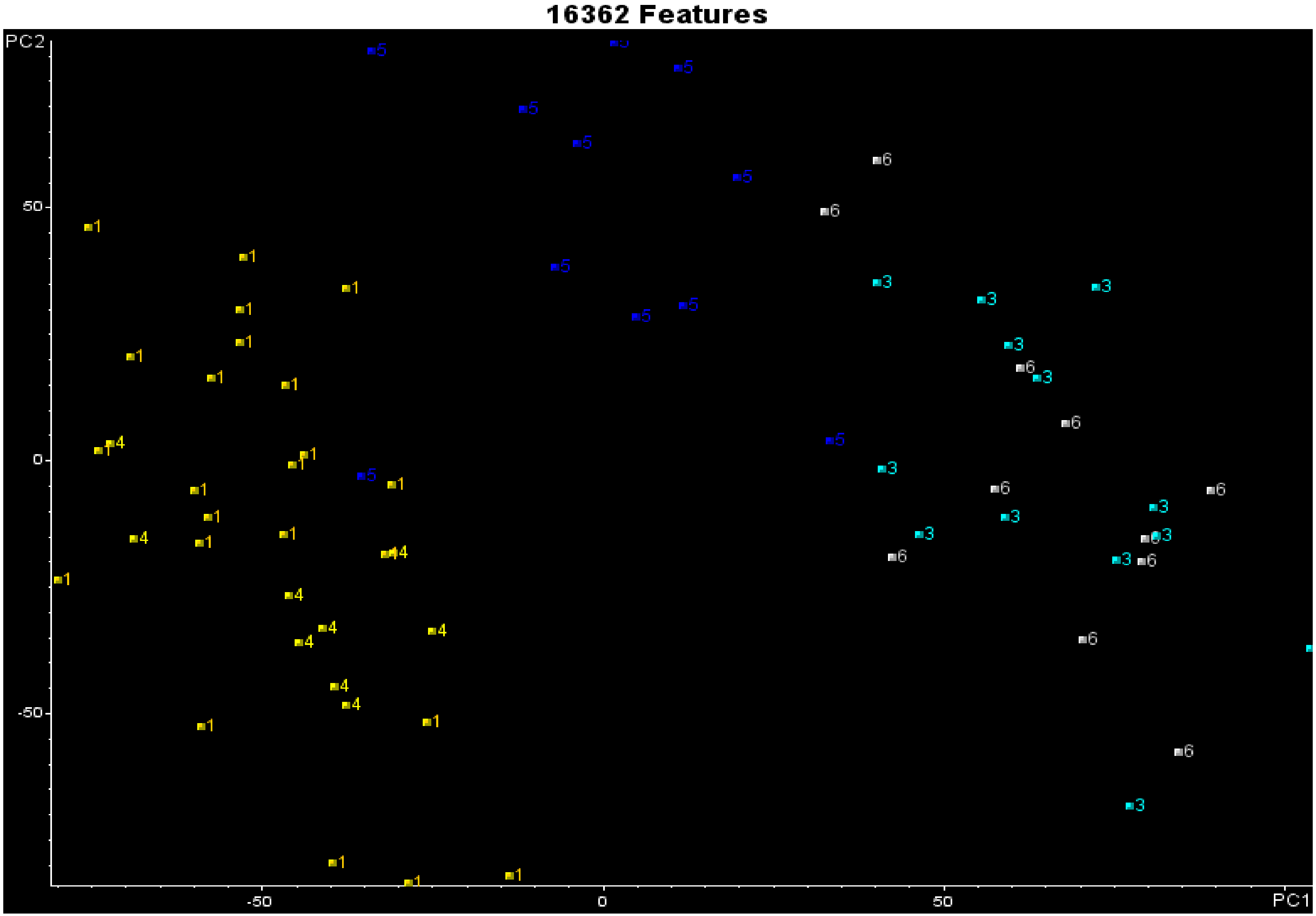
Take Home Points

- More powerful spectral preprocessing methods are needed to discriminate BRA, JFN, STH, SAL, and NEW paint spectra.
- For this reason, the wavelet packet transform will be applied to the IR data.
- Our goal is deconvolution, not data compression. Hence, the number of wavelet coefficients generated will be greater than the number of data points comprising the spectra in the wavelength domain.
- We will use the pattern recognition GA to identify informative wavelets.

Wavelets

- The Daubachies 12 mother wavelet was used to decompose each IR spectrum by the wavelet packet transform into 16362 wavelet coefficients.
- A plot of the two largest principal components of the 16362 wavelet coefficients proved uninformative.

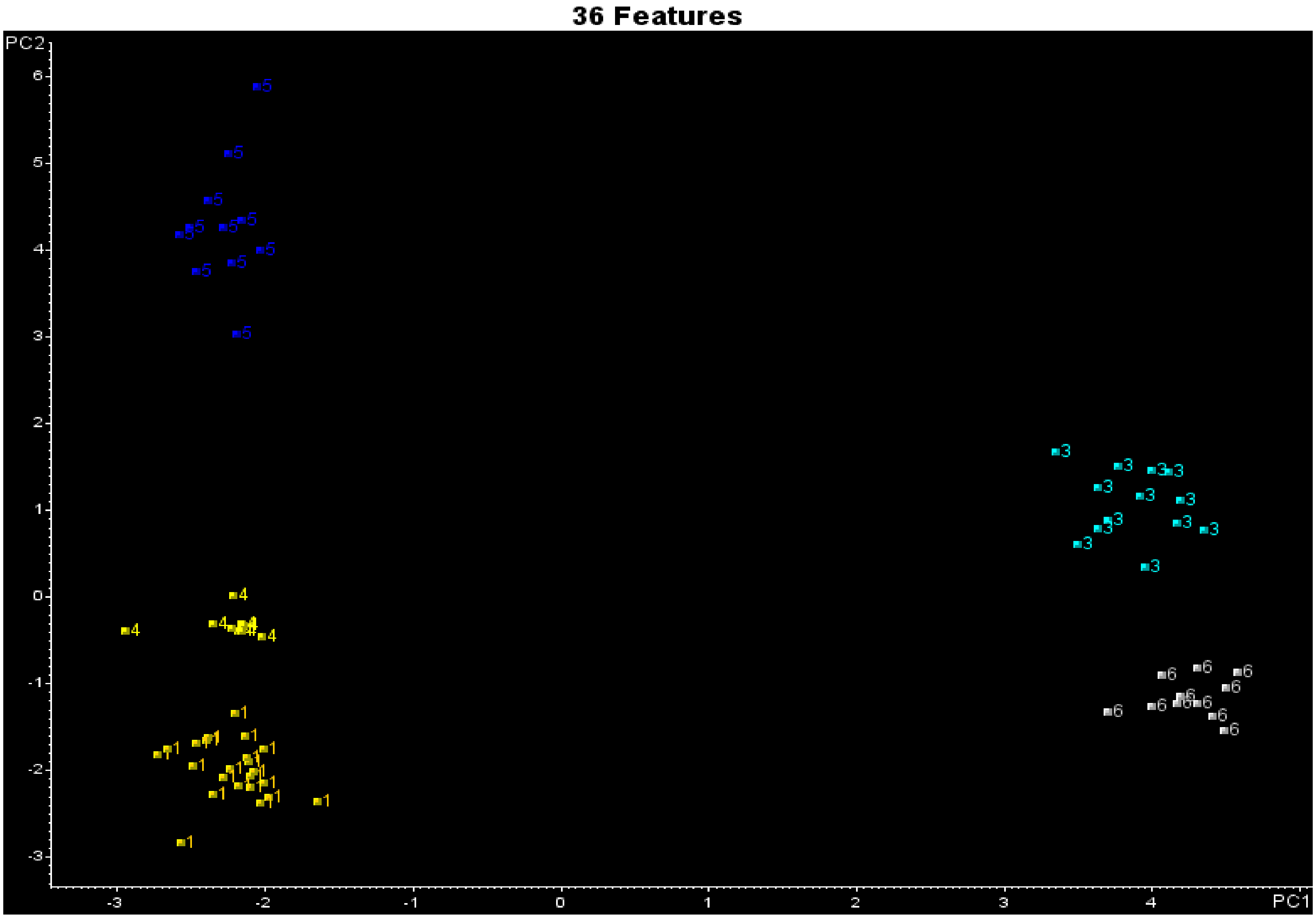
Score Plot developed from16362 wavelet coefficients used to represent the IR spectra in training set



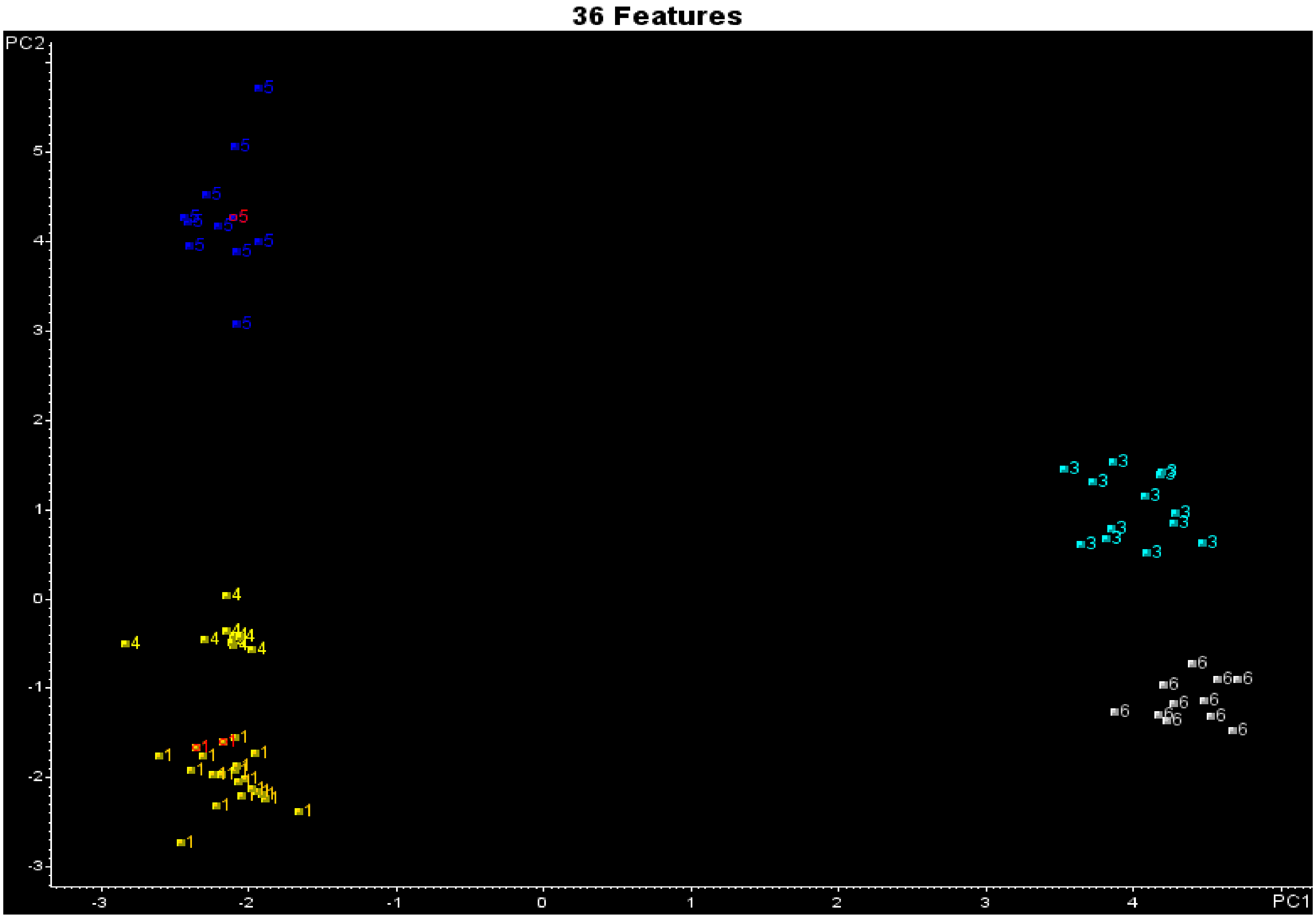
Searching the Coefficient Space

- The pattern recognition GA was then used to identify so-called informative coefficients.
- The GA identified these coefficients by sampling key coefficient subsets of the data, scoring their PC plots, and tracking those classes and samples that were difficult to classify.
- The boosting routine used this information to steer the population to an optimal solution.
- After 100 generations, the GA identified 36 wavelet coefficients which contained information characteristic of the manufacturing plant.
- Using transverse learning, 3 prediction set samples were correctly classified in the principal component plot developed from the 70 training set samples and 36 wavelet coefficients identified by the pattern recognition GA.

Score Plot developed from 36 wavelet coefficients identified by the Pattern Recognition GA for the training set



Prediction set samples (in red) projected onto the score plot developed from the 36 wavelet coefficients for the training set



Conclusion

- Pattern recognition methods can provide additional information about the class labels used to characterize IR spectra in the PDQ database.
- The wavelet packet transform when combined with the pattern recognition GA is able to identify fingerprint patterns in the IR spectra of paints characteristic of the manufacturing plant.
- Search prefilters developed from wavelet coefficients will simplify library searching.
- When combined with search algorithms that are more powerful but also more computationally intensive than the Euclidean distance, similarity searching will become feasible.

Acknowledgements

- Mark Sandercock and Tamara Hodgins, National Centre for Forensic Services – Alberta, Royal Canadian Mounted Police, Edmonton, Alberta, Canada

- Scott Ryland, Florida Department of Law Enforcement Regional Crime Laboratory, Orlando, FL

- Ayuba Fasasi, Jay Mehta, Collin White, Nuwan Perera, and Nikhil Mirjankar, Department of Chemistry, Oklahoma State University, Stillwater, OK.

- Financial support from the National Institute of Justice (2010-DN-BX-K217) is appreciated.