“By three methods we may learn wisdom: First, by reflection, which is noblest; second, by imitation, which is easiest; and third by experience, which is the bitterest.”

— Confucius

University of Alberta

# Alignment and Variable Selection Tools for Gas Chromatography – Mass Spectrometry Data

by

Nikolai Sinkov

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry

Spring 2013
Edmonton, Alberta

## Abstract

Each individual chromatogram is typically very sparse, especially when coupled with a technique such as mass spectrometry. Only a small portion of the recorded data actually contains potentially useful information. Since the inclusion of irrelevant information will have detrimental effects upon the resulting chemometric model, variable selection is an essential step in pre-processing of raw chromatographic data. Presented here is a model evaluation metric named cluster resolution which was developed to, when used in conjunction with variable ranking metrics such ANOVA or selectivity ratio, guide the variable selection process.

Retention time shifts are another problem that makes application of chemometric techniques to chromatographic data more challenging. While numerous alignment algorithms are available, alignment remains a challenge when dealing with highly dissimilar samples. To solve this problem, deuterated alkane ladder-based alignment was developed.

Finally, the tools developed in the course of this study were applied to pre-processing of casework fire debris data for the purpose of classification of fire debris based on ignitable liquid content. This project was performed in collaboration with Royal Canadian Mounted Police. The developed tools allowed effective chromatographic alignment and variable

selection, which permitted the construction of the first chemometric models able to reliably classify casework fire debris based on the presence or absence of gasoline.

**List of publications:**

J J. Harynuk, A.P. de la Mata, N.A. Sinkov, in *Chemometrics in Practical Applications.* K. Varmuza (ed.), InTech, Rijeka, 2012, 305-326

N.A. Sinkov, J.J. Harynuk, *Talanta* 83 (2011) 1079-1087.

N.A. Sinkov, B.M. Johnston, P.M.L. Sandercock, J.J. Harynuk, *Anal. Chim. Acta* 697 (2011) 8-15.

N.A. Sinkov, J.J. Harynuk, *Talanta* 103 (2013) 252-259.

## Acknowledgements

The work presented here could not have been accomplished without the support of the following individuals and funding agencies.

I would like to extend many thanks to my supervisor, Dr. James Harynuk for his leadership and guidance throughout my graduate studies. I would like to thank my parents for providing me with education, support and encouragement to pursue my degree; my brothers, sisters, nieces and nephews for supporting me. I would like to thank, in no particular order, my friends, Bob, Teague, Paulina, Katie, Brandon, Bryan, Reshma and many others for all their support.

I would like to thank Dr. Mark Sandercock for his collaboration, without which our work would not be possible.

I would also like to thank the personnel in the University of Alberta's Mass Spectrometry Laboratory and Technical Shops and Services for their assistance.

I would like to acknowledge the financial support provided by the University of Alberta, the Natural Sciences and Engineering Research Council of Canada (NSERC) and Alberta Ingenuity (now AITF).

Last but not least, I would like to thank Lei for her encouragement, understanding and care during my studies. I love you.

# Contents

# List of tables

## List of figures

## List of symbols

| | | |
|---|---|---|
| $\alpha$ | – | confidence limit |
| $D_{critical}$ | – | half of the distance between sum of two neighbouring points of each ellipse |
| $D_{min}$ | – | smallest distance between two points on different ellipses |
| $l$ | – | length of an ellipse axis |
| $L$ | – | eigenvalue for PC for a given axis |
| $\mathbf{L}$ | – | diagonal matrix of eigenvalues for principal components |
| m/z | – | mass-to-charge ratio |
| $n$ | – | number of samples or classes, depending on context |
| $p$ | – | number of components used in a model |
| Q | – | Q-residual |
| $\mathbf{S}$ | – | scores covariance matrix |
| $T^2$ | – | Hotelling $T^2$ value |
| $\mathbf{U}$ | – | Loading vectors matrix |
| $\mathbf{V}$ | – | Loading vectors matrix |
| $\mathbf{X}$ | – | matrix containing X-data for a dataset |

## List of abbreviations

| | | |
|---|---|---|
| AC | – | Affinity Chromatography |
| ANOVA | – | Analysis of Variance |
| BE | – | Backwards Elimination |
| BTEX | – | Benzene, Toluene, Ethylbenzene, Xylenes |
| COW | – | Correlation Optimized Warping |
| CR | – | Cluster Resolution |
| DTW | – | Dynamic Time Warping |
| EIC | – | Extracted Ion Chromatogram |
| FS | – | Forward Selection |
| GC | – | Gas Chromatography |
| GC×GC | – | comprehensive two-dimensional gas chromatography |
| GC-FID | – | Gas Chromatography – Flame Ionization Detector |
| GC-MS | – | Gas Chromatography- Mass Spectrometry |
| HDPE | – | High Density Polyethylene |
| HILIC | – | Hydrophilic Interaction Liquid Chromatography |
| IL | – | Ignitable Liquid |
| LC | – | Liquid Chromatography |
| LC×LC | – | comprehensive two-dimensional liquid chromatography |
| LV | – | Latent variable |
| MCCR | – | Mean Correct Classification Rate |
| MDF | – | medium density fibreboard |

| | | |
|---|---|---|
| MLR | – | Multiple Linear Regression |
| MS | – | Mass Spectrometry |
| NPLC | – | Normal Phase Liquid Chromatography |
| OPLS-DA | – | Orthogonal Partial Least Squares Discriminant Analysis |
| PC | – | Principal Component |
| PCA | – | Principal Component Analysis |
| PCR | – | Principal Component Regression |
| PLS | – | Partial Least Squares |
| PLS-DA | – | Partial Least Squares Discriminant Analysis |
| RMSECV | – | Root-Mean-Square Error of Cross Validation |
| SEC | – | Size Exclusion Chromatography |
| SIM | – | Selected Ion Monitoring |
| SR | – | Selectivity Ratio |
| TIC | – | Total Ion Current |
| TP | – | Target Projection |
| ROC | – | Receiver-Operator Characteristic |
| RPLC | – | Reversed Phase Liquid Chromatography |

# 1 Chapter One: Introduction[1]

Molecular separations are often used to analyze complex mixtures associated with highly complex challenges in fields such as metabolomics and forensics. Recently, there has been increased interest in applying chemometric techniques to separations data. Chemometrics provides powerful tools to assist in the interpretation and visualization of complex data so that these complex questions can be addressed. However, prior to the application of chemometric techniques, separations data must be pre-processed so that the resulting model is focused on relevant variations within the data. The goal of my work has been to develop tools to assist in the preparation of separations data for chemometric analysis. The primary contributions are a metric termed cluster resolution (CR) which has been developed to assist in the automation of variable selection routines and the implementation of an alignment algorithm that relies on a retention ladder of deuterated alkanes. These allow for the alignment of chromatographic data with highly variable compositions across samples. The techniques developed have been successfully applied to determine the presence of gasoline in casework arson debris samples.

[1] A version of this chapter has been published as J. J. Harynuk, A. P. de la Mata, N. A. Sinkov, Application of Chemometrics to the Interpretation of Analytical Separations Data, In *Chemometrics in Practical Applications.* K. Varmuza, InTech, Rijeka, 2012, 305-326.

## 1.1 Separation methods and chemometrics.

Analytical techniques are applied to a sample with the goal of answering some question about the sample based on the presence and/or abundance of one or more analytes. However, real-world samples are usually present as complex mixtures which, in addition to the analyte(s) of interest, contain matrix components which are largely irrelevant to answering the fundamental question at hand. These matrix components are often present in significantly higher amounts than analytes of interest, both in terms of the number of components and their concentrations.

While in some cases selective identification and quantification of analytes may be possible, often separations techniques such as liquid chromatography (LC), gas chromatography (GC), or electrophoresis must be applied. With individual components in the mixture separated, they can be detected and quantified individually to arrive at an answer to the question about the sample. Recently, more complex analytical questions are being asked, such as identifying the origin of a food product [1,2,3], identifying the presence or absence of ignitable liquids in a sample of debris from a fire [4,5,6], or determining the disease state of a patient based on analysis of a blood sample [7,8,9]. Answering these types of questions usually involves far more than simply quantifying a couple of analytes in the sample and requires the identification of complex patterns of variables in the data. For example, relative abundances of select alkanes and alkyl benzenes can be used for the purpose of identifying gasoline in an arson debris sample [10].

Chemometrics is a discipline that uses mathematical and statistical tools to extract useful chemical information from raw chemical data [11]. The extracted information can be used to answer questions about samples (e.g. classification or quantification) or optimize measurement procedures and experiments [12,13]. A wide variety of chemometric methods have been developed for the purpose of extracting chemical information from data. Some examples are the exploratory approaches such as principal component analysis (PCA) [14] and cluster analysis [15], the calibration approaches such as multiple linear regression (MLR) [16], principal component regression (PCR) [17,18], partial least squares regression (PLS) [17,19, 20], and the classification approaches such as soft independent modelling by class analogy (SIMCA) [21], partial least squares discriminant analysis (PLS-DA) [22], and orthogonal least squares discriminant analysis (OPLS-DA) [23]. The approaches listed above are just a few examples of the variety of tools today's analysts have at their disposal.

### 1.1.1 Chromatography

Chromatography is an approach widely used for the separation, purification, and analysis of mixtures. Generally, analytes contained in a mobile phase, which is usually a gas or a liquid, are flowed past a stationary phase, which is usually a viscous liquid or a solid supported inside a chromatographic column. Depending on the possible molecular interactions between analytes and the stationary and mobile phases, as well as other conditions, such as temperature, pH, etc., different compounds will partition between the two phases to varying degrees. As a result of this differential partitioning, separation will be observed, with analytes associating more strongly with the mobile phase passing through the column more quickly than those with a greater affinity for the stationary phase [24].

Among the numerous types of chromatography developed, the most common ones are liquid chromatography (LC), where the mobile phase is a liquid, and gas chromatography (GC), where the mobile phase is a gas. A block diagram of a chromatographic system is provided in Figure 1-1.

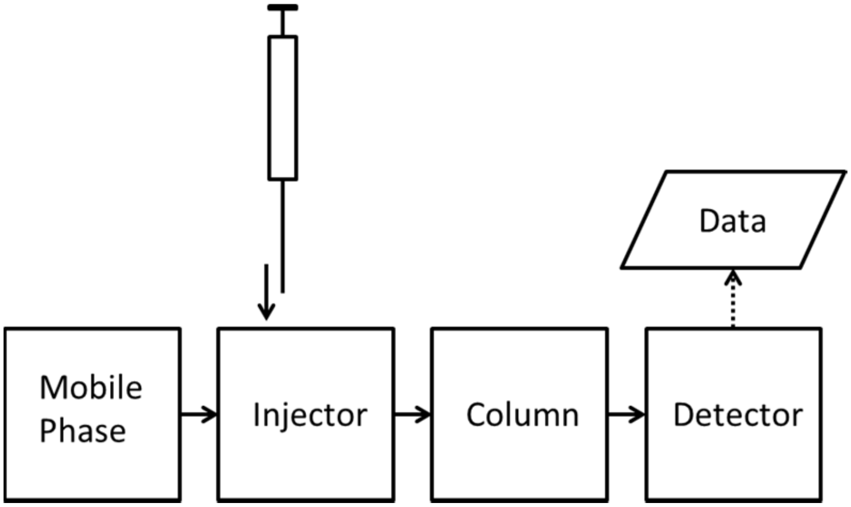


**Figure 1-1:** Block diagram of a chromatographic system

In GC, which was used for the bulk of the research in this work, stationary phases are either solid, or more often highly cross-linked polymers with liquid-like properties while the mobile phase is a gas, usually helium or hydrogen. There are many modes of LC, ranging from reverse-phase liquid chromatography (RPLC), normal-phase liquid chromatography (NPLC), ion chromatography (IC), size exclusion chromatography (SEC), hydrophilic interaction liquid chromatography (HILIC) and affinity chromatography (AC). Different types of chromatography have their advantages and disadvantages and are suitable for different kinds of samples. For example, GC requires volatile or semi-volatile analytes. Thus large biomolecules which would thermally degrade before volatilizing are not amenable to GC analysis. In general, the choice of a chromatographic technique depends on the type of sample being analyzed, required analysis time, resources available, etc. [24,25].

When dealing with highly complex samples, so-called comprehensive multidimensional separations such as GC×GC or LC×LC [26,27,28,29,30] are emerging as useful options in the analyst's toolbox. With these techniques, a mixture of analytes is sequentially separated by two different mechanisms. For example, in the case of GC×GC a sample might be separated first on a non-polar column, followed by a separation on a polar column. These techniques are capable of separating exceedingly complex mixtures comprising thousands of individual compounds. Due to the vastly improved separation power of these techniques, the resulting data are much more information-rich than data from one-dimensional separations. As a result of the power of these techniques, they are well-suited to addressing complex questions. However, to fully exploit their immense potential, the development of chemometric tools for processing their data is essential.

### 1.1.2 Chromatographic data.

Before discussing the application of chemometric techniques to chromatographic separations, the form of the data obtained from such chromatographic systems must be understood. The detector signal obtained from a separation represents the detector response as a function of time with a single analyte ideally providing a single Gaussian peak.

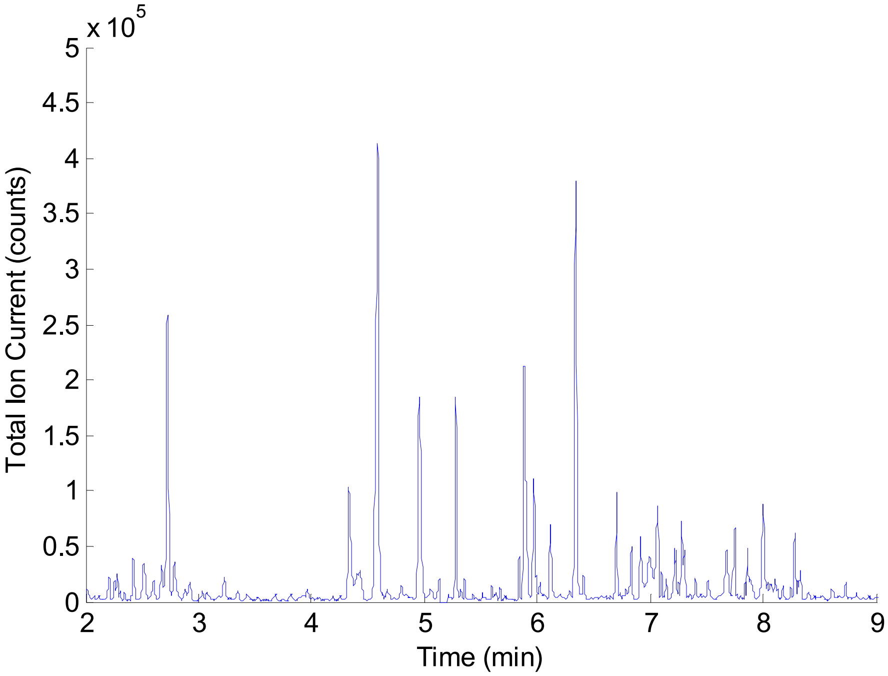


**Figure 1-2:** Segment of a GC-MS Total Ion Current chromatogram of gasoline

Figure 1-2 depicts a segment of a gas chromatography-mass spectrometry (GC-MS) chromatogram of a gasoline sample. The plot is of total ion current (the sum of the abundances of all ions in each scan) vs. time. This chromatogram was collected with a data rate of approximately 10 Hz. Thus, this 7-minute segment of data consists of about 4200 individual data points. To obtain an accurate peak profile, the instrument should acquire 10 points across each peak [31]. The data rate of 10 Hz is acceptable for this chromatogram where peaks are about 2 s wide at the base. In cases where there are narrower peaks, such as would be the case with a GC×GC separation, higher data rates, on the order of 50-200 Hz, would be required for quantitative analysis.

Another important consideration is the nature of the detector. The chromatogram in Figure 1-2 is an example of a univariate detector response. Even though the MS collected data for many ions, the TIC is the sum of these ions and the multivariate nature of the MS data is lost in this representation. Other examples of univariate detectors include a flame ionization detector (FID) in GC or a refractive index detector in LC. The difference between a univariate and a multivariate response is illustrated in Figure 1-3, which presents the same data as in Figure 1-2, with the full mass spectral dimension of information preserved.

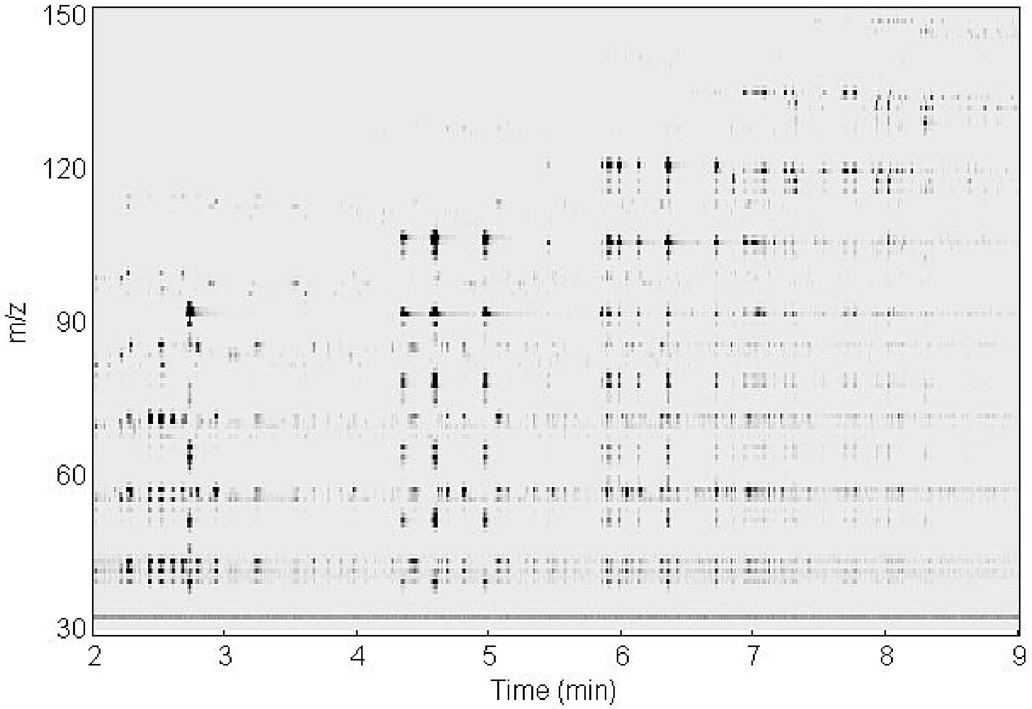


**Figure 1-3:** Section of a GC-MS chromatogram of gasoline showing MS information. Darker colour indicates stronger intensity.

In Figure 1-3, the intensity of an ion at certain mass/charge ratio (m/z) at a given time point (scan) is represented as an individual variable. The number of scans in Figure 1-3 is the same as in Figure 1-2 (4200 scans) but now individual ion intensities for ions with m/z ratios ranging from 30 to 150 are considered, a total of 121 ions. As a result, this section of the chromatogram consists of 508 200 (4200×121) variables, a much larger number than seen with a univariate detector. As a comparison, a 10-minute long GC-FID chromatogram with the detector operating at 10 Hz will provide 6000 variables while an hour-long GC×GC-MS analysis collecting ions over a range of 30-300 m/z at 200 Hz will provide 195 120 000 individual data points.

In the interpretation of separations data, there are two fundamental approaches that can be taken. The most common is to integrate the peaks to obtain a table of peak areas for each sample. Chemometric techniques may then be applied to the data which have been reduced to a simple peak table. The other approach, which is becoming more popular, is to apply chemometric techniques to the raw chromatographic data directly in an attempt to extract useful features from it. These approaches will be discussed further in this chapter.

## 1.2 Challenges posed by chromatographic data.

Chromatographic data poses unique challenges to the application of chemometrics. To obtain the best results when applying chemometric methods to chromatographic data, it is essential that the analyst understands both the chromatographic and chemometric techniques used. Analytical separations data, like data derived from any other analytical instrument, are based on chemical and non-chemical aspects of the analysis. While non-chemical variation is irrelevant, chemical variation can be both relevant and irrelevant to a given analytical question. Consider the case of determining whether an individual is healthy or sick based on an LC analysis of a blood sample. Chromatograms obtained from sick and healthy groups of individuals will contain signals from numerous metabolites. The responses of some metabolites will increase or decrease depending on the disease state, while other metabolites' variability will not be related to the disease state. Furthermore, many of the metabolites' responses will also be influenced by day-to-day activities (e.g. diet). The challenge is to identify relevant chemical variation in the sample (that which is indicative of a disease state) while ignoring irrelevant chemical variations. In addition to sample variability, there are often other components of the signal to deal with, such as stationary phase “bleed”, which contribute additional irrelevant chemical variation to the data.

Non-chemical variation includes retention time shifts of peaks due to fluctuations in operating conditions, baseline drift for non-chemical reasons, electronic noise, etc. When significant irrelevant variation is present, it interferes with the analysis, degrading the quality of the resultant model [32]. Therefore, prior to the application of chemometric tools to separations data, analysts must pre-process the data to ensure that the tools are focused on relevant chemical variation while ignoring irrelevant variations of both chemical and non-chemical natures. Utilizing integrated peak tables is one of the ways to correct for non-chemical variation to some extent. However severe non-chemical variation may make accurate integration and assignment of chromatographic peaks impossible.

### 1.2.1 Baseline and noise

Most chromatographic data will contain some degree of background noise. In the case of LC separations, the composition of the mobile phase will change during a gradient separation, possibly introducing baseline drift. Gradual build-up and slow, prolonged elution of strongly-retained matrix components can also cause baseline drift. Figure 1-4 presents an example of baseline drift in LC.

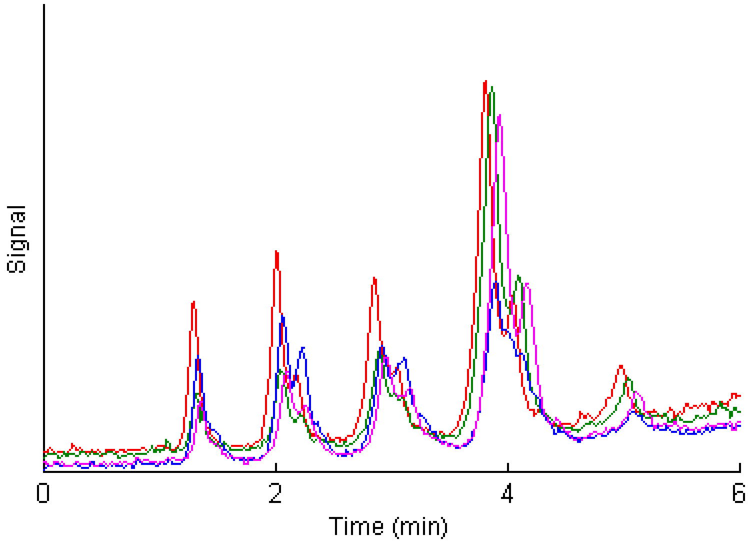


**Figure 1-4:** A gradient LC chromatogram of edible oils showing baseline drift.

Similarly, in temperature-programmed GC, increased levels of stationary phase bleed will be observed at higher temperatures. Noise and drift could also result from factors such as contamination of mobile phase or change in detector response with time.

### 1.2.2 Retention time shifts

Retention time shifts are a common problem in chromatographic separations. When chemometric analyses are performed on integrated peak tables, minor retention time shifts will not pose many problems as long as the peaks are identified correctly. However, when retention time shifts are severe, problems with peak identification may arise since identification is often based on retention time. When dealing with complex samples where the identity of some peaks may be unknown, retention time shifts can make comparisons between chromatograms challenging or even impossible. If analytes are misidentified, the resulting model will likely produce nonsensical output.

When raw chromatographic signals are used, even minor shifts in retention can pose severe problems for chemometric analysis even with relatively simple samples. Figure 1.5 presents a simulation of misaligned peaks.

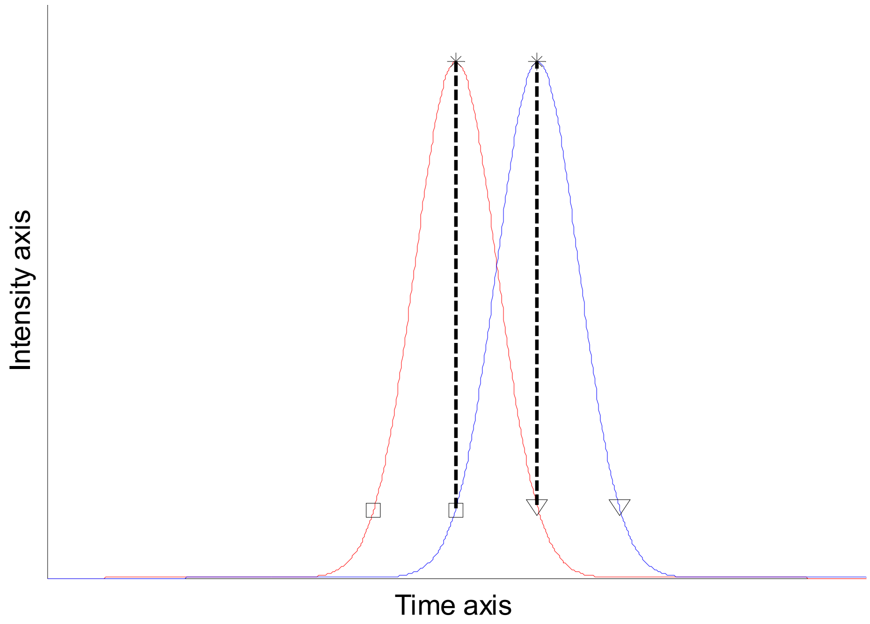


**Figure 1-5:** Misaligned simulated Gaussian peaks. Squares, asterisks and triangles show same data points on the two peaks. Dashed lines indicate difference in signal recorded for two peaks at the same time point. Widths and amplitudes of the peaks are identical.

In the figure, the two peaks represent the same compound and they should elute at the same time. However, they do not. As a consequence, the signal in the blue peak is not recorded in the expected locations in the signal vector. The data points indicated by the squares, triangles, and asterisks would occur at the same three times if the peaks were aligned. However, there is a significant difference in the intensities recorded at the same points in time in the two analyses. When attempting to use chemometric techniques on such misaligned raw data, variation in the data introduced by the misalignment will severely degrade the model.

There are many causes of retention time shifts. In GC, peaks may shift due to changes in the stationary phase through degradation, decreasing retention times over time; build-up of heavy matrix components which foul the column, effectively changing the chemistry of the stationary phase; minor gas leaks which may alter the flow rate of carrier gas; or hot or cold spots created on the column if portions of it come into contact with objects such as walls of the oven. In LC, peak shifts may be due to small fluctuations in mobile phase composition from one run to the next; temperature fluctuations which can in turn affect solvent viscosity as well as partitioning of compounds between mobile and stationary phases; or changes in the stationary phase of the column due to degradation and fouling.

Shifts in retention times are minimized by proper instrument maintenance, precise control of instrumental conditions as well as by using approaches such as retention time locking in GC to account for variations in instrument performance [33,34]. However, even with these efforts, some variability in retention time will remain. Additionally, researchers may wish to combine data from samples that were analyzed on different instruments by different people in different laboratories over a period of months or years.

### 1.2.3 Data overload

The over-abundance of data coming from the detector poses its own challenges. Chemometric analyses are generally performed on a data matrix where the number of columns is equal to number of variables and the number of rows is equal to the number of samples. Considering a data set comprising 200 samples analyzed by the GC×GC-MS example from Section 1.1.2 where a single chromatogram contained ~200 million individual variables, the resultant 200-chromatogram dataset will contain almost 40 billion individual variables. Such a massive dataset can exceed the capabilities of a computer system and, even if the system can handle the sheer number of variables, it will slow down the data analysis. Most importantly, only a very small fraction of the data matrix will contain meaningful data: most of chromatogram will contain only baseline and noise (Figure 1-3), especially in the case of multidimensional chromatography and chromatography involving multivariate detectors [35,36,37,38]. Inclusion of this massive quantity of irrelevant data will have a significant deleterious effect upon the resulting model, necessitating some kind of variable selection [13].

## 1.3 Preparation of chromatographic data for chemometric analysis

Due to the issues described in Section 1.2, it is necessary to process chromatographic data prior to chemometric analysis. As discussed before, one way of achieving this is to use an integrated peak table. Another way is to process raw data with the goal of addressing the problems of alignment and the size of the data set. Advantages and disadvantages of using raw data instead of integrated peak tables, as well as some ways of processing the raw data, will be discussed in this section.

### 1.3.1 Baseline correction

The goal of baseline correction is to separate the analyte signal of interest from signal which arises due to factors such as changes in mobile phase composition in LC or stationary phase bleed in GC, as well as from signal due to electronic noise. A variety of baseline correction methods are available in the literature, common approaches being to fit a curve though the baseline and subtract this curve from the chromatogram, yielding a baseline-corrected signal. Alternatively, the baseline can be modeled using factor models and then excluded from the analysis [39].

The curve fitting approach is used in most commercial software packages provided by vendors of separations equipment. The algorithms used in this approach will fit an equation (usually a first-order polynomial) across segments of the chromatogram using regions where no analyte signal is present to determine the coefficients of the polynomial. The background signal is then interpolated in regions where analytes are eluting. With the equation of the background signal determined, the fitted curve can then be subtracted from the signal [40,41,42,43,44]. An example of baseline correction using curve fitting is demonstrated in Figure 1-6.

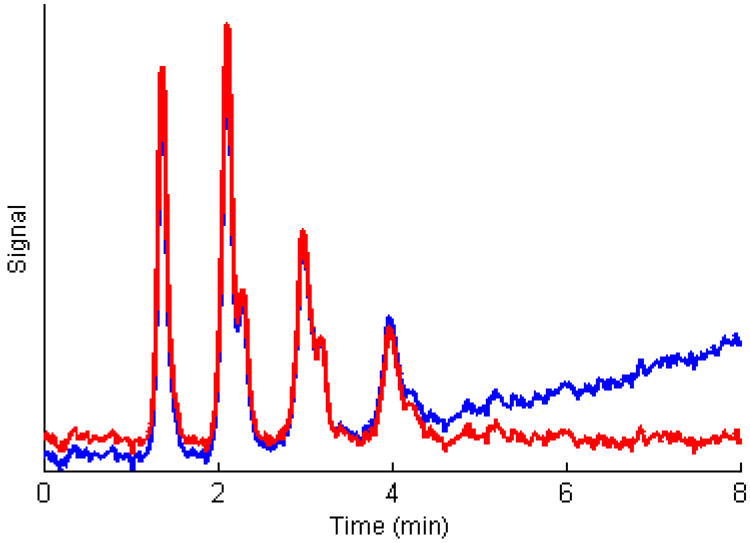


**Figure 1-6:** Baseline correction of one of the LC chromatograms from Figure 1-4. Blue line represents uncorrected chromatogram while red line represents a chromatogram corrected using a second-order polynomial.

### 1.3.2 Retention shift alignment

The retention times of analytes will fluctuate from one analytical run to the next. Before chemometric techniques can be applied to the data, these fluctuations must be corrected. This correction will ensure that the signal from each analyte in each chromatogram is correctly registered within the data matrix to be processed. There are essentially two approaches available to the analyst: the use of integrated peak tables, or the mathematical warping of the raw signal.

Use of integrated peak tables is a straightforward way to ensure that analytical separations data are properly aligned for chemometric processing. To utilize this approach, the analyst must be able to reliably assign a unique identifier to each peak in each sample of the data set. This will ensure that the same compound is identified with the same identifier in each sample, even if the exact identity of each compound is not determined. A series of labels such as “Unknown x”, where x is a numerical identifier would be acceptable, as long as the compounds were matched correctly. Rather than identifying peaks by retention time, an analyst could use relative retention times or retention indices to adjust for slight variations in the elution times. Algorithms for aligning peak tables generally perform well, as long as at least some peaks can be easily and reliably matched across all chromatograms [45]. One major downside of this approach stems from its reliance on integrated peak tables. Integration algorithms are not perfect and can provide integration errors due to poorly-resolved or tailing peaks or peaks that are missed due to falling outside of integration thresholds in the software [46,47]. These errors will impact any subsequent analysis.

Alignment of raw chromatographic signals prior to chemometric processing is more complex than the alignment of peak tables. When deciding which approach to use for raw signal alignment, one of the first questions to be answered is whether the analysis is to be qualitative or quantitative since some alignment methods can distort peaks, affecting their quantification. Some of the more common alignment algorithms include correlation optimized warping (COW) [48,49], correlation optimized shifting (coshift) [50], and a piecewise peak-matching algorithm [51], among others [52,53,54,55,56].

In instances where peak shifts are non-systematic, COW is a popular algorithm. COW relies on warping (stretching or compressing) segments of a sample signal such that the correlation coefficient between the sample and a reference signal is maximized for each interval. Care must be taken with the selection of the input parameters to avoid significant changes in peak shapes since this approach to the warping of the chromatogram may affect peak areas, potentially leading to poor quantitative conclusions [48,49].

Coshift is a fast and simple alignment algorithm [50]. This algorithm is useful when data only require a single left-right shift in retention time (decrease in retention times or increase in retention times). The entire data matrix is shifted in one or the other direction by a set amount, maximizing the correlation between the data matrix being aligned and a target. The single shifting value for the entire data matrix is a weakness, especially for chromatographic data where peaks in a single chromatogram can shift in different directions and to different extents. To address this problem, an algorithm termed icoshift (interval-correlation-shifting) has been derived from coshift [57]. Icoshift aligns each data matrix to a target by maximizing the cross-correlation between the sample and the target within a series of intervals defined by the user. Since multiple intervals are used, separation data where shifts vary in magnitude and direction can be aligned. These alignment algorithms have been successfully applied to one-dimensional data [1,58,59] and with some modifications to two-dimensional data as well [60]. It is important to note that coshift and icoshift algorithms do not distort peak shapes and, as a result, do not introduce errors into quantitative results.

The piecewise peak matching algorithm [51] provides another avenue for chromatographic alignment. In this approach, peaks are first identified in a target chromatogram. Then, for each sample chromatogram, the algorithm identifies peaks located within predetermined time windows of the peaks identified in the target chromatogram. Peaks within windows are then considered to come from the same compound, and are matched. As a result, the chromatograms are aligned by warping the regions between the peak apexes. A variant of this algorithm has been developed to utilize MS data, if available [61]. In this case, the mass spectrum at the apex of a given peak in the target chromatogram is compared to the mass spectrum of each peak within a set window on the sample chromatogram. If the peak apex spectra between the target and sample peaks have a high enough match quality, the peaks are aligned. Depending on the number and relative positions of the peaks in chromatograms matched using this approach, peak shapes may be altered due to chromatogram warping, possibly affecting quantitative results. A general scheme for peak alignment using this approach is described in Figure 1-7.

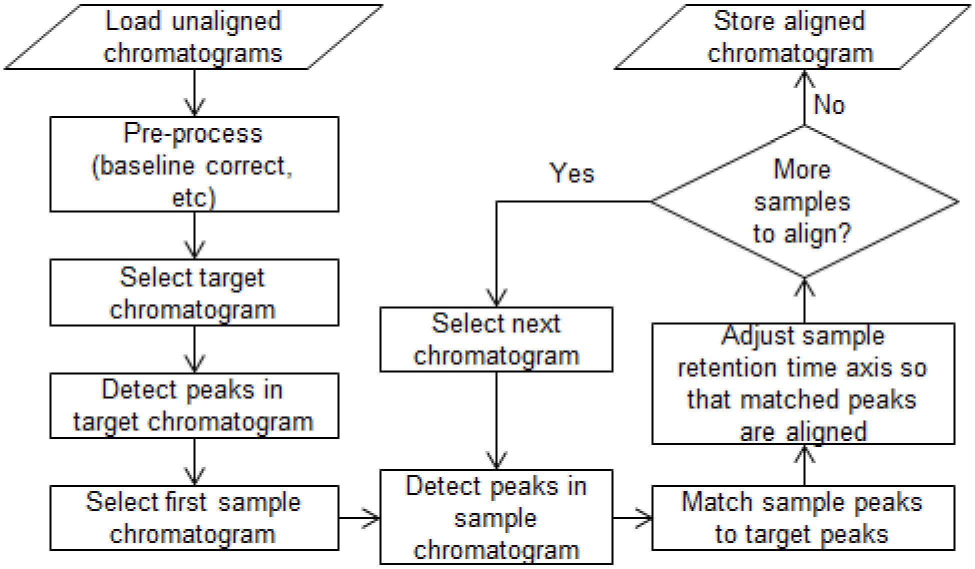

Load unaligned chromatograms
Pre-process (baseline correct, etc)
Select target chromatogram
Detect peaks in target chromatogram
Select first sample chromatogram
Detect peaks in sample chromatogram
Match sample peaks to target peaks
Adjust sample retention time axis so that matched peaks are aligned
More samples to align?
Yes
Select next chromatogram
No
Store aligned chromatogram

**Figure 1-7:** Flowchart for the piecewise peak matching algorithm, adapted from Johnson et al., 2003 [51]

One of the challenges for all alignment algorithms is that they perform best when the chemical composition of the samples being aligned is reasonably consistent. In cases where the chemical composition of the sample and matrix are highly variable, conventional alignment algorithms can fail. This situation is observed with arson debris where the matrix is extremely variable from one sample to the next. This challenge was addressed in this work, and will be described in Chapter 3.

### 1.3.3 Variable selection

High data acquisition rates combined with potentially long analysis times can result in a large number of data points collected for a given separation, as discussed in Section 1.2.3. In most separations, the majority of the data are collected by the detector at times when no analytes are eluting from the chromatographic column and contain only background and detector noise. The problem becomes more severe when multivariate detectors, such as spectroscopic and especially mass, are used. At a given point in time, many of the recorded data in the spectral dimension will not contain useful information, even at times when an analyte of interest is eluting. In addition to background and noise, many components in the mixture can be completely irrelevant to analysis [36,37,62]. Consequently, only a very small portion of a chromatographic signal has any potential utility. It is also well known that any chemometric model will be heavily influenced by the specific variables that are included in its construction [63].

The inclusion of irrelevant data will be detrimental to the model because the mathematics will attempt to account for variation observed in the data, both relevant and irrelevant. Consequently if the model is forced to model noise, that will result in a decrease in its predictive ability. Worse, given enough irrelevant variables, the model might actually fit the training data reasonably well and provide a seemingly useful prediction, but upon validation the model would be discovered to be fitting noise and generating poor predictions. Finally, the inclusion of extraneous variables would also increase the demands on the computer system being employed, making model construction slower, or, in the case of very large numbers of variables, outright impossible. Therefore, prior to chemometric analysis, the chromatographic data must be reduced to a manageable size.

One way to achieve data reduction is to use a table of integrated peaks instead of the raw chromatographic data. This has the advantage of reducing the number of variables to those compounds included in the peak list, which is usually a relatively small number, removing baseline noise and, if the analyst knows which exact peaks are relevant to analysis, removing signal from irrelevant compounds. Potential problems with this approach have been described in Section 1.3.2. Additionally, some peaks identified may be irrelevant to analysis, necessitating further variable selection. This can be true in the case of multidimensional separations where a peak table can easily contain hundreds, if not thousands of compounds [64].

In the case of multivariate detection, it may be advantageous to monitor only one or a few channels such as wavelengths, mass/charge ratios, etc. This will allow the analyst to selectively detect only a portion of the analytes, avoiding many interfering species, as well as greatly reducing the size of the data. However, in these cases the analyst must know exactly what signals to use or risk missing important features of the data encoded in the channels that were ignored. Further, using this approach destroys much of the multivariate advantage that can otherwise be realized through using these more complex detection strategies.

Objective feature selection techniques generally have two steps: variable ranking, and variable selection. Objective variable ranking techniques such as selectivity ratio (SR) plots [65,66], analysis of variance (ANOVA) [62], and informative vectors [67] have the distinct advantage that that variables are ranked based on a mathematically calculable "perceived utility" rather than a subjective analyst's perception. In essence, the data are given the chance to inform the user of what is likely relevant and what is likely noise, providing an approach that can be generalized to any set of analytical data.

ANOVA is an effective method when the goal is to discriminate between two or more classes of samples. ANOVA calculates the F ratio for each variable: the ratio of between-class variance to within-class variance. If the F ratio for a given variable is high, it is considered more valuable for describing the difference between classes. With F ratio calculated for every data point in the chromatogram, the variables can be ranked in order of decreasing F ratio. A chemometric model can then be constructed using a fraction of variables having the highest F ratio. One significant advantage of ANOVA is that the calculations are relatively simple and the algorithm can be written with memory conservation in mind, which allows it to be easily and directly applied to data sets with very large numbers of samples and variables (hundreds or thousands of samples, each containing millions of variables), something that is more challenging when using other feature ranking approaches.

Selectivity Ratio is another feature ranking technique that can aid feature selection prior to chemometric analysis [65,66]. This approach involves the creation of a PLS-DA model that includes all candidate variables. Regression coefficients from the PLS-DA model are then used to calculate scores and loadings for the single target-projected (TP) component, which provides a TP model [65]. From this, the ratio of explained variance to residual variance for each variable in the TP model will provide the SR for each candidate variable, upon which variables can be ranked [65,66,68,69]. SR produces a ranking that is slightly different than that produced by ANOVA, though a direct comparison of the two metrics on chromatographic data has not yet been published.

Once variables have been ranked, those to be included in the model must be selected. One approach is to use a certain number of top-ranked variables that would provide best class discrimination [36,62]. However, no variable ranking metric is perfect and it is possible that a variable, even if it is ranked relatively high, may end up harming the resulting model while a relatively lower ranked variable would provide a net positive effect. To address this problem, a model can be constructed using a forward-selection or backwards elimination approach in an attempt to maximize some metric of model quality [70,71,72].

Model quality can be assessed based on several metrics such as mean correct classification rates [66], receiver operating characteristic (ROC) curves [73] or the degree of separation between classes of samples in principal component (or latent variable) space, for example using either a Euclidian distance-based metric [35]. A new metric of model quality that accounts for size and shape of clusters [36,37] was developed as the heart of the research presented in this thesis. This metric, termed cluster resolution (CR) will be described in great detail in Chapter 2, and its application in model optimization will be demonstrated in Chapters 2, 3 and 4.

### 1.3.4 Validation

Validation is one of the most important steps in model construction, especially when processing raw separations data and if a feature ranking approach such as ANOVA was used. As discussed previously, a data set consisting of raw separations data may contain on the order of $10^5$ to $10^6$ data points for each sample. In the cases of such overdetermined systems it is entirely possible to select combinations of variables containing only noise that will, by random chance, indicate a difference between the samples. When handling raw separations data, a good approach to avoid this problem is to break the data into separate sets. For example, a training set to construct the model, an optimization set to optimize data processing parameters (such as alignment and feature selection), and finally a test set to determine if the optimized model has any meaning. Of course this does require that the data are collected for a large number of samples so that a representative population of samples can be provided for each of the subsets of data.

## 1.4 Scope of thesis

The goal of my work has been to develop tools for processing chromatographic data prior to chemometric analysis. The cluster resolution metric has been developed to guide the variable selection process. In response to the variability observed in GC-MS chromatograms of arson debris, a deuterated alkane ladder-based alignment approach was also developed. The techniques were then applied to the automated optimization of models for the identification of traces of gasoline in casework arson debris data in collaboration with the Royal Canadian Mounted Police (RCMP).

### 1.4.1 Cluster resolution metric

When modeling a system with a goal of discriminating between samples, clusters of points representing different classes of samples must be separated in the model space. When optimizing the population of variables to include in the model, a metric of model quality is required. In principle, the separation between clusters of points can be used as a model quality metric. A previously proposed metric for evaluating separation between clusters of points, called the degree-of-class-separation [62], measures the between-class variance relative to the within-class variance of two clusters of points on a scores plot. However, this metric represents clusters of points on a scores plot as circles (or spheres). Since clusters of points are often better represented by confidence ellipses (or ellipsoids) than circles (or spheres), a metric that addresses the shortcoming of considering clusters of points being perfectly symmetrical shapes was developed. The metric was termed cluster resolution (CR). Initially, cluster resolution was developed to consider projections of models into a two-LV space, but later was expanded to simultaneously consider data in a three-LV space. This is presented in Chapter 2.

### 1.4.2 Deuterated alkane ladder-based alignment of GC data

Alignment of highly dissimilar chromatograms in a dataset can be challenging since many chromatograms will not share sufficient common features to ensure reliable alignment. This can be especially true in forensic analyses such as the analysis of arson debris since the chemical nature of debris matrices is extremely diverse and unpredictable. To address this issue, an alignment method for GC-MS data relying on spiking the sample with a deuterated n-alkane ladder has been developed, and is presented in Chapter 3. The deuterated alkanes show unique fragmentation patterns in the MS dimension. As a result, elution times of the ladder components can be determined in each chromatogram even when some of the deuterated components are co-eluting with abundant matrix components. The deuterated alkanes then serve as anchors for the alignment algorithm.

### 1.4.3 Detection of gasoline in casework arson debris samples

Finally, CR and ladder-based alignment techniques have been utilized for alignment and variable selection in GC-MS chromatograms of real arson debris data. The data were obtained from actual casework samples at the Royal Canadian Mounted Police National Centre for Forensic Services in Edmonton, Alberta. They were collected by different people on different instruments over the course of several months. As a result, severe chromatographic alignment issues are present in the data. In Chapter 4, the deuterated alkane ladder alignment and CR-guided feature selection were applied successfully to develop a model that was capable of determining the presence or absence of gasoline in the debris samples.

## 1.5 References

[1] P. de la Mata-Espinosa, J.M. Bosque-Sendra, R. Bro, L. Cuadros-Rodriguez, *Talanta* 85 (2011) 177-182.

[2] P. de la Mata, A. Dominguez-Vidal, J.M. Bosque-Sendra, A. Ruiz-Medina, L. Cuadros-Rodríguez, M.J. Ayora-Cañada, *Food Control* 23 (2012) 449-455.

[3] M. Ferrand, B. Huquet, S. Barbey, S. Barillet, F. Faucon, H. Larroque, O. Leray, J.M. Trommenschlager, M. Brochard, *Chemom. Intell. Lab. Syst.* 106 (2011) 183-189.

[4] P.M.L. Sandercock, E. Du Pasquier, *Forensic Sci. Int.* 134 (2003) 1-10.

[5] P.M.L. Sandercock, E. Du Pasquier, *Forensic Sci. Int.* 140 (2004) 43-59.

[6] P.M.L. Sandercock, E. Du Pasquier, *Forensic Sci. Int.* 140 (2004) 71-77.

[7] M. M. Zeng, Y. Z. Liang, H. D. Li, B. Wang, X. A. Chen, *Anal. Methods* 3 (2011) 438-445.

[8] H. T. Lv, C. S. Hung, K. S. Chaturvedi, T. M. Hooton, J. P. Henderson, *Analyst* 136 (2011) 4752-4763.

[9] M. M. Zeng, Y. Xiao, Y. Z. Liang, B. Wang, X. Chen, D. S. Cao, H.D. Li, M. Wang, Z. G. Zhou, *Metabolomics* 6 (2010) 303-311.

[10] ASTM Standard C1618, 2006e1, "Standard Test Method for Ignitable Liquid Residues in Extracts from Fire Debris Samples by Gas Chromatography-Mass Spectrometry," ASTM International, West Conshohocken, PA, 2003, DOI: 10.1520/E1618-11, www.astm.org.

[11] B. R. Kowalski, *J. Chem. Inf. Comp. Sci.* 15 (1975) 201-203.

[12] M. Otto, *Chemometrics*, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2007.

[13] R.G. Brereton, *Applied Chemometrics for Scientists*, John Wiley & Sons, Inc. Toronto, 2007.

[14] S. Wold, *Chemom. Intell. Lab. Syst.* 2 (1987) 37-52.

[15] N. Bratchell, *Chemom. Intell. Lab. Syst.* 6 (1989) 105-125.

[16] B. K. Slinker, S. A. Glantz, *Am. J. Physiol.* 255 (1988) R353-R367.

[17] R. Marbach, H. M. Heise, *Chemom. Intell. Lab. Syst.* 9 (1990) 45-63.

[18] T. Næs, H. Martens, *J. Chemometrics.* 2 (1988) 155-167.

[19] B. M. Zorzetti, J. M. Shaver, J. J. Harynuk, *Anal. Chim. Acta* 694 (2011) 31-37.

[20] P. J. Brown, *Anal. Proc.* 27 (1990) 303-306.

[21] S. Wold, M. Sjöström, SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy, In *Chemometrics: Theory and Application.* B. R. Kowalski, American Chemical Society, Washington, DC, 1977, 243-282.

[22] M. Barker, W. Rayens, *J. Chemometrics* 17 (2003) 166-173.

[23] M. Bylesjo, M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes, J. Trygg, *J. Chemometrics* 20 (2006) 341-351.

[24] J. M. Miller, *Chromatography: concepts and contrasts*, (2nd ed.) Wiley, Hoboken, N.J., 2005.

[25] C. F. Poole, *The essence of chromatogpraphy*, (1st ed.), Elsevier, Amsterdam, 2003.

[26] T. Górecki, J. Harynuk, O. Panić, *J. Sep. Sci.* 27 (2004) 359-379

[27] H. J. Cortes, B. Winniford, J. Luong, M. Pursch, *J. Sep. Sci.* 32, (2009) 883-904.

[28] I. François, K. Sandra, P. Sandra, *Anal. Chim. Acta* 641 (2009) 14-31

[29] F. Erni, R. W. Frei, *J. Chromatogr.* 149 (1978) 561-569.

[30] Z. Liu, J. B. Phillips,. *J. Chromatogr. Sci.* 29 1991 227-23.

[31] E. Katz, *Quantitative analysis using chromatographic techniques*, Wiley, New York, N.J., 1987.

[32] P. De la Mata-Espinosa, J. M. Bosque-Sendra, R. Bro, L. Cuadros-Rodríguez, *Anal. Bioanal. Chem.* 399 (2011) 2083-2092.

[33] J. Mommers, K. Knooren, Y. Mengerink, A. Wilbers, R. Vreuls, S. van der Wal, J. Chrom. A, 1218 (2011) 3159-3165.

[34] N. Etxebarria, O. Zuloaga, M. Olivares, L. J. Bartolomé. P. Navarro, *J. Chromatogr. A* 1216 (2009) 1624-1629.

[35] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, *J. Chromatogr. A* 1096 (2005) 101-110.

[36] N.A. Sinkov, J.J. Harynuk, *Talanta* 83 (2011) 1079-1087.

[37] N.A. Sinkov, B.M. Johnston, P.M.L. Sandercock, J.J. Harynuk, *Anal. Chim. Acta* 697 (2011) 8-15.

[38] N.A. Sinkov, J.J. Harynuk, *Talanta* 103 (2013) 252-259.

[39] J. M. Amigo, T. Skov, R. Bro, *Chem. Rev.* 110 (2010) 4582-4605.

[40] R.G. Brereton, Chemometrics Data Analysis for the Laboratory and Chemical Plant, Wiley, Toronto, 2003.

[41] P. H. C. Eilers, *Anal. Chem.* 75 (2003) 3631-3636.

[42] F. Gan, G. Ruan, J. Mo, *Chemom. Intell. Lab. Syst.* 82 (2006) 59-65.

[43] K. Kaczmarek, B. Walczak, S. de Jong, B. G. M. Vandeginste, *Acta Chromatogr.* 15 (2005) 82-96.

[44] Z. M. Zhang,S. Chen, Y. Z. Liang, *Analyst* 135 (2010) 1138-1146.

[45] B. K. Lavine, D. Brzozowski, A. J. Moores, C. E. Davidson, H. T. Mayfield, *Anal. Chim. Acta.* 437 (2001) 233-246.

[46] B. J. Asher, L. A. D'Agostino, J. D. Way, C. S. Wong, J. J. Harynuk *Chemosphere* 75 (2009) 1042-1048.

[47] P. De la Mata-Espinosa, J. M. Bosque-Sendra, R. Bro, L. Cuadros-Rodríguez, *Talanta* 85 (2011) 183-196.

[48] N-P. Nielsen, J. M. Cartensen, J. Smedsgaard, *J. Chromatogr. A* 805 (1998) 17-35.

[49] G. Tomasi, F. Van den Berg, C. Andersson, *J. Chemom.* 18 (2004) 231-241.

[50] F. Van den Berg, G. Tomasi, N. Viereck, Warping: investigation of NMR pre-processing and correction, In *Magnetic resonance in food*

*science: the multivariate challenge*, S. B. Engelsen, P. S. Belton, H. J. Jakobsen, Royal Society of Chemistry, Cambridge, 2005, 131-138

[51] K. J. Johnson, B. W. Wright, K. H. Jarman, R. E. Synovec, *J. Chromatogr. A* 996 (2003) 141-155.

[52] P. H. C. Eilers, *Anal. Chem.* 76 (2004) 404-411.

[53] S. Toppo, A. Roveri, M. P. Vitale, M. Zaccarin, E. Serain, E. Apostolidis, M. Gion, M. Mariorino, F. Ursini, *Proteomics* 8 (2008) 250-253.

[54] A. M. Van Nederkassel, M. Dazykowski, P. H. C. Eilers, Y. Vander Heyden, *J. Chromatogr. A* 118 (2006) 199-210.

[55] W. Yao, X. Yin,Y. Hu, *J. Chromatogr. A* 1160 (2007) 254-262.

[56] J. W. H. Wong, C. Durante, H. M. Cartwright, *Anal. Chem.* 77 (2005) 5655-5661.

[57] F. Savorani, G. Tomasi, S. B. Engelsen, *J. Magn. Reson.* 202 (2010) 190-202.

[58] K. Laursen, S. S. Frederiksen, C. Leuenhagen, R. Bro, *J. Chromatogr. A* 1217 (2010) 6503-6510.

[59] Y. Liang, P. Xie, F. Chau, *J. Sep. Sci.* 33 (2010) 410-421.

[60] D. Zhang, X. Huang, F. E. Regnier, M. Zhang, *Anal. Chem.* 80 (2008) 2664-2671.

[61] N. E. Watson, M. M. VanWingerden, K. M. Pierce, B. W. Wright,R. E. Synovec, *J. Chromatogr. A* 1129 (2006) 111-118.

[62] K. J. Johnson, R. E. Synovec, *Chemom. Intell. Lab. Syst.* 60 (2002) 225-237.

[63] K. Kjeldahl, R. Bro, *J. Chemometrics* 24 (2011) 558-564.

[64] Y. Felkel, N. Dorr, F. Glatz, K. Varmuza, *Chemom. Intell. Lab. Syst.* 101 (2010) 14-22.

[65] T. Rajalahti,R. Arneberg, F. S. Berven, K. M. Myhr, R. J. Ulvik, O. M. Kvalheim, *Chemom. Intell. Lab. Syst.* 95 (2009) 35-48.

[66] T. Rajalahti,R. Arneberg, A. C. Kroksveen, M. Berle, K. M. Myhr, O. M. Kvalheim, *Anal. Chem.* 81 (2009) 2581-2590.

[67] R. F. Teofilo, J. P. A. Martins, M. M. C. Ferreira, *J. Chemometrics* 23 (2009) 32-48.

[68] O. M. Kvalheim, T. V. Karstang, *Chemom. Intell. Lab. Syst.* 7 (1989) 39-51.

[69] O. M. Kvalheim, *Chemom. Intell. Lab. Syst.* 8 (1990) 59-67.

[70] I. Guyon, A.J. Elisseeff, *Mac. Learn. Res.* 3 (2003) 1157-1182.

[71] R.R. Hocking, *Biometrics* 32 (1976) 1-49.

[72] D.E. Axelson, *Data Preprocessing for Chemometric and Metabonomic Analysis*, first ed., MRi Consulting, Kingston, Ontario, 2010.

[73] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, *Metabolomics* 4 (2008) 81-89.

# 2 Chapter Two: Cluster resolution, a robust metric of model quality[2]

When applying chemometric techniques directly to raw chromatographic data, some form of variable selection is necessary. This is especially true for GC-MS data [1,2]. A few examples of data reduction strategies were discussed in Section 1.3.3 including integrated peak tables [3,4], single ion monitoring [5] (SIM) or extracted ion chromatograms [6,7] (EICs). However, those techniques have severe limitations. First, *a priori* knowledge is required for an analyst to select the portion of data to be used. This information may be unavailable when the system is not well understood. Second, SIM and EIC approaches will select mostly background variables and signal due to irrelevant analytes. Using total ion current [8] (TIC) chromatograms is another option, but this results in the loss of MS information as well as the inclusion of irrelevant data. An alternative to the above methods is to utilize an objective variable ranking technique.

## 2.1 Variable selection process

After variables are ranked, the analyst will then need to choose an appropriate way of selecting the variables based on their rank. This requires the calculation of a metric of model quality. When modeling data using chemometric techniques, there are essentially two types of models:

[2] This chapter is based on the following papers: N.A. Sinkov, J.J. Harynuk, *Talanta* 83 (2011) 1079-1087 and N.A. Sinkov, J.J. Harynuk, *Talanta* 103 (2013) 252-259.

classification/exploration models and regression models [1]. In this thesis, we have focused on variable selection processes for classification and exploration models of data. Consequently, the discussion is limited to variable selection for these types of models.

### 2.1.1 Variable ranking

Variable ranking methods use a calculated metric to estimate the potential value of each variable in a data set. Variables are then ranked relative to each other by this metric, which allows the analyst to determine which variables should be used to construct the model. When deciding which exact variables to use, the analyst will either use variables ranked above some threshold (i.e. variables above a specific ranking score, or above a certain percentile score), excluding all variables that fall below the threshold or perform further variable selection. In either case, an evaluation of model quality is required before a decision can be reached on what threshold to use or what to do during a given selection step.

Examples of ranking metrics include, but are not limited to, ANOVA-based ranking [9,10,11,12,13], SR-based ranking [14,15,16], and informative vectors [17]. Apart from their inherent advantage of objectivity, objective ranking strategies also allow the user to consider more candidate variables without requiring *a priori* information. This allows for easier automation of the variable selection process.

Once variables have been ranked, the analyst determines which variables are to be used. Variables with a higher ranking are more likely to improve the class separation, while those with a lower ranking are more likely to be irrelevant. On one hand, as more variables are included, it is more likely that information useful for class discrimination will be included in the model. On the other hand as lower ranked variables are added to the model, each additional variable is likely to be less useful than the previous ones [1]. On the other hand, with each new variable that is added, more noise is inevitably added to the model, reducing the model's ability to discriminate between classes. At some point, the addition of new variables will result in an overall loss of model quality. Thus a strategy that balances the costs and benefits of adding variables to the model is required to guide the selection of variables prior to further analysis.

### 2.1.2 Variable selection strategy

A simple way to address the task of variable selection is to use variables ranked above a certain threshold [9,10,11,12,13,14,15,17]. To determine this threshold, an analyst would construct models using different thresholds and see which of the models performed best. When the number of variables that can be potentially included in the model is large and the analyst has little information as to which or how many variables will actually be useful, many models will need to be constructed and evaluated, requiring considerable time and effort. Additionally, when all variables above a specific threshold are selected without further testing, a large number of correlated variables may be included in the model. When those variables come from a single analyte, it can result in over-representation of that analyte signal in the model, potentially degrading predictive ability of the model.

Stepwise approaches such as forward selection (FS) or backward elimination (BE) (Figure 2-1) are also popular [18,19,20]. In FS, a model is constructed with either one or a small number of variables and the quality of the model is evaluated. Then, a new variable is added, a new model is constructed, and its quality is evaluated. If the model improved as a result of adding the variable, the new variable is permanently retained in the model. If not, the variable is excluded.

In BE, a model is first created using all the variables to be considered. The quality of the original model is calculated, a single variable is eliminated, and the quality of the new model is calculated. If the original model had a higher quality score, the eliminated variable is replaced. Otherwise, the variable is eliminated permanently.

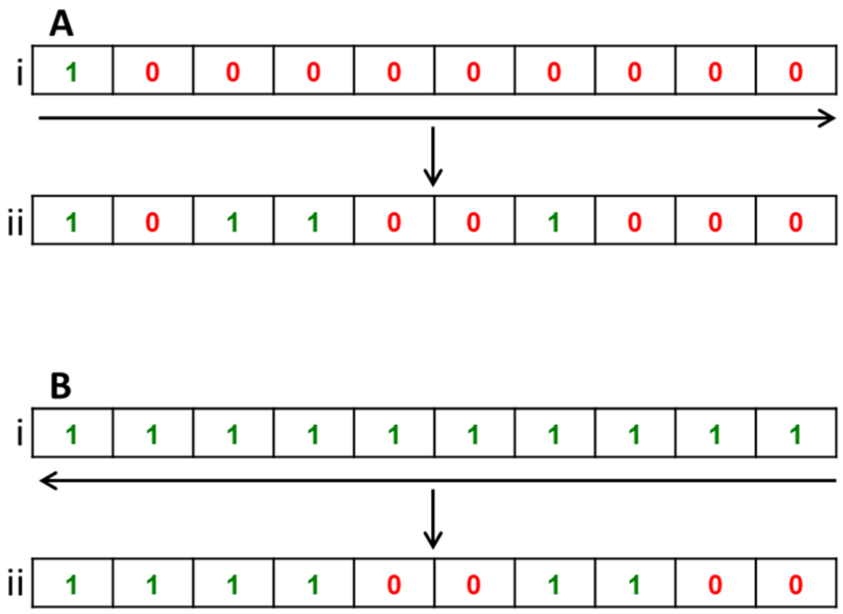


**Figure 2-1:** A schematic showing two of the stepwise variable selection techniques. **A** represents forward selection technique while **B** represents backwards elimination technique. “i” shows variables used in the first step of the selection while “ii” shows variables selected after the process is complete. “1” means that a variable is included in the model while “0” means that variable is excluded from the model.

Both approaches have advantages and disadvantages: FS is considered to be more computationally efficient, while BE considers each variable in the context of both variables permanently included in the model and unchecked variables. FS and BE are just two examples of stepwise approaches [18]. In the latter stages of this work a hybrid BE/FS approach was adopted as described in Section 2.3.2. Regardless of the variable selection strategy used, the quality of the model should be evaluated during each selection step.

### 2.1.3 Evaluation of model quality.

The evaluation of model quality is a central step in the optimization process. Some examples of approaches to model evaluation are the root mean square error of cross-validation (RMSECV) [17], the receiver operating characteristic (ROC) curves [21], or metrics based on Euclidian [2,11] or Mahalanobis [5,22] distances. When the goal is to distinguish between samples, it is desirable when scores for samples belonging to the same class are similar in PC or LV space, while scores for samples belonging to different classes are as different as possible. The degree of class separation metric defines the separation as the Euclidean distance in PC space between the centroids of two clusters of points, divided by the square root of the summed variances of each point in a cluster, relative to the centroid of the cluster [2,11]. Effectively, it will consider a cluster of points to be a circle (in 2-component space) or a (hyper)sphere (in 3+ component space). The calculation is fast; however, clusters of points usually do not resemble circles and are instead better defined by confidence ellipses. A metric termed cluster resolution (CR) has been developed in our group to quantify the separation between two clusters of points and is the topic of this chapter.

## 2.2 Cluster resolution metric in two dimensions

The CR metric is based upon determining at what maximum confidence limit a pair of confidence ellipses described around a pair of clusters will still not collide. In my first study of the CR metric, it was applied to the determination of the optimal number of variables to include in the construction of a PCA model with the goal of separating three classes of gasoline in the two-dimensional space described by PCs 1 and 2. Raw GC-MS data and ANOVA-based ranking were used.

### 2.2.1 Calculation of cluster resolution

The first step in the calculation of CR for a pair of clusters is to construct an ellipse at an initial confidence limit. A cluster is defined as the collection of scores of samples belonging to the same class on PCs (or LVs) 1 and 2. Considering only the data from a single class, the scores for the points in the cluster provide new variables from which a two-component PCA model is constructed upon mean-centered data, with no other scaling. The resultant loading vectors for PCs 1 and 2 provide the directions of the major and minor axes for the resulting confidence ellipse that describes the class. Then, using eigenvalues for the two components, the critical Hotelling $T^2$ value for a given confidence limit and the number of samples in the class will be used to calculate the lengths of each of the ellipse's axes. This process is repeated every time a new confidence ellipse needs to be constructed.

To construct a confidence ellipse for a class, the covariance matrix of the scores belonging to the samples in the class is first calculated:

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} \qquad (2\text{-}1)$$

Where **S** is the scores covariance matrix, *n* is the number of samples contained in the cluster being evaluated, and **X** is a matrix where the rows are the samples in the cluster and the columns are the scores of each sample on PCs 1 and 2.

Next, singular value decomposition (SVD) is performed on the covariance matrix:

$$\boldsymbol{S} = \boldsymbol{ULV^T} \qquad (2\text{-}2)$$

Where **U** and **V** in this case are identical and provide loading vectors for the model describing the confidence ellipse, while **L** is a diagonal matrix of eigenvalues for components 1 and 2 of the new model.

That information, along with the number of samples in the cluster, permits the determination of the Hotelling $T^2$ value for a given confidence limit [23]:

$$T^2 = \frac{p(n-1)}{n-p} F(\alpha, p, n-p) \qquad (2\text{-}3)$$

Where *p* is number of components in the model, *n* is the number of samples in the class, *α* is the confidence limit and F(*α*,*p*,*n*-*p*) is the F statistic for given values of *α*, *p* and *n*.

The length of each confidence ellipse axis (*l*) is given by:

$$l = \sqrt{T^2 \times L} \qquad (2\text{-}4)$$

When the length of the major axis is calculated, *L* is the eigenvalue for PC 1, and when the length of the minor axis is calculated, *L* is the eigenvalue of PC 2.

As a result, lengths and directions of the axes describing the ellipses at a given confidence limit are known. Next, a set of approximately 1000 evenly-spaced points are distributed along the circumference of each ellipse.

This is achieved by warping a circle comprising 1 000 000 points until it is superimposable on the confidence ellipse. To reduce the number of points in the ellipse, its circumference is calculated using Rajmanujan's approximation [24] and divided by 1000 to yield the desired distance between two adjacent points, *d*. Then, beginning at an arbitrary point on the circumference of the ellipse, the algorithm proceeds along the ellipse until a point a distance *d* along the ellipse is found. Points between the starting point and this second point are then discarded. This process is repeated around the entire ellipse with the result being an ellipse with approximately 1000 points distributed mostly evenly along its circumference.

The choice of 1000 points was made because it provides a balance between accurate representation of ellipses and computational speed. The obtained confidence ellipse is then rotated around the origin so that major and minor axes match the directions of the loadings of PC1 and

PC2 (of the model calculated for that cluster), respectively. Finally, the ellipse is moved in the original model space so that center of the ellipse matches center of the cluster. The process is then repeated to obtain the second ellipse in the pair that needs to be constructed.

To determine if two confidence ellipses overlap at a given confidence level, the Euclidean distances between all points on one ellipse and those on a second ellipse are calculated, a total of approximately 1 000 000 distances depending on the exact number of points along the circumference of each ellipse. The minimum of these distances ($D_{min}$) is then compared to half the sum of the distances between two neighbouring points on the circumferences of each ellipse ($D_{critical}$). If $D_{min}$, is less than $D_{critical}$, the two ellipses are deemed to overlap.

The $D_{critical}$ is given by:

$$D_{critical} = \frac{1}{2}(D_A \times D_B) \qquad (2\text{-}5)$$

Where $D_A$ is the average distance between neighbouring points on the confidence ellipse A and $D_B$ is the average distance between neighbouring points on the confidence ellipse B.

To determine the maximum confidence limit at which ellipses will not overlap, the algorithm begins with an arbitrary confidence limit for each pair of classes (in this work we chose to use 75%) and determines if there is any overlap. If overlap is detected, the algorithm decreases the confidence limit for both ellipses in a stepwise fashion until overlap is no longer detected. Conversely, if there is no overlap detected, the algorithm increases the confidence limits of the two ellipses until overlap is detected. The highest confidence limit at which there is no overlap detected is defined as the cluster resolution for that particular pair of classes. Cluster resolution is calculated for each pair of classes separately and will have a value between 0 and 1 (representing 0 and 100% confidence limits). It should also be noted that for a given pair of clusters, the algorithm will always converge to the same value of CR, regardless of the initial guess. The only consequence of a poor initial guess is increase in computational time the first time CR value is calculated for a given pair of classes.

During determination of the optimal number of variables to use, the highest-ranked unused variables are added to the model with each step. The first time that CR is determined for a given pair of classes, the algorithm starts with the arbitrary confidence limit (75%). However, after the first iteration, the cluster resolution for each pair of clusters is stored and used as the starting point for subsequent iterations with additional variables. This approach recognizes that addition or deletion of a variable will typically have a small effect on CR, so the number of calculations in each step is minimized, resulting in faster calculations. Figures showing the step-by step construction of a confidence ellipse are presented in Appendix 1. As shown in Figure 2-2, the measured CR value will depend on the orientations of clusters as well as their shapes.

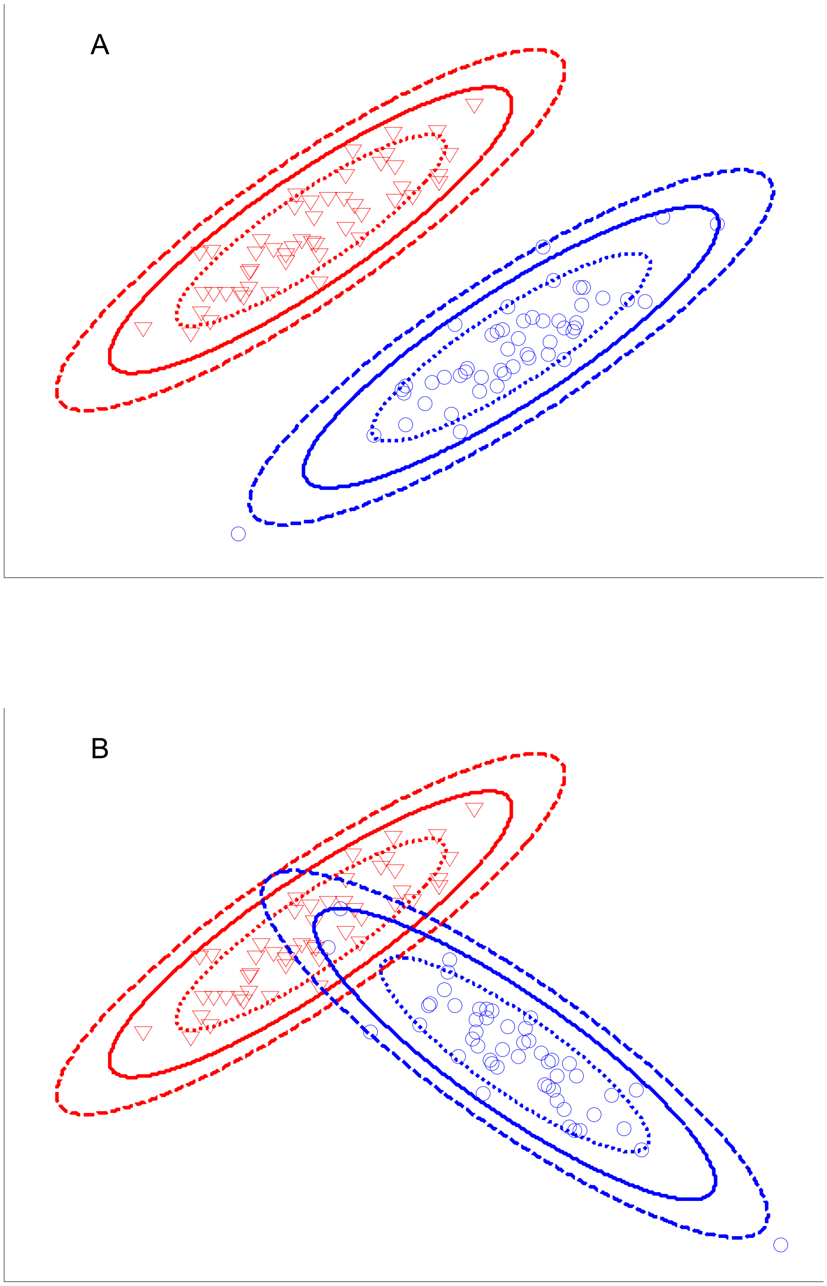


**Figure 2-2:** Two clusters of points with 75, 95 and 99 % confidence ellipses. **A** Ellipses are oriented parallel to each other, and **B** ellipses oriented such that they have some overlap. The centroids of the ellipses and the sizes of the ellipses have not changed during the rotation.

### 2.2.2 Application of two-dimensional cluster resolution.

The first step in automated variable selection is to rank the variables according to some metric, such as ANOVA or SR. The choice of ranking metric is up to the analyst, though it should be noted that in our work we have observed that different ranking metrics produce different optimal models exhibiting maximal cluster resolution. In this study ANOVA ranking was used, as it is computationally inexpensive and straightforward. The two main limitations in using ANOVA are that it assumes that the observed variance is normally distributed, and that when used on a data set where the number of variables vastly exceeds the number of samples (which is a very relevant concern when dealing with chromatographic data) it is entirely possible for ANOVA to find some features that can discriminate between classes based upon nothing other than random fluctuations in the data as opposed to meaningful variances. The later limitation can be addressed through splitting the data into a training set and a validation set.

The output of ANOVA is a series of F ratios for each variable. The F ratio is a measure of the ratio of between-class variance to within-class variance [9,10,11]. If a variable has an elevated F ratio, then it is deemed to be more valuable for describing the difference between classes. Once the F ratio is calculated for every data point in the chromatogram, the variables are ranked in order of decreasing F ratio. A PCA model is then constructed using a fraction of variables that have the highest F ratio.

Since the CR metric used during this study worked in two dimensions, the PCA model contained two components.

Once a model is constructed, scores from samples belonging to a given class will form a cluster in a certain region on the scores plot, and confidence ellipses can be described around the cluster formed by each class.  The cluster resolution between each pair of classes is then calculated. In this study, the process was repeated, including more and more top-ranked variables with every step, until the desired endpoint is reached. As the number of variables was increased in each step, the threshold above which ranked variables on the ranking would be introduced was decreased. There are different ways to define the endpoint. For example, it may be defined by a pre-set number of variables or the number of variables where the resolution is maximized (such as when the critical pair of classes shows the highest cluster resolution or when product of all cluster resolutions for all class pairs is maximized). It may also be defined by the minimum cluster resolution becoming greater than a threshold value, for example 0.95 (meaning that no confidence ellipses exhibit overlap at the 95% confidence level). In this study, the algorithm was allowed to run until a set number of variables had been checked.

### 2.2.3 Experimental

To demonstrate cluster resolution and its use in automated feature selection, a set of gasoline samples was used. Three samples of gasoline having octane ratings of 87, 89, and 91 were obtained from a single local gas station in Edmonton, Alberta, Canada. These samples were diluted 20:1 by volume in pentane and analyzed using GC-MS. The GC-MS used for these experiments was a 7890A GC with a 5975 quadrupole MS (Agilent Technologies, Mississauga, ON) equipped with a 30 m × 250 µm; 0.25 µm HP-5 column (Agilent). The carrier gas used was helium at constant flow rate of 1.0 mL·min$^{-1}$. The injector was held constant at 250 °C and a volume of 0.2 µL was injected with a split ratio of 100:1. The temperature program for the GC was 50 °C (3.5 min hold) with a 20 °C·min$^{-1}$ ramp to 300 °C. The transfer line and source temperatures were 185 and 230 °C, respectively. The total run time was 16 min. The initial solvent delay was 2.5 min and mass spectra were collected from m/z 30 to m/z 300 at the rate of 9.2 spectra/s.

A total of 24 chromatograms were collected for each of the gasoline samples over a period of two weeks. The entire raw mass chromatogram for each analysis was exported as a .csv file, which was then imported into MATLAB 7.10.0 (The Mathworks, Natick, MA) as a 7400×271 (scan number × m/z ratio) matrix using a lab-written algorithm. Data were then handled in MATLAB using lab-written algorithms. Chemometric models were constructed using PLS toolbox 5.2 (Eigenvector Research Inc., Wenatchee, WA). The calculations were performed on an Intel Core i5 750 2.76 GHz processor with 8 GB of RAM and 64-bit Microsoft Windows 7 Professional operating system.

### 2.2.4 Results and discussion.

As a demonstration, the CR metric was used in an algorithm to automatically select the features in the data which can be used to construct the PCA model having the greatest degree of separation between clusters for each class. In our example, the data consisted of 72 GC-MS chromatograms of gasoline samples having three different octane ratings. The 72 chromatograms were randomly split into a training set (containing 16 chromatograms from each class) and a test set (containing the remaining 8 chromatograms from each class). This was repeated 4 times to obtain a total of five different randomly chosen training and test sets to evaluate the stability of the solution to minor variations in the training data. Finally the procedure was performed on the complete set of data with no test set.

For data alignment, chromatograms were aligned using a homemade alignment function based upon the piecewise alignment algorithm developed by Johnson et al [25], with an additional mass spectral confirmation of features to be matched, though in principle any alignment algorithm could be used. The target used for data alignment was a composite chromatogram of a series of aligned gasoline samples of different octane ratings. This ensured that all components present in the samples were present in the alignment target, if not necessarily at the same abundances. The aligned matrices were then unfolded along the time axis to yield a series of vectors. ANOVA ranking was applied to the set of 48 chromatograms in each training set using a lab-written algorithm. For each set, this yielded a vector of F ratios that was used to rank the features. The test data sets were aligned as well, but were not used in calculation of F ratios. Baseline correction was not necessary as the ANOVA process automatically down-weights background ions which do not vary significantly from sample to sample.

With variables now ranked by their F ratios, the data in the training sets were autoscaled and for each step in the selection process, subsets of data containing all rows (samples) and the desired number of columns (features) were extracted and used to construct a two-component PCA model. The cluster resolution between each possible pairing of classes on the scores plot for PC1 vs. PC2 was then calculated on the basis of the training data set. This step was repeated sequentially, adding more and more variables at each step to find the optimal number of variables to include in the PCA model for each training set.

The original training data set comprised a matrix of 48 rows (representing samples), and 2 005 400 columns (representing variables). In all cases except one, the maximum number of variables to be included was limited to 100 000. In one case calculations were performed up to the total data set of 2×$10^6$ variables to demonstrate the problem with utilizing the entire raw data file, especially when the data are incredibly sparse (Figure 2-3 C). In terms of computational time, it may take a few minutes to calculate the initial cluster resolution. The exact time depends on the data, the initial confidence limit guess and the step size used for changing the confidence limits as the algorithm must incrementally adjust from the arbitrary initial value until a collision is observed. However, as the resolutions that are found in a given iteration are used as the starting points for the subsequent iteration, the speed is limited by how fast variables can be extracted from a dataset and the PCA model can be constructed. In practice the process took about two seconds per step. To efficiently determine the optimal number of variables to use, a large step size can be used in the first pass through the data to find the approximate location of the optimum. Then progressively smaller step sizes can be used in the vicinity of the optimum to locate its exact position. Additionally, when determining CR for a given pair of clusters, the incremental changes in confidence limit tested must become smaller as the confidence limit approaches 100%. The reason for that is that a relatively small change in confidence limit will result in large change in the size of the confidence ellipses.

Figure 2-3 depicts the results of the optimization process for the first of the five sets of data. The cluster resolution between pairs of ellipses is plotted on the y-axis vs. the number of features that are included in the model. It is apparent from Figure 2-3 that with few variables it is relatively easy to model the differences between 87- and 89-octane gasolines and 87- and 91-octane gasolines. Conversely, it is difficult to distinguish between 89- and 91-octane gasolines, which represent the critical pair of clusters in this case. This figure also highlights the advantage of using a metric that is bounded between 0 and 1. Overall model quality may be assessed by taking the product of individual cluster resolutions. In Figure 2-3, it is apparent that at a low number of included variables, 89- and 91-octane gasolines are not separated, a fact that is accurately reflected by the product of individual cluster resolutions.

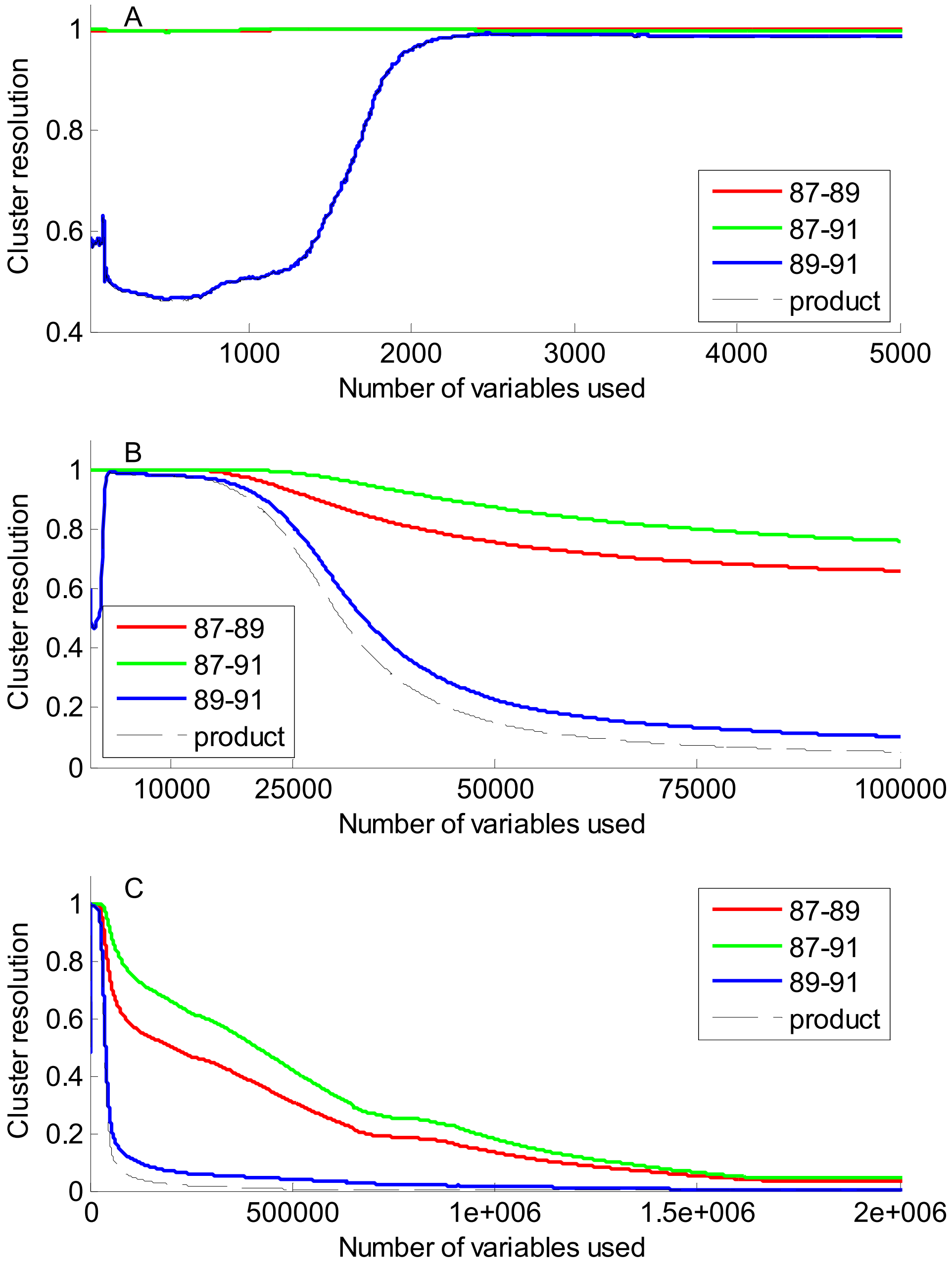
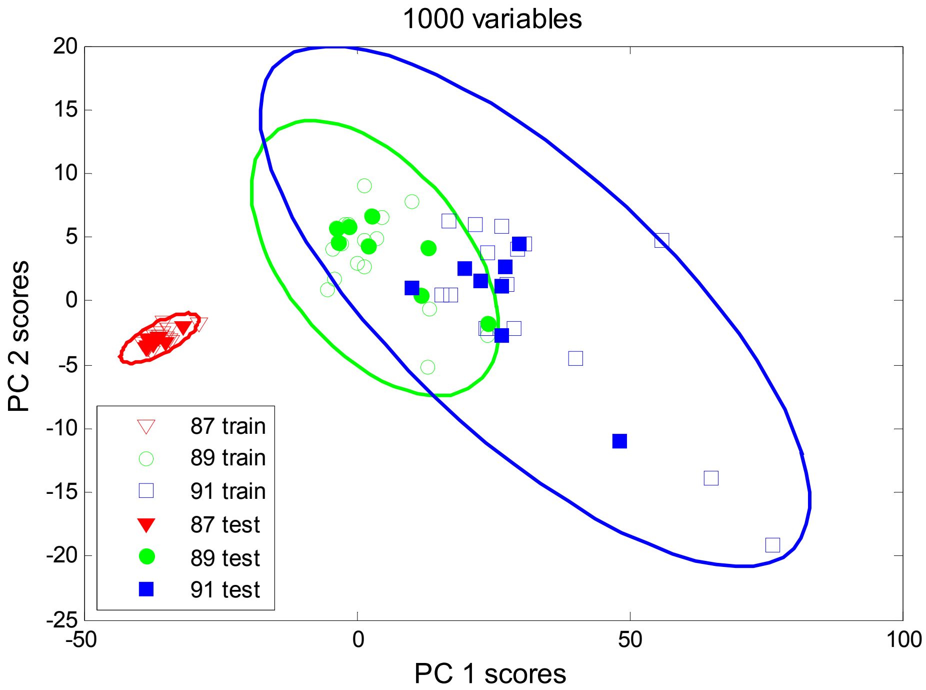
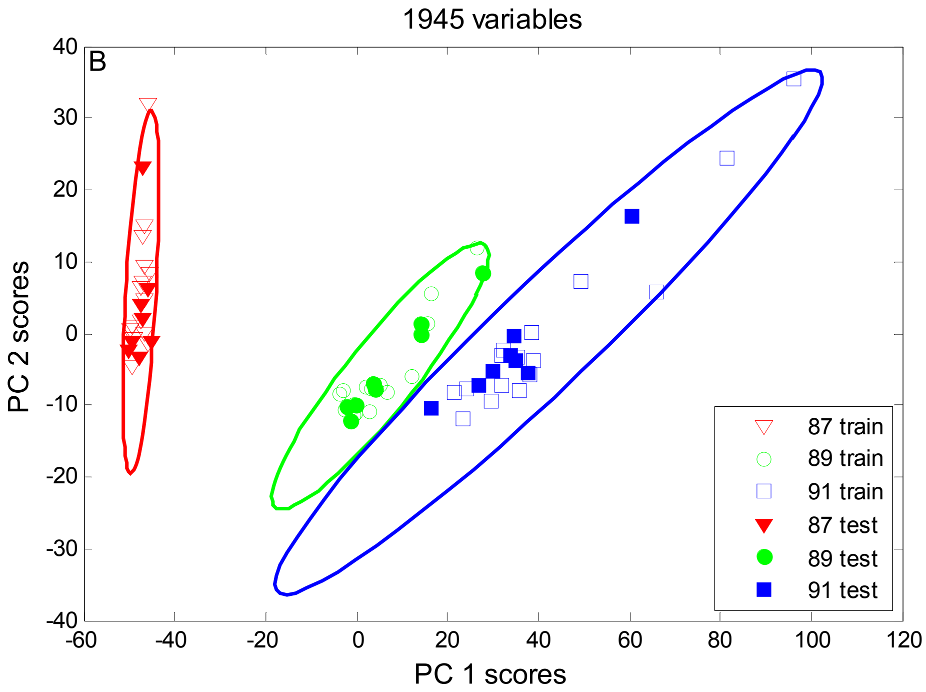
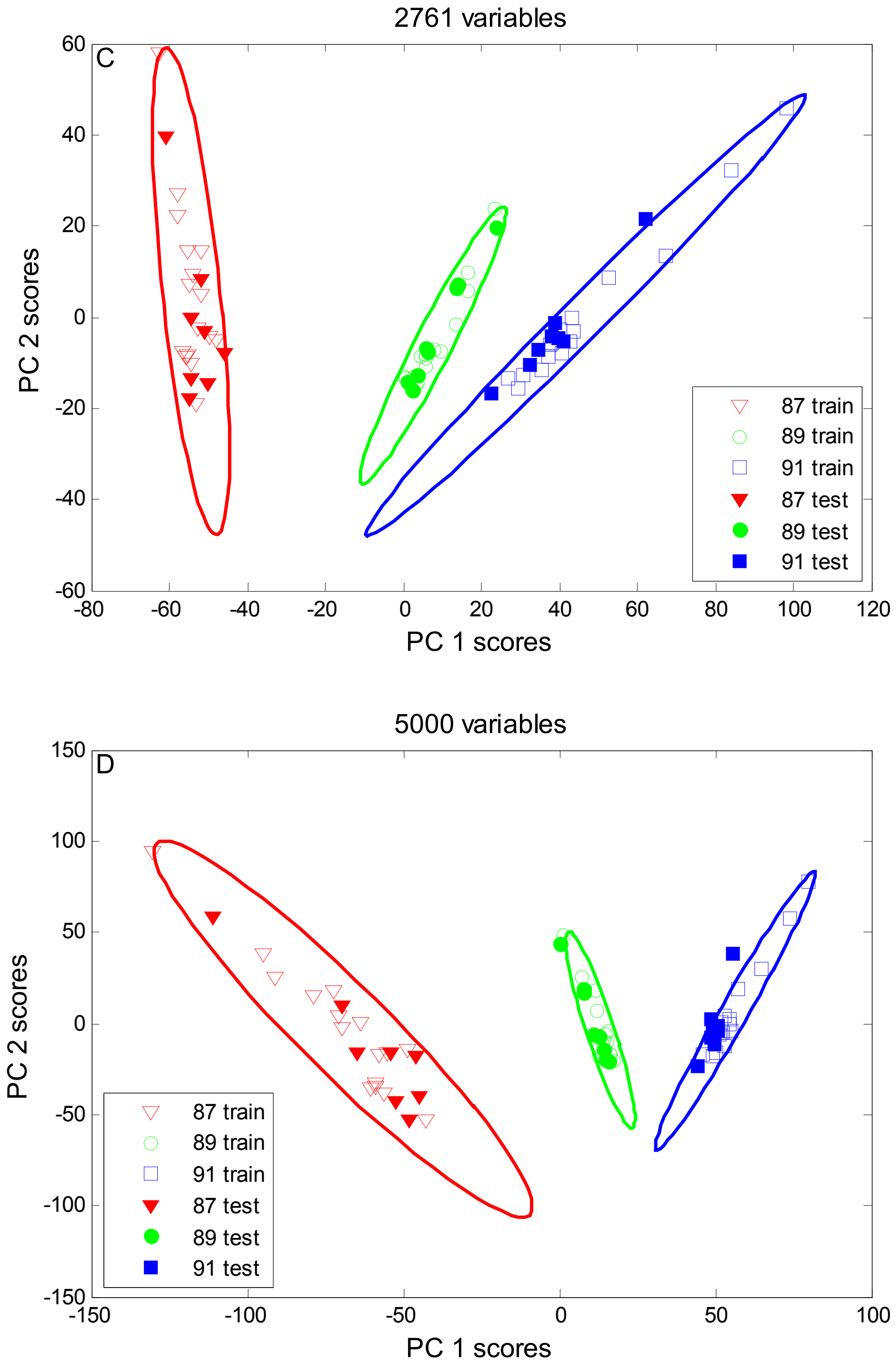


**Figure 2-3:** Resolution of gasoline clusters as a function of the number of variables used for Set 1. **A** close up of the region from 0 to 5000 included variables; **B** close up of the region from 0 to 100 000 included variables; **C** full resolution plot from 0 to 2 000 000 included variables.

As the number of variables included in the model increases from zero to the optimal value, the separation between the 89- and 91-octane gasolines shows marked improvement, with the 95 % confidence ellipses becoming separated when 1945 variables are used and reaching a maximum at 2761 variables when used on Set 1. As this pair of clusters was always limiting the quality of the model (as seen from Figure 2-4), the optimal number of features was determined based on the resolution of this pair. Investigating the trend in resolutions beyond this optimum, a gradual decrease in the resolution for the critical pair was observed until about 20 000 variables (Figure 2-3 B). Figure 2-3 C demonstrates the extreme degradation in resolution that is observed when a very large number of variables is included in the model.


1000 variables
PC 2 scores
PC 1 scores
20
15
10
5
0
-5
-10
-15
-20
-25
-50
0
50
100
▽ 87 train
○ 89 train
□ 91 train
▼ 87 test
● 89 test
■ 91 test


1945 variables
B
PC 2 scores
PC 1 scores
40
30
20
10
0
-10
-20
-30
-40
-60
-40
-20
0
20
40
60
80
100
120
▽ 87 train
○ 89 train
□ 91 train
▼ 87 test
● 89 test
■ 91 test


2761 variables

5000 variables

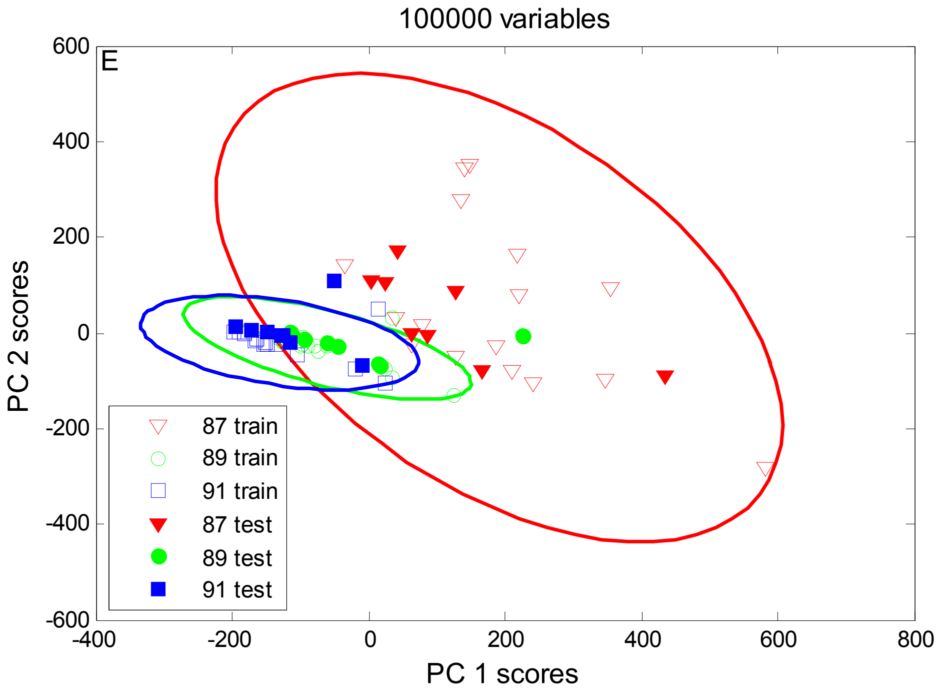


**Figure 2-4:** Scores plots for selected PCA models for set 1. Red triangles represent 87-octane gasoline, green circles represent 89-octane gasoline and blue squares represent 91-octane gasoline. Filled markers represent samples used for feature selection and model construction. Hollow markers represent test data to which model was applied. 95% confidence ellipses indicated for each class. **A**, **B**, **C**, **D**, **E** show plots for 1000, 1945, 2761, 5000, and 100 000 included variables, respectively.

Figure 2-4 depicts the scores plots from the two-component PCA models constructed using different numbers of variables to highlight the regions in Figure 2-3. As predicted by the plot in Figure 2-3 A, a 1000-variable model does not include sufficient features to separate all of the classes at the 95 % confidence level. Overlap of the 89- and 91-octane gasoline samples is observed, while the 87-octane gasoline is well separated from the other two. The model constructed using 1945 variables (Figure 2-4 B) should show that all ellipses are separated at the 95% confidence level, with 89 and 91 octane rating samples being barely separated. The model that is constructed using 2761 variables (Figure 2-4 C) exhibits the best overall resolution between the three classes. The addition of more variables (e.g. the 5000-variable model shown in Figure 2-4 D) somewhat decreases the resolution. In the extreme case that far too many variables are included (100 000, Figure 2-4 E) the quality of the model is degraded to a large extent, with all ellipses exhibiting significant overlap at 95 % confidence level.

The model constructed using the optimum number of points from the training set, shown in Figure 2-4 C, was applied to the test set data. It can be observed that all samples from the test set fall within their respective confidence ellipses. Similar results were also observed for other sets, as summarized in Table 2-1. Additionally, when different training sets were selected from the original data set, the number of variables required to reach the optimum did not show much variation. Moreover, the optimum that is indicated by the least-separated class is identical (or very similar) to the optimum determined from the product of the resolutions between all pairings, and in both cases most points fall into their respective 95% confidence ellipses (Table 2-1).

After important variables have been identified, their positions in the original data can be identified and a binary mask may be generated to visualize the relevant chromatographic information.

**Table 2-1:** Numbers of variables identified as optimum for a given set using a given metric as well as false positive and false negative rates. FP – False Positive. Samples in the test set that do not fall within 95 % confidence ellipse of their class. FN – False Negative. Samples in the test set that fall inside 95 % confidence ellipse of at least one other class. n – Number of variables at optimum. All Train – numbers of variables at optimum of a set that includes all data**.**

| | Cluster resolution | | | | | | Euclidian distance | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Critical pair | | | Product | | | Critical pair | | | Product | | |
| Set | *n* | FP | FN | *n* | FP | FN | *n* | FP | FN | *n* | FP | FN |
| 1 | 2761 | 0 | 1 | 2761 | 0 | 1 | 29900 | 8 | 3 | 97600 | 12 | 2 |
| 2 | 2461 | 0 | 1 | 2461 | 0 | 1 | 9300 | 0 | 4 | 66600 | 8 | 2 |
| 3 | 2265 | 0 | 1 | 2265 | 0 | 1 | 24900 | 0 | 1 | 56800 | 10 | 0 |
| 4 | 2985 | 0 | 1 | 2657 | 0 | 2 | 31800 | 12 | 2 | 98400 | 16 | 2 |
| 5 | 3189 | 0 | 2 | 3194 | 0 | 2 | 31100 | 6 | 10 | 92600 | 12 | 11 |
| Average | 2732 ± 376 | | | 2668 ± 350 | | | 25400 ± 9400 | | | 82400 ± 19300 | | |
| All train | 2027 | | | 2027 | | | 22900 | | | 62000 | | |

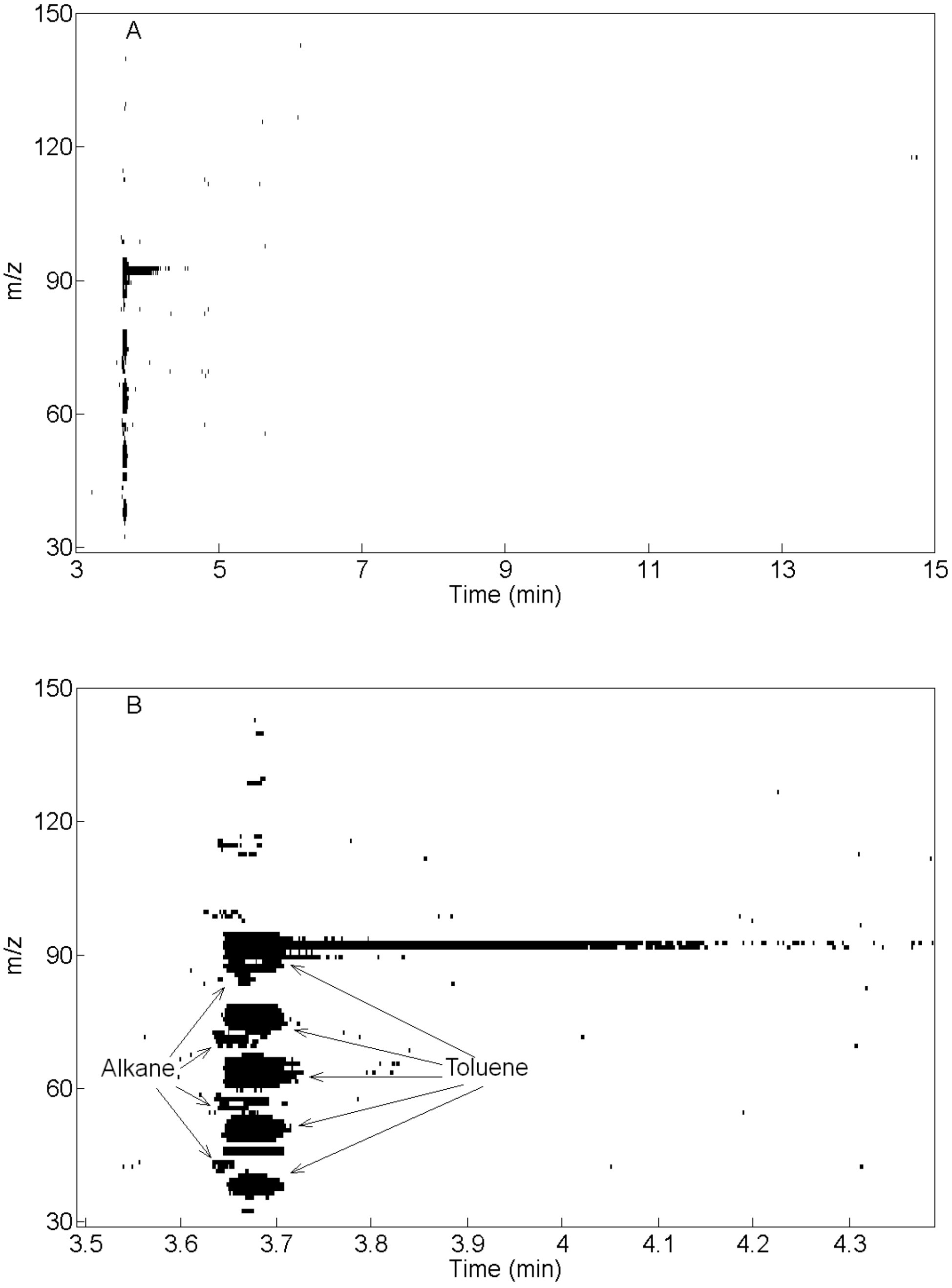


**Figure 2-5:** Mask representing variables selected by the algorithm. Black dots indicate variables selected, white space indicates variables not selected. **A** shows full time axis, **B** shows a close-up of the region of the chromatogram where most variables were selected

In the mask (Figure 2-5), the included variables are assigned a value of one and excluded variables are assigned a value of zero. Applying this mask to a chromatogram will result in only the relevant variables remaining, as shown in Figure 2-6.

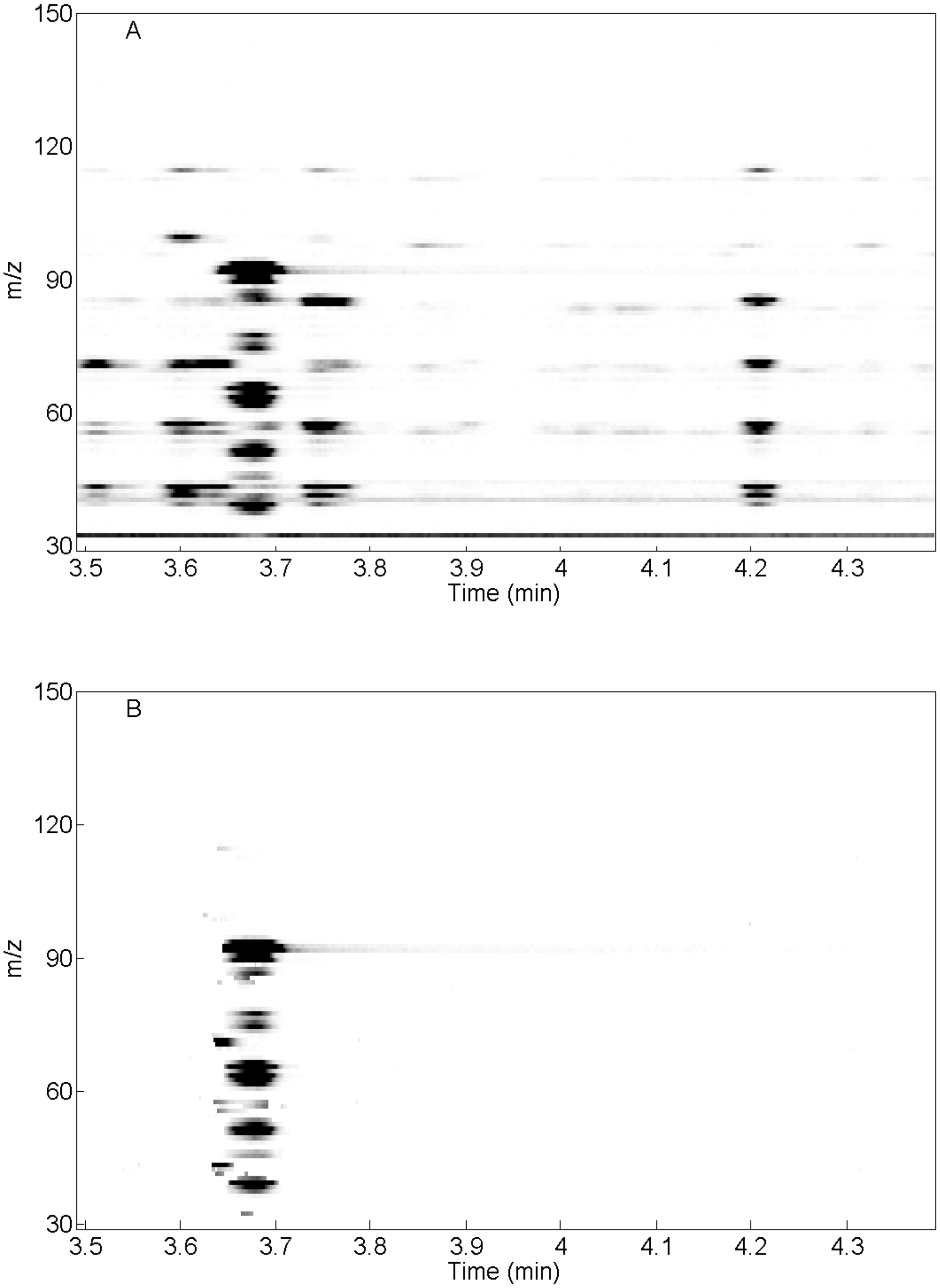
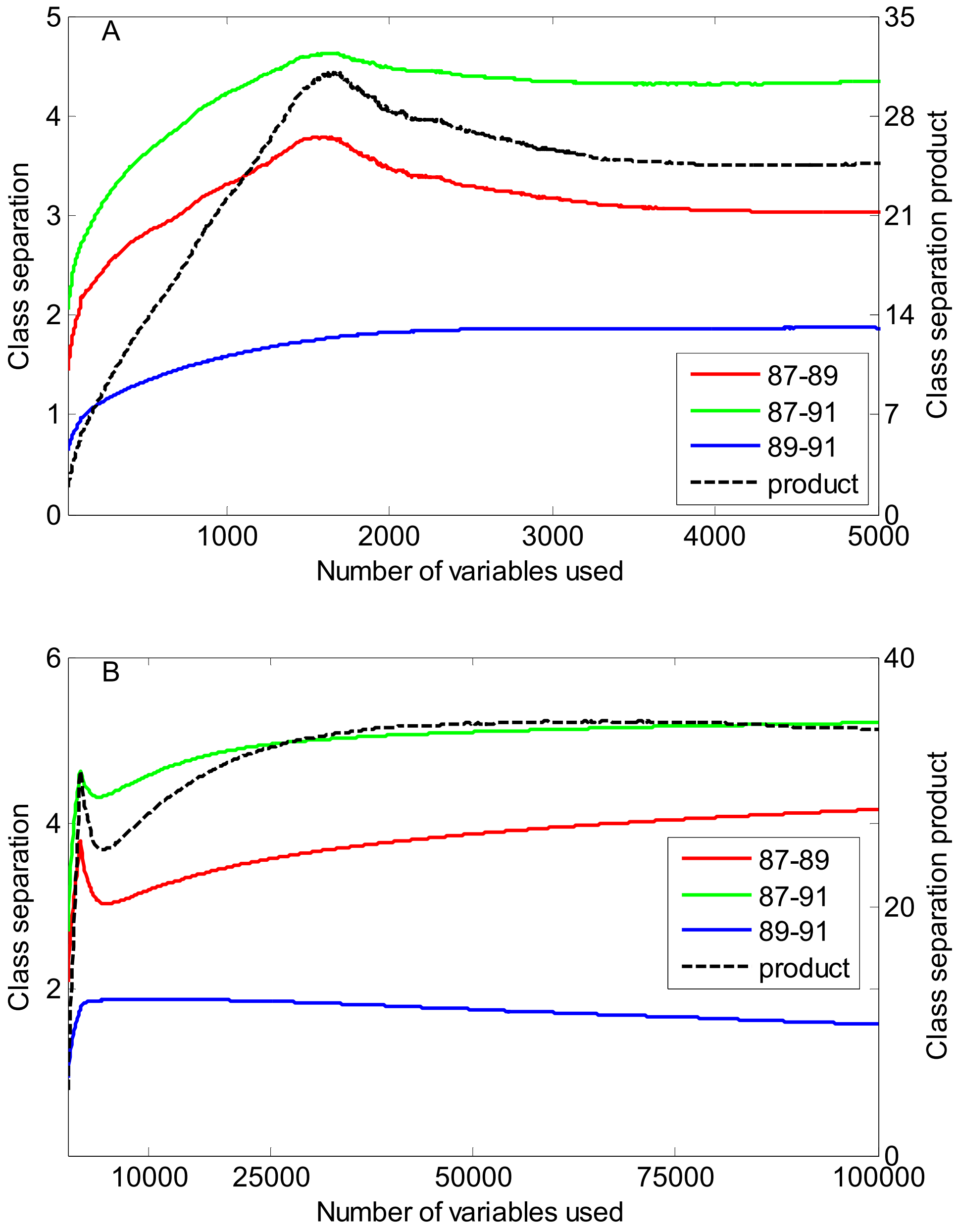


**Figure 2-6:** A segment of a GC-MS chromatogram of gasoline sample. **A** shows raw signal. **B** shows only signal due to the variables selected. Darker points represent stronger signal.

As can be seen, the process selected signal due to two coeluting compounds which allow discrimination between the three classes of gasoline. Investigation of the raw GC-MS data indicates that the compound responsible for the ions at m/z ratios of 91, 78, 65, 52, and 39 is toluene. These variables are added first. As can be seen in Figure 2-3 A, it is relatively easy to distinguish between 87- and 89-octane and 87- and 91-octane gasolines but difficult to distinguish between 89- and 91-octane gasolines. This is indeed what is observed in the GC-MS data. The 89- and 91-octane gasolines have similarly high concentrations of toluene and other aromatics while the concentrations of these compounds are relatively low in the 87-octane gasoline.

The second compound that was included by the algorithm, indicated by the points that elute slightly before toluene (m/z 43, 57, 71, 85, and 99) are due to a hydrocarbon that coelutes with toluene. Based on the mass spectrum of the compound it is a branched, saturated hydrocarbon having eight total carbons, likely 4-methyl heptane. Inspection of the chromatographic data shows that these features are in fact due to a compound which has a relatively high concentration in the 87- and 89-octane gasoline samples and a relatively low concentration in the 91-octane gasoline. This indicates that the selection process is able to automatically identify features in the data that have an actual chemical origin. Furthermore, it shows the power of using the raw chromatographic data over integrated peak tables: if integrated peak areas were used here, it is very likely that the alkane would not have been observed due to the coelution with the much larger toluene peak. The other observation is that a further data preprocessing step could be implemented to reduce the number of ions considered for the model. Here, most of the approximately 2500 variables were used to describe essentially two chromatographic peaks. This will be the subject of future research.

The cluster resolution metric was then compared to a previously described metric which is based on the Euclidean distance between the centroids of pairs of classes relative to square root of the sum of the variance within each group [2,11]. Figures 2-7 A and 2-7 B show the degree of class separation for each set of classes, as well as the product of the class separations (which has been suggested as a parameter for optimizing overall class separation [9]), using the same test data as the one used with the CR metric. When the product of class separations for three classes was considered, the optimal separation was predicted to be at 29 900 variables. A visual inspection of the scores plot in Figure 2-7 C and the one created using the optimum number of variables predicted by the algorithm using cluster resolution (Figure 2-4 C) shows that the cluster resolution metric provides a model with a significantly more distinct class separation.


A
Class separation
0
1
2
3
4
5
Class separation product
0
7
14
21
28
35
1000
2000
3000
4000
5000
Number of variables used
87-89
87-91
89-91
product
B
Class separation
2
4
6
Class separation product
0
20
40
10000
25000
50000
75000
100000
Number of variables used
87-89
87-91
89-91
product

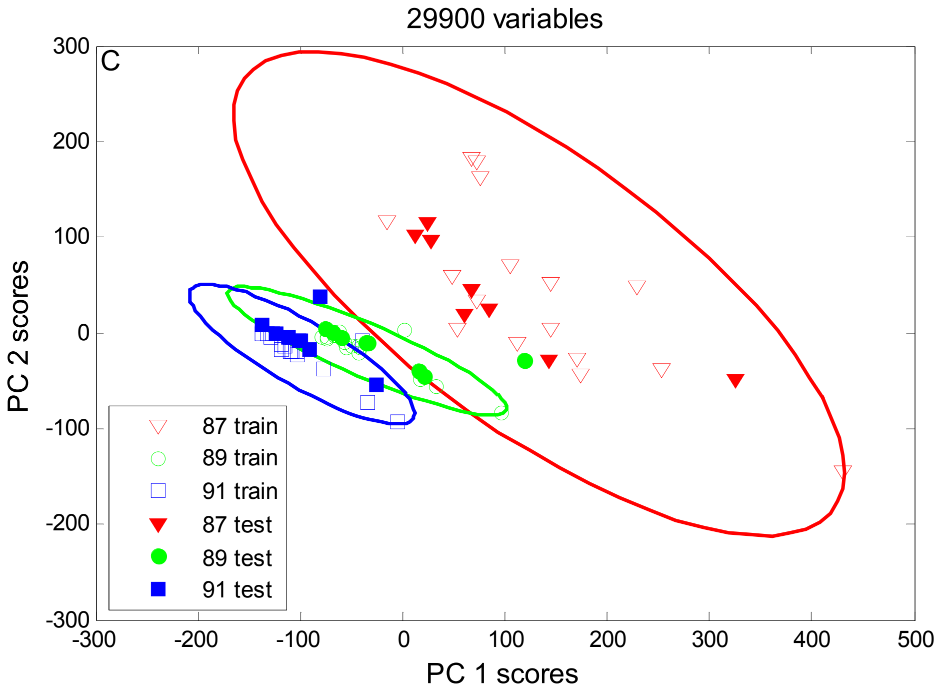


**Figure 2-7:** Degree of separation vs. number of variables when calculated using a Euclidean distance approach as well as the product of degree of separation for the three classes. **A** close up of the region from 0 to 5000 included variables; **B** close up of the region from 0 to 100 000 included variables; **C** Scores plot from PCA model using optimal number of variables found in **B**.

However, it should be noted that samples used in this study were relatively simple and there was little in-class variation. Also, the CR metric was only shown to work with two-dimensional models. Finally, using a more complex variable selection strategies compared to using a threshold for ranked variables is another point that needed to be addressed. Thus, the CR metric was extended to work in three dimensions and a hybrid backward elimination (BE) and forward selection (FS) approach was adopted.

## 2.3 Cluster resolution metric in three dimensions

The use of CR in the optimization of variable selection for a three-component PCA model is demonstrated in this study. To further increase the effectiveness of variable selection, a hybrid variable selection strategy involving both FS and BE was employed. To provide a greater challenge for the metric, the simultaneous optimization of a model that would distinguish between six classes of gasoline samples comprising three octane ratings from two different vendors was attempted. Furthermore, in-class variance was introduced into each class by weathering some gasoline samples within each class as well as introducing contaminants to samples within each class. Performance of the three-dimensional CR metric was compared to using three projections of the two-dimensional CR metric.

### 2.3.1 Changes in CR calculation

Confidence ellipsoids can, theoretically, be created in any number of dimensions. This can be achieved by constructing an *n*-component PCA model around a cluster of points, with resulting loading vectors defining the directions of the *n* ellipsoid axes. Then, the lengths of the axes at a given confidence limit can be calculated using equations provided previously (Equation 2-4). With the position of the ellipsoid center, as well as sizes and directions of the ellipsoid axes, a confidence ellipsoid with approximately evenly-spaced points covering its surface is constructed. To construct a three-dimensional ellipsoid, first a two-dimensional ellipse is constructed with its major and minor axes being defined by PCs 1 and 2 respectively. Next, a distance equal to the distance between the points of the 2-D ellipse is measured along the circumference of an ellipse which has its major and minor axes defined by PCs 1 and 3 respectively (since PCs are by definition orthogonal, this ellipse is perpendicular to the first one). With that distance measured, this then defines how much smaller an ellipse parallel to the PC1/PC2 ellipse but passing through the identified point on the PC1/PC3 ellipse must be and how much should it be offset from the original on PC 3. This new ellipse will maintain the same distance between points while containing fewer of them and will be offset from the plane defined by PC 1 and PC 2. The process is repeated until the apex of the minor axis is reached and half of the 3-D ellipse is constructed. Since the ellipsoid is symmetrical, the second half of the ellipsoid is a mirror image of the first half. With an ellipse of the correct size constructed, it is then rotated and moved to the cluster of points. A visual representation of ellipsoid construction is provided in Appendix 2.

With two such ellipsoids constructed around two clusters of samples, collision detection can be performed as described previously (Section 2.2.1). If a collision is detected, the confidence limit is reduced for the following iteration of collision detection. If a collision is not detected, the confidence limit is increased in the following iteration. The highest confidence limit at which confidence ellipses are still separated defines the CR. In a multi-class model, the calculation must be performed for each possible pairing of classes, and the product of the CRs for all possible pairs of classes is the overall quality metric for the model.

### 2.3.2 Variable ranking and selection strategy

In this study, the Selectivity Ratio (SR) was adopted as a variable ranking technique. SR has been previously described in the literature [14,15]. Briefly, it involves the creation of a PLS-DA model which is then used to calculate scores and loadings for the target projected (TP) model. Then the ratio of the explained variance versus the residual variance is calculated for each variable, providing the SR for that variable [14,15].

Once again, the CR was used to guide the variable selection process. However, the variable selection strategy employed was different. The selection strategy involved a combination of BE and FS. First, the 1000 top-ranked variables were selected and subjected to BE starting with the lowest-ranked variable among those (#1000) and ending with the top-ranked variable (#1). Subsequently, forward selection was performed starting with the highest-ranked unchecked variable (#1001) and ending when a set number of variables had been checked (3000 in this case). A schematic of the selection process is shown in Figure 2-8.

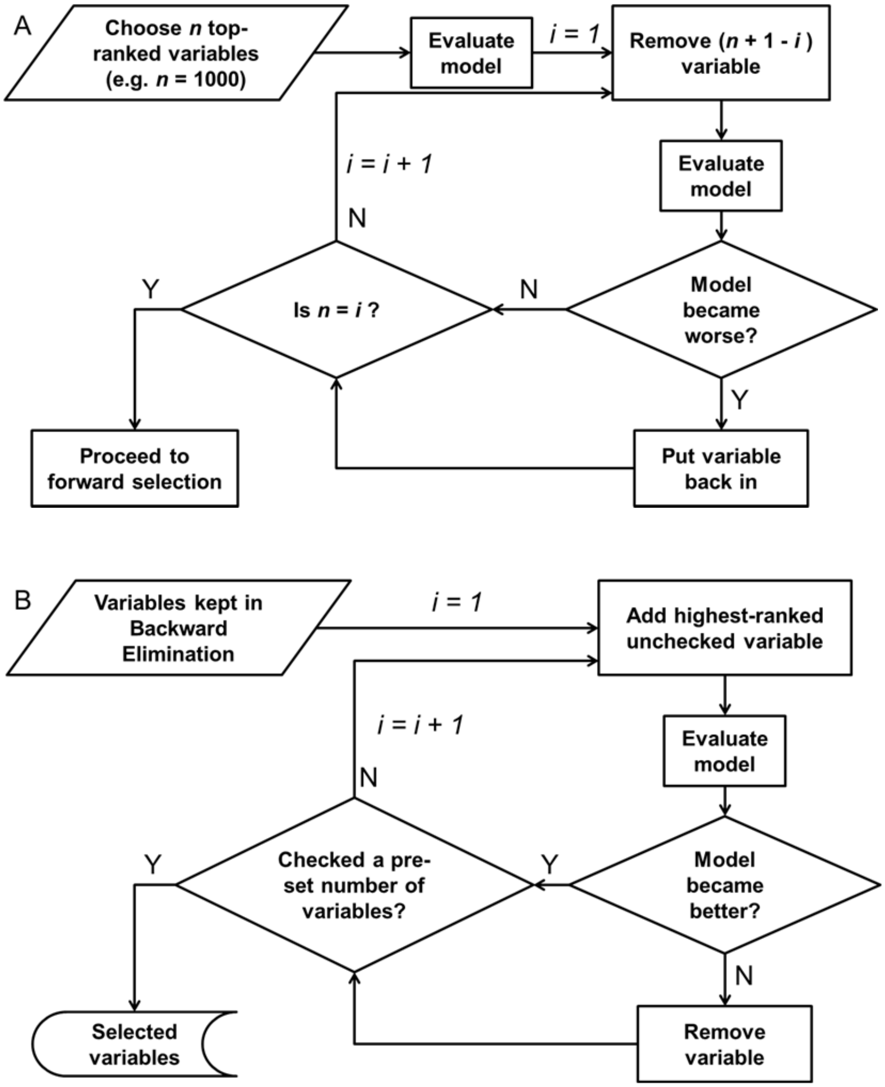

A
Choose *n* top-ranked variables (e.g. *n* = 1000)
Evaluate model
*i* = 1
Remove (*n* + 1 - *i* ) variable
Evaluate model
Model became worse?
N
Is *n* = *i* ?
*i* = *i* + 1
N
Y
Y
Proceed to forward selection
Put variable back in
B
Variables kept in Backward Elimination
*i* = 1
Add highest-ranked unchecked variable
Evaluate model
*i* = *i* + 1
N
Model became better?
Y
Checked a pre-set number of variables?
Y
N
Selected variables
Remove variable

**Figure 2-8:** Variable selection techniques used. **A** Backwards Elimination; **B** Forward Selection. CR metric used in the Evaluate Model step.

### 2.3.3 Experimental

Gasoline samples were obtained from two local gas stations in Edmonton, Alberta, Canada. The two stations belonged to different vendors and three different octane ratings of gasoline (87, 89 and 91) were obtained from each, providing a total of six classes. To introduce some challenge to the variable selection process, the datasets were made more complicated by introducing a higher degree of within-class variance. To this end, half of the samples within each class were weathered approximately 50% by volume using a gentle stream of clean, dry, compressed air. To introduce further in-class variance, some samples from each class were left uncontaminated, some were contaminated by adding either turpentine (~5% by volume), lacquer thinner (~5% by volume), kerosene (~5% by volume), or a mixture of turpentine, lacquer thinner and kerosene together (~5% by volume each). As a result, a total of 120 samples were prepared. Their compositions are shown in Figure 2-9.

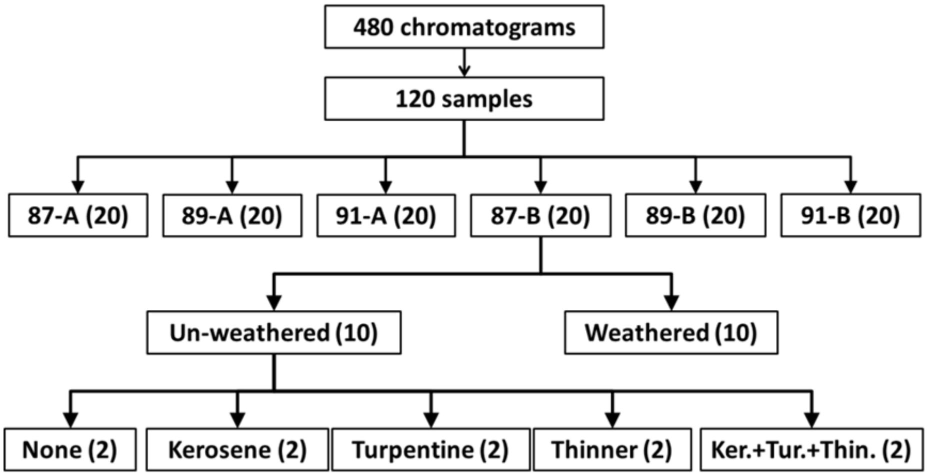


**Figure 2-9:** Schematic for sample preparation

The samples were then diluted 20:1 by volume in pentane and analyzed by GC-MS. The GC-MS used for these experiments was a 7890A GC with a 5975 quadrupole MS (Agilent Technologies, Mississauga, ON) equipped with a 30 m × 250 µm; 0.25 µm HP-5 column (Agilent). The carrier gas used was helium at constant flow rate of

1.0 mL·min$^{-1}$. The injector was held constant at 250 °C and a volume of 0.2 μL was injected using a split ratio of 100:1. The temperature program was 50 °C (3.5 min hold) with a 20 °C·min$^{-1}$ ramp to 300 °C. The transfer line and source temperatures were 185 and 230 °C, respectively. The total run time was 16 min. The initial solvent delay was 2.5 min and mass spectra were collected from m/z 30 to m/z 300 at the rate of 9.2 spectra·s$^{-1}$.

Four chromatograms were collected from each sample, providing a total of 480 chromatograms (80 for each class). Chromatograms were then assigned to training, optimization and validation sets. From each class, 40 chromatograms were assigned to the training set, 20 were assigned to the optimization set and 20 were assigned to the validation set. The training set contained a total of 240 chromatograms, optimization set contained 120 chromatograms and validation set contained 120 chromatograms.

Chromatograms from the training set were used to create the alignment target, rank variables and to obtain the loading vectors for the PCA model during each variable selection step. Chromatograms from the optimization set were used together with chromatograms from the training set, to obtain scores during variable selection (the model being generated using training data and tested against optimization data). Both training and optimization data were combined to create the final PCA model after variable selection was complete. Chromatograms from the validation set were only used to evaluate the final model.

For each analysis, an entire chromatogram was exported as a .csv file, which was then imported into MATLAB 7.10.0.499 (The Mathworks, Natick, MA) as a 7300×271 (scan number × m/z ratio) matrix using a lab-written algorithm. Data were then handled using lab-written algorithms. Chemometric models were constructed using PLS toolbox 5.8 (Eigenvector Research Inc., Wenatchee, WA). The calculations were performed on a Intel Core i5 750 2.76 GHz processor with 8 GB of RAM and 64-bit Microsoft Windows 7 Professional operating system. Chromatographic alignment was based upon the piecewise alignment algorithm developed by Johnson et al. [25] with an additional mass spectral confirmation to match features. First, an alignment target was created. The preliminary target was constructed by first randomly choosing a single chromatogram from the training set. Then, a second, randomly chosen chromatogram from the training set was aligned with the target chromatogram, after which the aligned chromatogram was added to the preliminary target. Then, the aligned chromatogram was discarded and the algorithm proceeded to the next chromatogram in the training set. After all chromatograms from the training set had been included in the target, the algorithm proceeded with alignment of all chromatograms in all sets to the composite target chromatogram.

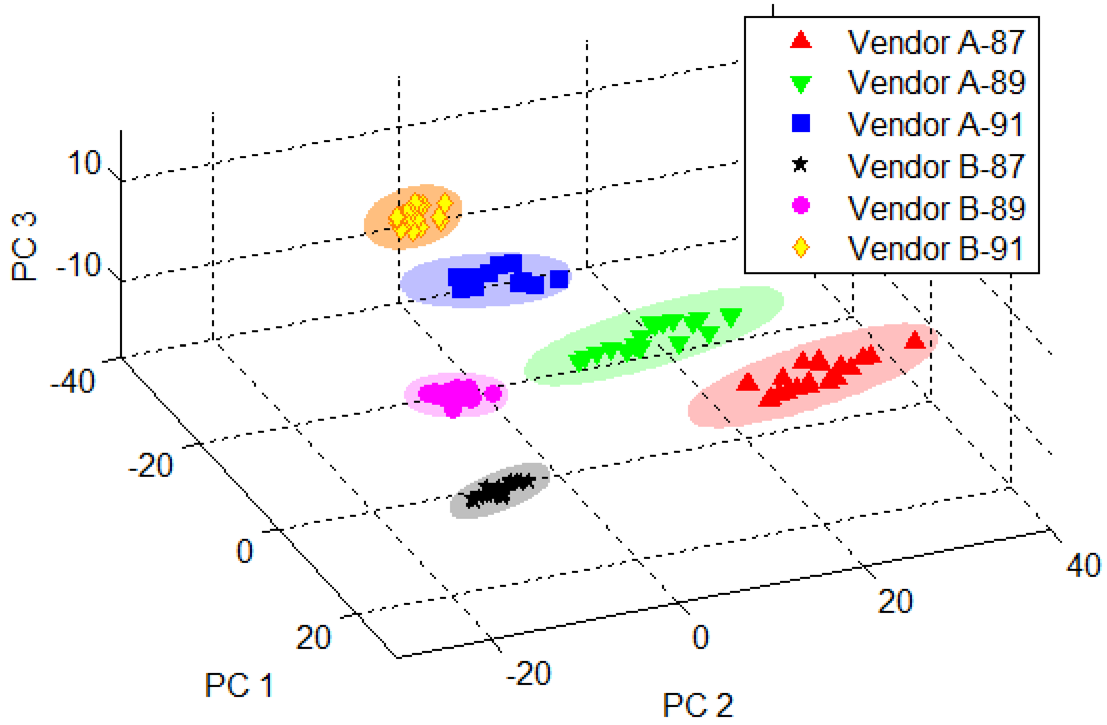
### 2.3.4 Results and discussion

CR was used to guide a combined BE/FS variable selection process with the goal of constructing a three-component PCA model with the greatest possible separation observed between clusters for each class. In this example, the data consisted of 80 GC-MS chromatograms for each of six types of gasoline (three octane ratings from each of two vendors), to the total of 480 chromatograms. The data were split into training, optimization and validation sets as described in Section 2.3.3, after which chromatographic alignment was performed.

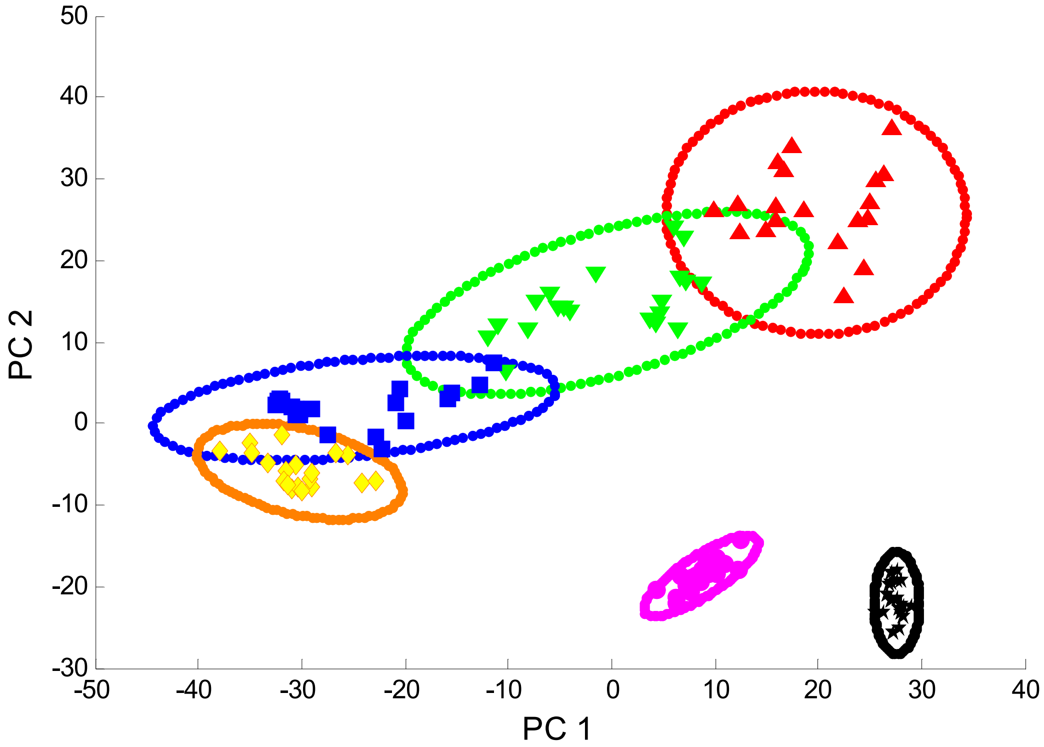
The aligned matrices were then unfolded along the time axis to yield a series of vectors. Each vector consisted of ~$2\times10^6$ variables. SR variable ranking was applied to the set of 240 chromatograms in the training set using a lab-written algorithm. This yielded a vector of selectivity ratios that was used to rank the features. The optimization and test data sets were aligned as well, but were not used during the calculation of selectivity ratios. Baseline correction was also not necessary as the variable ranking process automatically down-weights background ions which did not vary significantly from sample to sample. Computation of the SR ranking vector from the aligned data required approximately one minute.

After variable ranking, the 1000 top-ranked variables were selected and a three-component PCA model was created using the training set and applied to the optimization set. Prior to construction of the model, each vector containing selected variables was normalized to an area of 1. Using the scores from both the training and the optimization sets on the first 3 PCs, six clusters of points with 60 points per cluster were obtained. Overall model quality was calculated as the product of the individually determined CR measurements for each of the 15 possible of pairings. BE was performed on the 1000 top-ranked variables (Figure 2-8 A). The variables retained after BE were then passed to FS where variables ranked 1001 through 3000 were considered for inclusion (Figure 2-8 B). Variable selection took approximately 36 h to complete and selected a total of 644 variables from the 3000 variables checked.

A:

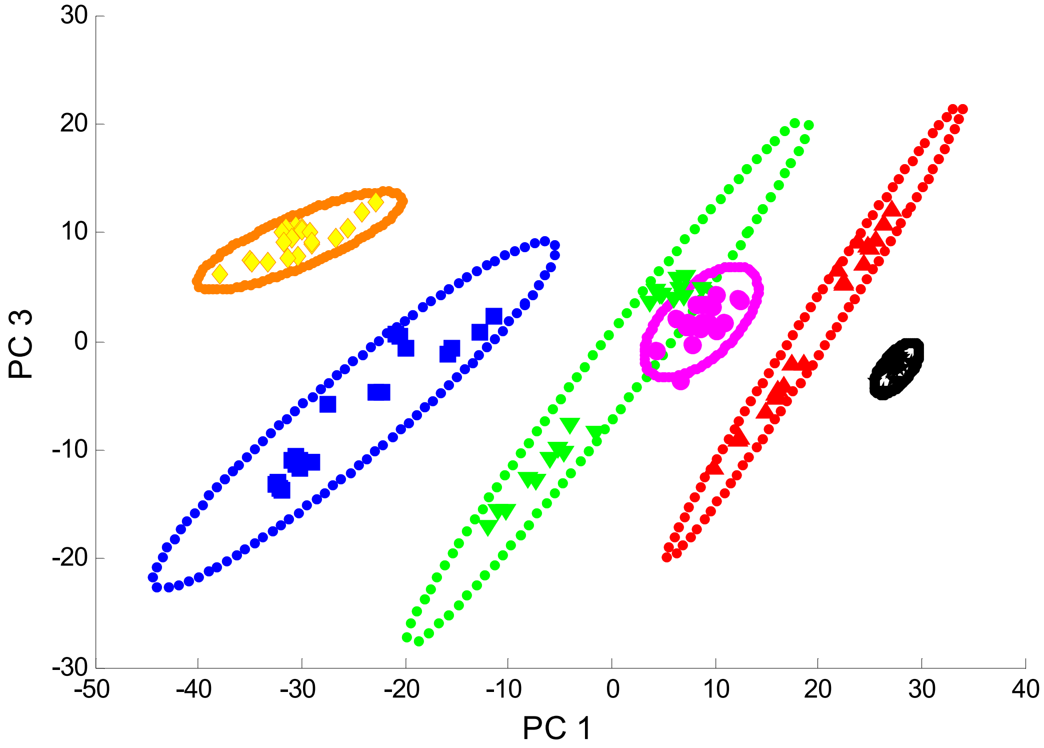


B:



50
40
30
20
10
0
-10
-20
-30
PC 2
-50
-40
-30
-20
-10
0
10
20
30
40
PC 1

C:



30
20
10
0
-10
-20
-30
PC 3
-50
-40
-30
-20
-10
0
10
20
30
40
PC 1

D:

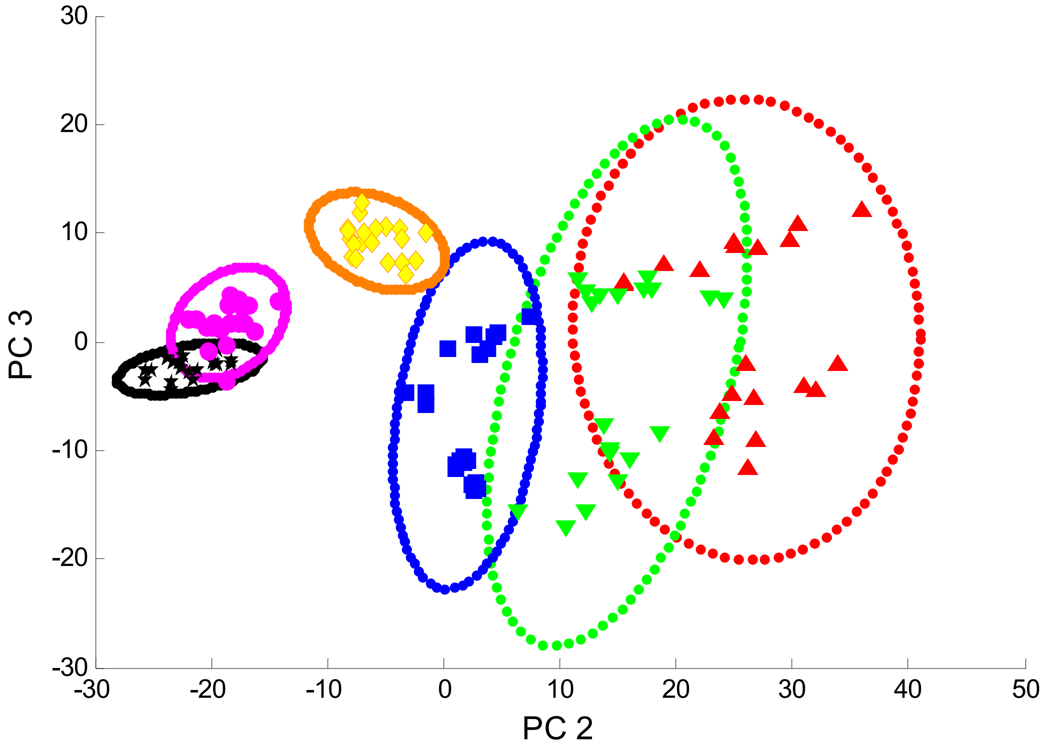


**Figure 2-10:** Scores plots for the final PCA model. Red, green, and blue represent 87-, 89-, and 91-octane gasoline from Vendor A, respectively. Black, magenta, and orange represent 87-, 89-, and 91-octane gasoline from Vendor B, respectively. Shaded regions represent three-dimensional confidence ellipsoids (98%) described around clusters of training set samples (individual points not shown), solid markers represent individual points for validation set samples. B, C, D dotted lines represent two-dimensional confidence ellipses at 98% confidence limit described around clusters of training set samples (individual points not shown) and markers represent individual points for validation set samples on projection of three-dimensional model onto components 1 and 2, 1 and 3, 2 and 3 respectively.

The training and optimization sets (360 chromatograms) were then combined to train the final three-component PCA model using normalization to an area of 1 and autoscaling as the only pre-processing methods. Subsequent projection of the validation set (120 as yet unused chromatograms) permitted evaluation of the final model. Figure 2-10 A depicts the resulting model where lightly shaded regions are three-dimensional 98% confidence ellipsoids described around clusters of training set samples (individual points are not shown) and markers represent individual points for validation set samples. Colours and shapes of the markers represent different classes. As can be seen from the figure, validation set samples projected into the same regions as the training set samples, and all classes were separated in the three-dimensional space. Overall, the final measured three-dimensional CR for this problem was calculated as 0.9997.

The three-dimensional CR was then compared with the previously developed two-dimensional CR metric. The same training, optimization and validation sets were used and, just as in the three-dimensional case, backwards elimination started with the 1000 top-ranked variables (as shown in Figure 2-8 A) and forward selection checked variables ranked 1001 through 3000 (as shown in Figure 2-8 B). To make the comparison fair, a three-component PCA model was constructed at each step and the two-dimensional CR metric was calculated for the three possible two-dimensional projections of the three principal components. For each pair of classes, the CR value retained to guide optimization was the highest value among the three projections (e.g. for discriminating the 91-octane gasolines from Vendor A and Vendor B (blue and orange classes) in Figure 2-8 B-D, the CR score on PC1 vs. PC3 was retained). Figure 2-8 B-D represent two-dimensional projections of the optimized model, showing 98% confidence ellipses based on the training set samples (individual points not shown) and individual markers for validation set samples. As can be seen, no individual two-dimensional projection was able to separate all classes of all samples. However, using the three projections together, separation was achieved using a total of 1009 variables. The final two-dimensional CR, calculated based on the best projection for each pairing, was 0.9985. When the three-dimensional CR was calculated for a model based on the 1009 variables selected using two-dimensional projections, it was found to be 0.9991, an improvement, but still not as high as the value obtained when 3-D ellipsoids were used during the variable selection. It should also be noted that there is a significant difference in sizes between sizes of ellipsoids as 99.91 % and 99.97 % confidence limit values. The advantage of the two-dimensional approach was in the computational time: 12 hours against 36 hours for the 3-D projection. This is due to the fact that the construction of two-dimensional ellipses is significantly faster than the construction of three-dimensional ellipsoids.

After variables have been selected, they can be traced back to the original data and tentative identities of the selected compounds can be determined. Here, mass spectral information was used in combination with linear retention indices [26] for the purpose of compound identification. Figure 2-11 depicts a binary mask where the black dots represent the 644 variables selected when using the three-dimensional CR-guided approach and white space represents excluded variables. As a comparison, three

89-octane samples are presented in Figures 2-12 and 2-13, with Figure 2-12 depicting a region of the raw GC-MS chromatograms and Figure 2-13 depicting the abundances of the selected variables in each chromatogram (Figure 2-13 masked by Figure 2-11).

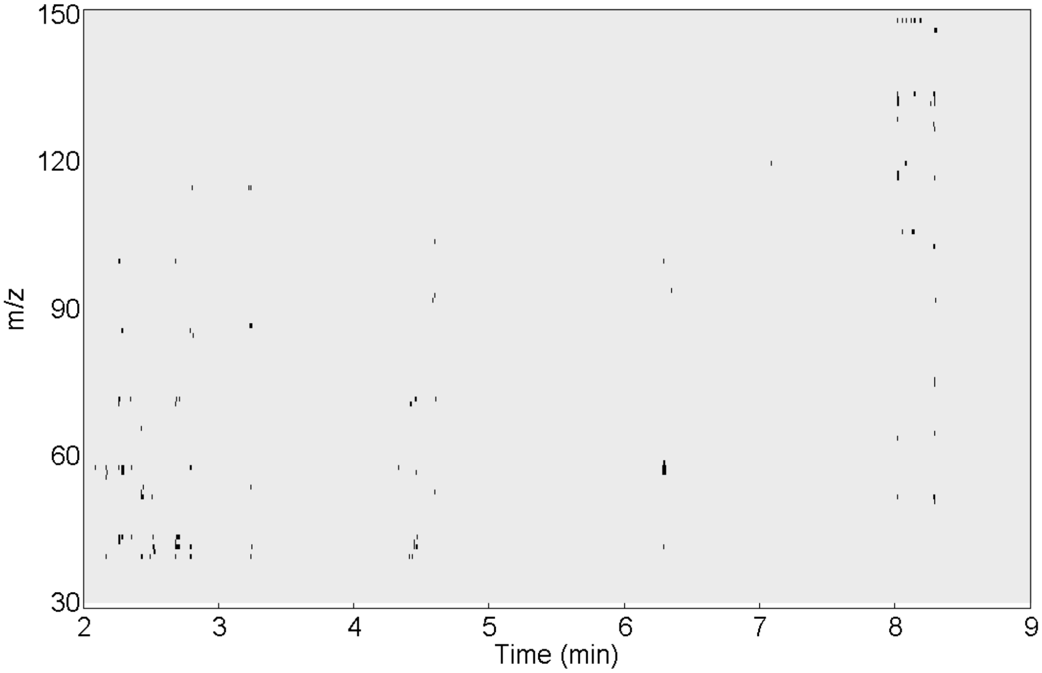
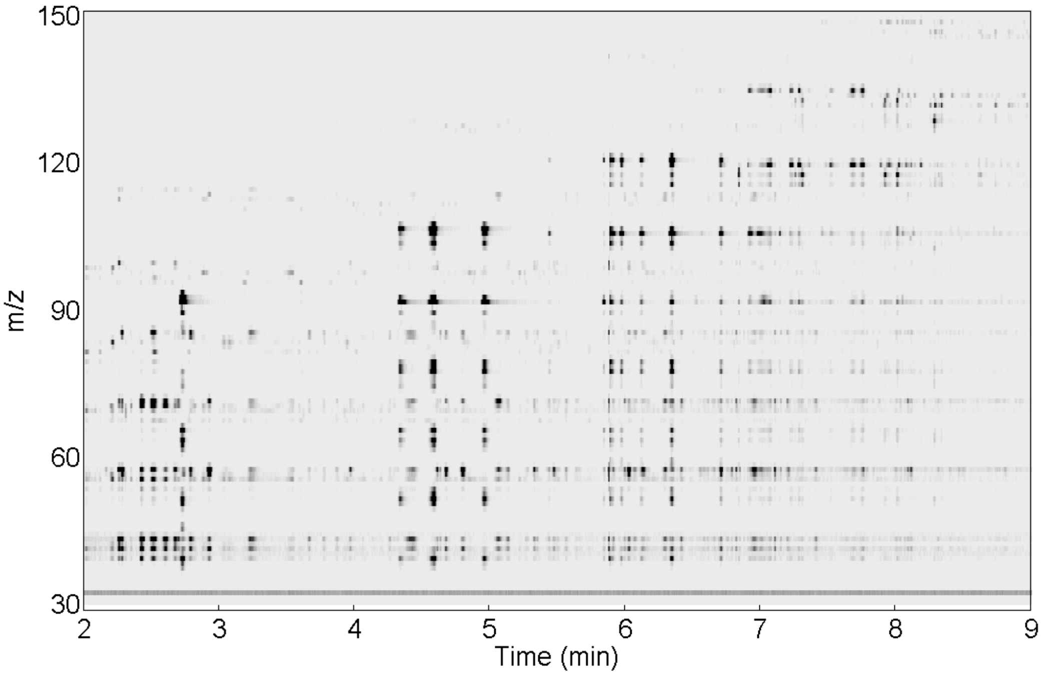


**Figure 2-11:** Variables selected from the original data by the algorithm. Black dots represent variables that were selected after BE/FS.

A


150
120
90
60
30
m/z
2
3
4
5
6
7
8
9
Time (min)

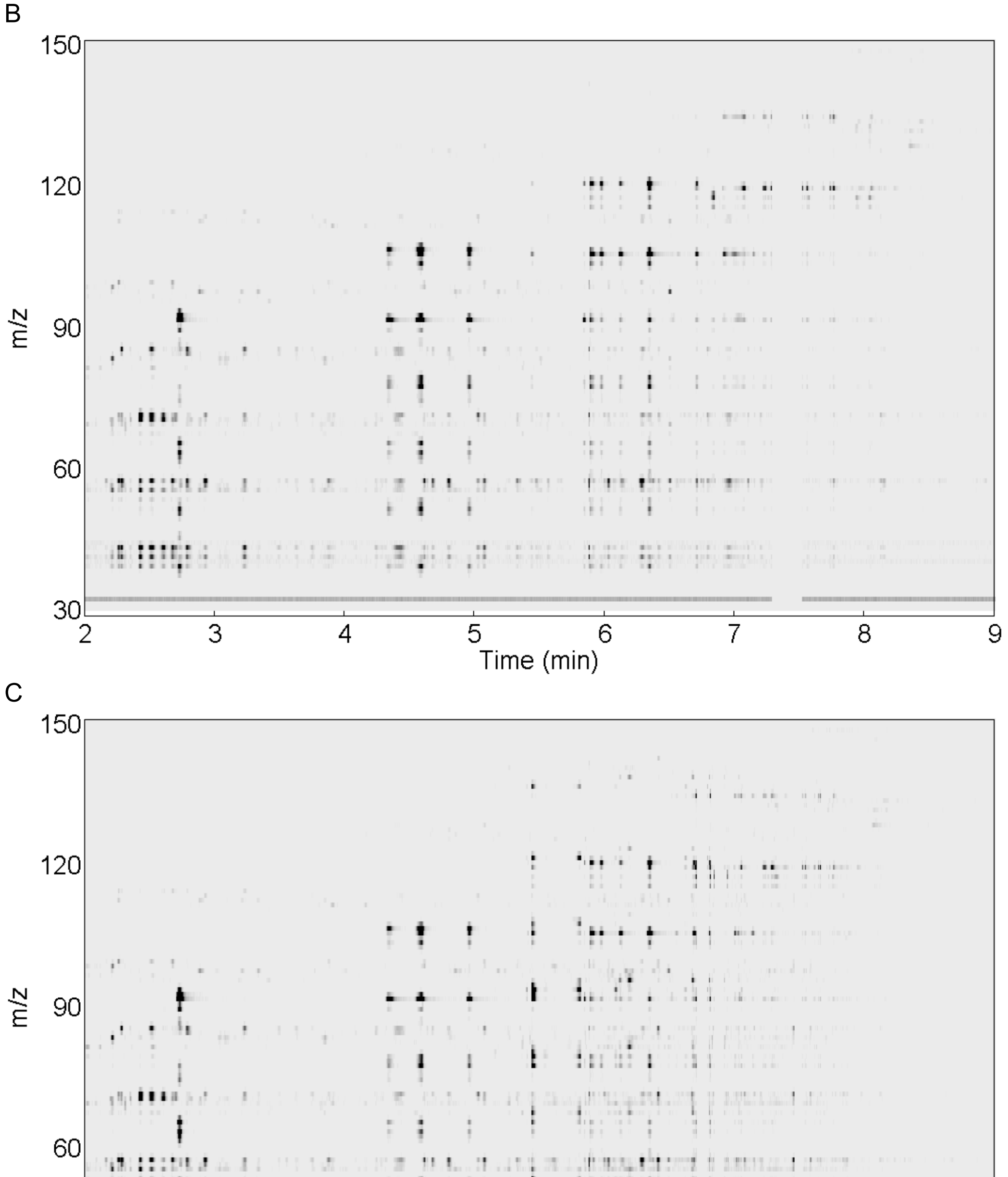
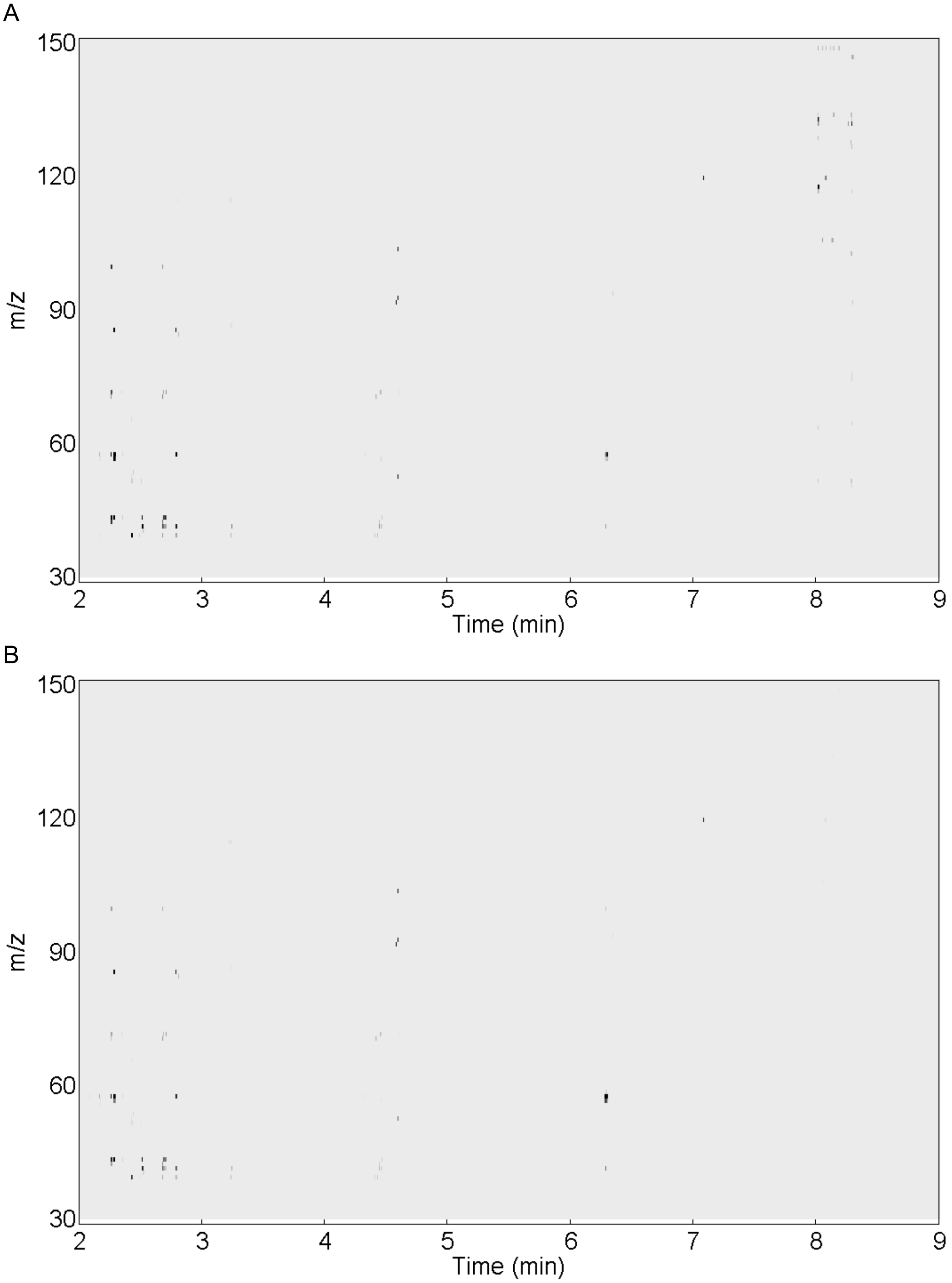


**Figure 2-12:** GC-MS chromatograms of selected gasoline samples after alignment was performed. Light grey indicates low signal while dark grey indicates high signal for a variable. **A)** Vendor A 89-octane weathered uncontaminated gasoline. **B)** Vendor B 89-octane weathered uncontaminated gasoline. **C)** Vendor B 89-octane unweathered gasoline contaminated with kerosene, turpentine and lacquer thinner.


A
150
120
90
60
30
m/z
2
3
4
5
6
7
8
9
Time (min)
B
150
120
90
60
30
m/z
2
3
4
5
6
7
8
9
Time (min)

C

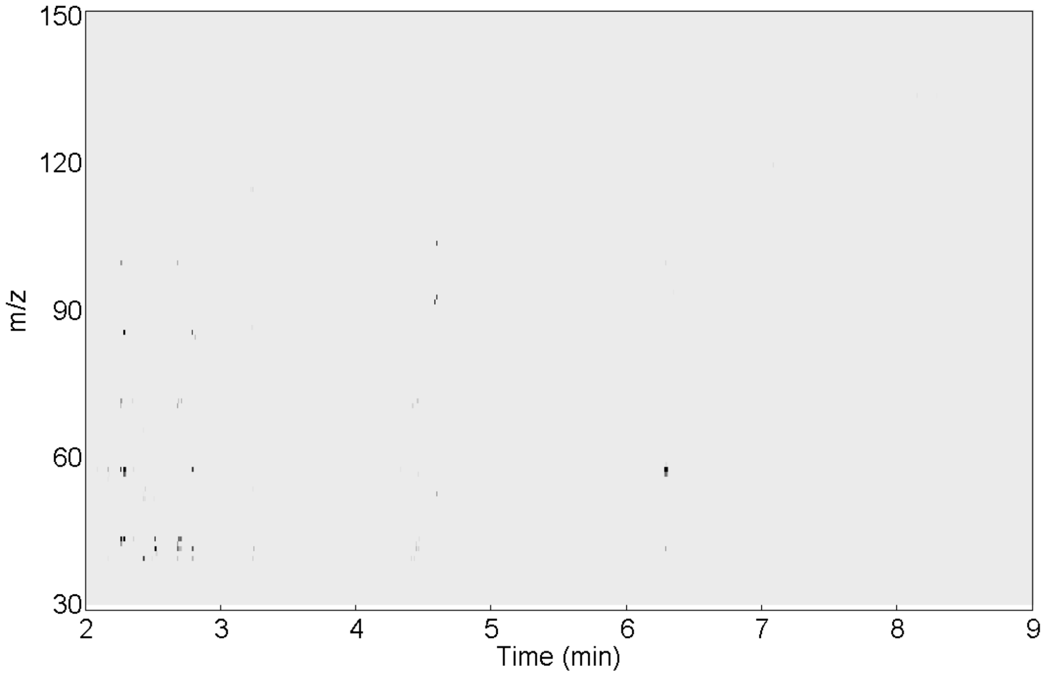


**Figure 2-13:** Variables selected for GC-MS chromatograms of selected gasoline samples after alignment was performed. Light grey indicates low signal while dark grey indicates high signal for a variable. Variables that were not selected were assigned value of zero. **A)** Vendor A 89-octane weathered uncontaminated gasoline. **B)** Vendor B 89-octane weathered uncontaminated gasoline. **C)** Vendor B 89-octane unweathered gasoline contaminated with kerosene, turpentine and lacquer thinner.

As seen from the figures, the gasoline samples from the two vendors can be distinguished on the basis of several compounds. First of all, Vendor A (Figure 2-13 A) has a relatively high abundance of C5 alkylbenzenes (eluting between 8 and 9 min) whereas Vendor B (Figure 2-13 B and 2-13 C) does not have an appreciable amount of these compounds. Additionally, Vendor A has a slightly increased abundance of a peak at 4.4 min, and a much lower abundance of a compound at 6.3 min. These two compounds have been tentatively identified as 4-methyl octane and 2,2,4,6,6-pentamethylheptane on the basis of their mass spectra and their linear retention indices (Table 2-1).

**Table 2-2:** Information used for identification of selected compounds. Calculated linear retention index was obtained from the data, literature linear retention index was obtained from reference 26, forward and reverse mass spectral matches were obtained by comparing obtained mass spectrum with NIST 2005 database spectrum for a given compound.

| Name | 4-methyl octane | 2,2,4,6,6-pentamethylheptane |
|---|---|---|
| Calculated linear retention index | 868 | 992 |
| Literature linear retention index | 864 | 997 |
| Forward mass spectral match | 850 | 876 |
| Reverse mass spectral match | 889 | 879 |

Considering two samples belonging to the same class but containing added within-class variance, we see that there is significant difference between unweathered contaminated Vendor B 89-octane, and weathered, uncontaminated gasoline from the same class (Figures 2-12 B and C, respectively). However, considering only the masked data, the two samples (which belong to the same class: Vendor B, 89-octane) are essentially identical (Figures 2-13 B and C). Upon closer inspection of Figure 2-11, it can be seen that between 2 and 3 minutes, some signals due to several alkanes were selected. This is consistent with the fact that light alkanes are present at very different levels in gasolines of different types. Toluene is commonly present in gasoline and, differences in toluene abundance have been previously shown to be useful in discriminating between different classes of gasoline (Section 2.2). Under the conditions of this experiment, toluene eluted at approximately 2.7 min, and is shown to be an abundant compound in gasoline. However, as seen in Figures 2-11 and 2-13, toluene was not selected as a useful variable. This is due to the high concentration of toluene in the lacquer thinner that was added as a contaminant. Thus, in this data set, toluene contributes significantly to within-class variation, decreasing its utility for the task of class discrimination. The feature selection algorithm was able to discover this automatically and correctly discard toluene from the model.

A final point for this discussion is computation time. It took approximately 36 hours to perform variable selection, with cluster resolution being the slowest step. However, it should be noted that with $n$ classes there are $\frac{n^2}{2} - n$ possible pairs of classes; here 15 pairs were considered. Thus it only required slightly more than 2 h of computation time for each pair of classes. Since the calculation of CR for each pair of classes is independent from the calculation for each other pair, this step could be distributed and calculated in parallel using multiple processors, greatly speeding up computation time. It should also be noted that variable selection was completely automated. Once the class assignments were made to the data files, the remainder of the process concluded with no user intervention or attention required.

## 2.4 Conclusions

CR has been shown to be an effective metric for guiding the variable selection process in well-controlled data sets. Both the two- and three-dimensional versions of the metric were used to optimize models for the classification of gasoline samples on the basis of GC-MS data. The metric was successful in the simultaneous optimization of feature selection on a dataset containing six classes and a significant degree of in-class variance. The metric is a generic “goodness measure” for classification models and should be applied in the future to data from other types of instruments and in the optimization of other types of classification problems. The end goal of this specific research project is to develop tools for determining the presence of gasoline in arson debris, which will be the subject of the following chapters.

## 2.5 References

[1] R.G. Brereton, Chemometrics Data Analysis for the Laboratory and Chemical Plant, Wiley, Toronto, 2003.

[2] K. J. Johnson, R. E. Synovec, *Chemom. Intell. Lab. Syst.* 60 (2002) 225-237.

[3] Weldegergis B. T.; Crouch A. M. *J. Agric. Food Chem.* 56 (2008) 10225-10236.

[4] R.B. Gaines, G.J. Hall, G.S. Frysinger, W.R. Gronlund, K.L. Juaire, *Environ. Forens.* 7 (2006) 77–87.

[5] J.H. Christensen, G. Tomasi *J. Chromatogr. A,* 1169 (2007) 1–22.

[6] C.R. Borges, *Anal. Chem.* 79 (2007) 4805–4813.

[7] L.J. Marshall, J.W. McIlroy, V.L. McGuffin, R. Waddell Smith *Anal. Bioanal. Chem.* 394 (2009) 2049–2059.

[8] M.D. Krebs, R.D. Tingley, J.E. Zeskind, M.E. Holomboe, J. Kang, C.E. Davis, *Chemom. Intell. Lab. Syst.* 81 (2006) 74–81.

[9] K.J. Johnson, R.E. Synovec, *Chemom. Intell. Lab. Syst.* 60 (2002) 225–237.

[10] N.E. Watson, M.M. VanWingerden, K.M. Pierce, B.W. Wright, R.E. Synovec, *J. Chromatogr. A* 1129 (2006) 111–118.

[11] K.M. Pierce, J.K. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, *J. Chromatogr. A* 1096 (2005) 101–110.

[12] N.A. Sinkov, J.J. Harynuk, *Talanta* 83 (2011) 1079-1087.

[13] N.A. Sinkov, B.M. Johnston, P.M.L. Sandercock, J.J. Harynuk, *Anal. Chim. Acta* 697 (2011) 8-15.

[14] T. Rajalahti,R. Arneberg, F. S. Berven, K. M. Myhr, R. J. Ulvik, O. M. Kvalheim, *Chemom. Intell. Lab. Syst.* 95 (2009) 35-48.

[15] T. Rajalahti,R. Arneberg, A. C. Kroksveen, M. Berle, K. M. Myhr, O. M. Kvalheim, *Anal. Chem.* 81 (2009) 2581-2590.

[16] N.A. Sinkov, J.J. Harynuk, *Talanta* 103 (2013) 252-259.

[17] R. F. Teofilo, J. P. A. Martins, M. M. C. Ferreira, *J. Chemometrics* 23 (2009) 32-48.

[18] I. Guyon, A.J. Elisseeff, *Mac. Learn. Res.* 3 (2003) 1157-1182.

[19] R.R. Hocking, *Biometrics* 32 (1976) 1-49.

[20] D.E. Axelson, Data Preprocessing for Chemometric and Metabonomic Analysis, 1st ed., MRi Consulting, Kingston, Ontario, 2010.

[21] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, *Metabolomics* 4 (2008) 81-89.

[22] J.H. Christensen, A.B. Hansen, U. Karlson, J. Mortensen, O. Andersen, *J. Chromatogr. A* 1090 (2005) 133-145.

[23] Srivastava M. S.; Khartri C. G.; *An Introduction to Multivariate Statistics*., Elsevier North Holland, Inc., New York, 1979.

[24] Almkvist G.; Berndt B. Gauss, *Amer. Math. Monthly,* 95 (1988) 585–608.

[25] Johnson K. J.; Wright B. W.; Jarman K. H.; Synovec R. E. *J Chromatogr. A* 996 (2003) 141-155.

[26] X. Xu, L.L.P. van Stee, J. Williams, J. Beens, M. Adahchour, R.J.J. Vreuls,U.A.T. Brinkman, J. Lelieveld, *Atmos. Chem. Phys.* 3 (2003) 665-682.

# 3 Chapter Three: Chromatographic alignment based on a deuterated alkane ladder[3]

Prior to chemometric interpretation of raw chromatographic data, the time axes must be precisely aligned so that the signal from each analyte is registered in the same column of the data matrix for each and every analysed sample [1,2]. A variety of alignment approaches exists in the literature and those approaches work well when the samples to be aligned have somewhat similar chemical compositions [3]. In cases where the samples and/or background matrix are highly variable, chromatographic alignment is more challenging. Presented here is an alignment approach that relies on a series of deuterated alkanes which act as retention anchors for alignment. This approach was then coupled with an automated feature selection routine based on the CR metric (Chapter 2) and applied to the identification of gasoline in a series of simulated arson debris samples analyzed by passive headspace extraction and GC-MS. Classification was performed based on partial least squares discriminant analysis (PLS-DA).

## 3.1 Chromatographic alignment

As discussed is Chapter 1, many factors can cause misalignment of chromatographic signals. The extent of misalignment depends largely on the stability of the system over the period of time for which data were

[3] This chapter based on N.A. Sinkov, B.M. Johnston, P.M.L. Sandercock, J.J. Harynuk, *Anal. Chim. Acta* 697 (2011) 8-15; highlighted in SeparationsNOW (May 2, 2011)

collected. In the case of GC, chromatographic peaks can be very narrow (2-3 s base width). In such cases, without correction a 3 s misalignment will result in a peak due to the same analyte in two chromatograms being considered as two completely different analytes. A number of alignment techniques have been proposed.

### 3.1.1 Alignment techniques

Chromatographic alignment relies on comparing a chromatogram with a target, which can be a chromatogram in the data set [3], or it can be a composite target containing information from most chromatograms in the data set [4,5,6]. After the comparison, the chromatogram's time axis is warped such that it is aligned with the target. Alignment approaches include piecewise alignment [4], which relies on identification of common peaks in the chromatogram and the alignment target. This approach has been adapted to include information from the MS dimension of the data when a MS detector is used [5,6]. Other approaches include correlation-optimized warping (COW) [1,7], dynamic time warping (DTW) [7], as well as many others [8,9,10,11]. These alignment techniques perform well when the chemical make-up of the samples remain similar across the series of chromatograms (such as when comparing fuel samples [4,5,6] or fungal cultures [1]). However, when the chemical profiles of different sample classes (or even samples within a given class) are highly dissimilar, these methods will often yield a poor alignment. When the background matrix of the samples is also highly variable, alignment is even more challenging since the alignment algorithm may become unable to lock onto features that must be aligned. A good example of such a system is arson debris, where the debris matrix and the observed chromatogram will depend on materials at the scene and conditions of the fire itself. To solve this problem, I developed an approach that relies on the addition of a deuterated alkane ladder.

### 3.1.2 Simulated debris

Reliable and automated chemometric analysis of arson debris is an inherently difficult task, which to date has not been reported in the literature for real data, and applications to realistic simulated debris are scarce [12]. The goal of the arson investigation is to determine whether the fire was intentional or accidental. Arson is one of the more difficult crimes to investigate because much of the evidence available at the scene is destroyed by fire and firefighting measures [13,14,15]. An important aspect of the investigation is the determination of the presence of an ignitable liquid (IL) in the fire debris. In most cases involving arson, a petroleum-based IL (most often gasoline) [16,17] is used because these are inexpensive, readily available, and highly effective [18]. Some common ILs, such as gasoline and kerosene, are complex mixtures containing hundreds of individual components. Additionally, the fire debris matrix is highly variable and complex, containing numerous combustion and pyrolysis products that may interfere with the analysis [14,15]. To identify traces of an IL in fire debris, techniques based upon the concentration of headspace vapours [19], often passive headspace sampling using activated carbon strips [20] is used. Extracted analytes are then separated and analyzed using GC-MS.

### 3.1.3 Classification tools

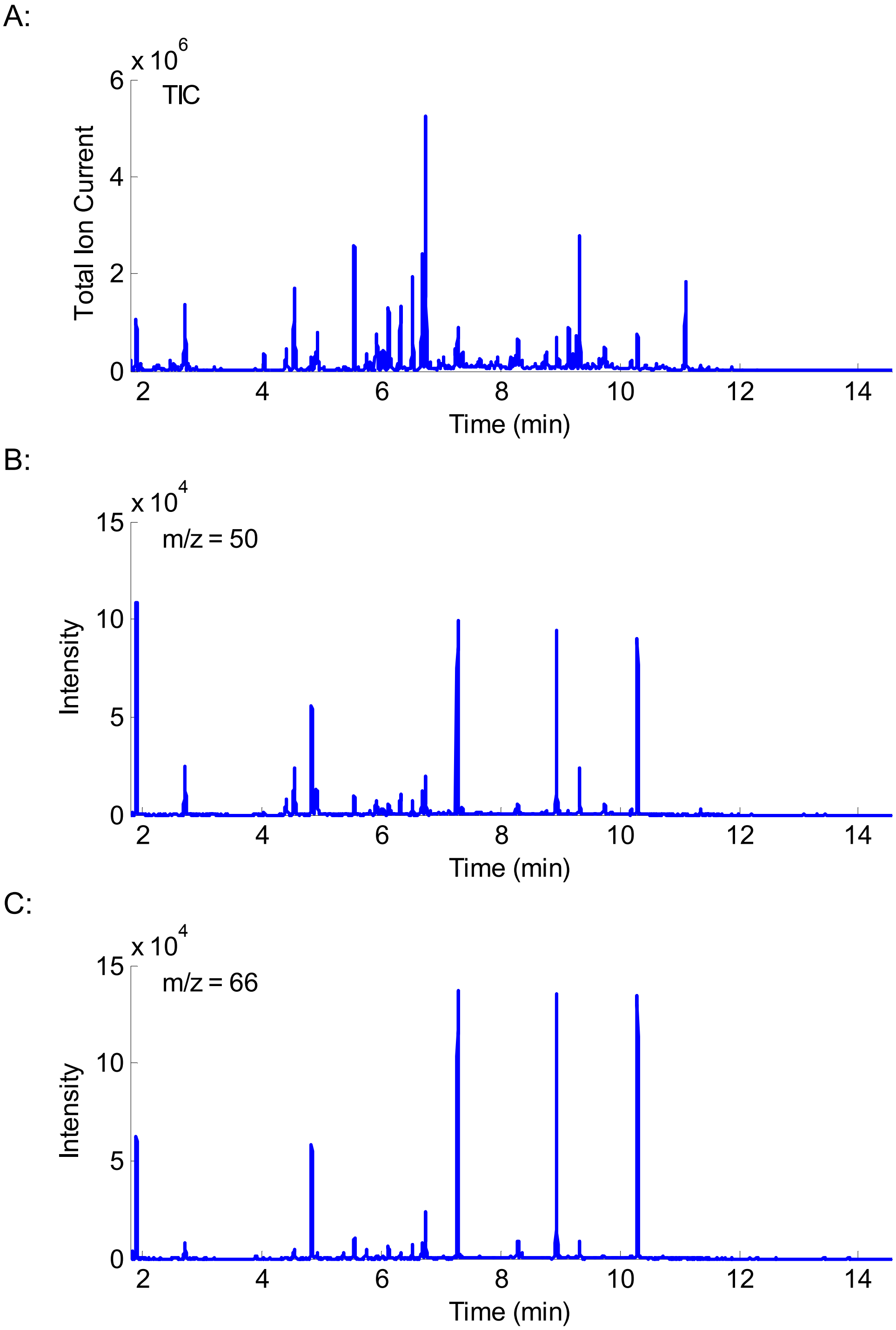
In Chapter 2, CR was used to optimize PCA models for data visualization. In the case of fire debris analysis, the goal is classification, which PCA alone does not provide [21]. Thus PLS-DA [22], a classification technique, was employed instead. It should be noted that despite efforts of variable ranking techniques to identify relevant variables, the risk of including irrelevant variables is present. The smaller the number of samples relative to the number of variables, as is the case with raw GC-MS data, the more severe the risk [23]. This is especially true if a classification technique such as PLS-DA is used where over-fitting of the training data is a real possibility. To ensure that a model has not been over-fit, validation is always necessary [24].

## 3.2 Application of deuterated alkane ladder alignment and CR-guided feature selection to the identification of gasoline in simulated fire debris

Deuterated alkane ladder alignment relies on the addition of small amounts of deuterated n-alkane standards to the solvent used for sample extraction/preparation prior to analysis by GC-MS. Deuterated alkanes provide a unique mass spectral signature that is unlike any other compound reasonably expected to be present in an arson sample. As a result, the position of deuterated n-alkanes can be reliably determined even in cases of severe co-elution with abundant matrix components. Therefore, deuterated alkanes can serve as “anchors” for alignment, even in cases of extremely dissimilar samples. Deuterated ladder-based alignment was compared to the piecewise alignment discussed in Chapter 2.

### 3.2.1 Alignment using deuterated n-alkane ladder

Chromatographic alignment requires a target to which chromatograms will be aligned. For deuterated alkane ladder-based alignment, the target peaks are present in all chromatograms. Thus, construction of a composite target is not necessary. Therefore, any chromatogram in a set (typically the first) may be chosen as the target. The target chromatogram is loaded and a vector containing the product of the responses for specific m/z ratios (deuterated alkane fragments) at each scan is calculated. The result is a vector with the alkane ladder peaks being the most intense signals, as shown in Figure 3-1.


A:
x 10^6
TIC
Total Ion Current
6
4
2
0
2
4
6
8
10
12
14
Time (min)
B:
x 10^4
m/z = 50
Intensity
15
10
5
0
2
4
6
8
10
12
14
Time (min)
C:
x 10^4
m/z = 66
Intensity
15
10
5
0
2
4
6
8
10
12
14
Time (min)

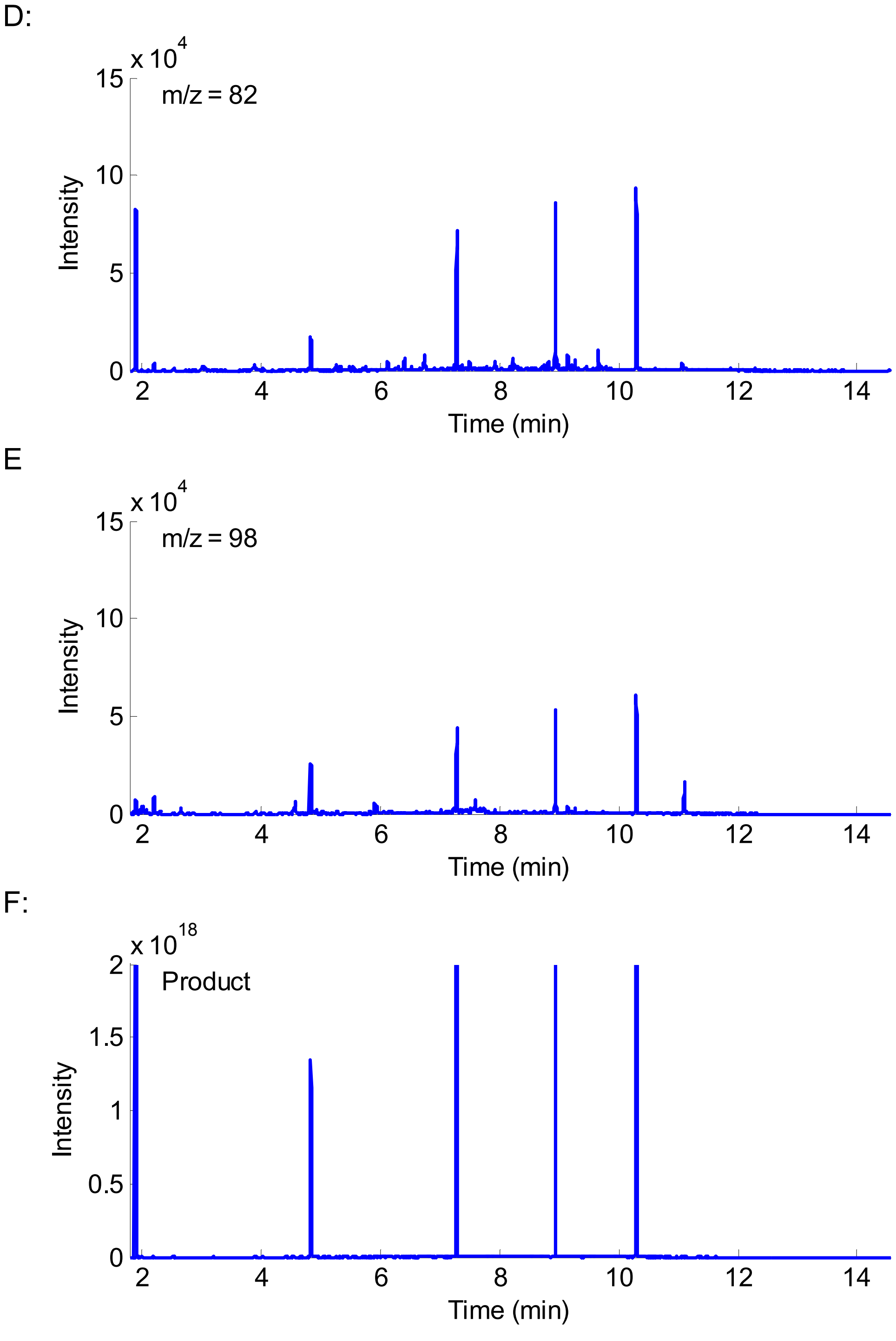


**Figure 3-1:** A chromatogram containing deuterated alkane ladder. **A** shows total ion current chromatogram, **B**, **C**, **D** and **E** show extracted ion chromatograms for ions common to the deuterated alkane ladder. **F** shows the product of intensities at every time point in figures **B**, **C**, **D** and **E**, calculated by multiplying intensities at each m/z channel for each time point in the chromatogram.

As shown in Figure 3-1 A, the location of the deuterated alkane ladder peaks cannot be determined from the TIC. EICs shown in Figure 3-1 B-E show the alkane ladder signal, but every individual EIC also contains signals from other components in the sample. However, when the product of the characteristic ion abundances is taken (Figure 3-1 F), the alkane ladder signal is easily distinguished from the rest of the sample.

After the product plots are obtained, locations of all deuterated alkanes are determined on both the target and sample chromatograms. In the next step, the sample chromatogram is aligned to the target by warping the signal between the deuterated alkane ladder components, providing alignment of a chromatogram between retention times of the ladder compounds. The alignment algorithm then proceeds to the next chromatogram, aligning it to the same target. This approach provides rapid and effective alignment of all chromatograms in the data set regardless of relative difference between chromatograms.

### 3.2.2 Experimental

Nine different gasoline samples were obtained from three local gas stations belonging to different vendors. These samples represented a range of octane ratings (87, 89, and 91) from each vendor. Samples of three varieties of perfume were also collected locally to be used as negative controls (assigned to non-gasoline class). Samples of fresh lumber (pine), painted scrap lumber, plywood, carpet, fabric (50:50 cotton polyester blend and pure cotton), glossy magazine pages, linoleum, vinyl siding, asphalt shingles, tar paper, Tyvek® building wrap, and polyethylene vapour barrier were all obtained locally and used in the generation of simulated debris. Carbon disulfide (Omnisolv; VWR, Mississauga, ON) was used as the solvent for extraction of analytes from activated carbon strips (8 mm × 20 mm; Albrayco Technologies, Cromwell, CT). The solvent was spiked with a deuterated alkane ladder consisting of n-heptane (d16), n-nonane (d20), n-undecane (d24), n-tridecane (d28), and n-pentadecane (d32) (CDN Isotopes, Pointe-Claire, QC) at concentrations of 1.3 $\mu L \cdot L^{-1}$ each.

Samples of weathered gasoline were prepared from aliquots of each gasoline sample that were evaporated to levels of 50, 75 and 90% by weight at room temperature using a jet of clean, dry compressed air. Unweathered samples of gasoline were also used. Weathering for each step was stopped when the flask reached its target weight for the evaporation. A Pasteur pipet containing a 3-cm bed of granular activated carbon (Fisher Scientific, Nepean, ON) held in place by plugs of glass wool was used as a nozzle for the air jet. The activated carbon prevented contamination of samples with any organics that could have been present in the air stream. The weathering procedure was repeated for each of the nine gasoline samples at the three levels of evaporation, to generate a total of 36 different IL samples, both weathered and unweathered.

Seven different combinations of materials for the debris generation, shown in Table 3-1, were placed into aluminum roasting pans and set on fire inside a fire-resistant fume hood using a propane torch to avoid leaving any undesired IL signature. Samples were allowed to burn until the contents were well-charred, at which point the fire was extinguished by suffocation and the debris was allowed to cool. After the debris had cooled, the samples were placed in PTFE-lined 1-gallon paint cans (General Paint, Edmonton, AB), and stored for later use.

**Table 3-1:** Compositions of arson debris. HDPE - high density polyethylene. MDF - medium density fibreboard

| Debris | Contents |
|---|---|
| 1 | Wood, carpet, cotton fabric, 50:50 cotton:polyester fabric, glossy magazine |
| 2 | Wood, polyethylene film, old linoleum, carpet |
| 3 | Wood, siding, shingles, tar paper, Tyvek™ |
| 4 | Wood |
| 5 | Wood, polyethylene film, newpaper |
| 6 | Wood, HDPE food container, new linoleum, MDF |
| 7 | Wood, carpet, cotton fabric, 50:50 cotton:polyester fabric, glossy magazine, linoleum, siding, shingle, tarpaper, Tyvek™ |

For passive sampling, a debris sample was placed in a 1 L mason jar and 1 µL of gasoline or perfume was spiked directly onto the debris, depending on whether the sample was designed to contain gasoline or be a negative control. For debris blanks, the debris was placed into the jar without the addition of gasoline or perfume. Jars were then capped with an activated carbon strip suspended from a safety pin on the inside of the jar lid as shown in Figure 3-2. Safety pins were held in place by magnets placed on the outside of the lid. The jars were placed in an oven at 60 °C for 16 h to equilibrate. For quality control, one empty jar containing nothing but a suspended activated carbon strip was included in each batch of jars placed into the oven. After the equilibration time, the carbon strips were removed and each coiled into a 1.8 mL GC vial (Chromatographic Specialties, Brockville, ON). 1.0 mL of $CS_2$ containing the deuterated alkane ladder was then added to each vial to extract the analytes.

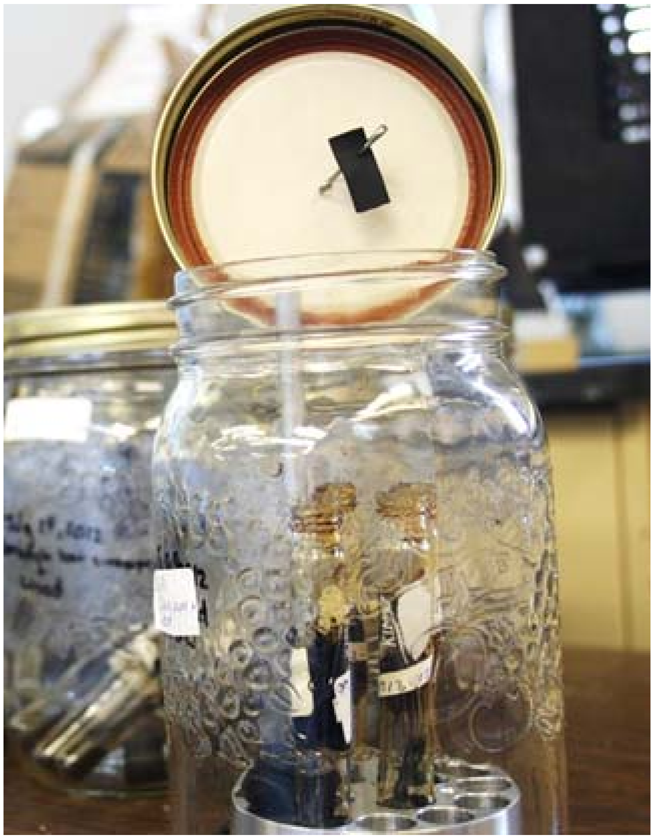


**Figure 3-2:** Photo of a setup for activated carbon strip extraction of volatile components from simulated arson debris.

Samples were analyzed using an Agilent Technologies 7890A gas chromatograph (GC) with a 5975 quadrupole mass spectrometer (MS) and a 7683 auto sampler (Agilent Technologies, Mississauga, ON). Data acquisition and automation were accomplished using MS ChemStation (Agilent). A 30 m × 250 μm; 0.25 μm HP-5 column (Agilent) was used for the separation. The oven program used was 50 °C (held for 3.5 min) followed by a ramp to 280 °C at a rate of 20 °C·min$^{-1}$. Samples were injected in split mode into an injector held at 250 °C. The injection volume was 1 μL, with a split ratio of 20:1. The transfer line and source temperatures were 185 and 230 °C, respectively. Mass spectral searching was performed against the 2005 edition of the NIST MS Database (NIST, Gaithersburg, MD).

Chromatograms obtained for 204 samples containing gasoline and 84 containing either no gasoline or containing perfume (negative controls) were exported from Chemstation as .csv text files and then imported into MATLAB 7.10.0 (The Mathworks, Natick, MA). Retention time alignment algorithms and chemometric analysis algorithms were performed using lab-built routines in MATLAB, using some chemometric analysis functions from the PLS Toolbox 5.2 (Eigenvector Research Inc, Wenatchee, WA). The calculations were performed on a Intel Core i5 750 2.76 GHz processor with 8 GB of RAM and 64-bit Microsoft Windows 7 Professional operating system.

### 3.2.3 Results and discussion

Two alignment routines were compared in this study. The first alignment technique was based on a piecewise alignment algorithm where the features to be aligned were automatically identified based on peak apexes [4], with an additional mass spectral comparison performed before peak matches were assigned [5,6]. Application of this method requires a composite target chromatogram to which all chromatograms are aligned. Application of this method using a single IL-containing chromatogram as the alignment target did not yield a data set from which a usable model could be generated (data not shown). The composite chromatogram was constructed by first selecting one chromatogram at random. A second chromatogram was then aligned to the initial target using the piecewise alignment algorithm. These two aligned chromatograms were summed, forming a new target chromatogram. A third chromatogram was aligned to this composite using the same alignment algorithm and then added into the composite target. This process was repeated until all 288 chromatograms were included in the target, at which point the intensities of all points on the composite chromatogram were divided by 288 to obtain an average value. The aligned chromatograms were discarded at this stage since the composite target was still being constructed. As the alignment routine does not depend on the absolute intensity of a peak for matching, skewing the abundances of peaks at this step does not affect the final result. Since all chromatograms must be treated the same way, all 288 chromatograms were, once again, aligned with the composite target, except this time target was not modified and the aligned chromatograms were saved. The resulting 288 aligned chromatograms were used further in this study. The deuterated alkane ladder alignment routine was used as described in Section 3.2.1.

For model construction and testing, the data set was separated into a training set containing 240 chromatograms and a test set containing 48 chromatograms. The training set was used for feature selection and model construction while the test set was used only for model validation. Assignment of chromatograms to either set was random, though the exact same sets were used for both alignment methods. The process of assigning the training and test sets and constructing a model was repeated ten times. It should be noted that during the piecewise alignment algorithm, chromatograms from the validation set were mistakenly used during construction of the composite alignment target. We believe that any effect that this could have had on the results would have improved the performance of the piecewise alignment algorithm, and therefore would not have resulted in the deuterated ladder-based alignment gaining an unfair advantage.

Each chromatogram contained 7 300 scans with m/z values from 30 to 300, providing approximately $2×10^6$ individual variables per chromatogram, where a variable is defined by abundance of a given ion in a given scan.  Each GC-MS chromatogram was then unfolded along the retention time axis, providing a single vector of approximately $2×10^6$ variables. ANOVA-based feature ranking [25] was subsequently performed (details provided in Section 2.2.2). After the variables were ranked, variable selection was performed using an automated forward-selection approach. The endpoint based on maximizing the cluster resolution (CR) [5] observed between the gasoline-containing samples and those that did not contain any gasoline on the PLS-DA scores plot. In the interest of time, a step size of 10 was used during variable selection, resulting in the 10 top-ranked unused variables being added to the model in every variable selection step. At the time that this research was conducted, the hybrid BE/FS approach had not yet been implemented.

PLS-DA models with two LVs were constructed for the ten data sets using both alignment methods. Two LVs were chosen because the two-dimensional CR metric was employed (three-dimensional CR had not yet been developed), and two LVs should, in principle, be sufficient for classifying samples between two groups. The endpoint chosen for the feature selection was a minimum number of top-ranked variables which would yield a CR value of 0.9999, or in cases where this degree of resolution could not be reached, the maximum achievable resolution. The results are summarized in Table 3-2.

This alignment algorithm is contrasted by the models generated with the piecewise alignment algorithm. Several of those sets are plagued by false positives, false negatives, and generally provide more ambiguous results. A careful inspection of the chromatograms after alignment revealed that the alignment based on the deuterated alkane ladder was often superior to the alignment performed by the peak matching algorithm, which exhibited numerous mismatched peaks. A likely reason for these mismatches was due to the variability in the matrix background.

**Table 3-2:** Results of cross-validation of PLS-DA models. Minimum number of variables required to reach a cluster resolution of 99.99%, or the maximum achievable resolution in cases where 99.99% was unachievable using the training. False positive indicates non-gasoline-containing sample being classified as a gasoline-containing sample. False negative indicates a gasoline-containing sample classified as non-gasoline containing sample. Minimum probability for true positive indicates the lowest probability value calculated for gasoline-containig sample to be assigned to gasoline-containing class. Maximum probability for true negative indicates the highest probability value calculated for non-gasoline-containig sample to be assigned to gasoline-containing class

| **Deuterated alkane ladder-based alignment** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Set number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Number of variables selected | 320 | 320 | 300 | 40 | 30 | 260 | 220 | 270 | 220 | 240 |
| False positive | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| False negative | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Minimum probability for true positive | 1 | .9998 | 1 | .9988 | 1 | .9999 | .9976 | .9991 | .9998 | .9966 |
| Maximum probability for true negative | .0001 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | | | | | | | | |
| **Peak matching-based alignment** | | | | | | | | | | |
| Set number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Number of variables selected | 460 | 570 | 400 | 390 | 200 | 400 | 430 | 520 | 520 | 500 |
| False positive | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| False negative | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| Minimum probability for true positive | .6592 | .5782 | .9199 | .7636 | 0 | .8852 | .7468 | .9968 | .8404 | .7509 |
| Maximum probability for true negative | 1 | 1 | 1 | .0355 | 1 | .1737 | .1866 | .7709 | .0759 | .0674 |

Conversely, the deuterated alkane ladder provided a series of unambiguous anchors for alignment, regardless of matrix present in a given sample. With the ladder, the alignment of individual peaks was not as precise; however, there were no mismatched peaks.

Consequently the overall alignment of the chromatograms by the ladder approach was superior. It should also be noted that ladder-based alignment can be followed by an additional alignment step (such as feature-based or COW alignment) with a small window size if fine-tuning is needed.

A further comparison of the results for the two alignment approaches is presented in Figures 3-3 through 3-5 which depict the scores, y-predicted value plots, and probability plots, respectively, for sets 5 and 10. Sets 5 and 10 were chosen since set 5 shows best performance for the deuterated ladder-based alignment against the piecewise alignment while set 10 shows best results for the piecewise alignment relative to the deuterated ladder-based alignment. For both of these sets, the ladder-based alignment delivers less ambiguous results.

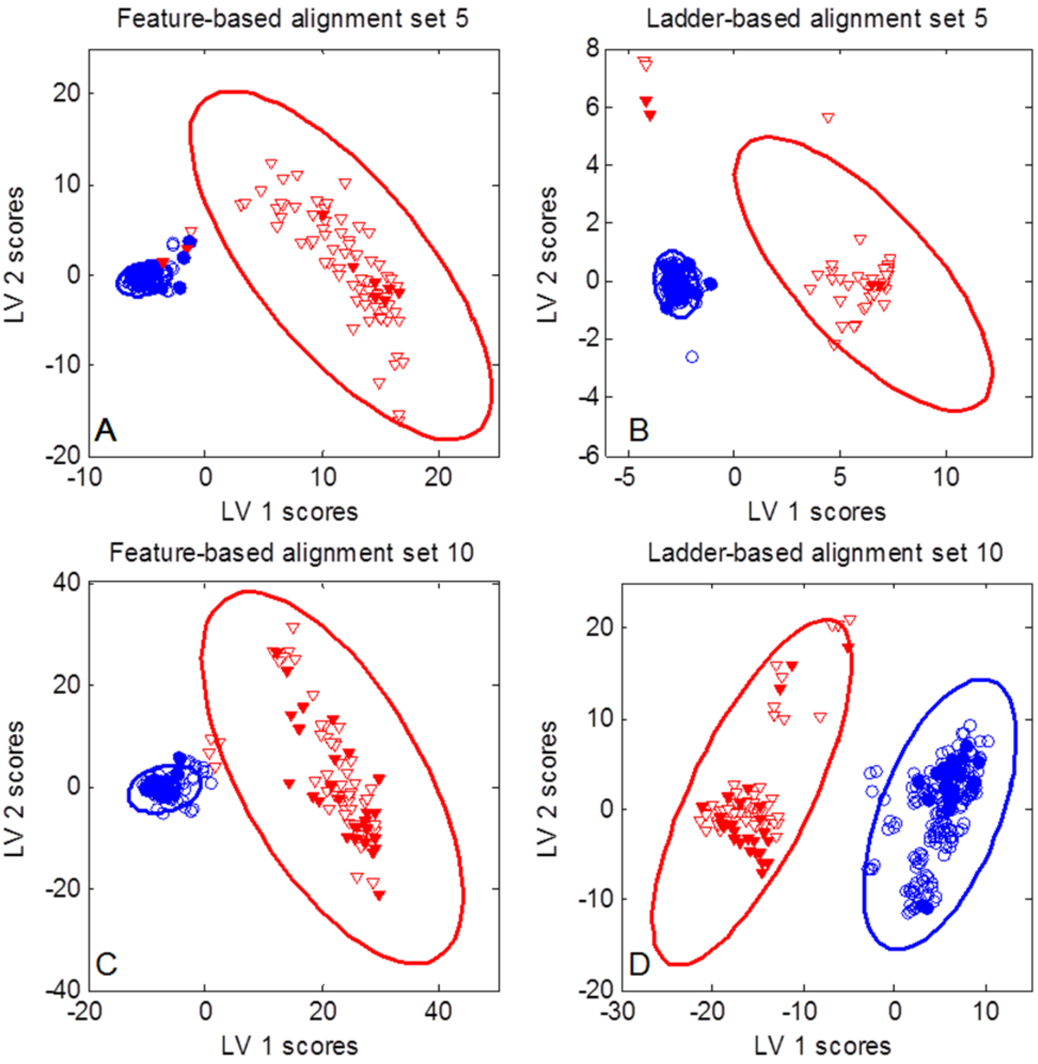


**Figure 3-3:** Score plots for PLS-DA models for the worst classification (Set 5) and one of the best (Set 10) when considering feature-based alignment and ladder-basedalignment. Samples containing an gasoline are indicated by blue circles; samples with no gasoline are indicated by red triangles. Samples used in the training set are indicated by hollow markers while samples used in the test set are indicated by filled markers. **A** Feature-based alignment, Set 5 **B** Ladder-based alignment, Set 5 **C** Feature-based alignment, Set 10 **D** Ladder-based alignment, Set 10.

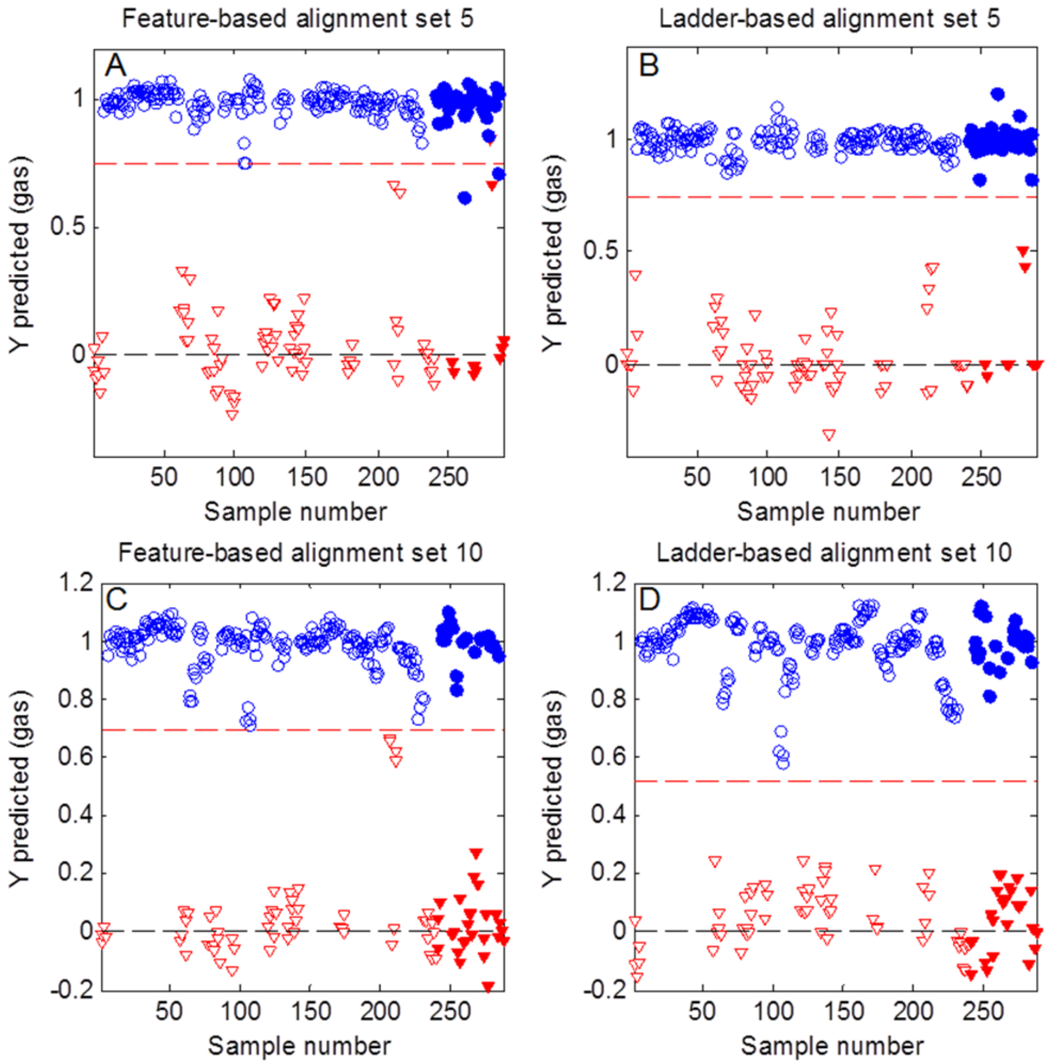


**Figure 3-4:** Predicted Y-value plots for PLS-DA models for Set 5 and Set 10 using both feature-based and ladder-based alignment. Samples containing an gasoline are indicated by blue circles; samples with no gasoline are indicated by red triangles. Samples used in the training set are indicated by hollow markers while samples used in the test set are indicated by filled markers. **A** Feature-based alignment, Set 5 **B** Ladder-based alignment, Set 5 **C** Feature-based alignment, Set 10 **D** Ladder-based alignment, Set 10.

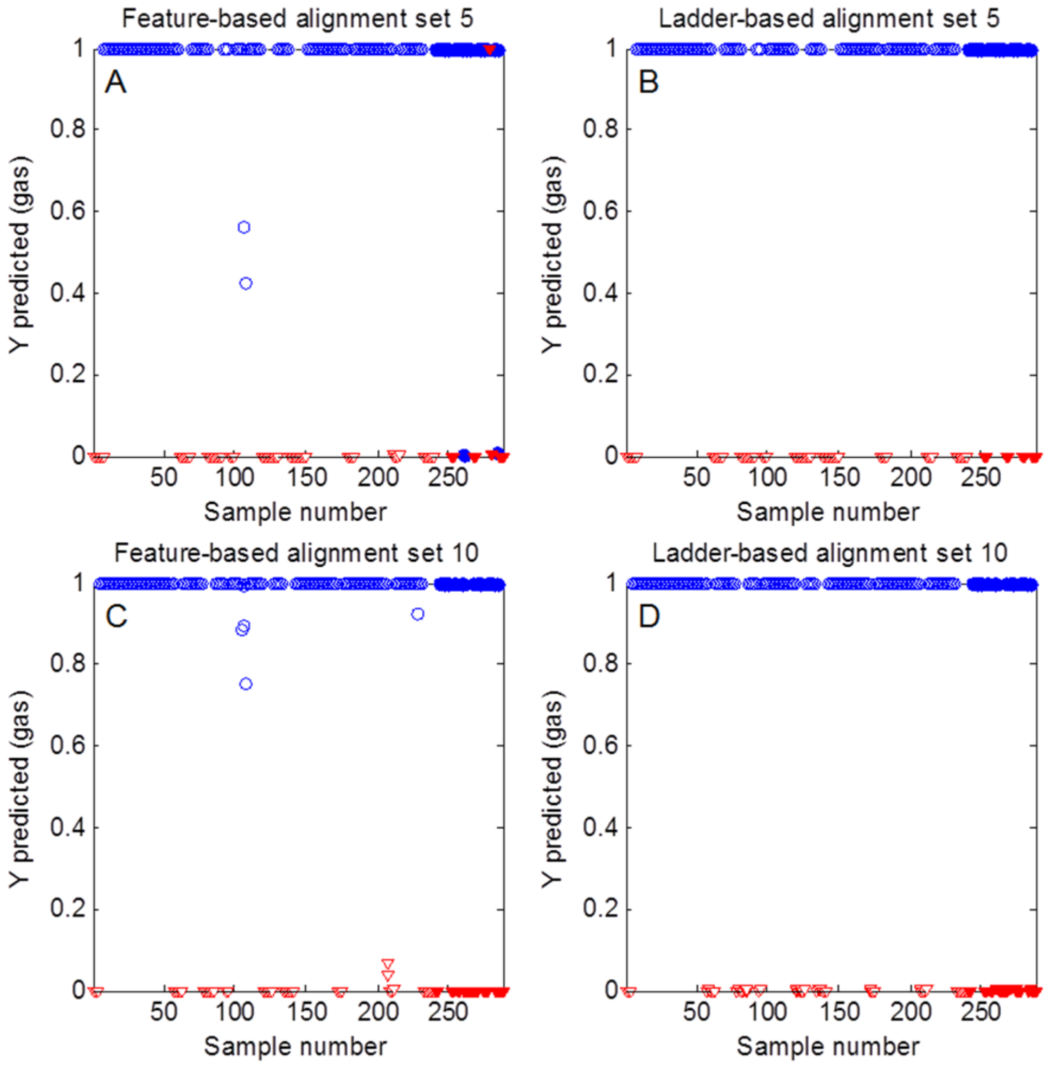


**Figure 3-5:** Predicted probability plots for identifying gasoline in samples for PLS-DA models for Set 5 and Set 10 using both feature-based and ladder-based alignment. Samples containing an gasoline are indicated by blue circles; samples with no gasoline are indicated by red triangles. Samples used in the training set are indicated by hollow markers while samples used in the test set are indicated by filled markers. **A** Feature-based alignment, Set 5 **B** Ladder-based alignment, Set 5 **C** Feature-based alignment, Set 10 **D** Ladder-based alignment, Set 10.

Considering the original size of the data (about 2×$10^6$ features) a relatively small number of variables (on average less than 300 or 0.015%) allow efficient discrimination between the two classes. To verify that the features selected by the algorithm were reasonable in a chemical sense, the selected features were traced backwards to identify the corresponding components in the gasoline-containing debris samples. Mass spectra of compounds corresponding to the identified features were then compared to the NIST MS Database and retention indices were estimated based on the deuterated alkane ladder signal and compared to literature values (Table 3-3).

**Table 3-3:** Tentative identification of compounds from which the features responsible for identifying gasoline are derived. *Literature values are for 6 °C·min$^{-1}$ taken from [26]

| Compound | Estimated Retention Index | Literature Retention Index* | Relevant Masses |
|---|---|---|---|
| Ethylbenzene | 870 | 857 | 51,77,91,106 |
| Para-xylene | 880 | 866 | 52,76,98,106 |
| Ortho-xylene | 908 | 891 | 105,106 |
| Propylbenzene | 980 | 952 | 62,89,91,120 |
| 1-ethyl-3-methyl benzene | 989 | 961 | 41,65,91,120 |
| 1,2,4-trimethyl benzene | 996 | 967 | 91,103,105,120 |
| 1,3,5-trimethyl benzene | 1022 | 992 | 41,89,116,119 |

It is worth noting that in GC analysis, deuterated alkane retention times will be slightly earlier than the corresponding non-deuterated alkanes. This will result in a slight over-estimate of the retention indices in this experiment. Additionally, it is well documented that retention indices are dependent on temperature programming rate, with the observed retention index increasing with temperature programming rate [26]. The literature values presented were collected at a temperature programming rate of 6 °C·min$^{-1}$ and the temperature programming rate used in our research was 20 °C·min$^{-1}$. Thus even though our estimates do not match exactly with the literature values, the positive variation of about 25-30 units is easily rationalized and the relative values for each compound are consistent. Furthermore, these compounds are consistent with the presence of gasoline in a sample. Some of the features included in the model are shown in Figure 3-6.

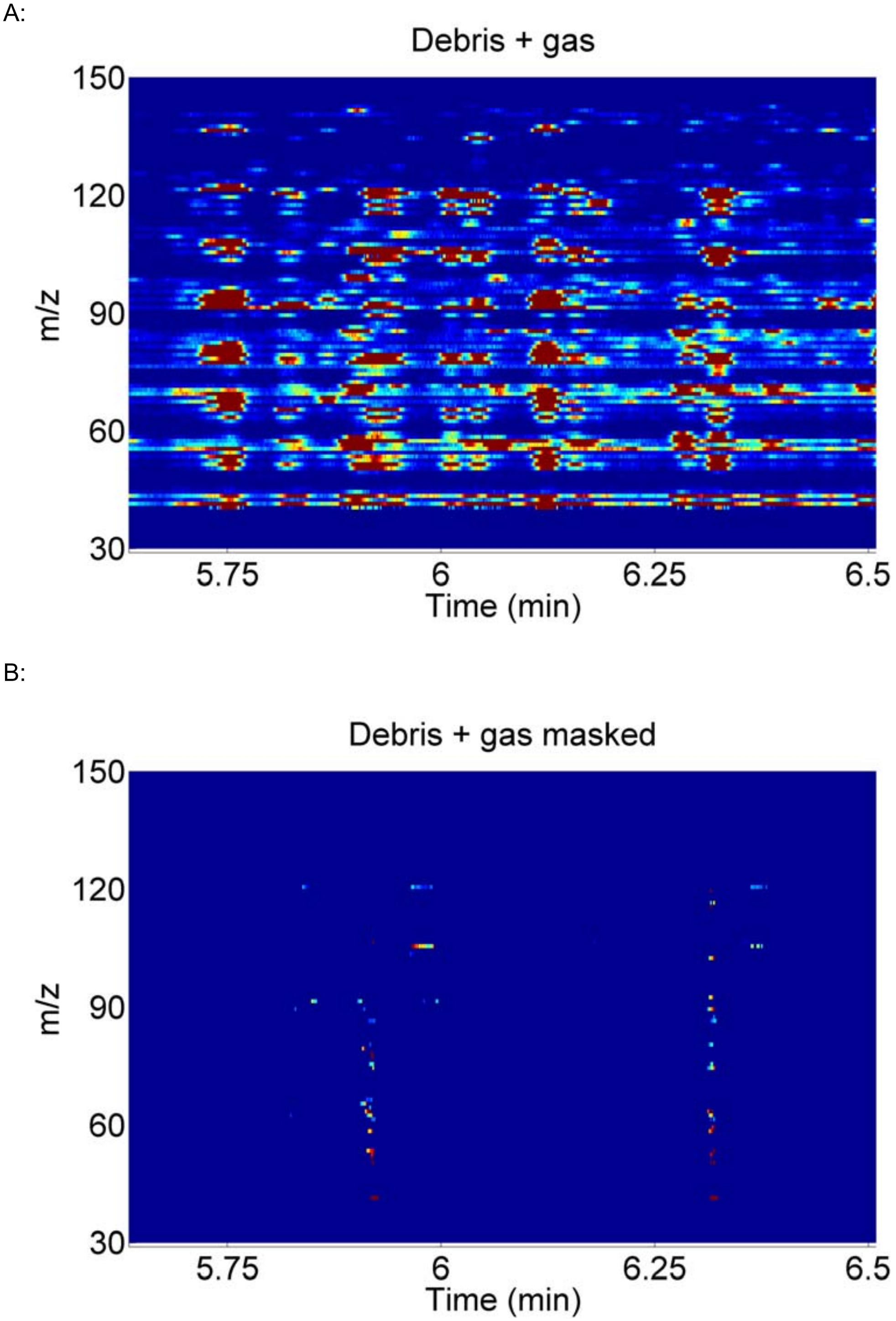


**Figure 3-6:** Region of GC-MS data containing features of interest, plotted as a contour plot. Blue colour indicates zero intensity. **A** Raw GC-MS chromatogram. **B** GC-MS chromatogram containing only selected variables.

## 3.3 Conclusions

Deuterated ladder-based alignment has been shown to effectively align chromatograms containing highly dissimilar features. Coupled with a CR-guided variable selection process, the deuterated alkane alignment allowed the automated optimization of a superior PLS-DA model for identifying gasoline in simulated arson debris than would have been possible using a peak-matching algorithm. However, the fire debris studied were created in a controlled environment. All samples were prepared by the same person and analyzed on the same instrument in a relatively short period of time. For real-world application, this approach must be able to provide similarly unambiguous results when applied to real casework samples. These samples will be much more variable. In addition to combustion products from a real fire and real matrix, artifacts will be introduced due to firefighting efforts, the individual officers and technicians preparing and analyzing the samples over the course of months, ideally using the same nominal methods on different instruments. This challenge will be the topic of Chapter 4.

## 3.4 References.

[1] N-P.V. Nielsen, J. M. Carstensen, J. Smedsgaard *J. Chromatogr. A* 805 (1998) 17-35.

[2] K.M. Åberg, E. Alm, R.J.O. Torgrip, *Anal. Bioanal. Chem.* 394 (2009) 151-162.

[3] N.A. Sinkov, B.M. Johnston, P.M.L. Sandercock, J.J. Harynuk, *Anal. Chim. Acta* 697 (2011) 8-15.

[4] K. J. Johnson, B. W. Wright, K. H. Jarman, R. E. Synovec, *J. Chromatogr. A* 996 (2003) 141-155.

[5] N.A. Sinkov, J.J. Harynuk, *Talanta* 83 (2011) 1079-1087.

[6] N.A. Sinkov, J.J. Harynuk, *Talanta* 103 (2013) 252-259.

[7] G. Tomasi, F. van den Berg, C. Andersson, *J. Chemometrics* 18 (2004) 231-241.

[8] P. H. C. Eilers, *Anal. Chem.* 76 (2004) 404-411.

[9] S. Toppo, A. Roveri, M. P. Vitale, M. Zaccarin, E. Serain, E. Apostolidis, M. Gion, M. Mariorino, F. Ursini, *Proteomics* 8 (2008) 250-253.

[10] A. M. Van Nederkassel, M. Dazykowski, P. H. C. Eilers, Y. Vander Heyden, *J. Chromatogr. A* 118 (2006) 199-210.

[11] W. Yao, X. Yin,Y. Hu, *J. Chromatogr. A* 1160 (2007) 254-262.

[12] M.R. Williams, M.E. Sigman, J. Lewis, K. M. Pitan, *Forensic Sci. Int*. 222 (2012) 373-386.

[13] P.M.L. Sandercock, *Forensic Sci. Int*. 176 (2008) 93-110.

[14] E. Stauffer, J. A. Dolan, R. Newman, F*ire Debris Analysis*, Elsevier, Inc. Amsterdam, 2008.

[15] E. S. Bodle, J.K. Hardy, *Anal. Chim. Acta* 589 (2007) 247-254.

[16] D. C. Mann, *J. Forensic Sci.* 32 (1987) 606-615.

[17] P. M. L. Sandercock, E. Du Pasquier, *Forensic Sci. Int.* 134 (2003) 1-10.

[18] B. Tan, J.K. Hardy, R.E. Snavely, *Anal. Chim. Acta* 422 (2000) 37-46.

[19] J. Dolan, *Anal. Bioanal. Chem.* 376 (2003) 1168-1171.

[20] K. Cavanagh, E. Du Pasquiera, C. Lennard, *Forensic Sci. Int.* 125 (2002) 22–36.

[21] J.M. Bosque-Sendra, L. Cuadros-Rodriguez,C. Ruiz-Samblas, A.P. de la Mata, *Anal. Chim. Acta* 724 (2012) 1-11.

[22] M. Barker, W. Rayens, *J. Chemometrics* 17 (2003) 166-173.

[23] R. Brereton, *Trends Anal. Chem.* 25 (2006) 1103-1111.

[24] K. Kjeldahl, R. Bro, *J. Chemometrics* 24 (2010) 558-564.

[25] K. J. Johnson, R. E. Synovec, *Chemom. Intell. Lab. Syst.* 60 (2002) 225-237.

[26] W.C. Lai, C. Song, *Fuel* 10 (1995) 1436 - 1451.

# 4 Chapter Four: Chemometric classification of casework arson samples based on gasoline content

The presence of an ignitable liquid (IL) in arson debris is one of the critical pieces of evidence in arson investigations. The presence of an ignitable liquid is typically ascertained through the use of headspace extraction coupled with GC-MS for analysis [1,2]. Interpretation of the resultant data sets is a time-consuming step which requires two highly trained analysts to manually examine the GC-MS data in order to reach a conclusion regarding the presence or absence of an ignitable liquid. If these two analysts do not agree on the interpretation, a third analyst will independently interpret the data. The three will then discuss and try to come to a consensus [3]. Thus, the interpretation of a single sample can require two to three person-hours of time and creates a very expensive bottleneck in the analytical process.

The goal of my research has been to develop tools for the automated optimization of classification models, particularly for GC-MS data. Collaboration with the RCMP Trace Evidence Operations Support arson laboratory has provided a highly challenging set of samples upon which to apply my tools. In Chapter 3, gasoline was successfully identified in simulated arson debris. However, the real challenge lies in the analysis of actual casework debris samples, which was the purpose of this study. Casework debris samples were analyzed by the RCMP in accordance with their standard methods, though a deuterated n-alkane ladder was added to their solvent. The raw data files were transferred to our laboratory and a model for the identification of gasoline in the fire debris samples was automatically optimized and constructed using a hybrid backward elimination (BE) and forward selection (FS) variable selection approach guided by the cluster resolution (CR) metric.

## 4.1 Fire debris analysis.

Arson is defined as "*the act of wilfully and maliciously setting fire to another man's house, ship, forest, or similar property; or to one's own, when insured, with intent to defraud the insurers*" [4]. Arson damage to residences, businesses, vehicles or other property is but one of the problems; arson also leads to loss of life, and feelings of insecurity in the community. Furthermore, financial costs extend beyond the price of the property damaged, leading to increased insurance rates, costs of fire protection, law enforcement, etc. [5]. Arson tends to be difficult to investigate since much of the evidence is inevitably damaged by the fire [6] as well as by the firefighting efforts, despite best efforts taken to minimize damage to the scene [5]. Important pieces of evidence during a fire investigation include ascertaining the presence of an ignitable liquid (IL) at the scene, as well as the determination of its identity [5].

Due to availability, efficacy, and low cost, petroleum-based accelerants, usually gasoline, are most often used by arsonists [7]. These

ILs contain hundreds of individual compounds with a specific composition that varies over time and depends on the vendor. Adding to the challenge, ILs undergo weathering during a fire due to the high temperatures and air flows [5,8] and, if debris is not recovered immediately after the fire, can also be affected by bacterial degradation [5,9]. All of These will lead to changes in the IL composition.

Further complicating the problem is the fact that debris matrices are highly variable, often complex and containing numerous precursor, pyrolysis and combustion products that interfere with the analysis [5,10]. Consequently, laboratory analysis of arson debris requires the extraction of potential IL traces from the debris. Extraction is then followed by separation and detection of the potential traces. This is followed by careful interpretation and analysis of the obtained data in the hope of identifying and classifying any ILs found in the debris.

### 4.1.1 Fire debris matrix

While the contents of the arson debris will vary from one fire scene to another, it is up to the investigator to determine which exact part(s) of the scene to sample. Investigators will normally select a location that is likely to contain an IL based on evidence such as burn patterns at the scene [5,11], or as indicated by aids, such as sniffer dogs [12,13,14]. Porous materials such as carpet or wood are generally good choices since they are more likely to retain traces of ILs [5,15]. Carpets and rugs are especially common, being over half of all debris samples collected [15]. This is due to their ability to retain significant quantities of ILs and keep retained ILs relatively intact due to flame-resistant coatings [5,15]. Additionally, carpets and rugs, being floor coverings, a common substrate to which ILs are delivered. Since carpets are made from a variety of natural (e.g. wool, cotton) and synthetic (e.g. polyolefin, nylon, polypropylene) fibres, there is a significant degree of chemical diversity between different types of carpets. Furthermore, carpets contain dyes, resins, and flame-resistant coatings, and generally have some form of underlay, which collectively add additional components to the debris. Other materials such as paper, plastics, paint, wool, cotton, leather (natural or synthetic), food [5], and even arsonists [16] which are present at the scene will further complicate the chemical make-up of the matrix.

Matrix components will also undergo chemical changes over the course of the fire. Temperature and oxygen levels will vary, meaning that a given location in a fire scene may undergo both combustion in the presence of oxygen and pyrolysis in the absence of oxygen over the course of a single fire [5,17,18]. Pyrolysis products from some materials are identical to some components in ILs, examples being ethylbenzene and toluene, which are common components in gasoline. Therefore, the investigator must be careful not to confuse matrix components with actual IL components.

### 4.1.2 Ignitable liquids

The word “accelerant” is properly reserved for ILs placed at the scene of a fire with the specific intent of causing the fire. ILs tend to be volatile liquids which can be easily delivered to the scene of fire, and which will provide enough energy to initiate and sustain the fire when ignited [5]. The choice of IL depends on ease of access and suitability of the IL to cause a fire. Thus, gasoline tends to be the most common IL used in arson since, in most parts of the world, it can be obtained easily and cheaply [5,6,7]. Gasoline is a petroleum product, containing alkanes, alkylbenzenes and condensed aromatics [1,5].

While ILs are generally fresh at the moment of delivery to the fire scene, the composition of the IL may change significantly over the course of the fire. Due to temperature and air flow, components of the IL will evaporate. However, due to differences in boiling points of various components within an IL, some components will evaporate to a greater degree compared to others, resulting in weathering. The extent of weathering will vary from one fire scene to the next [5,8]. Furthermore, ILs may undergo bacterial degradation if samples are not collected shortly after the fire [5,9]. This variability will pose additional challenges for IL detection and identification.

### 4.1.3 IL extraction and analysis

ILs tend to be volatile liquids and to identify their traces they are usually extracted from debris by concentration of headspace vapours [19]. Typical sampling methods include direct sampling of headspace vapors [20], dynamic headspace sampling using activated charcoal beds [21], passive headspace sampling using activated carbon strips [22], or techniques such as solid phase microextraction (SPME) [23]. Passive headspace extraction (other than by SPME) is typically followed by solvent extraction of the IL residues from the adsorptive medium using a solvent such as $CS_2$ or occasionally $Et_2O$ [2]. Extracts are then analyzed by GC-MS [1,5]. The method favoured by the RCMP Laboratory in Edmonton is a passive headspace extraction with activated charcoal strips, using a setup similar to that shown in Figure 3-2, followed by $CS_2$ extraction [3].

Once collected, chromatographic data are manually interpreted by two (or sometimes three) analysts to determine if there are traces of IL present in the debris, and if possible the identity of the IL [1]. This final step is currently an expensive bottleneck in arson debris analysis that we seek to address though the application of chemometric techniques.

## 4.2 Experimental

### 4.2.1 Casework Data

Debris were collected, stored, extracted, and analyzed according to RCMP protocols [1,2]. The only deviation from the standard protocol was the addition of a deuterated alkane ladder consisting of n-heptane (d16), n-nonane (d20), n-undecane (d24), n-tridecane (d28), n-pentadecane (d32), n- heptadecane (d36), n-nonadecane (d-40) and n- heneicosane (d-44) (CDN Isotopes, Pointe-Claire, QC) at concentrations of 16 $\mu L \cdot L^{-1}$ each to the solvent used to elute analytes from the activated carbon strips ($CS_2$).

Samples were analyzed using one of three Agilent Technologies 7890A gas chromatographs (GC) with 5975 quadrupole mass spectrometers (MS) and 7683 auto samplers (Agilent Technologies, Mississauga, ON). Data acquisition and automation were accomplished using MS ChemStation (Agilent). The GCs were equipped with 30 m × 250 µm × 0.25 µm HP-1MS columns (Agilent). The oven program used was 40 °C (held for 3.0 min) followed by a ramp to 250 °C at a rate of 8 $°C \cdot min^{-1}$, with a final hold of 0.75 min. Samples were injected in split mode into an injector held at 250 °C. Hydrogen carrier gas was used with flow rate of 1.1 mL/min. The injection volume was 1 µL, with a split ratio of 20:1. The transfer line and source temperatures were 300 and 230 °C, respectively.

Casework samples were processed in duplicate at the RCMP laboratories according to the ASTM 1618 protocol. The data provided to our laboratory was given dummy identifiers and with none of the attached identifying metadata that accompanies actual casework samples. This was done to ensure that no information that could compromise the confidentiality of an investigation was transmitted to our laboratory.

Overall, 232 casework chromatograms were provided by the RCMP. Identification of gasoline was performed by RCMP personnel. 65 samples were confirmed to contain gasoline, 155 samples were confirmed to contain no IL (but most did contain pyrolysis products) and 12 samples provided ambiguous results and could not be positively determined either way. Out of 65 gasoline-containing and 12 ambiguous samples, all were obtained as casework debris. Out of 155 gasoline-free samples, 79 were casework debris samples and 76 were gasoline-free debris samples simulated by the RCMP in accordance with a published protocol [24].

### 4.2.2 Chemometric treatment of arson data

Chromatograms were exported from Chemstation as .csv text files and then imported MATLAB 7.10.0 (The Mathworks, Natick, MA). Chromatograms were aligned on the basis of the deuterated alkane retention ladder (Chapter 3). Variable selection to optimize the chemometric models was performed using lab-written CR-guided BE/FS approach employing two-dimensional CR (Chapter 2). Final chemometric analysis of the optimized models was performed using lab-written MATLAB routines, and some chemometric analysis functions from the PLS Toolbox 5.2 (Eigenvector Research Inc, Wenatchee, WA). The calculations were performed on a Intel Core i5 750 2.76 GHz processor with 8 GB of RAM and 64-bit Microsoft Windows 7 Professional operating system.

## 4.3 Results and discussion

A potential solution to the high cost of data interpretation for arson investigations lies in the development of chemometric models for rapid, objective, and automated identification of ILs in fire debris samples. Should a successful chemometric solution be discovered, it would essentially remove the bottleneck in the analytical procedure, increasing the overall sample throughput for an arson laboratory. This would, by extension, permit fire investigators to increase the number of samples that are taken from a fire scene, while possibly decreasing the overall analysis time. As a result, more thorough, faster investigations of fire scenes would be possible.

Previous work has involved the application of exploratory techniques, such as PCA, to the identification of ILs [7,8,25]. SIMCA has also been used to classify ILs on a charred carpet sample [7]. In our work, PLS-DA was used to classify simulated arson debris based on the presence or absence of gasoline (Chapter 3) [26]. To date, there are no reported studies of the successful application of chemometric techniques to the interpretation of actual arson casework samples, to the best of our knowledge. Due to extreme conditions and variability of fire scenes, actual casework studies are crucial. It is likely that simulated debris will not accurately reflect debris obtained from real arson scenes.

Casework samples used in this study were collected over several months by a variety of arson investigators from fire scenes located across Canada (at the time of the study, the Edmonton Laboratory handled samples from all jurisdictions in Canada except for Ontario and Quebec). As most arsonists rely on gasoline as the IL, a sufficient number of debris samples could only be obtained for gasoline-containing and gasoline-free debris. Therefore, this initial test on real data focused on the classification of debris based on gasoline content.

With the use of real arson data, there was no control over the contents of the fire scenes, the nature, or amount of ILs being used, and the extent of variability in the data was staggering. The amount of gasoline remaining in the debris varied due to differences in the amount of IL used in a given arson, the substrate for the sample, and different extents of combustion and weathering in each fire. Additionally, the composition of gasoline varies depending on factors such as refinery, season, and region of the country. The matrix at the fire scenes was completely uncontrolled, and samples were prepared and analyzed by different analysts on one of three GC-MS systems with the same nominal operating conditions. No deviations were made from the standard analytical protocol, with the exception of the addition of the deuterated alkane ladder to the desorption solvent.

### 4.3.1 Chromatographic Alignment

Prior to the application of chemometric techniques to casework arson data, the raw chromatograms were aligned using the deuterated ladder-based alignment method presented in Chapter 3. The product of ions of m/z 34, 50, 66, 80, and 82 was used, with a randomly selected chromatogram from the training set as the alignment target. Ion 34 which is due to $C_2D_5^+$ was required to add selectivity to the generation of the alignment target for some samples of real debris. Due to the use of multiple GC-MS systems to collect the data, extreme shifts in retention times were observed, as shown in Figure 4-1.

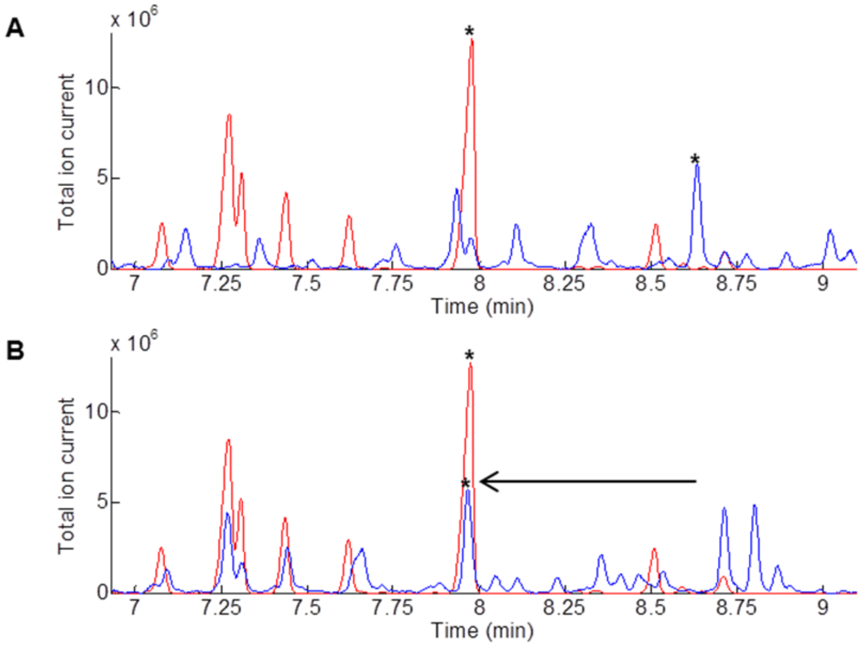


**Figure 4-1:** Segments of two chromatograms in the casework debris dataset collected on different instruments**. A** shows unaligned chromatograms. **B** shows aligned chromatograms. Asterisks indicate a pair of peaks that should be aligned.

As shown in Figure 4-1, extreme misalignment (~40 s) was observed for some samples. Additionally, the chromatographic profiles of the two debris samples were highly dissimilar with only a few peaks present in both chromatograms. Nevertheless, the deuterated alkane ladder alignment approach was able to successfully align all of the chromatograms.

### 4.3.2 Feature Selection

Once chromatograms were aligned, variables were selected using SR variable ranking followed by a hybrid BE/FS approach that relied on two-dimensional CR as the model evaluation metric (Chapter 2).

For model construction and testing, the data set was separated into three sets: training, optimization and validation. The 220 chromatograms with known class identities, were randomly split into a training set (110 chromatograms), optimization set (55 chromatograms), and validation set (55 chromatograms). All 12 unidentified samples were assigned to the validation set, bringing the total number of samples in that set to 67.

Chromatograms from the training set were used to create the alignment target, rank variables and to obtain loading vectors for the PCA model during each variable selection step. Chromatograms from the optimization set were used together with chromatograms from the training set, to obtain scores during variable selection (the model being generated using training data and tested against optimization data). Both training and optimization data were combined to create the final PCA model after variable selection was complete. Chromatograms from the validation set were only used to evaluate the final model.

Each chromatogram consisted of 16 000 scans with m/z values from 30 to 300, providing a total of 4 336 000 individual variables per chromatogram.  As was done in previous chapters, each GC-MS chromatogram was unfolded along the retention time axis, providing a single vector of 4 336 000 variables. SR-based feature ranking [27,28] was subsequently performed. After the variables were ranked, variable selection was performed using a hybrid BE/FS approach guided by the two-dimensional CR metric [37,29,30] (Chapter 2). The evaluation was performed upon a 2-component PCA model constructed using data from the training set and applied to data in the optimization set. The initial number of variables used in the BE approach was 10 000 and variables up to rank 25 000 were checked with the FS approach. The flowchart for variable selection is shown in Figure 2-7. A total of 1597 variables were selected. Variables selected are shown in Figure 4-2.

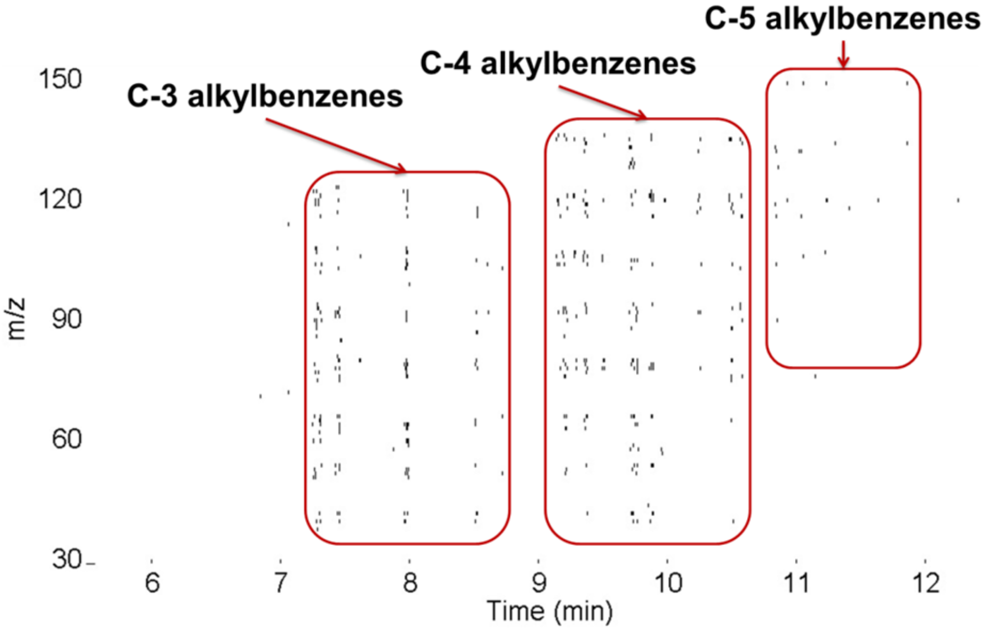


**Figure 4-2:** Features from GC-MS chromatograms included in optimized model for identification of gasoline in arson debris. Black dots represent variables used in model construction.

As seen from Figure 4-2, C-3, C-4 and C-5 alkylbenzenes were selected. As mentioned before, gasoline contains light alkanes, alkylbenzenes and condensed aromatics [1,19]. According to standard method ASTM E 1618, alkanes present in gasoline samples vary by brand, grade and lot. Furthermore, being relatively light molecules, they are more likely to evaporate during gasoline weathering. They are also generated by pyrolysis of some materials (e.g. polyethylene) [31]. Thus one would expect the alkanes to be of little diagnostic value for the purpose of identifying gasoline in arson debris. This explains the exclusion of light alkanes by the algorithm.

ASTM E 1618 also cautions against using BTEX (benzene, toluene, ethylbenzene, xylenes) and condensed aromatics such as naphthalene as markers for gasoline. These compounds are also natively present even in gasoline-free debris matrix as they can be formed by numerous pyrolysis processes. It is reassuring that the automated approach to variable selection also ignored this group of compounds. ASTM E 1618 recommends using the C-3, C-4, and C-5 alkylbenzenes as markers for gasoline as these compounds are characteristic of gasoline and do not generally have other sources in debris. As seen in Figure 4-2, variable selection guided by the CR metric selected variables originating from the compounds recommended by the standard method for identifying gasoline in fire debris. It is important to note that the selection was performed automatically without any direction as to which variables to focus on. In fact the only information provided was the binary class assignment (gasoline/no gasoline) of the chromatograms.

Following selection of relevant variables, chemometric models for classification of arson debris were constructed. All chemometric models involved the following pre-processing: the signal for each sample was normalized to an area of 1, followed by autoscaling of the combined training and optimization sets used to construct the final model. The autoscaling parameters determined in this step were then applied to the validation data.

Initially, a PLS-DA classification model was constructed (Figure 4-3). The number of LVs was chosen using venetian blinds cross-validation with 10 data splits and using number of LVs that provided the lowest misclassification rate. 3 LVs were used in the model construction.

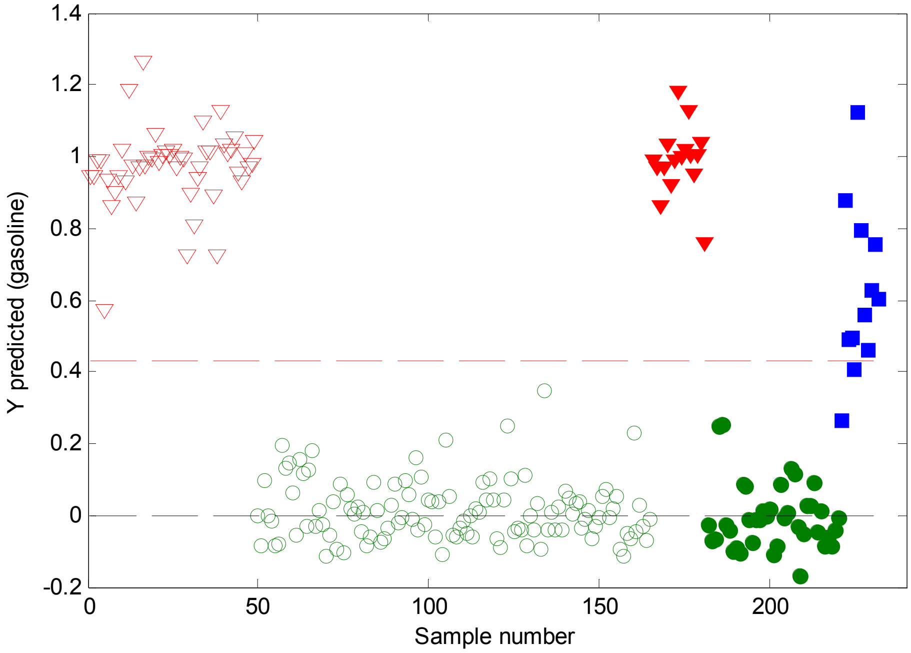


**Figure 4-3:** PLS-DA plot for arson data. Red triangles indicate gasoline-containing samples. Green circles indicate gasoline-free samples. Blue squares indicate ambiguous samples. Hollow markers indicate training and optimization set data. Filled markers indicate validation set data.

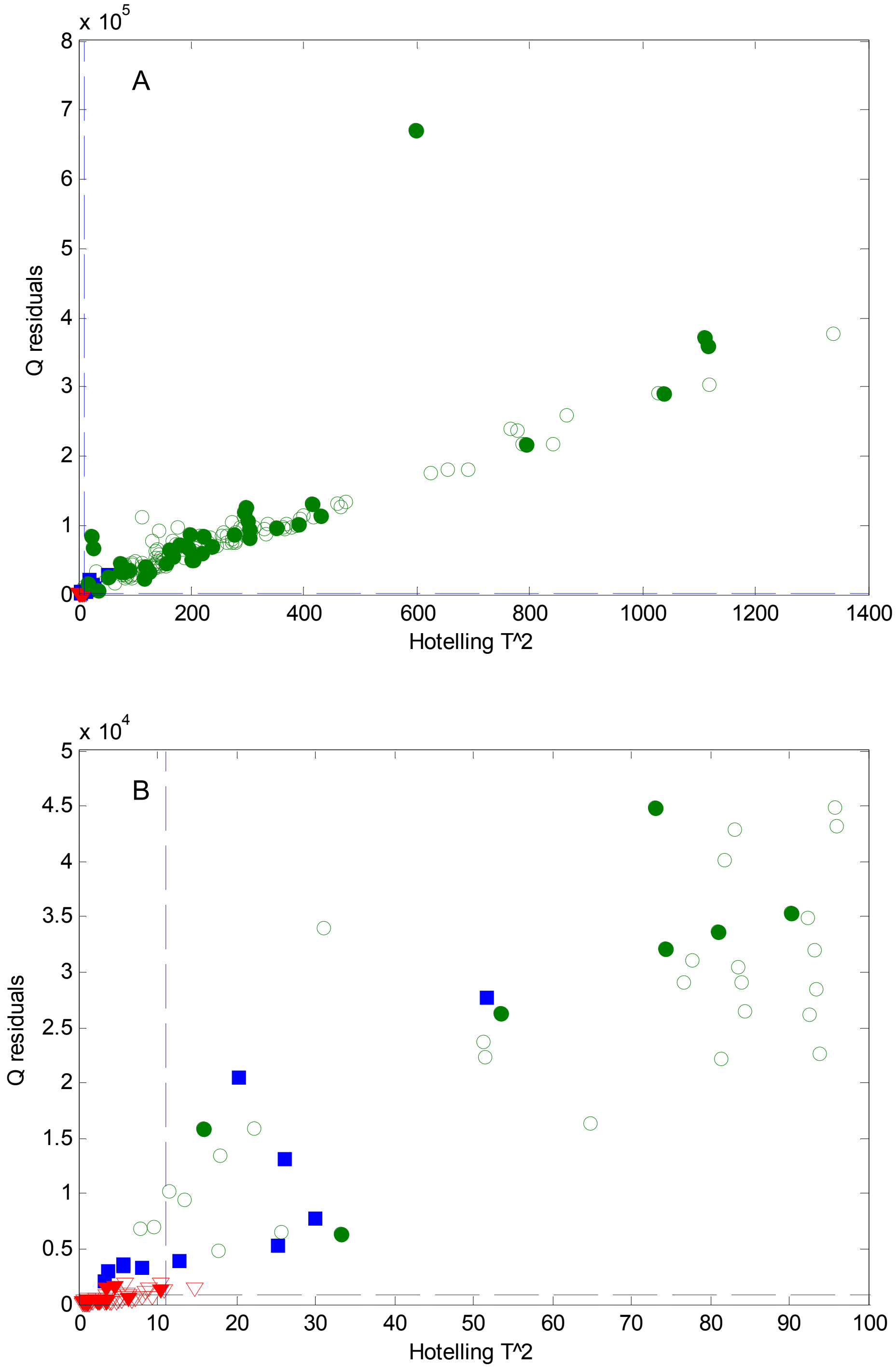
As seen from Figure 4-3, the PLS-DA model correctly classified all samples in the gasoline-containing and gasoline-free classes. Some of the ambiguous samples have fallen confidently in the gasoline-containing class while many remained near the classification border. However, PLS-DA is likely not the most appropriate technique for gasoline classification. The reason is that PLS-DA assigns value of zero for all samples in the gasoline free-class. However, assigning the same y value to all the samples in the gasoline-free class is not entirely correct: the only similarity between samples in the gasoline-free class is the lack of gasoline. The chemical composition of one non-gasoline containing sample can be completely alien to the chemical composition of another sample in the class.

To address this issue, SIMCA was tested as a modeling tool. SIMCA differs from PLS-DA in that it does not force a yes/no decision on a sample. Instead SIMCA creates a PCA model for one or more selected classes or groups of classes [32]. The samples are then projected into the collection of PCA models and $T^2$ and Q residuals for the samples are calculated for each class with each model. Class assignment is made on the basis of residual scores: as residual scores for a sample in a class model increase, the likelihood of class membership for the sample in the particular class decreases. Unlike PLS-DA, SIMCA allows a sample to be a member of none, one, or multiple classes. In the case of fire debris, it is possible that a mixture of ILs was used; making application of a technique that allows multiple class membership more appropriate and classification of a sample as IL-free would only result as a lack of fit into any of the models for ILs contained in the SIMCA model.

A SIMCA model was built for the gasoline-containing debris class. The number of PCs was chosen using venetian blinds cross-validation with 10 data splits and using the number of PCs that provided the lowest error of cross-validation. 4 PCs were used in the model construction.

Classification in SIMCA is made on the basis of residuals. If a sample has a pattern in the selected variables that is similar to gasoline, then it will have very low values for its Q and $T^2$ residuals. On the other hand, if a sample contains no gasoline, it will not fit the gasoline model well and it will have high residual values. The Q vs. $T^2$ plot for the gasoline data set is presented in Figure 4-4 at several magnifications. Gasoline-containing samples should lie in the bottom left corner of this plot, and as samples become less gasoline-like, they should drift towards the top right corner of the plot, as observed.

Comparing the results in Figure 4-4, to 4-3, SIMCA was also able to reliably classify arson samples based on gasoline content. The results for ambiguous samples will probably prove more useful and/or reliable for arson investigators.


A
x 10^5
Q residuals
0
1
2
3
4
5
6
7
8
0
200
400
600
800
1000
1200
1400
Hotelling T^2
B
x 10^4
Q residuals
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
0
10
20
30
40
50
60
70
80
90
100
Hotelling T^2

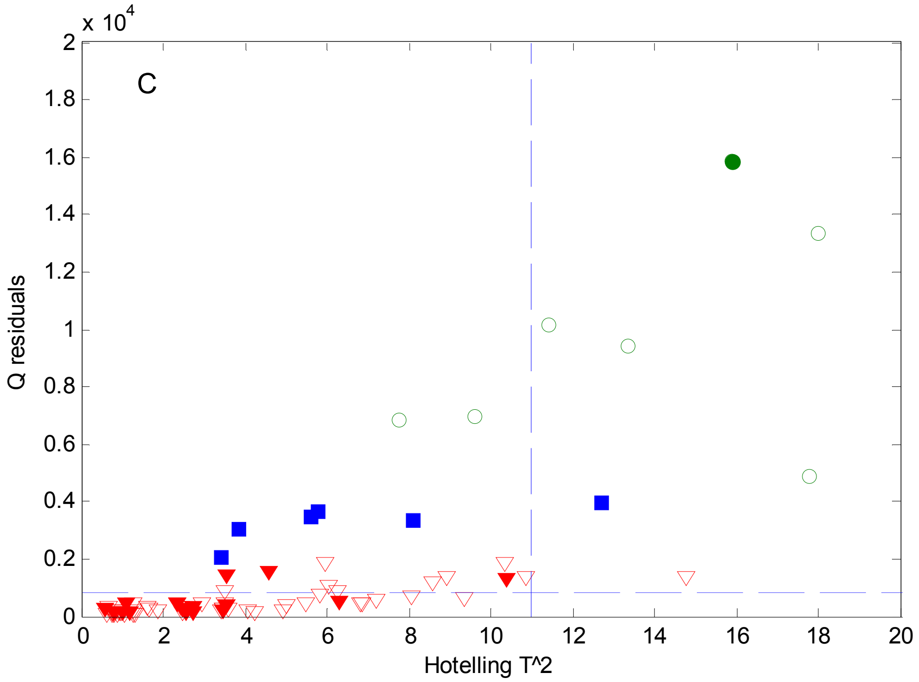


**Figure 4-4:** SIMCA plot for arson data. Red triangles indicate gasoline-containing samples. Green circles indicate gasoline-free samples. Blue squares indicate ambiguous samples. Hollow markers indicate training and optimization set data. Dashed lines indicate 95% confidence levels for Hotelling $T^2$ and Q residuals Filled markers indicate validation set data. **A**, **B** and **C** show different zoom levels for the plot.

## 4.3 Conclusions

Deuterated alkane ladder-based alignment and the CR-guided automated approach to variable selection have been applied to generate PLS-DA and SIMCA models for the classification of casework arson debris samples on the basis of gasoline content. The alignment was able to account for extreme retention time shifts (~40 s). The variable selection algorithm automatically selected a suite of variables derived from compounds identified in the standard ASTM method as being reliable markers for gasoline, while successfully ignoring compounds known to be unreliable markers of gasoline. The final PLS-DA and SIMCA models were able to reliably classify the samples as being either gasoline-containing or gasoline-free, with no false positives or false negatives.

## 4.4 References.

[1] ASTM Standard E1618, 2006e1, "Standard Test Method for Ignitable Liquid Residues in Extracts from Fire Debris Samples by Gas Chromatography-Mass Spectrometry" ASTM International, West Conshohocken, PA, 2006, DOI: 10.1520/E1618-11, www.astm.org.

[2] ASTM Standard E1412, 2007, " Standard Practice for Separation of Ignitable Liquid Residues from Fire Debris Samples by Passive Headspace Concentration With Activated Charcoal" ASTM International, West Conshohocken, PA, 2007, DOI: 10.1520/E1412-07, www.astm.org.

[3] P.M.L. Sandercock. Manager, Trace Evidence Operations Support, Royal Canadian Mounted Police, Edmonton, AB. Personal communication, February, 24, 2009.; September 10, 2009.

[4] Arson. *Oxford English Dictionary* [Online]; Oxford University Press, Second edition, 1989; online version June 2012, http://www.oed.com/view/Entry/11118 (accessed June 18, 2012).

[5] E. Stauffer, J. A. Dolan, R. Newman, *Fire Debris Analysis*, Elsevier, Inc. Amsterdam, 2008.

[6] P.M.L. Sandercock, *Forensic Sci. Int.* 176 (2008) 93-110.

[7] B. Tan, J.K. Hardy, R.E. Snavely, *Anal. Chim. Acta* 422 (2000) 37-46.

[8] P.M.L. Sandercock, E. Du Pasquier, *Forensic Sci. Int.* 140 (2004) 43-59.

[9] K.P. Kirkbride, S.M. Yap, B. App, S. Andrews, P.E. Pigou, G. Klass, A.C. Dinan, F.L. Peddie, *J. Forensic Sci.* 37 (1992) 1585-1599.

[10] E. S. Bodle, J.K. Hardy, *Anal. Chim. Acta* 589 (2007) 247-254.

[11] J.F. O’Donnell, *Fire and Arson Investig.* 35 (1985) 18–20.

[12] M.E. Kurz, M. Billard, M. Rettig, J. Augustiniak, J. Lange, M. Larsen, R. Warrick, T. Mohns, R. Bora, K. Broadus, G. Hartke, B. Glover, D. Tankersley, J. Marcouiller, *J. Forensic Sci.* 39 (1994) 1528-1536.

[13] M.E. Kurz, S. Schultz, J. Griffith, K. Broadus, J. Sparks, G. Dabdoub, J. Brock, *J. Forensic Sci.* 41 (1996) 868-873.

[14] S.R. Katz, C.R. Midkiff, *J. Forensic Sci.* 43 (1998) 329-333.

[15] W. Bertsch, Q-W Zhang, *Anal Chim Acta* 236 (1990) 183–95.

[16] 1999 Darwin Award: Firefighters Ignite! http://www.darwinawards.com/darwin/darwin1999-24.html (accessed Nov 8, 2012).

[17] E. Stauffer, Basic concept of pyrolysis for fire debris analysts, *Sci. Justice* 43 (2003) 29–40.

[18] J.D. DeHaan, D.J. Brien R. Large, *Sci. Justice* 44 (2004) 223–36.

[19] J. Dolan, *Anal. Bioanal. Chem.* 376 (2003) 1168-1171.

[20] ASTM Standard E1388, 2005, "Standard Practice for Sampling of Headspace Vapors from Fire Debris Samples" ASTM International, West Conshohocken, PA, 2005, DOI: 10.1520/E1388-05, www.astm.org.

[21] ASTM Standard E1413, 2007, "Separation and Concentration of Ignitable Liquid Residues from Fire Debris Samples by Dynamic Headspace Concentration" ASTM International, West Conshohocken, PA, 2007, DOI: 10.1520/E1413-07, www.astm.org.

[22] K. Cavanagh, E. Du Pasquiera, C. Lennard, *Forensic Sci. Int.* 125 (2002) 22–36.

[23] K. G. Furton, J. R. Almirall, M. Bi, J. Wang, L. Wu, *J. Chromatogr. A* 885 (2000) 419-432.

[24] P.M.L. Sandercock, *J. Forensic Sci.* 57 (2012) 738-743.

[25] P. Doble, M. Sandercock, E. Du Pasquier, P. Petocz, C. Roux, M. Dawson, *Forensic Sci. Int.* 132 (2003) 26-39.

[26] N.A. Sinkov, B.M. Johnston, P.M.L. Sandercock, J.J. Harynuk, *Anal. Chim. Acta* 697 (2011) 8-15.

[27] T. Rajalahti,R. Arneberg, F. S. Berven, K. M. Myhr, R. J. Ulvik, O. M. Kvalheim, *Chemom. Intell. Lab. Syst.* 95 (2009) 35-48.

[28] T. Rajalahti,R. Arneberg, A. C. Kroksveen, M. Berle, K. M. Myhr, O. M. Kvalheim, *Anal. Chem.* 81 (2009) 2581-2590.

[29] N.A. Sinkov, J.J. Harynuk, *Talanta* 83 (2011) 1079-1087.

[30] N.A.Sinkov,J.J.Harynuk, *Talanta* 103 (2013) 252-259.

[31] J. A. Onwudili, N. Insura, P. T. Williams, *J. Anal. Appl. Pyrolysis* 86 (2009) 293-303.

[32] R.G. Brereton, Chemometrics Data Analysis for the Laboratory and Chemical Plant, Wiley, Toronto, 2003.

# 5 Conclusions and future work

## 5.1 General conclusions and other studies.

In the course of my research, a novel metric for evaluating classification models, termed *cluster resolution* was invented [1,2]. An approach to aligning highly dissimilar GC-MS data that relies on the addition of a deuterated alkane ladder to the samples was also developed [3].

Both two- and three-dimensional versions of the CR metric have been developed and applied to guide automated variable selection processes that optimize chemometric models for the classification of gasoline samples by type as well as simulated and real arson debris. The deuterated alkane ladder-based alignment was essential for the alignment of highly dissimilar data and was successfully applied even in the case of arson casework samples that were analyzed by multiple analysts on multiple GC-MS instruments over the course of several months.

Combined, these tools allowed for the first time, the construction of effective classification models for casework arson debris samples based on gasoline contents with 100 % of studied samples being classified correctly.

Most importantly, it should be noted that the CR-guided approach to model optimization is neither limited to arson data, nor to analytical separations data. When variable selection is necessary for other kinds of data (e.g. spectroscopic), CR is a powerful tool that can be used to guide the variable selection process or any other process that requires the evaluation of a classification model. Recently, the CR metric has been successfully applied in the classification of edible oils on the basis of ATR-FTIR spectra [4]. In this study, the CR-guided BE/FS hybrid approach was shown to outperform models that had no feature selection, and other automated approaches such as iPLS and genetic algorithms (GA) in terms of optimization speed, final model quality, and ease of use (Table 5-1). Furthermore, it was shown in this application that the CR-guided approach successfully avoided regions of the spectrum that obviously contained no significant information, and regions of the spectrum that were known to be unreliable due to phonon bands of the diamond ATR cell used in the study.

**Table 5-1.** Comparison of the performance of variable selection methods

| Method | Raw data | PCA | iPLS-DA (1variable) | GA all variables | ANOVA | SR |
|---|---|---|---|---|---|---|
| Computational time for variable selection | - | < s | 628 min | 3434 min | 2.7 min | 2.5 min |
| Sensitivity (CV) for olive oil class | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Specificity (CV) for olive oil class | 0.94 | 1.00 | 0.97 | 1.00 | 1.00 | 1.00 |
| # of samples misclassified | 1 | 0 | 1 | 0 | 0 | 0 |
| # of LVs | 2 | 2 | 4 | 2 | 6 | 4 |
| # of variables (details) | 3320 (All variables) | 9 (PCs) | 11 intervals | 248 | 26 | 30 |

## 5.2 Future work.

Further advancement of the arson debris analysis project is required. The development and testing of methods for the simulation of chemometrically-realistic arson debris are ongoing. This will allow for the generation of training data for the detection of less-common ILs. Constructing such a library using only casework data would require great patience while waiting for sufficient arsons using these other ILs to be committed.

Another avenue of development lies in the commercialization of the tools created. With a classification model developed for the arson debris samples, a model can be applied to new data collected for the arson debris. The classified debris can then be added to the training set, updating the classification model over time. Since model application takes seconds, the SIMCA model, presented in Chapter 4, can be utilized as a screening tool, where samples that are confidently classified as gasoline-containing or gasoline-free would require little further analysis, while samples which lie close to the border will require manual interpretation, as shown in Figure 5-1. Since more than 90 % of the samples should be unambiguously classified, this will result in a considerable decrease in investigators' workloads and an increase in the throughput of the laboratory.

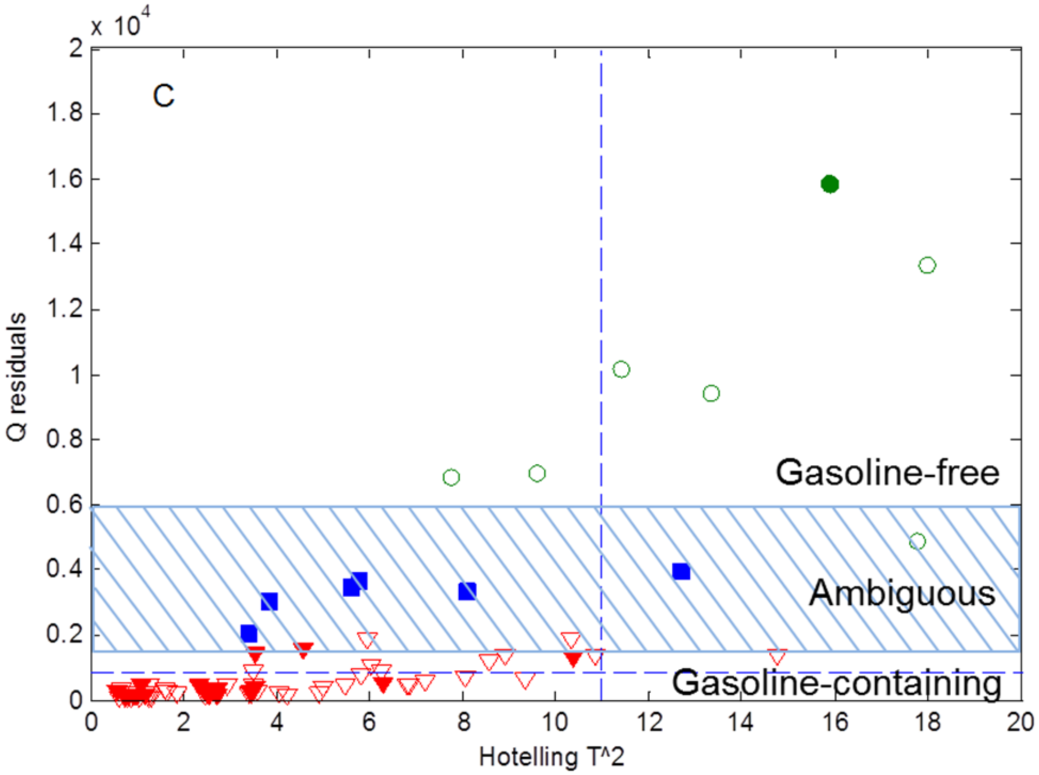


**Figure 5-1:** SIMCA plot for arson data. Red triangles indicate gasoline-containing samples. Green circles indicate gasoline-free samples. Blue squares indicate ambiguous samples. Hollow markers indicate training and optimization set data. Dashed lines indicate 95% confidence levels for Hotelling $T^2$ and Q residuals Filled markers indicate validation set data. **A**, **B** and **C** show different zoom levels for the plot.

In terms of applying chemometric techniques to raw GC-MS data, additional work should be focused towards the variable selection problem. While FS/BE variable selection guided by CR metric was able to successfully select relevant variables, hundreds or thousands of variables have been selected in applications presented in Chapters 2, 3 and 4. Many of the selected variables were result of signal from the same compounds and, as a result, were highly correlated and redundant. An intelligent approach towards selecting unique features for each analyte will further assist analysts with both extent of variable reduction and quality of the resulting models.

The CR metric could benefit from code optimization. Currently, the metric is able to evaluate models in a reasonable amount of time, however further code optimization, particularly in searching confidence ellipse sizes for collision detection, will allow acceleration of the variable selection process. The general approach to model optimization should be applied to other types of data such as metabolomics data (either raw or processed peak tables) and the CR metric should also be tested in other types of computational tools, for example as the "goodness metric" in support vector machines or neural networks.

It should also be noted that CR in its current form cannot be directly applied to feature selection for regression models. Adapting the approach for application in regression models would be a topic for future studies.

## 5.3 References.

[1] N.A. Sinkov, J.J. Harynuk, *Talanta* 83 (2011) 1079-1087.

[2] N.A.Sinkov,J.J.Harynuk, *Talanta* 103 (2013) 252-259.

[3] N.A. Sinkov, B.M. Johnston, P.M.L. Sandercock, J.J. Harynuk, *Anal. Chim. Acta* 697 (2011) 8-15.

[4] N. A. Sinkov, A. P. de la Mata, A. Dominguez-Vidal, A. Ruiz-Medina, M. J. Ayora-Cañada, J. J. Harynuk, Submitted to *Anal. Chem.*