Calibration, Feature Extraction and Classification of Water Contaminants Using a Differential Mobility Spectrometer

by

Bobby Ren

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of Master of Engineering in Electrical Engineering and Computer Science
at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY
May 2006

© Bobby Ren, MMVI. All rights reserved.
The author hereby grants to MIT permission to reproduce and distribute publicly paper and electronic copies of this thesis document in whole or in part.

Author .................................................................
Department of Electrical Engineering and Computer Science May 26, 2006

Certified by ............................................................... Nirmal Keshava
Head of Technical Staff
Thesis Supervisor

Certified by ............................................................... Dennis Freeman
Associate Professor
Thesis Supervisor

Accepted by ............................................................ Arthur C. Smith
Chairman, Department Committee on Graduate Students

ARCHIVES
Calibration, Feature Extraction and Classification of Water Contaminants Using a Differential Mobility Spectrometer

by

Bobby Ren

Submitted to the Department of Electrical Engineering and Computer Science on May 26, 2006, in partial fulfillment of the requirements for the degree of Master of Engineering in Electrical Engineering and Computer Science

Abstract

High-Field Asymmetric Waveform Ion Mobility Spectrometry (FAIMS) is a chemical sensor that separates ions in the gaseous phase based on their mobility in high electric fields. A threefold approach was developed for both chemical type classification and concentration classification of water contaminants for FAIMS signals. The three steps in this approach are calibration, feature extraction, and classification. Calibration was carried out to remove baseline fluctuation and other variations in FAIMS data sets. Four feature extraction algorithms were used to extract subsets of the signal that had high separation potential between two classes of signals. Finally, support vector machines were used for binary classification. The success of classification was measured both by using separability metrics to evaluate the separability of extracted features, and by the percent of correct classification ($P_{CC}$) in each task.

Thesis Supervisor: Nirmal Keshava
Title: Head of Technical Staff

Thesis Supervisor: Dennis Freeman
Title: Associate Professor
Acknowledgments

I would like to thank Nirmal Keshava, my Draper Advisor, for his guidance in the research and feedback on the manuscript, and Heidi Perry for her support. I would like to acknowledge the members of the Bioengineering Group at Draper, including Melissa Krebs and Angela Zapata, for technical insights into FAIMS and help with data collection. Also, thanks to Dennis Freeman, my MIT Faculty Advisor, for giving feedback on the manuscript.
Contents

1 Introduction ........................ 15
   1.1 Motivation .......................... 15
   1.2 Previous Work ........................ 15
   1.3 Project Goal ........................ 16

2 Project Overview ................... 17
   2.1 Motivation .......................... 17
   2.2 Threefold Approach ................... 17

3 FAIMS Overview and Data Collection 19
   3.1 Apparatus Setup ...................... 20
   3.2 Data Collection ...................... 22
      3.2.1 Acquisition ....................... 22
      3.2.2 Chemical Basis ................... 23

4 Methodology .......................... 25
   4.0.3 Two Classification Tasks ........... 25
   4.1 Step 1: Calibration ................... 27
      4.1.1 Preprocessing ....................... 27
      4.1.2 Calibration ........................ 29
      4.1.3 Alternative Calibration Techniques 31
   4.2 Step 2: Feature Extraction ........... 32
      4.2.1 Pixel Domain vs. Wavelet Transform 32
List of Figures

3-1 Ion movement due to asymmetric field mobility. .......................... 21
3-2 Unprocessed FAIMS signal. .................................................. 22

4-1 Removal of nitrogen and ammonium artifacts in FAIMS signals. ... 28
4-2 Marginal FAIMS spectra before calibration. .............................. 29
4-3 Boxplot of FAIMS intensities before calibration. ........................ 30
4-4 Features measured for background intensity in boxplot. .............. 30
4-5 Peak to noise ratio for benzene 250ppm runs before baseline calibration. 30
4-6 The Haar mother wavelet. .................................................. 33
4-7 Coefficients from 2-level wavelet decomposition. ........................ 34
4-8 Wrapper feature extraction vs. filter feature extraction. .............. 36
4-9 SVM Mapping to a higher dimension ..................................... 45
4-10 k-Nearest Neighbors in 2-D space, k=3 ................................. 46

5-1 Mean FAIMS Spectrum for benzene 125ppm. ............................ 48
5-2 Mean FAIMS Spectrum for benzene 250ppm. ............................ 48
5-3 Mean FAIMS Spectrum for benzene 500ppm. ............................ 49
5-4 Mean FAIMS Spectrum for DCM 125ppm. ................................. 49
5-5 Mean FAIMS Spectrum for DCM 250ppm. ................................. 50
5-6 Mean FAIMS Spectrum for DCM 500ppm. ................................. 50
5-7 Mean FAIMS Spectrum for water. .......................................... 51
5-8 Benzene 125ppm and 500ppm signal comparison. ........................ 51
5-9 Marginal FAIMS spectra before and after calibration. ................. 52
5-10 Boxplot of FAIMS intensities before and after calibration. .......... 52
5-11 Peak to variance ratio in benzene 250ppm before and after calibration. 53
5-12 $P_{CC}$ for benzene 250ppm vs DCM 250ppm in the pixel domain. . . 55
5-13 $P_{CC}$ for benzene 500ppm vs DCM 500ppm in the pixel domain. . . 57
5-14 Features selected for benzene 250ppm and DCM 250ppm in the pixel domain. .......................................................... 58
5-15 Features selected for benzene 500ppm and DCM 500ppm in the pixel domain. .......................................................... 59
5-16 $P_{CC}$ for benzene 125ppm vs benzene 500ppm in the pixel domain. . 60
5-17 Features selected for benzene 125ppm and benzene 500ppm in the pixel domain. .......................................................... 61
5-18 $P_{CC}$ for DCM 250ppm vs DCM 500ppm in the pixel domain. . . 62
5-19 Features selected for DCM 250ppm and DCM 500ppm in the pixel domain. .......................................................... 63
5-20 $P_{CC}$ for benzene 125ppm vs DCM 125ppm in the wavelet domain. . 64
5-21 Reconstruction of benzene 125ppm and DCM 125ppm with wavelet features. .......................................................... 65
5-22 $P_{CC}$ for DCM 125ppm vs DCM 500ppm in the wavelet domain. . . 66
5-23 Reconstruction of DCM 125ppm and DCM 500ppm with wavelet features selected by RFS. .......................................................... 67
5-24 Mean FAIMS Spectrum for DCM 500ppm. ................................. 68
5-25 NUCS score vs $P_{CC}$ for all feature sets in the pixel domain. . . 70
5-26 NUCS score vs $P_{CC}$ for all feature sets in the wavelet domain. . . 70
5-27 SVM margin vs $P_{CC}$ for all feature sets in the pixel domain. . . 71
5-28 SVM margin vs $P_{CC}$ for all feature sets in the wavelet domain. . . 71
A-1 $P_{CC}$ for benzene 125ppm dcm 125ppm, pixel domain. ............... 78
A-2 $P_{CC}$ for benzene 250ppm dcm 250ppm, pixel domain. ............... 78
A-3 $P_{CC}$ for benzene 500ppm dcm 500ppm, pixel domain. ............... 79
A-4 $P_{CC}$ for benzene 125ppm benzene 250ppm, pixel domain. ........... 79
A-5 $P_{CC}$ for benzene 125ppm benzene 500ppm, pixel domain. ........... 80
A-6 $P_{CC}$ for benzene 250ppm benzene 500ppm, pixel domain. . . . . . . . . . . . 80
A-7 $P_{CC}$ for dcm 125ppm dcm 250ppm, pixel domain. . . . . . . . . . . . 81
A-8 $P_{CC}$ for dcm 125ppm dcm 500ppm, pixel domain. . . . . . . . . . . . 81
A-9 $P_{CC}$ for dcm 250ppm dcm 500ppm, pixel domain. . . . . . . . . . . . 82
A-10 $P_{CC}$ for benzene 125ppm dcm 125ppm, wavelet domain. . . . . . . . . . . . 82
A-11 $P_{CC}$ for benzene 250ppm dcm 250ppm, wavelet domain. . . . . . . . . . . . 83
A-12 $P_{CC}$ for benzene 500ppm dcm 500ppm, wavelet domain. . . . . . . . . . . . 83
A-13 $P_{CC}$ for benzene 125ppm benzene 250ppm, wavelet domain. . . . . . . . . . . . 84
A-14 $P_{CC}$ for benzene 125ppm benzene 500ppm, wavelet domain. . . . . . . . . . . . 84
A-15 $P_{CC}$ for benzene 250ppm benzene 500ppm, wavelet domain. . . . . . . . . . . . 85
A-16 $P_{CC}$ for dcm 125ppm dcm 250ppm, wavelet domain. . . . . . . . . . . . 85
A-17 $P_{CC}$ for dcm 125ppm dcm 500ppm, wavelet domain. . . . . . . . . . . . 86
A-18 $P_{CC}$ for dcm 250ppm dcm 500ppm, wavelet domain. . . . . . . . . . . . 86
List of Tables

3.1 Benzene and Dichloromethane ........................................... 23
3.2 Available data samples by class ......................................... 23

5.1 Classification tasks performed .......................................... 54
5.2 Feature scores for benzene 250ppm vs. DCM 250ppm, pixel domain . 56
5.3 Feature scores for benzene 500ppm vs. DCM 500ppm, pixel domain . 56
5.4 Feature scores for benzene 125ppm vs. benzene 500ppm, pixel domain 61
5.5 Feature scores for DCM 250ppm vs. DCM 500ppm, pixel domain . 61
5.6 Feature scores for benzene 125ppm vs. DCM 125ppm, wavelet domain 66
5.7 Feature scores for DCM 125ppm vs. DCM 500ppm, wavelet domain . 68
Chapter 1

Introduction

1.1 Motivation

High-Field Asymmetric Waveform Ion Mobility Spectrometry (FAIMS) is a recently developed differential mobility spectrometer that separates ions in the gaseous phase based on their mobility in high electric fields. Traditional ion mobility spectrometry (IMS) is useful for separating ions based on size-to-charge at low field strengths. However, at low field strengths, ion mobility does not serve as a highly differentiating characteristic, whereas at high field conditions ion mobility can be used to characterize different chemicals[17]. Typical mass spectrometers, are also expensive and relatively large, and require operation at low pressures that cannot be easily reproduced in the field[10]. FAIMS, which operates at atmospheric pressure, is designed to separate ions based on differences in ion mobility in a high asymmetric radio-frequency electric field. With the right algorithms and software techniques, it is a tool that may pave the way for handheld tools that can be used in a wide variety of clinical and environmental applications.

1.2 Previous Work

Much previous work has been done to explore the utility of FAIMS. FAIMS was first developed at the Charles Stark Draper Laboratories in 1998 as a smaller, portable
tool for mass spectroscopy without compromising sensitivity or performance [6]. It has since been used in many applications, ranging from monitoring water quality to classifying potential biological weapons for homeland security [10]. This thesis is a logical extension of previous work of applying statistical algorithms toward the detection of impurities in water [3]. We emphasize classification between different contaminants rather than detection of a single contaminant in an otherwise clean background.

1.3 Project Goal

The goal of this thesis is the classification of FAIMS spectra involving two chemicals, benzene and dichloromethane. A three step algorithm will be developed: calibration, feature extraction, and classification. Calibration is a preanalysis step necessary because of observations of sensor drift. Feature extraction, on both pixel intensities and wavelet transform coefficients, will involve filter and wrapper methods to capture the most distinguishing features from the FAIMS spectra. Finally, classification will be performed with support vector machines (SVMs), a popular classification technique for bioinformatics. We also prefer SVMs because it is nonparametric and thus does not require class statistics to be effective. This threefold approach for classifying simple compounds may one day be extended to solving more complex classification problems with the FAIMS apparatus.
Chapter 2

Project Overview

2.1 Motivation

The FAIMS sensor is an alternative tool to traditional mass spectroscopy, and has potential for field and clinical applications because of its size and low cost. However, techniques for field applications of FAIMS sensors are not perfect. Furthermore, sensor drift and other variations of the instrument over time can make the FAIMS output difficult to interpret. We seek a method to extract reliable features from FAIMS spectra for classification that will still be robust to day to day changes in instrument sensitivity and field conditions.

The development of a general algorithm for performing classification with FAIMS data involves a threefold approach: calibration, feature extraction, and classification. This research will focus on the development of such a method for classifying between two chemical contaminants, benzene and dichloromethane. The process may be extended for other pairs of chemicals, and may be adapted to a large range of applications in the field.

2.2 Threefold Approach

First, the collection and processing of raw FAIMS data involves an investigation into calibration methods for the FAIMS sensor. The goal of calibration is to prepare
raw FAIMS data for the classification process by removing artifacts resulting from
sensor drift, excessive noise, and FAIMS operational noise that may negatively affect
the training of a robust classifier. There have been no literature on calibrating the
FAIMS sensor for classifying chemicals, but initial experiments show that calibration
is needed in order to accurately train classifiers.

Second, feature extraction using various algorithms is applied to extract features
that can be submitted to a classifier. These features should be robust subsets of the
spectra that distinguish between two classes of signals. We will explore both filter
and wrapper algorithms. Filter algorithms produce features that have no specificity
to a particular classification technique, whereas wrapper algorithms use the results
of classification algorithm to improve on the feature selection process. Features will
either come from raw pixel data of the FAIMS spectra or from wavelet coefficients (the
transform domain). Thus, feature extraction will also involve exploring the wavelet
decomposition of FAIMS data and how that contributes to the feature set.

The third step of the threefold approach is to be able to classify contaminants
using FAIMS spectra. We focus on support vector machine classification between
two chemical contaminants. SVMs are a popular classification tool in the bioinfor-
matics field and have been used at Draper Labs to classify stem cell images[8]. This
nonparametric approach allows for flexibility of input data; SVMs do not assume
specific distribution of data, and are also powerful in dealing with both linearly and
nonlinearly separable data.

Finally, the performance of the three step algorithm will be measured by its per-
centage of correct classification ($P_{CC}$). The $P_{CC}$ is the percentage of a test data
set that is correctly classified by the SVM classifier. The effectiveness of the feature
extraction algorithms will also be evaluated using various measures of distance.
Chapter 3

FAIMS Overview and Data Collection

FAIMS first emerged in the 1990s as a new tool for mass spectrometry. It is a relatively new development in a field that has sought instruments capable of separating compounds into their components based on chemical properties. In classic mass spectrometry, a mass spectrometer is coupled with gas chromatography to produce chemical fingerprints for compounds that can be matched in a library. In the search for better mass spectrometry instruments, ion mobility spectrometry (IMS) was developed in the 1970s by combining radioactive ionization with drift tube ion separation[18]. Successful applications of IMS include hand held chemical agent monitors (CAM) used for detection of chemical warfare agents[13].

However, at low field strengths, ion mobility does not serve as a highly differentiating characteristic, whereas at high field conditions ion mobility can be used to characterize different chemicals. Typical mass spectrometers are also expensive and relatively large, and require operation at low pressures that cannot be easily reproduced in the field. Extensions of IMS include the addition of high electric fields produced between two parallel plates[2]. A cylindrical configuration, named the Field Ion Spectrometer (FIS), was introduced by the Mine Safety Appliances Company[6]. Finally, in 1998 the FIS was combined with a mass spectrometer, capable of atmospheric ion separation[17], and became the FAIMS. FAIMS has the advantages of
being more portable, inexpensive and faster in operation than traditional tools of mass spectrometry.

Miniaturization of IMS has been a long time goal in the mass spectrometry field because of the potential to develop handheld, field-applicable tools. In 1998 researchers at Draper studied FAIMS, also called the differential mobility spectrometer, as a portable tool for chemical detection, and developed a micromachined (MEMS) version of the parallel plate FAIMS[13, 14]. This miniaturized technology is currently being used in the Draper laboratory for a wide variety of applications. Using pyrolysis to prepare and introduce biological agents, signature biomarkers for chemical weapon agents such as anthrax spores were studied[10]. To facilitate FAIMS applications in the field, methods for statistical detection of contaminants using FAIMS outputs were explored in 2005[3].

3.1 Apparatus Setup

The FAIMS apparatus at Draper Labs used in this research consists of headspace sampling, gas-chromatography, ionization with Ni-63, and a MEMS differential mobility sensor (DMS), which is capable of detection at the part-per-trillion scale[10]. First, a sample of water with contaminants is heated and vapor from the headspace of each sample is injected into the gas chromatography column. Temporal separation results from how long the chemical sample takes, relative to water vapor that enters with the chemical, to pass through the column.

Analytes exiting the column are ionized by a radioactive nickel source before entering the DMS, where a pair of parallel plates apply an asymmetric waveform of alternating high and low electric fields perpendicular to the flow of the carrier gas. This field causes the analyte ions to drift perpendicular to the direction of the carrier gas.

As seen in Figure 3-1, the electric field is applied with an asymmetric high and low waveform. The net electric field over time is zero. An ion’s mobility can be approximated by $K(E) = K_0(1 + \alpha(E))$ where $K_0$ equals the low field mobility[14].
When $\vec{E}$ is large, the ion mobility will change depending on the factor $\alpha(\vec{E})$. Because of this unequal mobility in the alternating high and low electric fields, an ion will have a nonzero net drift toward one of the plates. Its movement in the direction of the electric field is $\Delta Y = K(\vec{E})Et$. Because $E_{lo}t_{lo} = E_{hi}t_{hi}$, the amount of drift depends on the low and high field mobilities for that ion. Figure 3-1 shows the direction of drift for two ion species that have different mobilities.

For a certain average field strength, some compounds will flow right through the channel, while other compounds will be absorbed into the plates. By applying a range of compensation DC fields, all ions can be made to either drift through to the sensor or into the parallel plates. Thus, FAIMS is capable of detecting and visualizing, at high temporal and spatial resolutions, ion intensities for samples that may have very low chemical concentrations.

Figure 3-2 shows a sample FAIMS spectrum. Time is on the X axis, and $t = 0$ is located where the main cluster for the ammonium artifact ends. The 100 values of compensation voltage are on the Y axis. The figure shows a raw FAIMS sample before processing (see Section 4.1.1).
3.2 Data Collection

3.2.1 Acquisition

Data samples were gathered by Melissa Krebs of Draper Labs from November 3, 2005 through November 15, 2005. All samples were taken with the same machine parameters. The headspace sampler, where the vials of liquid are stored and sampled, is in an oven at 60 degrees C. Samples of headspace are injected into the gas chromatography oven at 40 degrees C and held at that temperature for 30 seconds. The oven is then ramped up to 100 degrees C at 10 degrees/minute. Nitrogen is used as the carrier gas, at 1.5 ml/min through the GC column, and 300 ml/min in the main FAIMS flow. Each run spanned 100 compensation voltage values that ranged from -40V to 10V, for approximately 500 scans over approximately 10 minutes. The radio frequency of the alternating field is 1200 Hz.
<table>
<thead>
<tr>
<th>Chemical</th>
<th>Benzene</th>
<th>DCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td><img src="benzene_structure.png" alt="Chemical Structure" /></td>
<td><img src="dcm_structure.png" alt="Chemical Structure" /></td>
</tr>
<tr>
<td>Molecular weight</td>
<td>78.1134</td>
<td>84.9328</td>
</tr>
<tr>
<td>Water solubility</td>
<td>0.18 g/100 mL</td>
<td>1.32 g/100 mL</td>
</tr>
<tr>
<td>EPA Maximum Contaminant Level</td>
<td>5 ppb</td>
<td>5 ppb</td>
</tr>
</tbody>
</table>

Table 3.1: Benzene and Dichloromethane.

<table>
<thead>
<tr>
<th></th>
<th>Benzene</th>
<th>DCM</th>
<th>Water (control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>125 ppm</td>
<td>99</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>250 ppm</td>
<td>101</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>500 ppm</td>
<td>100</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>0 ppm (control)</td>
<td>-</td>
<td>-</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 3.2: Available data samples by class

### 3.2.2 Chemical Basis

Benzene and dichloromethane (DCM) are government regulated solvents that have a health risk if found at high levels in drinking water[4, 5]. Table 3.1 shows the chemical structure and various characteristics about each chemical.

It was found empirically that the chemicals were difficult to detect below concentrations of 50 ppm, and that 1000 ppm was too large to produce clean data. Thus, the three concentrations used were 125, 250, and 500 ppm for both chemicals. The data were gathered over a period of two weeks using the batch collection ability of the FAIMS setup. A total of 15 datasets were taken, with roughly 44 samples in each set consisting of about 21 benzene at various concentrations, 21 DCM at various concentrations, and 2 of pure water. In each set, the order of chemicals and concentrations was randomized so that drift and machine error would not bias the data; see Section 4.1 for more insight into the effects of FAIMS sensor drift. After data collection, carried out by Krebs, the datasets were consolidated into 7 data sets, consisting of benzene at 125 ppm, 250 ppm, and 500 ppm, DCM at 125 ppm, 250 ppm, and 500 ppm, and water as a control. Table 3.2 shows the number of samples available in these sets.
Chapter 4

Methodology

4.0.3 Two Classification Tasks

In our threefold approach, we seek to develop a binary classifier that can classify two classes of FAIMS data. The data is composed of two chemicals at three different concentrations each. This creates two classification tasks that will be examined. In the first task, we are classifying two different chemical contaminants that are at the same concentration. In the second task, we classify two concentrations of the same contaminant. Thus, the two tasks are:

1. Chemical Classification: chemical A vs chemical B, both at concentration $\alpha$.

2. Concentration Classification: chemical A at concentration $\alpha$ vs chemical A at concentration $\beta$.

In developing a algorithm to classify between two classes of FAIMS data, we consider a signal model based around a binary hypothesis test:

\[ H_1 : r(V_c, t) = s_1(V_c, t) + n_1(V_c, t) \]  
\[ H_2 : r(V_c, t) = s_2(V_c, t) + n_2(V_c, t) \]

In every classification task we assume that the output of the FAIMS sensor $r(V_c, t)$ consists of the signal $s_i(V_c, t)$ corresponding to a class of contaminant added to a
background of noise $n_i(V, t)$. This model can be further simplified by assuming that the noise is statistically identical in both cases, but this is not rigorously proven. Given a set of training samples, a classifier must be able to make a decision on whether an unknown test sample $r_{\text{unknown}}$ fits hypothesis $H_1$ or $H_2$. The threefold approach that has been developed for calibration contains the following three steps:

1. Calibration
2. Feature Extraction
3. Classification

**Chemical Classification**

First, we want to classify between the two types of chemical contaminants. The support vector machine will be trained on a set of benzene at concentration $\alpha$ and a set of DCM at the same concentration $\alpha$. The goal is to then be able to distinguish what type of chemical contaminant is in a sample of unknown chemical at that same concentration. We want to determine whether a classification algorithm can be developed that will classify between the datasets, and whether concentration affects the ability to classify.

Classifying between benzene and DCM at the same concentration may be very straightforward. First, each contaminant is pure; both are single symmetric molecules that do not have isomers - left or right handed configurations that can behave differently. In biomarker detection, FAIMS inputs consist of more complex molecules and mixtures that produce multiple-peaked signatures[10]. Benzene and DCM should both only produce one peak each, and will be detected by the FAIMS at different temporal ($t$) or spatial ($V_c$) locations, thus their signature spectra $s(V_c, t)$ should be easy to identify and distinguish.

**Concentration classification**

The second classification task we consider is to distinguish between samples of the same chemical contaminant at different concentrations. A set of samples at con-
centration $\alpha$ and a set of samples of the same chemical at concentration $\beta$ will be used as training data for the concentration classifier. We want to determine whether it is possible to classify between the same chemical at different concentrations, and whether the type of chemical or the relative concentrations of that chemical affect classification outcome.

Despite being the same chemical, FAIMS spectra of two classes at different concentrations may have features that are different in both temporal-spatial location and intensity. Depending on whether the FAIMS sensor responds linearly to concentration, concentration classification may involve both intensity and location. There is evidence that the FAIMS output for many chemicals does not vary linearly over large ranges of concentration[3].

4.1 Step 1: Calibration

As with many sensor instruments, the sensitivity and response of the FAIMS sensor can vary with time. No literature was found to discuss calibration techniques for FAIMS applications. For FAIMS spectra, we consider three sources of variation:

1. Temporal variation
2. $V_c$ or spatial variation
3. Intensity or baseline variation

In our analysis of experimental data, we observed that readings from a FAIMS measurement can fluctuate from day to day, or even over the course of an hour. Because the collection occurred over the course of about two weeks, with each batch collection taking several hours, sometimes overnight, it is necessary to address the variations in the raw data before doing any further analysis.

4.1.1 Preprocessing

Before any calibration, feature extraction, or any other analysis was performed, the data was processed to remove nonessential spectra. Due to the carrier gas, there
are several artifacts in the FAIMS spectra that would make feature extraction and calibration difficult. The nitrogen carrier gas produces a large, high intensity streak along the time axis at around compensation voltage $V_c=-15V$. The presence of this line can be used for spatial calibration by making sure it shows up in the same $V_c$ position over all samples.

When the initial amount of water vapor emerges, the FAIMS signal will show an ammonium artifact in which the nitrogen line disappears and there is a “dip” at $V_c=-21V$. This dip is visually helpful in locating the start time of each sample, where $t$ is set to 0. If locating the dip is accurate, using the ammonium artifact as a temporal reference point serves as a form of temporal calibration.

However, because the intensity of these two artifacts is so great, they are likely to completely outweigh any measured ion intensity from the contaminant. Thus, before calibration and feature extraction, the nitrogen line and ammonium dip are removed from the data. Figure 4-1 shows the spectra directly from the data, and after preprocessing, in which the signature of the measured chemical is much easier to see.
Variation in baseline intensity occurs due to drift of baseline values in FAIMS sensitivity. In preliminary studies, test datasets that were taken on a different day than training sets of the same class were misinterpreted as a different class. Thus, this baseline fluctuation has a significant impact on the success of classification between classes and must be removed. Figure 4-2 shows fluctuation of baseline values throughout the collected data in the benzene 125ppm set as seen by the changes in their marginals. The marginal of a FAIMS sample is calculated by integrating the FAIMS signal over time (Eq. 4.3). Figure 4-3 shows a boxplot of the intensities of a small subset of 200 randomly chosen feature for each run in the benzene 125ppm dataset, in which samples 86-96 show significant differences in mean intensity. The features comprise mostly of background noise, as represented by the white pixels in Figure 4-4.

\[ R_{Vc}(V_c) = \sum_t r(V_c, t) \] (4.3)

Another type of intensity variation is the change in signal to noise ratio (SNR). Baseline fluctuations do not simply add a constant value to the intensities in a signal. When the FAIMS sensitivity changes, it is possible that the strength of the measured signal changes, and thus the SNR will be different from sample to sample. This change may cause a classifier to wrongly distinguish subsets of data that belong to
Figure 4-3: Boxplot for benzene 250ppm intensities before baseline calibration.

Figure 4-4: Features measured for background intensity in boxplot.

Figure 4-5: Peak to noise ratio for benzene 250ppm runs before baseline calibration.
the same class, simply because they have lower measured peaks. This effect can be seen in the difference between the size of the largest peaks in Figure 4-2. Since the nitrogen artifact should be independent of time and chemical type, the nitrogen peak \( V_C = -16V \) should remain constant, but instead those runs with more negative baseline values seem to have smaller nitrogen peaks. As seen in Figure 4-5, these smaller peaks result in a smaller signal to noise ratio. Calibration will also try to address the fluctuations in peak to noise ratio.

To calibrate all the data, the wavelet transform was used for baseline removal. The wavelet transform is capable of decomposing an image into approximation and detail levels at different scales. At a high scale (N=4), the approximation is close to the DC value, but may also include other low frequency artifacts. Thus wavelet calibration involves simply removing the level 4 approximation of each signal, and keeping the detail signal. Wavelets are discussed in more detail in Section 4.2.1.

Thus, in order to calibrate FAIMS, the following four steps are taken:

1. Temporal calibration in reference to ammonium dip
2. Spatial calibration in reference to nitrogen line
3. Removal of nitrogen and ammonium artifacts
4. Intensity calibration by wavelet transform baseline removal

### 4.1.3 Alternative Calibration Techniques

There are other possible calibration techniques that were not explored due to time constraints. There is an additional set of FAIMS spectra for diluted water with no chemical added. It is possible that these samples could have helped in determining some baseline level of noise for baseline calibration. However, this type of calibration reference may be more suitable for a detection algorithm in which there is need to quantify the background statistics.

The use of the ammonium dip as a temporal reference is not ideal, because the method of determining exactly when the dip starts and ends is not accurate. Discussions at Draper brought up the use of an internal standard in the FAIMS. Such
standards would be well studied and known to be consistent in their emergence time. One such standard could be the chloride ion studied by Viehland, whose mobility was calculated using FAIMS to agree with well published drift tube mobility[19].

4.2 Step 2: Feature Extraction

Feature extraction seeks to isolate and extract features in FAIMS spectra for use in classification. A single FAIMS signal has up to 50,000 dimensions (500 temporal values, 100 $V_C$ values). The goal of feature extraction is to reduce the dimensionality as much as possible while capturing as much of the discriminating features of the FAIMS signal as possible. This final feature set should create the largest “distance” possible between the two classes that are to be classified. There are several steps to feature extraction: wavelet domain transformation, filter prescreening, and finally, feature extraction in both the pixel and wavelet domains.

4.2.1 Pixel Domain vs. Wavelet Transform

The most important features in a set of data are not always obvious from the pixel representation. Single pixels are not adequate features if, for example, a chemical appears in the FAIMS spectrum over a large region, or if temporal variations remain after calibration. However, in the wavelet transform domain, it is possible to isolate and extract features that are larger than one pixel in size, and therefore more robust to system imperfections.

Wavelets[11] are compactly supported, time limited signals that can be dilated and translated to form a multiresolutional basis, which is then used to decompose the image in both the time and the frequency domains. The Haar wavelet, also known as the Daubechies wavelet $db1$, is shown in Figure 4-6. It is a step function, the simplest of the mother wavelet functions[15].

Wavelet coefficients are produced by correlating the pixel image with scaled and translated versions of the mother wavelet:
$\Psi_{a,b} = a^{\frac{1}{2}} \Psi \left( \frac{\lambda - b}{a} \right)$ \hfill (4.4)

Whereas in the pixel domain, features selected are single pixels with high differentiation potential, the wavelet transform coefficients correspond to areas of different sizes, matching up with the basis functions at different scales.

In image processing, the 2D discrete wavelet transform is used to filter an image both horizontally and vertically into an approximation signal $\tilde{A}$ and three detail signals $\tilde{D}h$, $\tilde{D}v$ and $\tilde{D}d$. $\tilde{D}h$ corresponds to filtering for the high horizontal and low vertical frequencies of the image, $\tilde{D}v$ are the low horizontal and high vertical frequencies, and $\tilde{D}d$ are the high horizontal and vertical frequencies. In multiple level deconstructions, the approximation signal is then decomposed again to produce a smaller approximation signal and three corresponding detail signals. Figure 4-7 shows the pyramid structure of the deconstruction. The original image can be completely reconstructed using the highest level approximation plus every detail coefficient.

By using the wavelet transform to analyze the data, an alternate basis is provided from which to select features for classification. For each set of classification tasks, both pixel and wavelet data will be used for feature extraction and classification. Note that in this paper “data” and “features” will refer to either wavelet coefficients or pixel intensities if not specified, as an orthonormal wavelet decomposition yields the same number of transform domain coefficients as the pixel representation.
4.2.2 Feature Prescreening

In both the pixel and wavelet domains, FAIMS spectra are still represented as 50,000 coefficients. We would like to screen out a large number of features that have no value for classification. If features that are unlikely to be distinguishing are first removed, the feature selection process will be much faster with little or no loss of performance. An initial stage of prescreening is done to select the top 1000 features that have the greatest distance between the two training classes and from which the final subset of features will be chosen. We use a simple metric for gauging the separability of two classes that is related to Fisher’s Linear Discriminant (4.5):

\[ D(i) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \]  

We will refer to this separability measure as the normalized univariate class separability (NUCS). For the \( i \)th feature, \( \mu_1 \) and \( \sigma_1 \) are its mean and variance in class 1, and \( \mu_2 \) and \( \sigma_2 \) are its mean and variance in class 2. This measure is applied to both the pixel and wavelet coefficients and selects for features that have a large distance measure between the two classes. The remaining 1000 features provide a much smaller search space for the main feature extraction algorithms.
4.2.3 Feature Extraction Algorithms

Filters vs. Wrappers

Four feature extraction algorithms have been developed to extract features from training data. In all feature selection algorithms, features are selected by evaluating their effectiveness using various criterion functions. If the set of all features is $X$ and the criterion function for a subset of features $S$ is $J(S)$, feature extraction algorithms will try to find the set $S \subseteq X$ of length $d$ to maximize its criterion value:

$$J(S) = \max_{S \subseteq X, |S|=d} J(S)$$

(4.6)

The criterion used divides the algorithms into two classes: wrappers and filters. Wrappers are feature selection algorithms in which the induction algorithm, or the classifier, is used as part of the evaluation criterion for the features[9]. In training the classifier, the training data is divided into subsets that are used to cross validate the feature sets selected. The feature set with the highest classification score is selected. Filter algorithms, on the other hand, use a separability measure to select features independently of the classifier. They are independent of a specific classification algorithm and thus are very versatile, but wrappers can have better performance if they select features relevant to the classification algorithm[9]. Figure 4-8 shows the process in which features are extracted using a wrapper method, compared with the filter method.

The four feature extraction algorithms implemented are:

- Magnitude Signature
- Sequential Forward Selection (SFS)
- SVM-aided Recursive Feature Elimination (RFE)
- SVM-aided Recursive Feature Selection (RFS)

The first feature extraction algorithm is a filter method called the magnitude signature filter. Its criterion function is simply the magnitude of the coefficients, and the largest coefficients are ranked higher. Sequential forward selection is a filter
method that was first proposed by Whitney [20]. We combine its forward selection algorithm with the divergence scoring metric (see Sec. 4.2.3). The last two methods, Recursive Feature Elimination and Recursive Feature Selection, are wrapper methods that use SVMs as the criterion. SVM-aided Recursive Feature Elimination (RFE), which eliminates the lowest ranked feature at each iteration has been used in gene expression feature selection[7]. SVM-aided Recursive Feature Selection (RFS) is a technique derived from RFE, and selects features in a forward manner similar to SFS.

**Magnitude Signature Filter**

The magnitude signature feature extraction method assumes that in the absence of chemical contaminant, the noise signature consists of white noise of very low intensity. When a chemical contaminant does emerge, its signature can be identified by finding the features with the largest magnitude. However, the overall feature set of each training class \( X \) is selected among the sample runs \( x \) by finding the features that are selected the most often by each sample. For each class, half the total number of features are selected, then the features from both classes are combined. All features
are ranked individually, thus the criterion function (Eq. 4.8) is an individual score, and feature scores are not evaluated as a group.

For each data run, the top $K$ features with the greatest squared magnitude are saved, represented by $Features(X)$. Define a function $SelectedBy(x)$ to equal 1 when a feature $i$ is one of the features selected by $x$:

$$SelectedBy(x) = \begin{cases} 1 & \text{if } i \subseteq Features(x) \\ 0 & \text{otherwise} \end{cases}$$

(4.7)

To find the signature set for the class as a whole, the features are ranked by how many times they have been selected in that set, so the scoring function for each feature is:

$$J(i) = \sum_{x \subseteq X} SelectedBy(x)$$

(4.8)

It follows that the higher the value of this criterion function, the more important the feature is for the set.

When selecting for $N$ features to distinguish two classes, each class contributes half of the feature set, or $\frac{N}{2}$ features. This way, features prominent in either class will be chosen as a distinguishing feature. It may be possible to adjust the number of features each class is allocated, if any knowledge exists about the density of features in each class. Other parameters that can be adjusted include the number $K$ of top ranking features each run selects, which can affect the stringency of the feature selection process. In this thesis, $K = 100$. A lower number of features selected can result in a feature set that appeared in only a low percentage of runs, but a higher number may result in features that are not distinguishing being selected.

**Sequential Forward Selection**

The sequential forward selection algorithm (SFS), first put forward by Whitney in 1971, is a simple feature selection algorithm that uses the divergence criterion to score the effectiveness of a feature set. Divergence (Eq. 4.10) is a measure of the distance between two classes based on their mean and variances, and was first used by Marill
and Green[12] in 1963 for their sequential backward selection, a similar but slower algorithm.

SFS assumes that the classes of data are normal variables with unequal covariance. The divergence of two such classes, \( i \) with density \( p_i \) and \( j \) with density \( p_j \), is defined as the difference in log likelihood of the two classes given feature set \( S \) [12]:

\[
\text{Div}(i, j|S) = E \left[ \log \left( \frac{p_i(S)}{p_j(S)} \right) \right]_i - E \left[ \log \left( \frac{p_i(S)}{p_j(S)} \right) \right]_j \tag{4.9}
\]

If \( p_i \) is normal with mean \( M_i \) and covariance matrix \( \Sigma_i \), and \( p_j \) is normal with mean \( M_i \) and covariance matrix \( \Sigma_j \), the divergence can be calculated as:

\[
\text{Div}(i, j|S) = \frac{1}{2} \text{tr}[(\Sigma_1 - \Sigma_2)(\Sigma_1^{-1} - \Sigma_2^{-1})] + \frac{1}{2} \text{tr}[(\Sigma_1 + \Sigma_2)(M_1 - M_2)(M_1 - M_2)^T] \tag{4.10}
\]

For SFS, the criterion function for a set is the divergence, but features are added based on their incremental score. The incremental score of a feature \( i \) is the difference between the score of the existing feature set \( S \) and the new feature set including the new feature, \( S' = S \cup i \):

\[
\Delta J(i) = J(S') - J(S) \tag{4.11}
\]

To find the best feature set, SFS starts with an empty feature set \( S \). In each iteration, it finds the one feature that has the greatest incremental score from the set of unselected features, and adds it to the existing feature set. The process is repeated with the new feature set until either the desired number of features is reached or the score can no longer increase.

The number of features is limited to the size of the training set, because the criterion function, divergence, requires invertible covariance matrices. Sequential forward selection is also a suboptimal algorithm because once a feature is selected, it cannot be removed. The optimal feature set of size 5, using the divergence criterion, may not include all the features that are in the optimal feature set of size 4, but SFS does
not provide a way to reevaluate features once they are selected, and thus may get stuck in local maxima. Pudil[16] has proposed a floating method that will reevaluate features to find better subsets, but the algorithm is beyond the scope of this thesis.

**SVM-aided Recursive Feature Elimination**

RFE, first proposed by Guyon et al.[7], uses the weights of a support vector machine classifier trained on the two sets to be classified. This is a wrapper method that includes the classifier as feedback for iteratively selecting the feature set. A SVM-aided feature selection algorithm is ideal because the classification step of the threefold algorithm will also utilize SVMs. For a set of features, the RFE criterion function is 

\[ J(S) = \frac{1}{2}||\vec{w}||^2 \]

so the incremental score for each feature is:

\[ \Delta J(i) = w_i^2 \]  \hspace{1cm} (4.12)

In SVM classification, the vector \( \vec{w} \) is a weight vector from the linear combination of the training data that corresponds to the weight that each feature should have in the decision function. See section 4.3 for a detailed review of SVMs.

The set criterion function \( J(S) \) is related to the size of the margin between the two classes as calculated by their support vectors. The contribution of each feature to this distance is simply the square value of the weight vector for that feature. The selection process for recursive feature elimination is to start with the full feature set, find the feature that has the lowest incremental criterion score, and eliminate it by removing it from the feature set. When the desired number of features is reached, the set consists of features that should have the greatest weight on classification.

**SVM-aided Recursive Feature Selection**

Recursive feature selection was developed to demonstrate a forward selection version of the RFE algorithm. RFS is similar to the elimination algorithm except feature selection is performed in a forward manner. Thus, after each iteration, the highest weighted feature is added to the feature list, and the criterion function is evaluated
This method differs from RFE in several ways. First, elimination of a low ranked feature in RFE ensures that the remaining feature set is optimal as a feature subset. These features are not individually most relevant, but only relevant because the feature that was removed was the least weighted. RFS selects features more individually, selecting the highest weighted feature from a group that does not include the features already selected. Thus, individually each feature is significant, but whether the features are a distinguishing as a set is unknown.

However, there are advantages to RFS. First, it is much faster. Even with pre-screening, there are still a thousand individual features left to select from; RFE would rank every one before it selects the top 50. Empirical results also show that features selected by RFS perform better in some cases than RFE.

### 4.2.4 Quantifying Degree of Separation

The success of the calibration, feature extraction and classification can be evaluated by the percent of correct classification ($P_{CC}$). However, it is also possible to quantitatively score the features selected before submitting them to the SVM classifier. Good feature selection algorithms will pick features that enable a classifier to easily distinguish between two classes. Without specifying any particular classifier, good features maximize the distance between the two classes, and this distance can be measured by various objective functions.

**Normalized Univariate Class Separability**

One way to measure separation is to use the normalized univariate class separability (NUCS), described under section 4.2.2. Equation 4.5, repeated below, was used to screen features before feature extraction. It can be used again to measure the distance between extracted features:

$$D(i) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$  \hspace{1cm} (4.13)
For a feature \( i \), Eq. 4.13 measures the separation between the two classes based on the signal to noise ratio for each particular feature. NUCS assumes normal distributions with means \( \mu_1 \) and \( \mu_2 \) and variances \( \sigma_1^2 \) and \( \sigma_2^2 \) for these variables. This measure can also be applied to the set of features by summing the individual distances for each feature, but it is important to notice that this technique does not use a covariance for the feature set, and hence does not account for correlation among features.

**SVM Margin Distance**

In a support vector machine trained on two datasets, there exists a natural measure of distance between the two classes that is derived from the greatest distance between its support vectors:

\[
D(S) = \frac{1}{||\vec{w}||} \tag{4.14}
\]

\[
\vec{w} = \sum_{x \in X} \alpha \cdot \vec{y} \cdot x \tag{4.15}
\]

In equation 4.14, the distance is the inverse of the magnitude of \( \vec{w} \), which is a weighted summation of the training samples in training set \( X \). Each sample has a weight \( \alpha \) associated with it, and a category in the vector \( \vec{y} \in \{-1, +1\} \). The margin between the hyperplanes that separate the two classes is \( \frac{2}{||\vec{w}||} \) [1] so the measure of SVM distance is a measure of the maximum separation between the two classes. The support vector machine classifier is discussed in more detail in section 4.3.

**4.3 Step 3: Classification**

The third step and final goal of this thesis is to provide a classification tool that will distinguish between two sets of contaminants based on the features selected from the calibrated FAIMS data. The previous section outlined algorithms that are able to extract a set of features that distinguish one class from another as much as possible. These features are now submitted to the support vector machine (SVM) classifier for
training, and subsequently, for classification with test data.

4.3.1 Motivation

SVMs were a powerful development in machine learning that was invented by Vladimir Vapnik in the late 1970s. They are commonly used in pattern recognition applications, and have been used for face detection, object recognition, and even stem cell culture classification[8]. Unlike classic Bayesian classifiers, SVMs are effective on non-parametric data, in which no data model or distribution is assumed. Furthermore, a trained support vector machine will be able to classify data only based on its support vectors, training data that fall on the optimal hyperplane boundaries (see following section), so its speed scales with the training data, not the dimensionality of the data. Thus, it is also possible to use various kernels to map the data to higher dimensions, which assists in classification and gives SVMs a large amount of flexibility and power.

4.3.2 Classifier Formulation

Given the samples from two classes, represented as vectors $\mathbf{x}_i \subseteq \mathbb{R}^d$ and their corresponding class labels $y_i \subseteq \{-1, +1\}$, the support vector machine will find the hyperplane with normal vector $\mathbf{w}$ that will linearly separate the samples in the two classes. Namely, the normal $\mathbf{w}$ is found so that:

$$
\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \quad \text{for } y_i = +1
$$

$$
\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \quad \text{for } y_i = -1
$$

or

$$
\hat{y}(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \quad \text{for all } i
$$

The detector output of an SVM classifier is a class label for each test data sample, generated from the sign of its dot product with the discriminating hyperplane:
\[ h(\vec{x}) = \vec{w} \cdot \vec{x} + b \]  
\[ f(\vec{x}) = \text{sign}(h(\vec{x})) \] (4.18)

There also exist two hyperplanes that contain the training points such that the equality in 4.17 holds. These two hyperplanes are parallel on either side to the discriminating hyperplane, and no training points fall between them. The points \( \vec{x}_i \) that lie on either of the planes are called the support vectors of the machine. The SVM classifier is found so that the distance between these two planes is maximized. This is the margin of the SVM, given by the difference in their perpendicular distance from the origin[1]:

\[ d = d_+ - d_- = \frac{2}{\|\vec{w}\|} \] (4.21)

Maximizing 4.21 is the same as minimizing \( \|\vec{w}\|^2 \). This corresponds with using \( J(S) = \frac{1}{2} \|\vec{w}\|^2 \) as the ranking criterion for features. In RFE, by removing features that have the lowest incremental score, the feature set is being reduced while maintaining as much of the original hyperplane as possible, which has been optimized for margin. In RFS, selecting features that have a large score attempts to attain the optimized hyperplane at each iteration.

Finding the optimal hyperplane is an optimization problem that can be solved using the Lagrange formulation of the problem[1]. The Lagrange multipliers \( \alpha_i \) give the following condition for the hyperplane:

\[ \vec{w} = \sum_i \alpha_i y_i \vec{x}_i \] (4.22)

\[ \sum_i \alpha_i y_i = 0 \] (4.23)

This allows the output (Eq. 4.18) to be rewritten[8]:

43
\[ h(\vec{x}) = \sum_i y_i \alpha_i (\vec{x} \cdot \vec{x}_i) + b \]  
\[ f(\vec{x}) = \text{sign}(h(\vec{x})) \]

\section*{4.3.3 Kernel Selection}

Eq. 4.25 assumes that all training points are linearly separable in their original space, and thus a hyperplane may be found by the linear combination of training data. However, if the data is not separable, that is, a decision function that is a linear function of the data does not exist, SVM kernels can be used to map the data into a higher dimension. If there is a mapping such that\([\text{I}]:\)

\[ \Phi : \mathbb{R}^d \rightarrow H \]  
\[ K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j) \]

then a discriminating hyperplane may be found in the new higher dimensional space \(H\). For example, if a set of points in 2D space are not separable, but are mapped to a sphere in 3D space, then the plane may be found that will cleanly separate the points. The kernel used for the mapping \(\mathbb{R}^2 \rightarrow \mathbb{R}^3\) in the example in Figure 4-9 is:

\[ \Phi = \left\{ \begin{array}{c} x \\ y \\ \sqrt{2 - x^2 - y^2} \end{array} \right\} \]  

Because of the kernel mapping, SVMs are very versatile. The output function becomes the sign of the kernel-mapped data (Eq. 4.29) where the set \(X\) are the support vectors of the SVM.

\[ f(\vec{x}) = \sum_{i \in X} \alpha_i y_i K(\vec{x}_i, \vec{x}) \]  

Three kernels are to be used for classifying features extracted from Section 4.2.3. For
each classification task, all three kernels are used. The kernels are the linear kernel (Eq. 4.30), polynomial kernel (Eq. 4.31), and gaussian radial basis function kernel (Eq. 4.32). For the polynomial kernel, the default order for parameter \( p \) is 3.

\[
K(\bar{x}_i, \bar{x}_j) = \bar{x}_i \cdot \bar{x}_j \tag{4.30}
\]

\[
K(\bar{x}_i, \bar{x}_j) = (\bar{x}_i \cdot \bar{x}_j)^p \tag{4.31}
\]

\[
K(\bar{x}_i, \bar{x}_j) = e^{-\frac{||\bar{x}_i - \bar{x}_j||^2}{2\sigma^2}} \tag{4.32}
\]

### 4.3.4 Evaluation of Classification

The success of the classifier is measured by the percent of correct classification \((P_{CC})\), which is a ratio of correctly classified test samples to total test samples. In order to provide training and test datasets, each classification task is carried out using 3-fold crossvalidation. For each trial, one third of the available data in each set are set aside as “unknown” test data, and the other two thirds are kept for feature extraction and classifier training. For each crossvalidation trial, the \( P_{CC} \) is calculated for the test data by finding the ratio of correctly classified test data to the size of the test data.
Figure 4-10: k-Nearest Neighbors in 2 dimensional space, using three nearest neighbors for class identification.

For each classification task, the three crossvalidation trials are averaged to produce the overall percent of correct classification.

A k-Nearest Neighbors (kNN) classifier was also used as a comparison for the classifier. The kNN classifier is a very simple, nonparametric classifier, in which the k neighbors with the smallest distance from a sample data are used to determine that sample's identity. kNN is simple and quick, and requires neither a signal model for a dataset, nor feature extraction algorithms to lower the dimension. We use a kNN classifier which has k=3, and uses the Euclidean distance metric, to act as a baseline for classification.

In a 2-dimensional classification task, the kNN algorithm can be visualized as locating the k neighbors that are closest to a test data point, then deciding the identity of the test point based on the class of the majority. The kNN algorithm uses Euclidean distance as the distance metric, and no feature extraction is performed, so all 50,000 or so points are used. Figure 4-10 shows the decision making process in which three neighbors are being used to identify the class of a test point (the square).
Chapter 5

Results and Discussion

We wish to evaluate the effectiveness of each step in the threefold approach. First the effectiveness of calibration will be discussed. Then, both chemical and concentration classification will be discussed, and the ability for classification in both pixel and wavelet domains will be examined. Because feature extraction algorithms can also be measured by feature set scores using the NUCS Score and SVM margin, we will also compare $P_{CC}$ values to feature set scores to see if any correlation can be found.

5.1 Calibration

Figures 5-1 to 5-6 show the mean spectra for each class of chemical after calibration: benzene 125ppm, benzene 250ppm, benzene 500ppm, DCM 125ppm, DCM 250ppm, and DCM 500ppm. The mean plot for pure water is also shown (Figure 5-7). For the first five classes, the chemical signal is visibly different from the pure water plot. Furthermore, in benzene, it seems that with increased concentration, the region of ion intensity becomes less uniform. Figure 5-8 shows two samples, one from benzene 125ppm, and one from benzene 500ppm, in which the signal seems to have changed or split in the more concentrated sample. Thus, benzene at lower concentrations seems to have a stronger or more consistent signal than at higher concentrations.

A similar change with concentration occurs in the DCM samples. In the 125ppm set (Figure 5-4), the mean signal is solid around $V_C = -15$, but at 250ppm, the
Figure 5-1: Mean FAIMS Spectrum for benzene 125ppm.

Figure 5-2: Mean FAIMS Spectrum for benzene 250ppm.
Figure 5-3: Mean FAIMS Spectrum for benzene 500ppm.

Figure 5-4: Mean FAIMS Spectrum for DCM 125ppm.
Figure 5-5: Mean FAIMS Spectrum for DCM 250ppm.

Figure 5-6: Mean FAIMS Spectrum for DCM 500ppm.
Figure 5-7: Mean FAIMS Spectrum for water.

Figure 5-8: Benzene 125ppm and 500ppm signal comparison.
Figure 5-9: Marginal spectra for benzene 250ppm before and after baseline calibration. Variation in baseline fluctuations are removed.

Figure 5-10: Boxplot of 200 pixel intensities in benzene 125ppm before and after baseline calibration. Variation in baseline fluctuations are removed.

signal is more spread out over the compensation voltage axis. There is also very little visible signal in the 500ppm dataset (Figure 5-6). These qualitative changes in signal shape and location are useful features that can hopefully be exploited by the threefold approach.

As discussed in Sec. 4.1.2, the baseline intensity was calibrated by removing the approximation coefficients from the four level wavelet decomposition of each sample run. This removed baseline drift throughout the datasets.

Figure 5-9 shows the marginal $V_C$ intensities for all runs in the benzene 250ppm dataset before and after baseline drift calibration. The colors indicate samples that
Figure 5-11: Peak to variance ratio in benzene 250ppm before and after calibration.

were collected in the same batch on the same day; the baseline variation from day to
day before calibration is removed after calibration. Thus, class-differentiating effects
of relative intensity from sample to sample are minimized. Figure 5-10 shows the
box plot of a set of 200 arbitrarily selected points in each sample before and after
calibration. Again, variation in the mean intensity disappears.

Calibration also improves the uniformity of signal to noise ratio (SNR). Figure 5-
11 shows the ratio of the measured nitrogen peak magnitude to noise before and after
calibration. Although the nitrogen artifact is removed for actual data processing, it is
a good peak for demonstrating signal to noise ratio changes because it is independent
of the FAIMS sample. Figure 5-11 shows that before using wavelets to remove the
baseline fluctuations, the peak to noise ratio varies greatly over time. After wavelet
calibration, the peak to noise ratios are more consistent.

5.2 Classification Results

First, we review the two main classification tasks, which are:

1. Classify two chemicals at the same concentration
2. Classify two concentrations of the same chemical
We will compare the classification of both wavelet and pixel data by the four feature selection algorithms. Table 5.1 summarizes the classification tasks that were performed. In each classification task, three-fold crossvalidation was carried out by withholding 30 runs from each sample set for testing, and using the sixty remaining samples as training data for each class. This was repeated three times so that a third of the 90 runs used in each dataset would act as test data in each trial. For each task, feature extraction was carried out with each of the four algorithms, and classification was done with each of the three kernels.
5.2.1 Classification Tasks

Chemical Classification in Pixel Domain

The percent of correct classification between two chemicals at the same concentration varied from just above 50% in a few cases to around 90%. k-Nearest Neighbors (kNN) was performed as well as a baseline comparison for classification. In most cases, the $P_{CC}$ from extracted features performed as good as or better than kNN, which used the nearest 3 neighbors ($k = 3$) and always used the complete feature set. See Appendix A for full $P_{CC}$ results for every task.

For chemical classification in the pixel domain, we will look at two examples: benzene vs. DCM at 250ppm, and benzene vs. DCM at 500ppm. Figure 5-12 shows the $P_{CC}$ for benzene and DCM at 250 ppm, which was typical of the chemical classification tasks. The kNN classifier performed worse than all SVM classifiers. The highest $P_{CC}$ was 84.4% using a linear SVM kernel on 50 RFE extracted features. With
Table 5.2: Feature scores for classification of benzene 250ppm and DCM 250ppm in the pixel domain, along with the mean $P_{CC}$ over the kernels in crossvalidation 3.

<table>
<thead>
<tr>
<th>Extraction Algorithm</th>
<th>NUCS Distance</th>
<th>SVM Margin</th>
<th>Mean Kernel $P_{CC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude</td>
<td>1.160</td>
<td>0.271</td>
<td>90.6</td>
</tr>
<tr>
<td>SFS</td>
<td>0.399</td>
<td>0.126</td>
<td>88.9</td>
</tr>
<tr>
<td>RFE</td>
<td>0.345</td>
<td>0.257</td>
<td>86.7</td>
</tr>
<tr>
<td>RFS</td>
<td>0.735</td>
<td>0.278</td>
<td>97.8</td>
</tr>
<tr>
<td>Noise</td>
<td>0.021</td>
<td>0</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 5.3: Feature scores for classification of benzene 500ppm and DCM 500ppm in the pixel domain, along with the mean $P_{CC}$ over the kernels in crossvalidation 3.

<table>
<thead>
<tr>
<th>Extraction Algorithm</th>
<th>NUCS Distance</th>
<th>SVM Margin</th>
<th>Mean Kernel $P_{CC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude</td>
<td>0.150</td>
<td>0.055</td>
<td>65.0</td>
</tr>
<tr>
<td>SFS</td>
<td>0.170</td>
<td>0.052</td>
<td>53.3</td>
</tr>
<tr>
<td>RFE</td>
<td>0.148</td>
<td>0.022</td>
<td>64.4</td>
</tr>
<tr>
<td>RFS</td>
<td>0.191</td>
<td>0.125</td>
<td>75.0</td>
</tr>
<tr>
<td>Noise</td>
<td>0.024</td>
<td>0</td>
<td>n/a</td>
</tr>
</tbody>
</table>

the threefold algorithm, it is thus possible to classify between the two contaminants with a relatively high rate of success.

Benzene and DCM generally have strong FAIMS signals that emerge in different temporal and spatial locations. However, in Figure 5-13 the $P_{CC}$ for classification is low for kNN and all feature extraction algorithms except for RFS. A possible hypothesis is that the signal strength in both chemicals at 500ppm is much weaker compared with signal strengths in other concentrations, and thus classifying those two chemicals is more difficult.

To test this hypothesis, we examine the features extracted for these two classification tasks and the distance measures for those feature sets. In the three crossvalidation trials, the third trial generally had the highest $P_{CC}$, so all features in following analyses are selected from the samples in crossvalidation 3. Figure 5-14 shows the features that were selected for classification at 250ppm, superimposed on a combined benzene 250ppm and DCM 250ppm spectrum obtained by averaging the two datasets.

56
Figure 5-13: $P_{cc}$ for benzene 500ppm vs DCM 500ppm in the pixel domain.
Figure 5-14: Features selected for benzene 250ppm and DCM 250ppm in the pixel domain. 

together. The bright white pixels represent the location of selected features. In all feature extraction algorithms, a good number of the extracted features include the large region of benzene signal around $(V_C, t) = (-10, 75)$. It seems that a high $P_{CC}$ results from being able to identify features in the benzene region in particular.

Table 5.2 has the feature set scores, as calculated by the NUCS score and the SVM margin. The scores for magnitude signature selection and RFS are both relatively high, as reflected in their $P_{CC}$. Also in the table are scores for 50 points of noise selected at random from throughout the signal; the low score for those features reflects that a higher score in both categories corresponds to higher $P_{CC}$. Also shown in the table is the mean $P_{CC}$ of the third crossvalidation trial, averaged over the three kernels. The NUCS scores for magnitude signature and RFS reflect their high $P_{CC}$. However, scores do not always correspond to classification success; they are only a qualitative measure (see Sec. 5.3).

Figure 5-15 shows the features selected for classification at 500ppm superimposed over a combined benzene 500ppm and DCM 500ppm signal. Both the benzene and
Figure 5-15: Features selected for benzene 500ppm and DCM 500ppm in the pixel domain.

DCM signals were weak at 500ppm and lacked a large central ion intensity region that were present in lower concentrations, so the $P_{cc}$ is lower. In all four extraction algorithms, the features selected were from a similar region. Table 5.3 show the feature scores for these features, as calculated by the quantifiers. RFS had a slightly higher NUCS distance, and a much higher SVM margin than the other three algorithms, and thus it performed better in classification. However, Table 5.2 had higher overall scores than Table 5.3, reflecting the higher rate of classification for 250ppm than 500ppm.

**Concentration Classification in Pixel Domain**

In general, $P_{CC}$ was much lower for concentration classification. In fact, for benzene, all three comparisons performed worse than k-Nearest Neighbors, and the feature extraction did not seem to find features that improved classification. For both tasks that classified adjacent concentrations (benzene 125ppm vs. 250ppm, benzene 250ppm vs. 500ppm), the $P_{CC}$ averaged 50%, so classification was not possible. Figure 5-16 shows the $P_{CC}$ for benzene 125ppm and 500ppm, which fared better. Figure 5-17 show the
features selected by RFS, which had the highest result of the four feature extraction algorithms. Table 5.4 shows the scores for those selected features. RFS had slightly higher NUCS distance, but the table scores do not really reflect the $P_{CC}$.

For DCM concentration classification, both tasks classifying between DCM 500ppm and another concentration averaged around 70% correct classification, and performed about as well as kNN. Figure 5-18 shows the $P_{CC}$ for DCM 250ppm vs. 500ppm, with an average of 70% correct classification. This may be because DCM’s signal in the FAIMS at 500ppm is significantly different than at other concentrations. Figure 5-19 show the features selected for that classification task, and Table 5.5 show the scores for those selected features.

In Table 5.5, the NUCS score for magnitude signature features were slightly higher. This may have contributed to the slightly higher $P_{CC}$ of the magnitude signature features. Figure 5-19 shows the features that were selected by all four algorithms. The magnitude signature features are strongly concentrated in the expected region.
Figure 5-17: Features selected for benzene 125ppm and benzene 500ppm in the pixel domain.

<table>
<thead>
<tr>
<th>Extraction Algorithm</th>
<th>NUCS Distance</th>
<th>SVM Margin</th>
<th>Mean Kernel $P_{CC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude</td>
<td>0.183</td>
<td>0.032</td>
<td>74.4</td>
</tr>
<tr>
<td>SFS</td>
<td>0.166</td>
<td>0.044</td>
<td>58.9</td>
</tr>
<tr>
<td>RFE</td>
<td>0.181</td>
<td>0.046</td>
<td>72.2</td>
</tr>
<tr>
<td>RFS</td>
<td>0.260</td>
<td>0.023</td>
<td>68.3</td>
</tr>
<tr>
<td>Noise</td>
<td>0.020</td>
<td>0</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 5.4: Feature scores for classification of benzene 125ppm and benzene 500ppm for pixel data, along with the mean $P_{CC}$ over the kernels in crossvalidation 3.

<table>
<thead>
<tr>
<th>Extraction Algorithm</th>
<th>NUCS Distance</th>
<th>SVM Margin</th>
<th>Mean Kernel $P_{CC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude</td>
<td>0.268</td>
<td>0.004</td>
<td>83.3</td>
</tr>
<tr>
<td>SFS</td>
<td>0.147</td>
<td>0.030</td>
<td>67.8</td>
</tr>
<tr>
<td>RFE</td>
<td>0.178</td>
<td>0.040</td>
<td>76.1</td>
</tr>
<tr>
<td>RFS</td>
<td>0.238</td>
<td>0.017</td>
<td>78.9</td>
</tr>
<tr>
<td>Noise</td>
<td>0.019</td>
<td>0</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 5.5: Feature scores for classification of DCM 250ppm and DCM 500ppm in the pixel domain, along with the mean $P_{CC}$ over the kernels in crossvalidation 3.
Figure 5-18: $P_{CC}$ for DCM 250ppm vs DCM 500ppm in the pixel domain.
Figure 5-19: Features selected for DCM 250ppm and DCM 500ppm in the pixel domain.

for DCM ion intensity, and thus those features have a higher NUCS score. However, the SVM margin is very low and does not represent its high $P_{CC}$. Because magnitude signature is a filter instead of a wrapper, the features may not be optimal for a SVM score, because they do not correspond to the most relevant features for SVM scoring. The NUCS score, in this case, reflects the relative success of the four algorithms better than the SVM margin.

**Chemical Classification in Wavelet Domain**

We will now look at the results of classification using wavelet data. In several classification tasks, classification was significantly better for wavelet data than for pixel data. In particular, chemical classification between benzene and DCM at 125ppm was in the high 90’s, and concentration classification of DCM 125ppm vs. 500ppm was also in the 70% to 80% range, much higher than in the pixel domain.

Figure 5-20 shows the $P_{CC}$ results for benzene and DCM at 125ppm. RFE features had an average of 90% correct classification. Figure 5-21 shows a sample reconstruction of the FAIMS signal using the selected features. In the reconstruction, all wavelet
Figure 5-20: $P_{CC}$ for benzene 125ppm vs DCM 125ppm in the wavelet domain.
coefficients were set to 0 except for the selected features. This type of visualization serves as a good representation of the temporal-spatial location of features, as they are no longer single pixels but compactly supported regions in the spectrum.

For the RFE wavelet extracted features, the reconstruction shows that the features are located near expected benzene and DCM signal regions. For SFS, which performed much lower, only one of those areas is present (the feature around \((V_c, t) = (-17, 40)\) is gone) and many other features are scattered about the rest of the spectrum. Table 5.6 shows the feature set scores for each feature extraction algorithm. SFS does have the lowest score in both distance metrics, and RFE had a high SVM margin. The Mean Kernel \(P_{CC}\) column in the table, specific to the third crossvalidation trial, reflects that magnitude signature and RFS, which had the highest NUCS and SVM scores, did have the highest \(P_{CC}\) in this task.

For other chemical classification tasks, \(P_{CC}\) went down with increasing concentration. It is interesting that for all four feature extraction algorithms there was a consistent change in \(P_{CC}\) with concentration.
Table 5.6: Feature scores for classification of benzene 125ppm and DCM 125ppm in the wavelet domain, along with the mean $P_{CC}$ over the kernels in crossvalidation 3.

<table>
<thead>
<tr>
<th>Extraction Algorithm</th>
<th>NUCS Distance</th>
<th>SVM Margin</th>
<th>Mean Kernel $P_{CC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude</td>
<td>1.5</td>
<td>0.143</td>
<td>98.3</td>
</tr>
<tr>
<td>SFS</td>
<td>0.173</td>
<td>0.077</td>
<td>81.7</td>
</tr>
<tr>
<td>RFE</td>
<td>1.18</td>
<td>0.225</td>
<td>96.7</td>
</tr>
<tr>
<td>RFS</td>
<td>1.21</td>
<td>0.343</td>
<td>98.3</td>
</tr>
<tr>
<td>Noise</td>
<td>0.017</td>
<td>0</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Figure 5-22: $P_{CC}$ for DCM 125ppm vs DCM 500ppm in the wavelet domain.
Concentration Classification in Wavelet Domain

Figure 5-22 shows a relatively high $P_{CC}$ for concentration classification between DCM 125ppm and 500ppm. For concentration classification using wavelet data, all benzene classification tasks and DCM 125ppm vs. 250ppm performed relatively poorly much like the pixel data, but both DCM 125ppm vs. 500ppm and DCM 250ppm vs. 500ppm had higher performance of around 70%. Thus, DCM classification is possible but benzene concentrations are not classifiable with the current algorithm.

Figure 5-23 shows the wavelet features selected by RFS for classifying DCM 125ppm and 500ppm. A large number of the features, when reconstructed, are found to cover regions where DCM 125ppm is expected to have a strong signal (near $(V_C, t) = (-14, 30)$). There are also some larger features around $V_C = -4V$, and a closer look at the mean spectra of DCM 500ppm (Figure 5-24) reveals that this region can also be expected to have measured ion intensity. RFS has much higher scores in both metrics than the other algorithms in Table 5.7.

Magnitude signature features also performed well, but again had a low SVM mar-
Figure 5-24: Mean FAIMS Spectrum for DCM 500ppm.

<table>
<thead>
<tr>
<th>Extraction Algorithm</th>
<th>NUCS Distance</th>
<th>SVM Margin</th>
<th>Mean Kernel $P_{CC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude</td>
<td>0.170</td>
<td>0.008</td>
<td>92.8</td>
</tr>
<tr>
<td>SFS</td>
<td>0.166</td>
<td>0.034</td>
<td>69.4</td>
</tr>
<tr>
<td>RFE</td>
<td>0.177</td>
<td>0.022</td>
<td>68.3</td>
</tr>
<tr>
<td>RFS</td>
<td>0.301</td>
<td>0.175</td>
<td>92.2</td>
</tr>
<tr>
<td>Noise</td>
<td>0.016</td>
<td>0</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 5.7: Feature scores for classification of DCM 125ppm and DCM 500ppm in the wavelet domain, along with the mean $P_{CC}$ over the kernels in crossvalidation 3.

gin score. This leads us to believe that SVM margin is not necessarily a good score metric for features selected by filter methods, or at least by the magnitude signature algorithm.

Overall, for concentration classification using wavelet domain features, there were some improvements in DCM classification but classifying benzene at different concentrations remains difficult. The magnitude signature and RFS feature extraction algorithms are good for this task, and the NUCS score is a better metric than the SVM margin for the extracted feature sets.
5.3 Feature Set Score Metrics

The percentage of correct classification has been used throughout this paper for quantitatively evaluating both a feature set and the classifier. However, there have also been quantitative scores used purely on a feature set as a distance metric. These feature set score metrics are useful in that they may be able to evaluate the effectiveness of a feature set in classifying or separating two classes of data, without actually performing classification.

Throughout the results, we have seen the score qualitatively reflect the relative usefulness of feature sets. There were also cases in which the scores, such as SVM margin, are not useful in describing the feature set’s distinguishing potential. We want to know if there is some sort of quantitative relationship between the scoring metrics and $P_{CC}$.

5.3.1 Normalized Univariate Class Separability

The first feature set score metric is the NUCS measure. The metric is shown again in Eq. 5.1:

$$D(i) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$  \hspace{1cm} (5.1)

Figure 5-25 shows a scatter plot of NUCS vs. $P_{CC}$. This was done by calculating the $P_{CC}$ and score for every set of features extracted, including each crossvalidation trial of each feature extraction algorithm. For most classification algorithms, the NUCS score clustered around 0.2 without regard to $P_{CC}$. For magnitude signature features, however, there are points that had scores above 1, and at higher $P_{CC}$s there were fewer low scores. Figure 5-26 shows the same scatter plot but for all features in the wavelet domain. Both magnitude signature features and RFS features have scores above 1. However, there seems to be little correlation between $P_{CC}$ and NUCS score, because high NUCS scores for magnitude signature features could correspond to 60% just as much as 80%.
Figure 5-25: NUCS score vs $P_{CC}$ for all feature sets in the pixel domain.

Figure 5-26: NUCS score vs $P_{CC}$ for all feature sets in the wavelet domain.
5.3.2 SVM Margin Score

Equation 5.2 is the SVM margin between two classes of data. The SVM margin measures the distance between the two hyperplanes that separate the two classes, so if margin is larger, classification should be better.

\[
D(S) = \frac{1}{||\vec{w}||} \tag{5.2}
\]

Figure 5-27 shows the SVM margin for all feature sets from all classification tasks on pixel data. Figure 5-28 shows the SVM margin for wavelet data. The SVM margin
only goes above 0.25 with $P_{CC}$ of 80% or higher, so for the RFE and RFS extraction algorithms, a higher score could mean a better $P_{CC}$. This is consistent with the two algorithms being wrapper methods that use SVM weight as a feature selection metric. Another observation is that magnitude signature tended to have lower SVM margins even though their $P_{CC}$ are high. This trend was seen several times in the data, and seems to suggest that features selected by magnitude signature are better scored with NUCS score.
Chapter 6

Conclusion

6.1 Summary

The goal of this thesis was to classify FAIMS data using a threefold approach that involved calibration, feature extraction and classification. We also examined the advantages of wavelet transforms and their effect on classification. The percent of correct classification (PCC) was used to quantify the ability to classify between two data sets. The effectiveness of feature sets on increasing the separability of two classes was also measured with two distance metrics.

A method of calibrating FAIMS spectra has been developed to remove variations in signal location and intensity. The intensity, temporal and spatial variations inherent in sensor instruments can be removed by using various artifacts such as the nitrogen line and the ammonium dip, as well as using the wavelet transform to remove baseline fluctuations. After calibration by wavelet approximation removal, both the baseline intensity of all samples and their peak to noise ratios were brought to about the same level. This is important in classification applications, because if baseline fluctuations are not removed, two data samples that are of the same class can be classified wrongly. Other methods of calibration would improve the accuracy of the data even more, and include using an internal standard as a reference for the movement of ions in FAIMS.

The performance of each feature extraction algorithm in isolating discriminating features varies with the classification task. Often, the best way to evaluate the features
is to use a feature set scoring metric. The normalized univariate class separability score and SVM margin score were used to evaluate the separation distance of two classes based on the feature set. For both metrics, a larger value indicates greater separation between classes. However, SVM margin proves to be a better indication of how SVM based wrapper algorithms will perform, whereas the NUCS score was better for rating feature extraction algorithms such as magnitude signature. Neither scoring metric is perfectly adapted to score any particular feature extraction algorithm, but both served to show that feature extraction algorithms always had higher scores than arbitrarily selected features.

In classification, the following two tasks were attempted. First, chemical classification sought to classify two different chemicals at the same concentration. Chemical classification worked well in general, resulting in $P_{cc}$ values above 90% for some combinations of feature extraction algorithms and SVM kernels. We suggest that the shape and location of the two chemical intensities in a FAIMS spectra are different enough that they can be easily selected as features.

However, for the second task of classifying two concentrations of the same chemical, we noticed that both benzene and DCM had spectra that either faded or fragmented as the concentration increased from 125ppm to 500ppm. This negatively affected classification of two chemicals at that concentration. Concentrations of benzene were difficult to distinguish and resulted in having only about 50% classification accuracy. DCM, on the other hand, still performed well against different concentrations of itself, especially when one of the classes being classified was 500ppm.

Of the feature extraction algorithms, there were no examples of one algorithm performing drastically better or worse. Both magnitude signature and RFS had higher $P_{cc}$ than the other algorithms in many cases, though there were instances where RFE did perform the best on average. The SFS algorithm performed lowest of all four in many cases. This can be attributed to its suboptimality, which can be remedied by floating algorithms that allow removal of already selected features if a better set can be achieved. Also, the divergence metric used by SFS assumes a signal distribution that may not have been accurate, and was also limited by the low
number of training data.

The wavelet transform is a useful tool to assist in feature extraction. Its robustness comes from the ability to select regions as features instead of single pixels. Chemical classification performed far better in the wavelet domain, resulting in $P_{CC}$ of 90% or more in many cases. For concentration classification, there are still very low $P_{CC}$s when classifying in the wavelet domain, but DCM concentration classification did slightly better with the wavelet domain as well.

The three-step algorithm has been used to perform classification on FAIMS data. If better methods of calibration and feature extraction can be developed, FAIMS has the potential to become a powerful yet portable tool for in-field chemical analysis.
Appendix A

$P_{CC}$ Results
Figure A-1: $P_{CC}$ for benzene 125ppm dcm 125ppm, pixel domain.

Figure A-2: $P_{CC}$ for benzene 250ppm dcm 250ppm, pixel domain.
Figure A-3: $P_{CC}$ for benzene 500ppm dcm 500ppm, pixel domain.

Figure A-4: $P_{CC}$ for benzene 125ppm benzene 250ppm, pixel domain.
Figure A-5: $P_{CC}$ for benzene 125ppm benzene 500ppm, pixel domain.

Figure A-6: $P_{CC}$ for benzene 250ppm benzene 500ppm, pixel domain.
Figure A-7: $P_{CC}$ for dcm 125ppm dcm 250ppm, pixel domain.

Figure A-8: $P_{CC}$ for dcm 125ppm dcm 500ppm, pixel domain.
Figure A-9: $P_{CC}$ for dcm 250ppm dcm 500ppm, pixel domain.

Figure A-10: $P_{CC}$ for benzene 125ppm dcm 125ppm, wavelet domain.
Figure A-11: $P_{CC}$ for benzene 250ppm dcm 250ppm, wavelet domain.

Figure A-12: $P_{CC}$ for benzene 500ppm dcm 500ppm, wavelet domain.
Figure A-13: $P_{CC}$ for benzene 125ppm benzene 250ppm, wavelet domain.

Figure A-14: $P_{CC}$ for benzene 125ppm benzene 500ppm, wavelet domain.
Figure A-15: $P_{CC}$ for benzene 250ppm benzene 500ppm, wavelet domain.

Figure A-16: $P_{CC}$ for dcm 125ppm dcm 250ppm, wavelet domain.
Figure A-17: $P_{CC}$ for dcm 125ppm dcm 500ppm, wavelet domain.

Figure A-18: $P_{CC}$ for dcm 250ppm dcm 500ppm, wavelet domain.
Bibliography


[5] Epa ground water and drinking water. EPA.


