Detection of Contaminants Using a MEMS FAIMS Sensor

by

Kristin Carr

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

[June 2005]
May 2005

Author_____
Department of Electrical Engineering and Computer Science
May 19, 2005

Certified by_____
Nirmal Keshava, Ph.D.
Draper Laboratory Thesis Advisor

Certified by_____
Julie Greenberg, Ph.D.
M.I.T. Thesis Advisor

Accepted by_____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Detection of Contaminants
Using a MEMS FAIMS Sensor

by
Kristin Carr
Submitted to the Department of
Electrical Engineering and Computer Science

May 19, 2005

In partial fulfillment of the requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Detecting the presence of contaminants in water is a critical mission, but thorough testing often requires extensive time at a remote facility. A MEMS implementation of a FAIMS (High-Field Asymmetric-Waveform Ion Mobility Spectrometry) sensor has recently been developed, and is capable of promptly analyzing water on-site. In this thesis, we apply two well-established statistical target detector algorithms to the detection of chlorite in water. The matched filter and the adaptive cosine estimator (ACE) are subspace detectors that possess complimentary geometric properties. We address several significant challenges in implementing these detectors, including the estimation of the covariance given the limited amount of data available and the design of a target signature subspace in response to the fact that the signature does not scale linearly with the contaminant concentration. In addition, we consider the need for dimension reduction through the use of wavelets. We evaluate each of the detectors on FAIMS data of pure and chlorite-contaminated water.

Draper Thesis Advisor: Nirmal Keshava, Ph.D.
Title: Principal Member of Technical Staff

M.I.T. Faculty Thesis Advisor: Julie Greenberg, Ph.D.
Title: Principal Research Scientist

Acknowledgements

May 19, 2005

I would like to thank everybody at Draper Laboratory for their help on the work presented in this thesis. I would like to specifically thank the biomedical engineering group for their work in developing the FAIMS sensor and all of the data acquisition techniques documented in this paper. Cristina Davis, Melissa Krebs and Julie Zeskind were responsible for the idea of water analysis using FAIMS. Heather Clark and Marianna Shnayderman were critical in the testing and development of the FAIMS instruments and were very helpful in their immense assistance in data acquisition. In addition, there were many people who helped in collecting data, developing instrumentation techniques, teaching and explaining FAIMS and biological concepts, and being willing to discuss ideas and research paths. These include those individuals named above, in addition to Angela Zapata, Sarah Cohen, Will Merrick, Daniel Traviglia.

I am also grateful for the guidance, support, knowledge, and patience of Nirmal Keshava, Melissa Krebs, Cristina Davis and Heidi Perry, and my advsior, Julie Greenberg, all of without whom this thesis would not be possible.

This thesis was prepared at The Charles Stark Draper Laboratory, Inc., under Internal Research and Development Project Number 12591-001.

Publication of this thesis does not constitute approval by Draper or the sponsoring agency of the findings or conclusions contained herein. It is published for the exchange and stimulation of ideas.

Kristin Carr

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and Motivation

The ability to accurately determine the safety level and contaminant concentration of water is critical, but thorough testing often requires extensive time and sophisticated facilities for chemical tests to be run a great distance away from where the water sample was taken [15]. This serves as the motivation for an on-site water monitoring tool through which water can be sampled and then processed by a statistical detector to give an indication of the presence or lack of contaminants in the water.

The FAIMS (High-Field Asymmetric-Waveform Ion Mobility Spectrometry) sensor is a recently-developed analytical sensor that has been used to measure and analyze ion mobility properties of biological and chemical materials [16, 17, 18, 19, 20, 21, 22, 26, 27]. It can be used to analyze water samples, and does not require the addition of any chemical reagents nor extensive handling and processing of the sample that many other water monitoring methods require [15]. Further, a MEMS (Micro-Electro-Mechanical System) device has been developed at Draper Laboratories to analyze samples and produce a pair of three-dimensional signals indicative of the concentration levels of different types of ions in the sample [1, 16].

Because the spectrometer has been implemented in a MEMS device, it can potentially be used to acquire and analyze water samples at the site as a small, portable device capable of almost instantly determining water safety levels. Researchers at Draper Laboratory have previously acquired data demonstrating the ability of FAIMS to detect changes in water quality and the presence of contaminants [37]; however, the initial study only provided the basis of proof-of-concept classification of signals evaluated post-acquisition. The primary objective of this thesis is to introduce and develop analytical techniques for the statistical detection of contaminants in water using FAIMS with the goal of operating with the lowest probability of false alarm (P_{FA}) and highest probability

of detection (P_D).

1.2 Project Overview

This thesis aims to develop and evaluate several advanced methods for the detection of contaminants in water using FAIMS. We employ techniques well-established on other types of data and examine their application to the FAIMS data. With a goal of minimizing the error rate of the detector, we devise an optimal strategy assuming an infinite quantity of FAIMS pure and contaminated water data for use in real-time analysis. Given our limited number of data samples and a desire for low computational requirements, we examine the various assumptions that allow us to proceed to a series of different statistical analysis methods, hereafter referred to as detectors. Note that the term "detector" will be used in this thesis to refer to the statistical detection algorithm implemented in software as the step following FAIMS data acquisition; this term will not be used to refer to the FAIMS hardware.

We analyze and evaluate two complimentary subspace detectors: a matched filter detector [9], which is based on the relative energy[1] in the received signal, and an adaptive cosine estimate (ACE) detector [9], which is based on the relative angle in the received signal. Both are based upon a linear signal model and Gaussian additive noise, and have been derived to be statistically optimal for distinct signal models. ACE also has the desirable property of maintaining a constant false alarm probability (CFAR) under certain variations in the noise model.

We also consider the case of a multi-dimensional target subspace for both the matched filter and ACE detectors. This will be relevant if the target signals span a multidimensional subspace as they vary with concentration instead of simply increasing in magnitude [9]. Algorithms developed for radar data have been employed to detect radar target signatures for the conditions and environments in which radar systems operate.

[1] In this thesis, energy is defined to be the sum of the squares of each ion intensity point.

Although FAIMS collects data in a significantly different environment, our goal is to investigate how well contaminants (i.e., targets) can be modeled and detected under the same assumptions.

In addition, we employ a wavelet transform [10] to reduce the dimensionality of the original data and thus improve the covariance estimates on a smaller number of coefficients. We then implement the corresponding matched filter and ACE detectors on a limited number of those wavelet coefficients.

In order to properly juxtapose the different detector performances, we create a metric based on the level of separation between classes yielded by each of the detectors. Without enough data to sufficiently generate a receiver operating curve (ROC), this metric allows us to compare detectors for the purpose of minimizing the rate of error. A more rigorous analysis of detector outputs would utilize the exact parametric form of the detector test statistic, but requires significantly more experimental data than was at our disposal.

1.3 Related Works

There has been no published work using a statistical detector with FAIMS sensor data to detect contaminants in water. The few related papers utilize the FAIMS technology, but concentrate on results that visually discern different ion species [14, 18, 19, 20, 21, 23, 26] and do not employ any statistically optimal approaches. In addition, because FAIMS can be used in a variety of different setups (e.g., using a different sensor or sample introduction technique), there is a vast array of FAIMS data that is very different, and yields different types of datasets indicative of different quantities [12, 13, 19, 20, 27, 29, 30]. As a result, the processing required on such datasets is different.

Despite the lack of published results involving statistical detectors on FAIMS data, the field of classification and detection is rich in similar work done on other types of data. These papers include the development and evaluation of a large class of subspace detectors, including the matched filter and adaptive cosine estimator (ACE) detectors

[9], both of which will be utilized in this thesis [35, 36].

We address the notion of a wavelet transform to gain parsimonious representations of FAIMS data; this concept as applied to other types of data is one which has been widely publicized [10, 11, 31, 32, 33, 34]. We will draw on these concepts as the building blocks for the wavelet section of this thesis.

2 FAIMS

2.1 Description

The technology in use is a FAIMS sensor which measures the abundance of ions arriving at the sensor after travelling through a channel with an externally applied electric field. Conventional ion mobility spectrometers operate in the low field regime where the applied electric field strength is less than 1000 V/cm and the mobility is essentially constant [1]. However, it has been demonstrated that the mobility of an ion is field dependent and can change significantly as the field strength increases [2]. Different ion species will have particular mobility dependencies on an electric field, and the FAIMS technology utilizes this differential mobility of ions in an electric field to identify the different ion species.

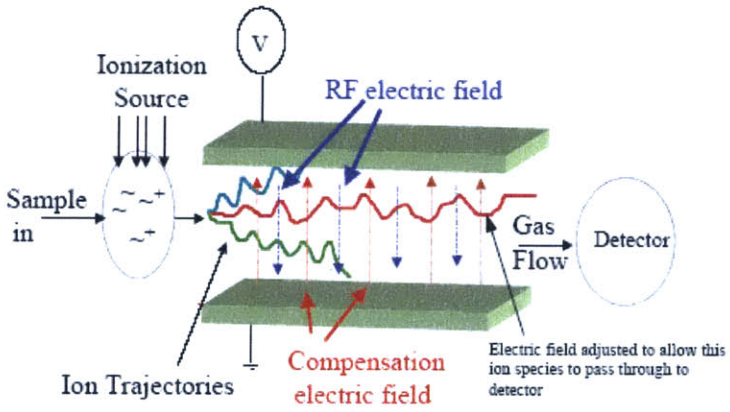


Figure 1: FAIMS schematic & ion flow path [1]

The FAIMS sensor operates as follows: a gas sample derived from the headspace above a water sample (described in Section 2.2) is introduced to the FAIMS spectrometer; after entering the spectrometer, the ions are transported by a carrier gas between a pair of parallel plates in which an asymmetric electric field, $E_{RF}(t)$, is applied at radio frequencies [1, 2] over time as shown in Figure 1 [1]. The ions travel along a path between

the parallel plates, and a sensors at the end of the plates measure the voltages resulting from the stream of positive and negative ions arriving at the sensors. The electric field alternates at radio frequencies with an asymmetric duty cycle between a high-magnitude positive electric field (E_{max}) and a low-magnitude negative electric field (E_{min}) so that the net electric field is zero, as shown in Figure 2 [1].

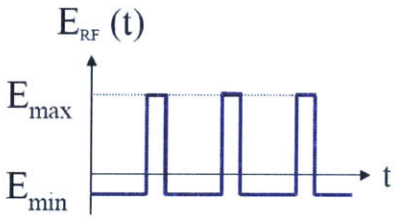


Figure 2: $E_{RF}(t)$: FAIMS asymmetric electric field [1]

To better understand how the FAIMS technology works, we can consider the forces that act on a single ion. The ion experiences a constant force from the carrier gas flow (z-directed) which transports it through the parallel plates.

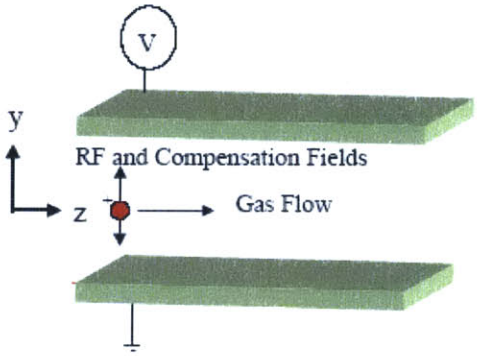


Figure 3: FAIMS force directions [1]

A transverse force (y-directed), produced by the RF electric field ($E_{RF}(t)$) and a DC compensation voltage (V_c), also acts on the ion, as shown in Figure 3 [1]. These fields are generated by applying voltages to the parallel plate electrodes. The resulting ion

velocity in the y-direction is given by [3]:

$$V_y = KE \tag{1}$$

where K is the coefficient of ion mobility for the ion species and E is the electric field intensity. The dependence of the mobility on the electric field intensity can be represented by the following expression [3, 4]:

$$K(E) = K_0[1 + \alpha_2 E^2 + \alpha_4 E^4 + ...] \tag{2}$$

where α_2 and α_4 are coefficients of a series expansion, and K_0 is the mobility coefficient in a vanishingly small field [6]. As the electric field strength increases (above 5000 V/cm), the second and higher order terms in the series become significant and the mobility coefficient can change substantially (10-15%) from its low field value.

If we take K_1 to be the mobility of a particular ion at E_{max} and K_2 to be the mobility of the ion at E_{min}, then the average displacement of the ion in the y-direction as a function of time can be expressed as [1]:

$$y(t) = \beta(K_1 - K_2) * t \tag{3}$$

where β is a constant determined by the strength and duty cycle of the applied electric field. Thus, the overall trajectory will not be straight if the mobilities K_1 and K_2 are not equal. To compensate for this and to allow the ion to arrive at the sensor, a low-field compensation electric field can be applied to the parallel plates.

A range of compensation voltages is repeatedly swept through linearly for each time sample, and the positive and negative sensors measure the voltage generated at the detector for each particular compensation voltage at each time sample, which yield a quantification of ions present. The ions strike a Faraday plate which generates a voltage based on the charge transfer. The FAIMS sensor measures separately the voltage

generated by both the positive ions and the negative ions, and this voltage is referred to as the ion abudance or ion intensity value for each point in time and compensation voltage. The resulting output of the FAIMS sensor is a pair (one for each polarity) of three-dimensional signals of ion intensity that are dependent on the time, t, and compensation voltage V_c.

2.2 Sample Introduction

The FAIMS setup discussed above requires specialized sample introduction and ionization methods. The data used in this thesis was taken with a headspace sampler and used gas chromatograph (GC) as a pre-separation step prior to the gas sample entering the FAIMS unit. The headspace sampler heats up the given fluid sample, which has the effect of transferring substances that are volatile at that temperature into the air in the sample vial above the water. The headspace sampler then removes a sample of the air in the vial and GC is used to provide a separation in time of the substances presented to the FAIMS unit. When the substance enters the FAIMS device, the material is ionized using a beta-particle emitting radioactive nickel source (^{63}Ni).

The motivation behind this comes from the fact that we are interested specifically in detecting chlorite; since chlorite is a particularly volatile substance, ^{63}Ni is well suited to ionizing the analyte. The initial headspace gas chromotograph (HS-GC) separation yields a first pass at extracting the substance of interest, thus yielding a higher ratio of contaminant signal to water background.

2.3 Sensor Quantization

There are limitations on the analog-to-digital converter attached to the FAIMS sensor that converts the analog voltage values into a digital output as read in by accompanying software. As a result, the sensor output is quantized to discrete values. This sets an upper bound on the level of resolution, which is empirically found to be approximately 10

uV. Figure 4 shows a histogram of points from the entire collection of pure and chlorite-contaminated data samples. This figure illustrates the fraction of data points that fell into each range of ion intensity values. The ionization intensity range was divided into sections, or "bins", 2 μV wide, and the histogram shows the fraction of the data fell into each section. As can be seen in Figure 4, there are many empty bins between each full bin; this indicates that the ion intensity values do not exist in a continuum, but instead have been quantized to discrete intervals. This can lower the overall performance of the data analysis methods and is one of the sources of noise that we will analyze in subsequent chapters.

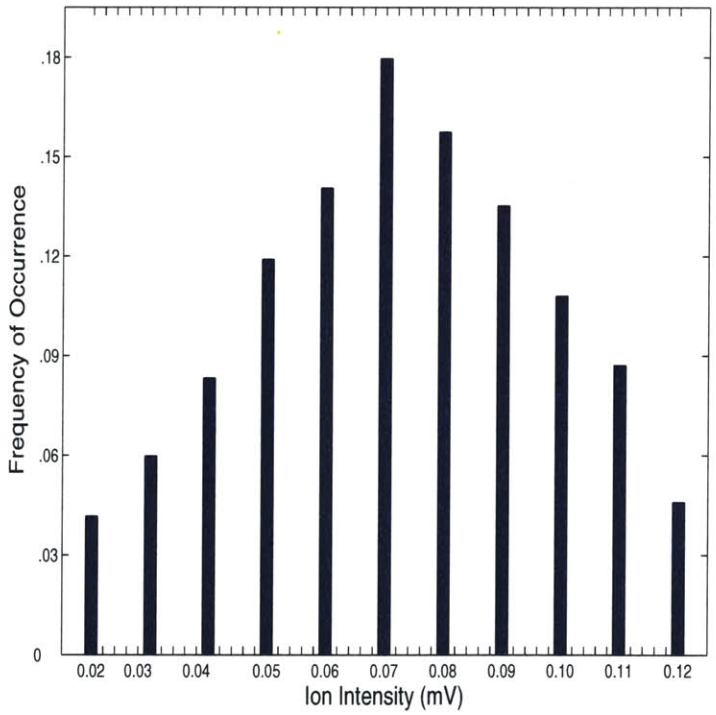


Figure 4: Histogram of noise points illustrating sensor quantization

2.4 Signal Model

In order to properly develop a statistical detector, it is necessary to model the received signal. Toward this end, we develop a signal model to describe the process that results in the measured signal from the FAIMS sensor.

The water sample without contamination possesses a characteristic background signal, and we model the contaminated water as the sum of the background water signal with an additional signal due to the contamination. This model assumes linearity in the addition of the water background and contaminant signals.

The received signal is dependent on the contaminant type, concentration level, E_{max} and E_{min} (the high and low electric field magnitudes), and the background water source.

In its most general form, the received 3-dimensional positive and negative ion signals will be:

$$r(t, V_c) = f(contaminant, concentration, E_{max}, E_{min}, water) + n(t, V_c) \tag{4}$$

where r, the received signal, is the voltage (mV) measured at the FAIMS sensor, and t and V_c are the time (seconds) and compensation voltage (V), respectively, at which the voltage was measured. The signal, $n(t, V_c)$, is the random additive noise accounting for the variation in the signal that we model as Gaussian and will examine further in section 5.

If we do not vary E_{max}, E_{min} and the deionized water source, as is the case in this project, we can simplify this to:

$$r(t, V_c) = f(contaminant, concentration) + n(t, V_c) \tag{5}$$

If we assume that the ion quantities in the contaminated sample add linearly to the background, then the function f will be the addition of the signal due to the contaminant and the signal due to the water background, denoted by s and w, respectively.

$$r(t, V_c) = s(t, V_c) + w(t, V_c) + n(t, V_c) \tag{6}$$

Thus, in the context of a statistical detector to differentiate between the hypotheses H_0 and H_1, we have the binary hypothesis test:

$$H_0 : r_0(t, V_c) = w(t, V_c) + n(t, V_c) \tag{7}$$

$$H_1 : r_1(t, V_c) = s(t, V_c) + w(t, V_c) + n(t, V_c) \tag{8}$$

where the signals r_0 and r_1 correspond to the received signals under the null hypothesis H_0 and the test hypothesis H_1, respectively.

3 Data Acquisition

3.1 Overview

We focus on chlorite as the contaminant used in the training and testing of a suitable statistical detector algorithm. Concentrations of 2.5 ppm, 5 ppm, 10 ppm, 20 ppm, and 40 ppm of chlorite-contaminated water, in addition to pure water, were submitted to the FAIMS sensor and used as training and testing data.

The contaminated samples were obtained by mixing deionized water with household bleach (5.25 % NaClO). 152 μL of bleach was mixed with 199.848 mL of deionized water to obtain 200 mL of a chlorite-contaminated solution with a concentration of 40 ppm. Half of this solution was removed and added to 100 mL of deionized water to obtain 200 ml of a chlorite-contaminated solution with a concentration of 20 ppm. This process was repeated to obtain new solutions with half the concentration of the previous solution until the lowest concentration of 2.5 ppm had been created.

A total of 76 vials were created, with each 20 mL vial containing 10 mL of pure or contaminated water. The 76 vials were comprised of 16 pure water samples and 12 samples of each concentration of the chlorite-contaminated water. The vials were were split over two runs (with 38 samples each). A total of 76 data files were obtained, with 16 pure water samples, and 12 samples of each concentration of chlorite solution.

3.2 FAIMS Setup & Parameters

The FAIMS setup consisted of a headspace sampler[2] connected to the inlet of a gas chromatograph[3] (GC) with a FAIMS sensor[4] connected to the detector outlet of the GC. The GC used a 10 meter HP VOC fused silica column with an inner diameter of 0.32 mm. Nitrogen was used as a carrier gas to direct the flow of ions from the headspace

[2]Agilent 7694 Headspace Sampler. Agilent Technologies. Palo Alto, CA
[3]HP 5890 II gas GC. Agilent Technologies. Palo Alto, CA
[4]microDMx^{TM}. Sionex Corporation. Waltham, MA

sampler through a transfer line into a silica column and carry it into to the FAIMS sensor. The sample carrier flow was regulated by the headspace sampler and was joined by a second flow of Nitrogen at 300 mL/min regulated by a mass flow controller[5]. The carrier gas and sample were ionized with 5 mCi of ^{63}Ni. Tables 1, 2, and 3 below show the parameters used for the headspace sampler, GC, and FAIMS sensor, respectively.

Oven temperature	60°C
3 mL sample loop temperature	75°C
Transfer line temperature	85°C
Sample vial heating time	1 minute
Sample vial heating temperature	40°C
Pressure	15.0 psi
Pressure duration	0.10 minutes
Loop fill time	0.02 minutes
Loop equilibrium time	0.05 minutes
Injection time	0.05 minutes

Table 1: Headspace sampler parameters

Inlet temperature	250°C
Oven ramp initial hold time	0.5 minutes
Oven ramp initial hold temperature	40°C
Oven ramp rate	30°C/min
Final hold time	2 minutes
Final hold temperature	170°C
Detector heating block temperature	140°C

Table 2: GC parameters

E_{max}	1200V/cm
V_c range	[-35,5]V
Number of V_c samples	100
Sample duration	1.6 seconds
Number of time samples	500

Table 3: FAIMS parameters

[5]MKS Instruments. Andover, MA

3.3 Detector Implementation

All statistical detectors were implemented in MATLAB, with the input being a single FAIMS signal, a matrix 500×100 in size. Each sample signal, denoted by $r(t, V_c)$, consisted of 500 scans in time, t, and 100 points in compensation voltage, V_c, yielding a matrix of ion intensities with 50,000 entries.

As noted previously, the FAIMS sensor outputs a pair of 3-dimensional signals corresponding to the positive and negative ion quantities. However, the negative ion signal was not found to yield any difference among the means of various concentrations of chlorite and water for this particular application (data not shown). As a result, we focus our work on the positive ion signal, and all received signals, $r(t, V_c)$, refer to the positive ion FAIMS output signal.

As we empirically determined from various FAIMS spectral traces, there is an element of time variance that exists within the data; two data samples from the exact same solution will look slightly different if they were run at different times. As a result, it is difficult to acquire large quantities of consistent data, and we have chosen to focus all of our testing and training on the two runs of data comprising 76 total samples. Because of the limited number of samples, we have both trained and tested on the same data. Although this is a limitation of our analysis, we were able to establish proof-of-concept for applying statistical detection algorithms to this complex data type.

For the wavelet transforms, the MATLAB Wavelet Toolbox was used for the function `'wavedec'` to implement the wavelet transforms on the data.

3.4 Data Description

As mentioned previously, each data sample signal, $r(t, V_c)$, consists of 100 scans over a compensation voltage range of [-35, 5] V, and 500 scans over a time range of [0, 800] seconds.

Shown below is Figure 5, which illustrates the mean signature (over all runs in

a particular class) of pure water and chlorite-contaminated water at 2.5 ppm and 40 ppm. Water and chlorite at 2.5 ppm are relatively visually indistinguishable, while the difference becomes much more obvious at 40 ppm of chlorite. A small peak arises in the 40 ppm chlorite mean while another peak disappears as compared to the 2.5ppm concentration of chlorite. Arrows indicate the areas in the data that are visually different between concentrations.

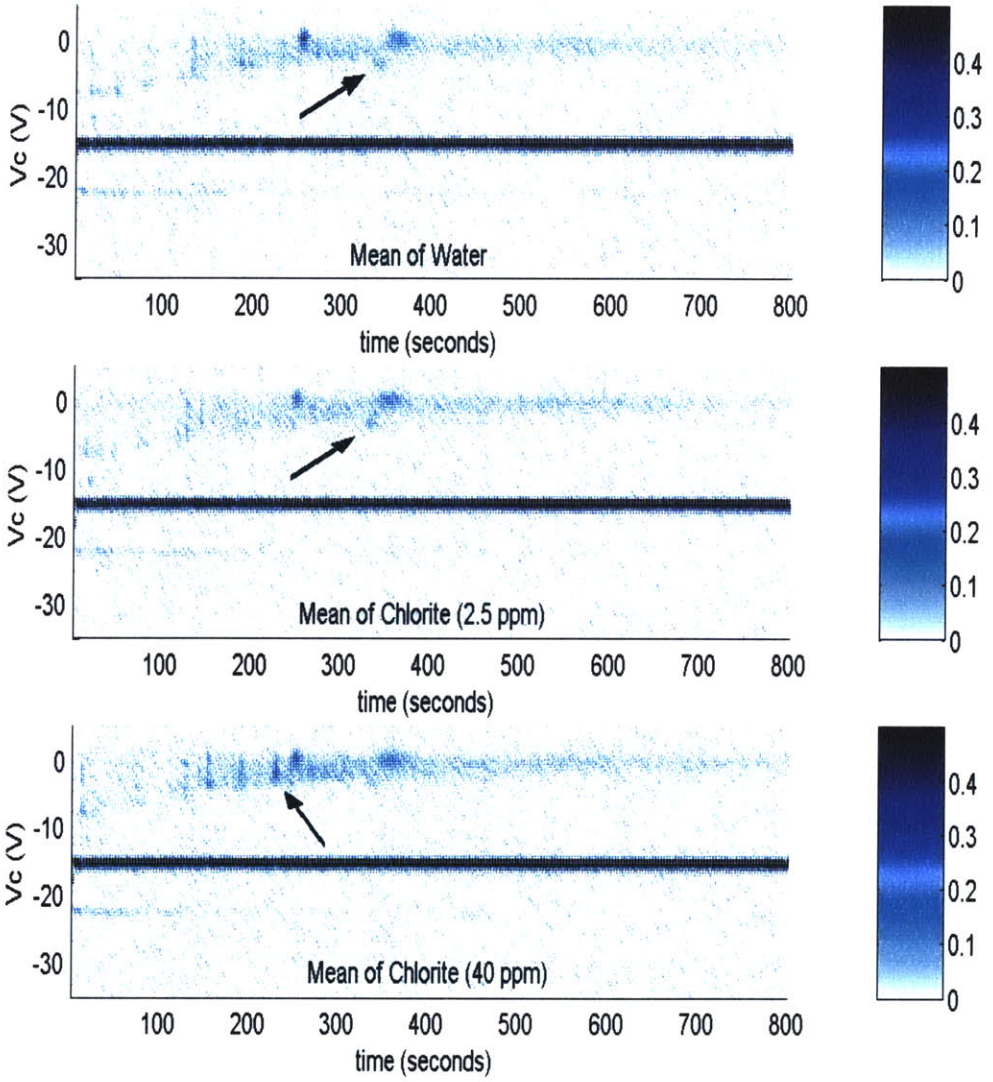


Figure 5: Top-View Plots of FAIMS Spectra for Water and Chlorite-Contaminated Water at 2.5 ppm and 40 ppm

4 Problem Formulation and Development

4.1 Problem Description

We wish to develop a statistical detector to determine with the lowest possible rate of error whether or not chlorite is present in the water; the detector output is a binary decision and we wish to minimize the rate of missed detections and false alarms.

While the final output of the detector is binary, the unthresholded detector test statistic is a random variable indicative of how much of the input sample lies in the domain of the contaminated water and how much of the input sample falls in the domain of pure water. This random variable will have a different mean and variance for a given training concentration and a given testing concentration, but the goal is for the means to be as far apart as possible for pure water and contaminated water, and for the variance to be as small as possible. Thus, the distribution of this random variable serves as an important tool for a particular detector as the probability distribution function (PDF) and parameters under each hypothesis dictate how far apart the classes are under a given detection algorithm.

4.2 Bayesian Derivation

We have the null and test hypotheses, H_0 and H_1, for each single observation, r; given this, we wish to estimate which hypothesis occurred and minimize the probability of error associated with this estimate. Thus, for every received signal, r, we wish to devise a mapping $\hat{H}(\mathbf{r})$ between the received signal and the estimated hypothesis. The problem becomes, for each particular r, to determine which hypothesis minimizes the probability of error. The probability of error will be the probability that the actual event is different from the hypothesis we choose.

For any given received signal r, the probability of error associated with choosing H_0 will be the probability that the actual hypothesis was H_1, conditioned on receiving r:

$$Pr[H = H_1|\mathbf{r} = r] \tag{9}$$

Similarly, for any given received signal r, the probability of error associated with choosing H_1 will be the probability that the actual hypothesis was H_0, conditioned on receiving r:

$$Pr[H = H_0|\mathbf{r} = r] \tag{10}$$

Since we are attempting to minimize the probability of error, the optimal decision rule is to choose H_0 when the probability of error for choosing H_0 is smallest, or when:

$$Pr[H = H_0|\mathbf{r} = r] > Pr[H = H_1|\mathbf{r} = r] \tag{11}$$

By using Bayes' Rule [8], we recall that:

$$Pr[H = H_m|\mathbf{r} = r] = \frac{p_{r|H}(r|H_m)P_m}{p_{r|H}(r|H_0)P_0 + p_{r|H}(y|H_1)P_1} \tag{12}$$

where P_0 and P_1 denote the a priori probabilities of hypotheses H_0 and H_1, respectively.

By plugging this into the expression for the optimal decision rule in (11), we see that we should choose H_0 when:

$$p_{r|H}(r|H_0)P_0 > p_{r|H}(r|H_1)P_1 \tag{13}$$

or, equivalently, when:

$$\frac{p_{r|H}(r|H_0)}{p_{r|H}(r|H_1)} > \frac{P_1}{P_0} \tag{14}$$

If we further assume that $p_{r|H}(r|H_0)$ and $p_{r|H}(r|H_1)$ are Gaussian distributions (which will be the case when the variation on the received signal is due entirely to Gaussian

noise, as we will discuss in the next section), then we can simplify this expression for the decision rule further. Here we denote m_0 as $E[\mathbf{r}|H_0]$, m_1 as $E[\mathbf{r}|H_1]$, Σ_0 as $E[(\mathbf{r} - m_0)(\mathbf{r} - m_0)^T|H_0]$, and Σ_1 as $E[(\mathbf{r} - m_1)(\mathbf{r} - m_1)^T|H_1]$, where $E[\cdot]$ indicates the expected value of the given random variable. This yields the following decision rule to choose H_0 when:

$$\frac{\frac{1}{\sqrt{(2\pi)^n|\Sigma_0|}}e^{-\frac{1}{2}(\mathbf{r}-m_0)^T\Sigma_0^{-1}(\mathbf{r}-m_0)}}{\frac{1}{\sqrt{(2\pi)^n|\Sigma_1|}}e^{-\frac{1}{2}(\mathbf{r}-m_1)^T\Sigma_1^{-1}(\mathbf{r}-m_1)}} > \frac{P_1}{P_0} \tag{15}$$

where n is the dimension of the received signal.

By simplifying, we obtain:

$$-\frac{1}{2}(\mathbf{r} - m_0)^T\Sigma_0^{-1}(\mathbf{r} - m_0) + \frac{1}{2}(\mathbf{r} - m_1)^T\Sigma_1^{-1}(\mathbf{r} - m_1) > ln(\frac{P_1}{P_0}\sqrt{\frac{|\Sigma_0|}{|\Sigma_1|}}) \tag{16}$$

This general form of the likelihood ratio test yields two problems. This first is that we cannot determine with great accuracy the covariance matrix of either the null or test hypothesis, H_0 and H_1. There are 500 x 100 = 50,000 total points in each received signal, $r(t, V_c)$, and therefore the covariance matrix we wish to estimate would be 50,000 x 50,000 in dimension. As discussed in [7], if we wish to estimate this covariance matrix, we would require at least 50,000 distinct data samples for each of the hypotheses, H_0 and H_1. To maintain an average loss ratio of better than one-half, we would require at least 2 x 500 x 100 = 100,000 samples of each case (with and without chlorite) to sufficiently calculate the covariance [7]. Given our limited number of data samples, this is not feasible.

The other issue that arises is that even if we were to have the appropriate covariance matrix estimates, to invert such a large matrix would require substantial computing power. While such a feat can be accomplished on a standard personal computer, to implement on a microcontroller or FPGA, the most cost-effective portable computation systems, would easily require too much time or space to compute, and would thus not

be feasible to implement in a cost-effective portable water monitoring tool.

These issues lead to a need of a reduction in dimensionality or a simplified covariance matrix. If we can find a way to represent the relevant information in the signal with only a few coefficients, then we will have enough data samples to form a valid estimate of the covariance matrix of those coefficients and consequently perform detection on a small number of coefficients. Alternatively, we can make simplifications and assumptions about the covariance matrix of the original data sample. If we can support the notion that the covariance matrix is diagonal (i.e., a constant variance multiplied by the identity matrix), then the inversion of the covariance matrices requires only a division of the constant variance (since the inverse of the identity matrix is itself). By using these simplifications, we can perform detection on the original 50,000 points in each data signal.

5 Noise Analysis

In order to properly characterize the received signal, $r(t, V_c)$, we need to analyze the noise and variation on the signal. Specifically, we seek to gain an understanding of Σ_0 and Σ_1, the covariances of the signal under hypotheses H_0 and H_1. The covariance is an indication of how correlated different random variables are; if the covariance is diagonal, then the random variables are completely uncorrelated. If the covariance is not diagonal, then the different random variables are not independent, but are instead influenced and affected by one another. As derived previously, Σ_0 and Σ_1 are necessary for use in the optimal detector, and thus we would like to make a valid estimate. Toward this end, we are looking to characterize the covariance with time, t, and the covariance with voltage, V_c, for the purpose of gaining information about the interdependence of points within the received signal $r(t, V_c)$.

5.1 Overview of Analysis and Estimation

As mentioned earlier, we assume that all variation on the signal comes from the additive random noise and is independent of the water background or chlorite signal. As a result, we can begin a preliminary characterization of the noise by generating a histogram of the points produced by the FAIMS sensor that occur before the sample has been injected into the FAIMS setup. During this time, the carrier gas is flowing through the FAIMS setup and the sensor is measuring the ion intensity variation, but there is no water background or chlorite signal.

The choice to use the points before the sample has been introduced (and thus evident at the sensor output) comes from the need to analyze inherent sensor variation without the presence of any signal, either from water or chlorite. The introduction of the signal would distort the underlying distribution of the noise, preventing us from adequately characterizing it. In addition, this allows for an accurate histogram of the noise at all

data points instead of requiring a histogram for each of the 50,000 data points in the sample.

In addition to the distribution of the individual noise points, we are interested in their interdependency upon one another. In order to estimate an accurate covariance matrix of size 50,000 x 50,000 entries, it would be necessary to have more than 100,000 samples, which is impractical. At a rate of approximately 11 minutes per sample, 100,000 samples would require a constant data acquisition of more than two years. As a compromise, we can accept the lack of of a full covariance estimate and instead make an approximation through the covariance in time and the covariance in voltage. These estimates are discussed below.

5.2 Covariance in Time

In order to address the issue of dependency between noise samples, it is necessary to estimate a covariance of the noise. Unfortunately, it is unrealistic to estimate a 50,000 x 50,000 sample matrix with such a small number of datasets. Thus, we are focusing first on the covariance with respect to time, and then on the covariance with respect to voltage. To estimate the covariance in time, the sample time for a given voltage scan was used to calculate the expected product as a function of time scan. A total of 76 different signals ($r(t, V_c)$) with each contributing 15 rows of V_c scans was used to estimate the covariance. The covariance estimate, $\Sigma_t(i, j)$ was calculated as given below:

$$\Sigma_t(i,j) = E[(r(i, V_c*) - \bar{r}(i, V_c*))(r(j, V_c*) - \bar{r}(j, V_c*))] \tag{17}$$

where V_c* indicates one of the first 15 V_c points in the signal before any contaminant was introduced.

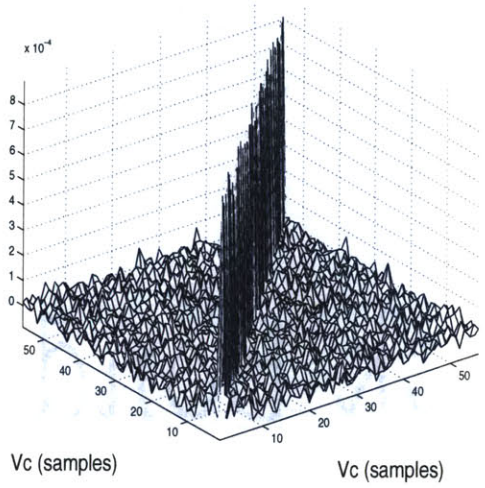


Figure 6: Covariance in V_c (mesh view)

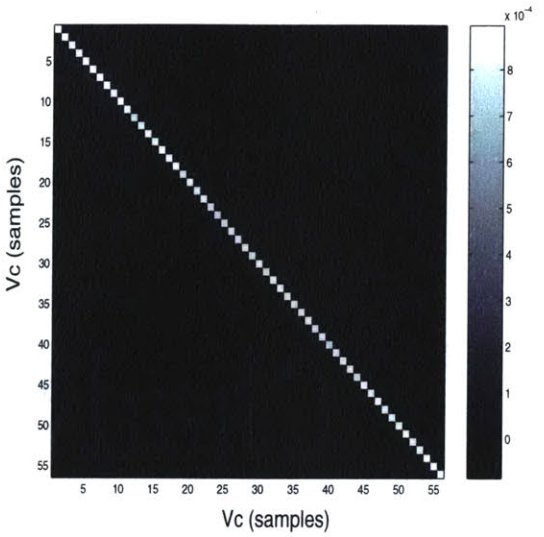


Figure 7: Covariance in V_c (top view)

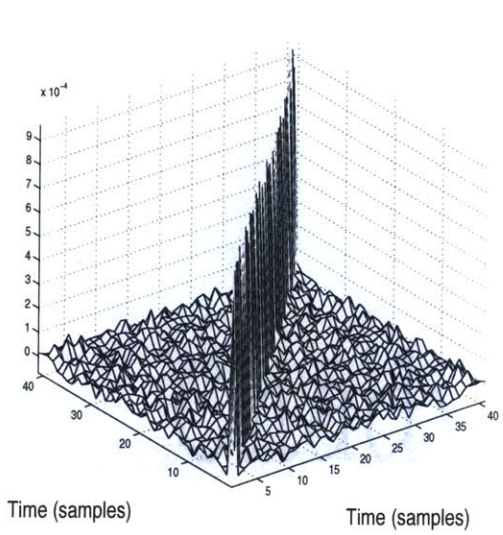
5.3 Covariance in Vc

Similarly to the estimate of the covariance in time, it was necessary to estimate the covariance in compensation voltage. To estimate the covariance with respect to compensation voltage, the sample voltage lag for a given time scan was used to calculate the expected product as a function of compensation voltage. A total of 76 different signals $(r(t, V_c))$ with each contributing 15 rows of t scans was used to estimate the covariance. The covariance estimate, $\Sigma_{V_c}(i, j)$ was calculated as given below:

$$\Sigma_{V_c}(i,j) = E[(r(t*,i) - \bar{r}(t*,i))(r(t*,j) - \bar{r}(t*,j))] \tag{18}$$

where $t*$ indicates one of the first 15 t points in the signal before any chlorite contaminant was introduced. Figures 6 and 7 show the estimated covariance matrix with respect to compensation voltage, while figures 8 and 9 show the estimated covariance matrix with respect to time.

As can be seen in Figures 6, ??, ??, and 9, the noise is roughly uncorrelated as the off-diagonal entries are at least a factor of 10 below the diagonal entries. In addition,



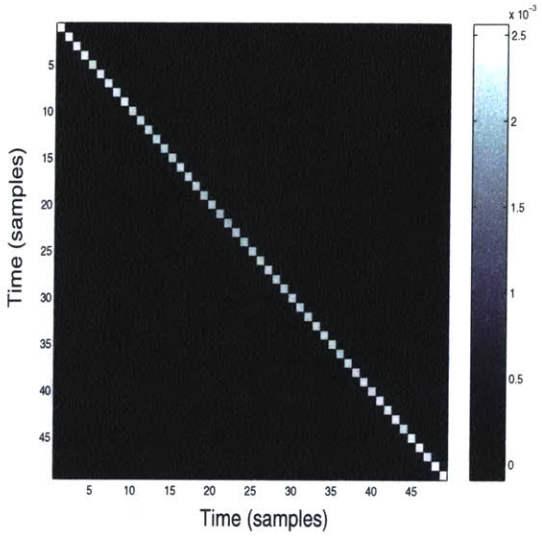


Figure 8: Covariance in time (mesh view) Figure 9: Covariance in time (top view)

the diagonal entries are all approximately 9.27×10^{-4} $(\text{mv})^2$ for both covariances, with a variation of less than 5% between diagonal entries in the same covariance and between the two covariance matrices. This provides a very strong indication that the noise is uncorrelated both in time and in voltage, and further that variance is equal for all points. Given that the noise is uncorrelated in both dimensions, it is reasonable to assume that the noise will be uncorrelated for all coordinate points (t, V_c) in the measured data and thus that the overall covariance matrix will be white and diagonal as well. We will employ this key assumption, based on these experiments, in the implementation of statistical detectors.

5.4 Histograms

The histogram for all data points of the baseline sensor output is shown in Figure 10, along with a fitted Gaussian distribution with the mean and variance obtained from the data. As mentioned previously in Section 2.3, there is a level of quantization that makes it difficult to estimate with a high resolution the histogram of data points, and

thus the resolution shown is the highest attainable.

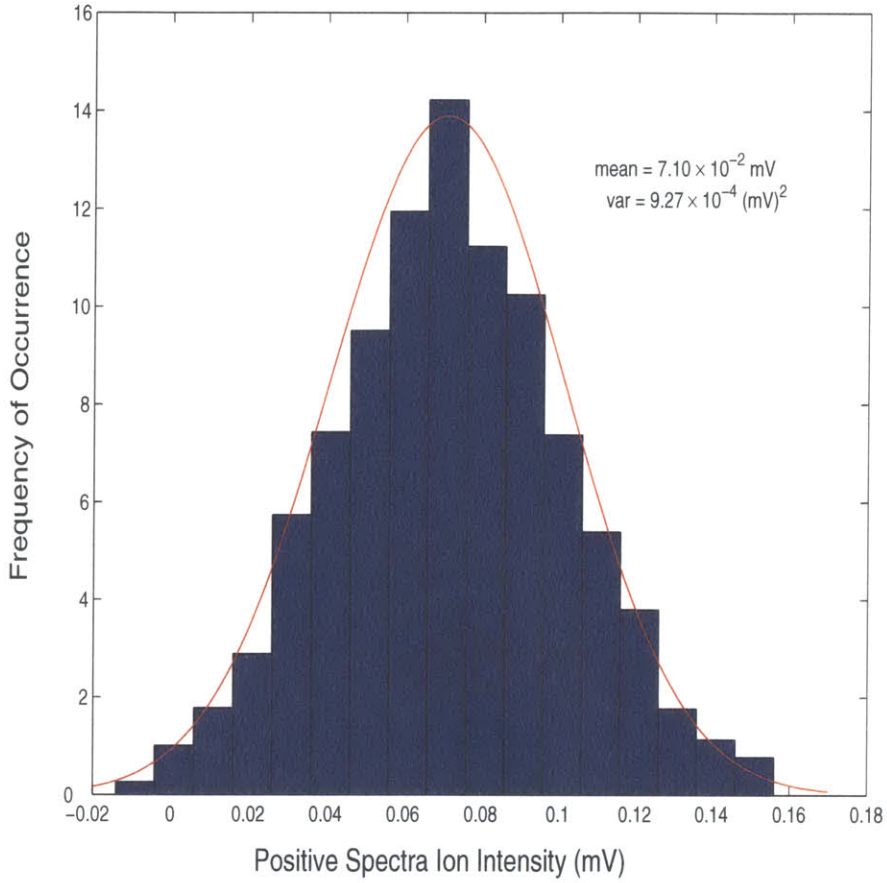


Figure 10: Histogram of noise with fitted Gaussian probability distribution function

The Gaussian distribution shown in Figure 10, when combined with the indication of decorrelation as seen in the covariance estimates, allows us to make the assumption that the overall covariance matrix is white and that each of the noise points is independent and identically distributed. The mean, μ_n, and variance, σ_n^2, of the noise was found to be 7.10×10^{-2} mV and 9.27×10^{-4} $(\text{mv})^2$, respectively.

6 Matched Filter and ACE Detectors

6.1 Matched Filter

6.1.1 Overview

As mentioned previously, there are two statistical analysis routes which we have chosen to take. The first route is a simplifed assumption on the covariance estimate which allows us to utilize noise that is independent and identically distributed (IID) and therefore has a diagonal covariance matrix that is equal for both H_0 and H_1 and does not require 100,000 samples to obtain a good estimate. Since we naively assume that the covariance matrix is white (i.e., the data is uncorrelated), the inversion is trivial and we can perform detection on the raw data. Thus, for a received signal, $r(t, V_c)$, we use all of the 50,000 points of a given sample in the detection, but the detector algorithm assumes a simple covariance matrix. This assumption leads us to two detectors: matched filter and ACE (adaptive cosine estimate) which we investigate further.

The optimal binary detection test was derived in Section 3.2. Through the assumption that the covariance matrices for both H_0 and H_1 are the same and diagonal as discussed above, we have $\Sigma_0 = \Sigma_1 = \sigma^2 I$ (where I denotes the identity matrix), and we can simplify the decision rule to:

$$\frac{\mathbf{r}^T(m_0 - m_1)}{\sigma^2} > ln(\frac{P_1}{P_0}) + \frac{1}{2\sigma^2}(m_0^T m_0 - m_1^T m_1) \tag{19}$$

This leads to our definition of the matched filter raw detector output:

$$\Lambda_{MF}(\mathbf{r}) \triangleq \frac{\mathbf{r}^T(m_0 - m_1)}{\sigma^2} \tag{20}$$

where the received signal $\mathbf{r}$ is taken as an inner product with the difference between the mean signals; this measure is compared against a threshold to produce a decision

output.

6.1.2 Assumptions

This detector makes the assumption that the noise is white and consequently that the covariance matrix of the noise is $\sigma^2 I$, where σ^2 denotes the variance. As mentioned previously, this assumption is supported in the previous section of noise analysis, but is not rigorously proven.

6.1.3 Implementation

The matched filter detector was trained separately on each of the five chlorite concentrations: 2.5 ppm, 5 ppm, 10 ppm, 20 ppm, and 40 ppm. This detector was then tested on each of the data samples for pure water and the five concentrations of chlorite. The likelihood ratio random variable is denoted by $\Lambda_{MF}(\mu_{train}, \mu_{test})$, where μ_{train} denotes the training concentration and μ_{test} denotes the concentration of the received signal. This random variable is characterized and documented with detector performance in section 7.2.

6.2 Adaptive Cosine Estimator

6.2.1 Overview

The adaptive cosine estimatator (ACE) [9] detector is similar to the matched filter detector, but relies on a measure of the angle between the received signal and the target signatures instead of a measure of the energy overlap. This is particularly important when there is a scaling factor of the received signal that varies with time or data collection. Unlike with the matched filter detector, the ACE detector will maintain a constant false alarm rate (CFAR) even with scale variations in the received signal, which is an important property for a deployed sensor. While the matched filter detector examines

the inner product of the received signal and the difference of the means, the ACE detector examines the angle between the received signal and the difference of the means as shown below:

$$\Lambda_{ACE}(\mathbf{r}) \triangleq \frac{\mathbf{r}^T(m_0 - m_1)}{||\mathbf{r}|| * ||m_0 - m_1||} \tag{21}$$

where $||\cdot||$ denotes the Euclidean norm.

This value is then compared against a threshold for a binary detector decision.

6.2.2 Assumptions

This detector makes the assumption that the noise is white and consequently that the covariance matrix of the noise is $\sigma^2 I$. As mentioned previously, this assumption is supported in the previous section of noise analysis.

6.2.3 Implementation

The ACE detector was trained separately on each of the five chlorite concentrations: 2.5 ppm, 5 ppm, 10 ppm, 20 ppm, and 40 ppm. This detector was tested on each of the data samples for pure water and the five chlorite concentrations. The likelihood ratio random variable is denoted by $\Lambda_{ACE}(\mu_{test}, \mu_{train})$, where μ_{test} denotes the testing concentration and μ_{train} denotes the training concentration. This random variable is characterized and documented with detector performance in section 7.2.

6.3 Performance Metrics

In order to evaluate different detectors, it is necessary to determine a set of metrics with which to compare relative performance. The likelihood ratio output, Λ_{ACE} and Λ_{MF}, for the ACE and MF detectors respectively, serves as an indicator of how much separation there is between contaminated water and pure water. To this end, it

is necessary to calculate the means of Λ for each contaminant concentration and detector. However, the mean itself does not indicate the degree of separation; the variance must also be used. We denote μ_{train} and μ_{test} as the training concentration and testing concentration, respectively. We define the following measures: $m_{detector}(\mu_{train}, \mu_{test})$ to denote the mean of $\Lambda_{detector}(\mu_{train}, \mu_{test})$, and $\sigma_{detector}(\mu_{train}, \mu_{test})$ to denote the standard deviation of $\Lambda_{detector}(\mu_{train}, \mu_{test})$.

We propose a metric to define the level of separation between two classes:

$$\gamma_{detector}(\mu_{train}, \mu_{test}) \triangleq \frac{m_c - m_w}{\sqrt{\sigma_c \sigma_w}} \tag{22}$$

where m_w and σ_w indicate the mean and standard deviation of Λ for the pure water class of a single type of detector and m_c and σ_c indicate the mean and standard deviation of Λ for the testing concentration, μ_{test}.

This quantity, γ, serves as a measure of detector performance evaluation that is independent of scaling factors of Λ and only depends on relative separation of the distributions between two classes. To illustrate this point, consider the case of multiplying a particular detector output random variable by a constant α. If the previous means were given by m_c and m_w, the new means will be $\alpha * m_c$ and $\alpha * m_w$. Similarly, if the previous standard deviations were given by σ_c and σ_w, the new standard deviations will be $\alpha\sigma_c$ and $\alpha\sigma_w$. Consequently, the new γ will be given by:

$$\gamma = \frac{\alpha m_c - \alpha m_w}{\sqrt{\alpha^2 \sigma_c \sigma_w}} \tag{23}$$

All factors of α cancel out, leaving the previous γ as before. This is important because it demonstrates that γ gives an accurate measure of how separated two classes are and does so without regard to the absolute values of Λ.

We recognize that m and σ completely characterize the distributions of the matched filter output, which is Gaussian. ACE, however, produces a T-distributed statistic which

is parameterized by a number of degrees of freedom and a non-centrality parameter [9]. However, we will focus on the first and second moments by utilizing our metric, γ, to compare the relative separation power of the detectors.

6.4 Results

As discussed in the previous section, it is important to examine the statistics for the likelihood ratio test, Λ_{ACE} and Λ_{MF}, to properly evaluate detector performance. Shown below in Figures 11 and 12 are the median, lower quartile, upper quartile, range of data and outliers [38] for Λ_{ACE} and Λ_{MF} respectively for each of the different classes (water and chlorite at 5 different concentrations). In both cases, the detectors were trained on the lowest concentration of 2.5 ppm of chlorite. As can be seen, by choosing a threshold of -40 for the ACE detector or -1.2 for the matched filter detector, the classes could correctly be separated for all of the data, thus yielding perfect detection on the limited available data.

In order to more easily compare the two detectors, we calculate γ as a function of the testing concentration for the training concentration of 2.5ppm; the plot of γ for both the matched filter and ACE detector is shown below in Figure 13.

As can be seen from the comparative γ values, the ACE detector performs better than the matched filter detector for all testing concentrations when trained at 2.5 ppm. Since the difference between the matched filter and ACE detectors lies in the measure of signal magnitude versus signal angle, it is evident that the received signal angle is a better indication of its class. This may be due to the fact that the data is scaled by an unknown and random factor which affects the relative energy in the signal. We can conclude that the target signature is best characterized by the angle it maintains instead of the energy and magnitude of relative points within the signal.

Also note that γ is not monotonically related to concentration. Since the detectors were trained at 2.5 ppm, this gives an early indication that the signals at higher con-

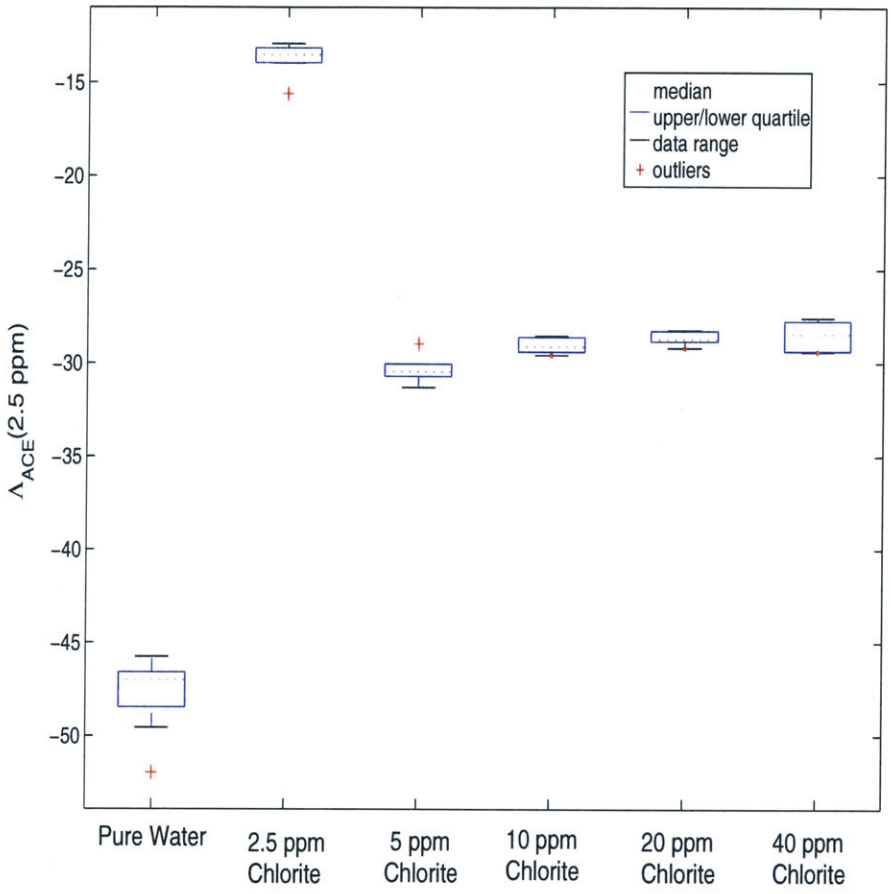


Figure 11: ACE detector output (Λ_{ACE}) statistics (μ_{train}=2.5 ppm)

centrations are not simply stronger versions than those at lower concentrations. This motivates the need for more sophisticated techniques to analyze the target signals.

In Figures 14 and 15, γ is plotted for the ACE and Matched Filter Detectors, respectively, for all training and testing concentrations. As can be seen, γ is uniformly higher when the testing concentration matches the training concentration (i.e., when the signature the detector is looking for is the one it was trained on). This makes intuitive sense, as the detector performance should be higher when the testing concentration is the same as the training concentration. However, we might have expected γ_{MF} to be increasing with the testing concentration if the chlorite signal were increasing in magnitude with concentration. Since this is not the case, we can infer that the chlorite signal

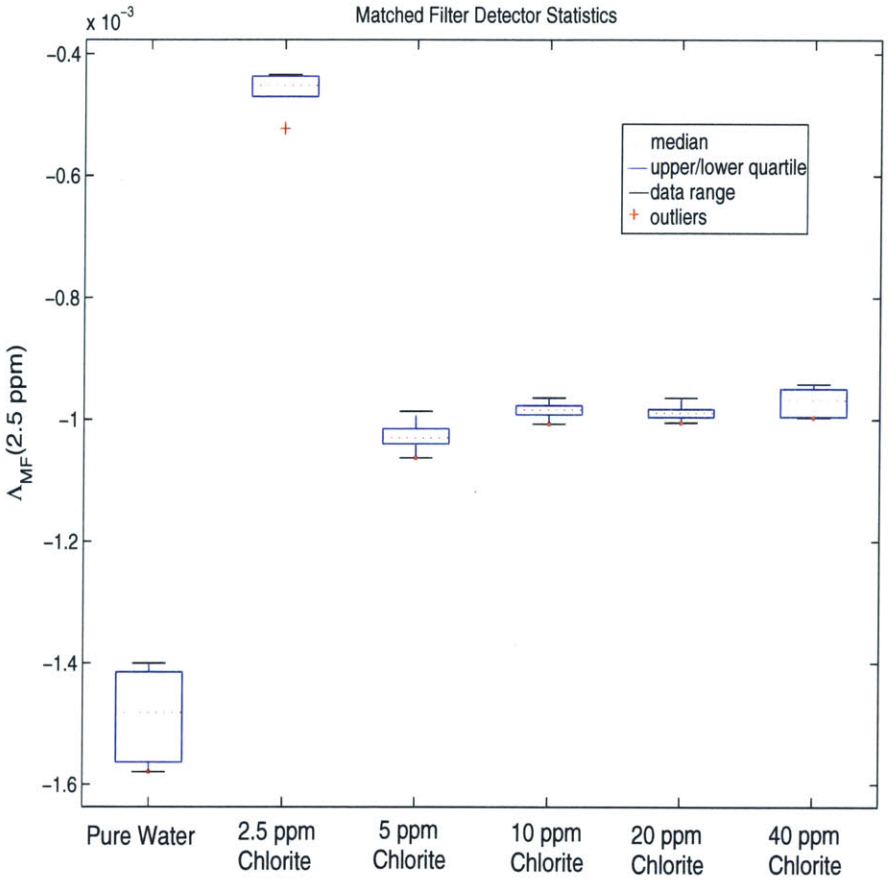


Figure 12: Matched filter detector output (Λ_{MF}) Statistics (μ_{train}=2.5 ppm)

varies nonlinearly. This gives another motivating factor for the need to analyze the target signals of various concentrations not simply as an increase in signal, but perhaps as a subspace inhabited by the class of chlorite-contaminated signals.

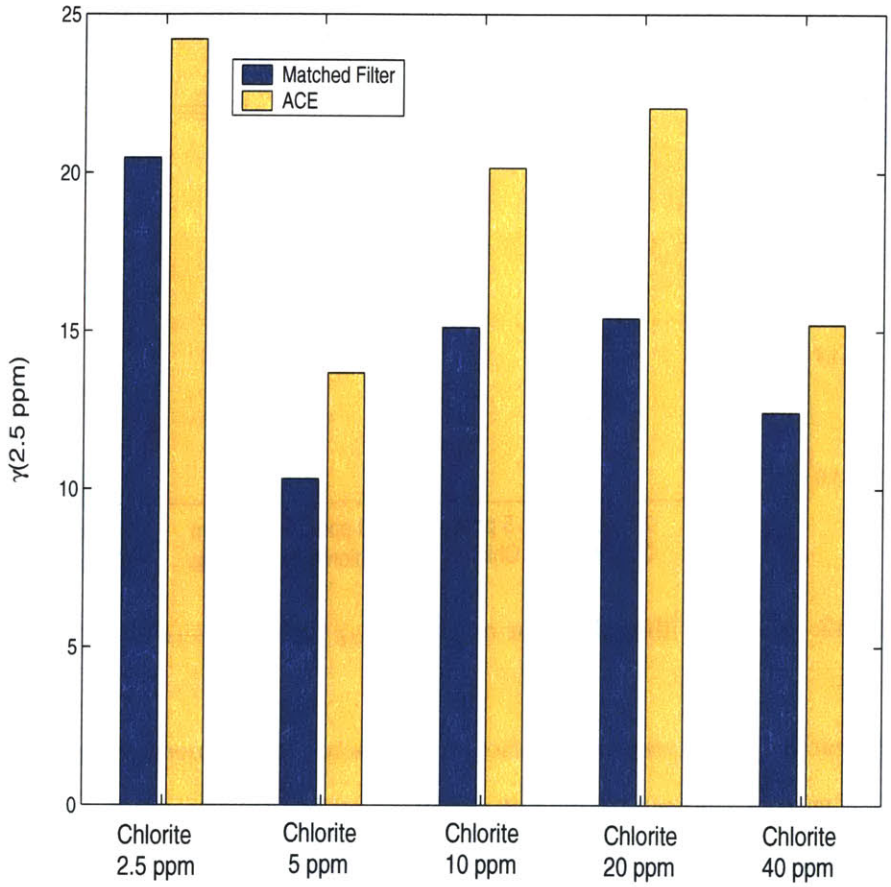


Figure 13: γ for matched filter and ACE detector (μ_{train}=2.5 ppm)

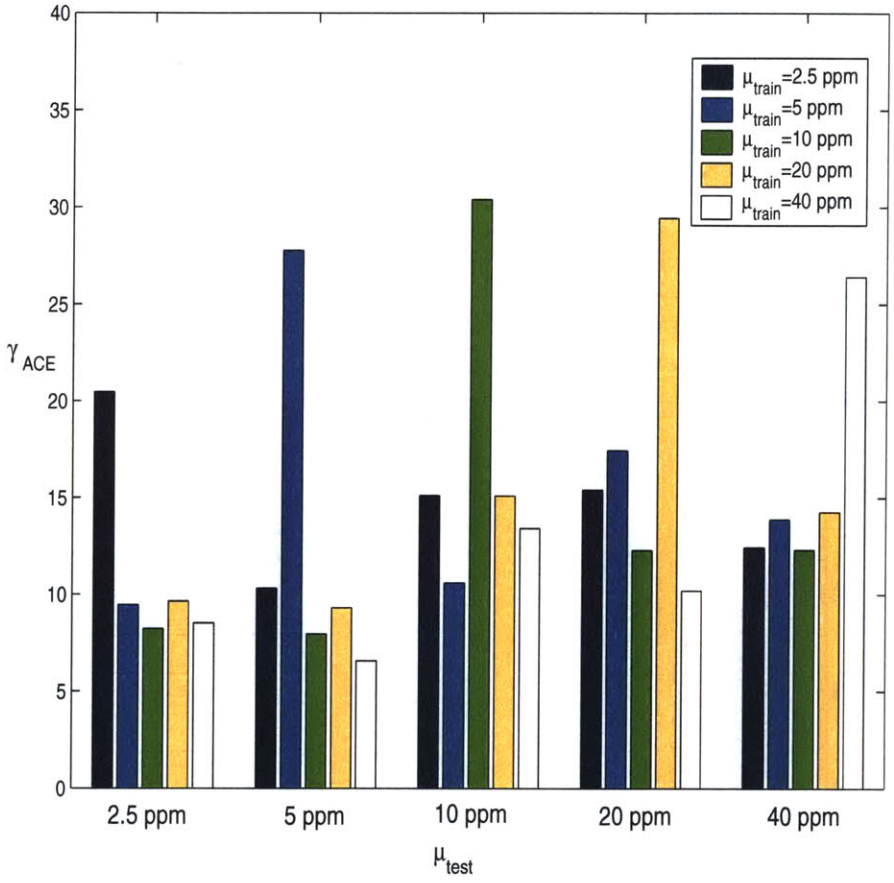


Figure 14: γ for ACE detector at all testing and training concentrations

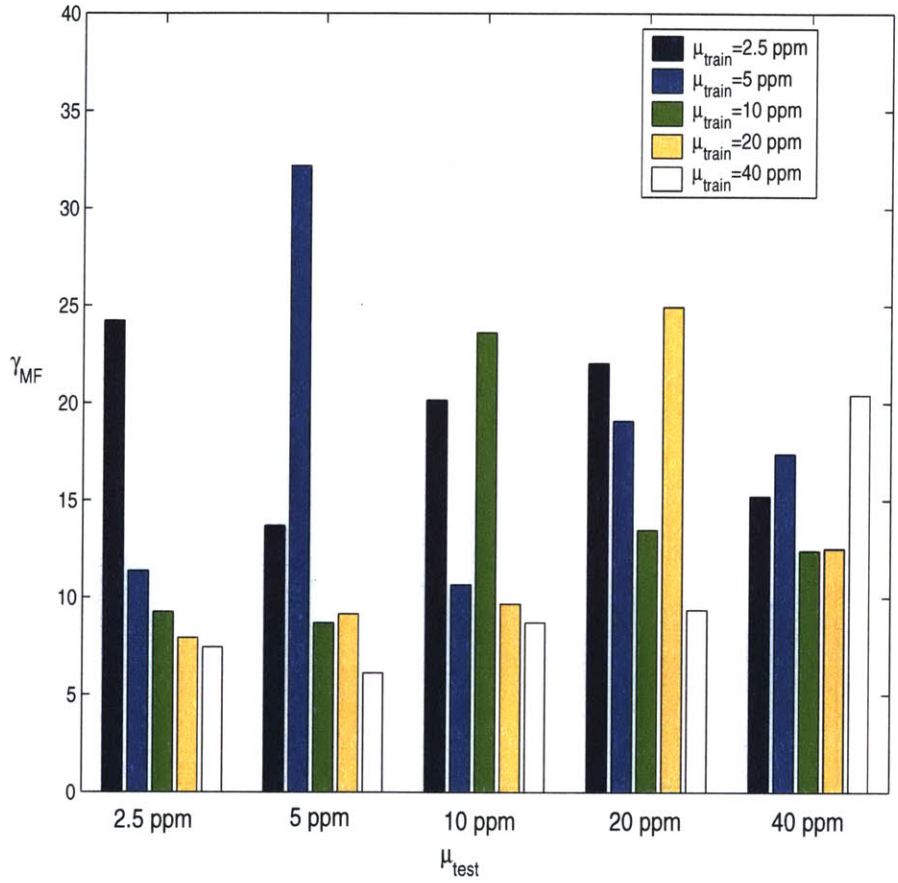


Figure 15: γ for matched filter detector at all testing and training concentrations

7 Multidimensional Target Subspaces

7.1 Overview & Motivation

The matched filter and ACE detectors discussed in previous sections assume that the target signatures (means) of various chlorite concentrations are scaled versions of one another. The detectors treat the training concentration mean as a vector and look for the received signal to be in the same direction as that vector, and expect that if the received signal concentration is higher, the received signal magnitude will increase, but the direction will be the same.

However, we have empirical evidence to show that the signatures of various concentrations of chlorite-contaminated water do not behave as discussed above. In Figures 16 and 17, we have plotted the mean energy as a function of concentration and the mean angle between higher concentrations of chlorite and chlorite at 2.5 ppm.

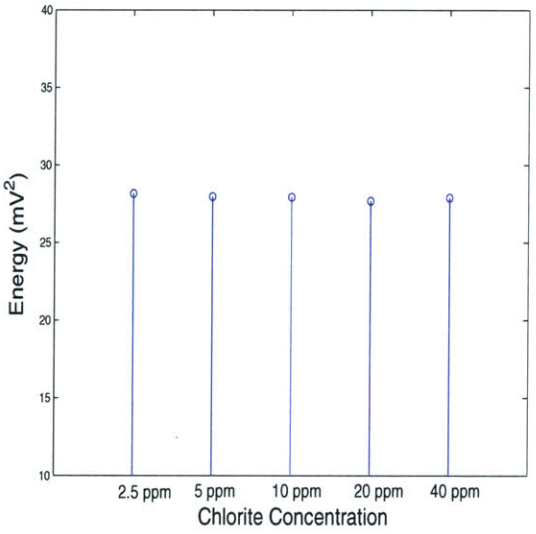


Figure 16: Mean energy for chlorite

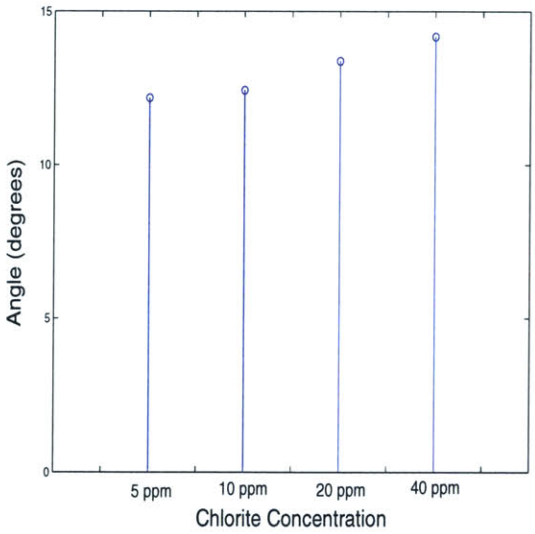


Figure 17: Angle between concentrations

As can be seen, the energy does not increase with concentration, as we would expect if the signal were scaling with concentration. Instead, the angle of the signature between

the signature of another concentration and chlorite at 2.5 ppm is significant and becomes more so as the concentration increases. It is evident that the signature is not simply scaling with concentration. Instead, higher concentrations of chlorite-contaminated water develop new features in varying locations, (t, V_c), and, in reference to the vector analogy, vary their direction as the concentration increases.

In a practical implementation of this detector, we will be looking to monitor a given water source and sound an alarm the first time chlorite is detected in the water. However, using this method, we are unable to predict at what concentration level the contaminant will first appear, making it difficult to best train the detector on the appropriate concentration.

In an effort to address this point and create a set of detectors that are more robust to varying levels of chlorite, we aim to create a subspace that is spanned by multiple levels of chlorite. As we recall from the single dimensional matched filter and ACE detectors, we projected the received signal onto the difference between the mean of chlorite and the mean of water. This quantity served as a measure of the energy and made up the basis for the matched filter detector. When this quantity was normalized by the energy in the signal, it became a measure of the angle of the signal and served as the basis for the ACE detector.

In the multidimensional case, we wish to extend the dimension of the target subspace onto which we project the received signal. Instead of consisting of a single vector composed of the mean difference of water and a single concentration of chlorite, we will create a subspace spanned by multiple vectors composed of the difference of water and varying concentrations of chlorite. The amount of energy in the received signal that lies in this subspace will be the basis for the matched filter multidimensional subspace detector. The angle that the received signal makes with this subspace will be the basis for the ACE multidimensional subspace detector.

It is important to note that when we begin looking for multiple dimensions that the

received signal lies in, we also open up the possibility of including more of the subspace that is spanned by pure water. This inclusion of multiple subspaces will increase the amount of the received signal that lies in those multiple subspaces, but will do so for both chlorite-contaminated water and pure water. Thus, a multidimensional subspace detector might perform better than a single dimensional detector compared to testing on a concentration different than that it was trained on, but we do not necessarily expect it to perform as well as the single dimensional detector when testing on the same concentration it was trained on.

For this section, we continue to maintain the previous assumption that the covariance is white and the noise is uncorrelated in both time and compensation voltage.

7.2 Implementation

We define a projection matrix, P_Φ, whose subspace, Φ, is spanned by the differences of the mean of water and means of the the five chlorite concentration signatures, denoted here by m_{cl1}, m_{cl2}, m_{cl3}, m_{cl4}, and m_{cl5}, where each is a column vector of length 50,000. The water mean is denoted by m_w. The subspace, Φ is given by:

$$\Phi = \begin{bmatrix} | & | & | & | & | \\ (m_{cl1} - m_w) & (m_{cl2} - m_w) & (m_{cl3} - m_w) & (m_{cl4} - m_w) & (m_{cl5} - m_w) \\ | & | & | & | & | \end{bmatrix}$$

The orthogonal projection matrix, P_Φ, which projects a received signal onto the subspace is given by [16]:

$$P_\Phi = \Phi(\Phi^T\Phi)^{-1}\Phi^T \tag{24}$$

Thus, the corresponding matched filter detector measures the energy of the received signal in the subspace, given by $\mathbf{r}^T P_\Phi \mathbf{r}$. The ACE detector measures the cosine-squared of the angle of the received signal and the subspace, given by $\frac{\mathbf{r}^T P_\Phi \mathbf{r}}{\mathbf{r}^T \mathbf{r}}$. We arrive at

the following definitions for the matched filter and ACE multi-dimensional subspace detectors:

$$\Lambda_{MFMD}(\mathbf{r}) = \frac{\mathbf{r}^T P_\Phi \mathbf{r}}{\sigma^2} \tag{25}$$

$$\Lambda_{ACEMD}(\mathbf{r}) = \frac{\mathbf{r}^T P_\Phi \mathbf{r}}{\mathbf{r}^T \mathbf{r}} \tag{26}$$

The multidimensional matched filter will produce a chi-squared statistic, while the multidimensional ACE filter will produce an F-statistic, both of which have numerous degrees of freedom [9]. However, as before, we will focus on the first and second moments by utilizing our metric, γ, to compare the relative separation power of the detectors.

7.3 Results

Shown in Figure 18 is the performance measure, γ, for the matched filter and ACE multidimensional detectors. The ACE detector does uniformly better than the matched filter detector, implying that the angle of a received signal is a better indication of its class than the relative energy in that signal. Recall that this result is similar to that of the single dimension case, in which the ACE detector also outperformed the matched filter. By examining the definition for the matched filter and ACE detectors, it is clear that the ACE detector normalizes the output by the energy in the received signal. Since the ACE detector performs better than the matched filter for this data, it suggests that the data may possess a varying scale multiplier that operates upon the energy in the received signal. Since ACE accounts for this, it performs better than the matched filter which bases its measurement off the energy in the signal.

The goal of implementing a multidimensional subspace detector is to improve the performance of the detectors across a range of concentrations. In the case of the multidimensional subspaces, the training of the detector occurs on a subspace consisting of all concentrations, while in the single dimensional case, the training occurs on a single

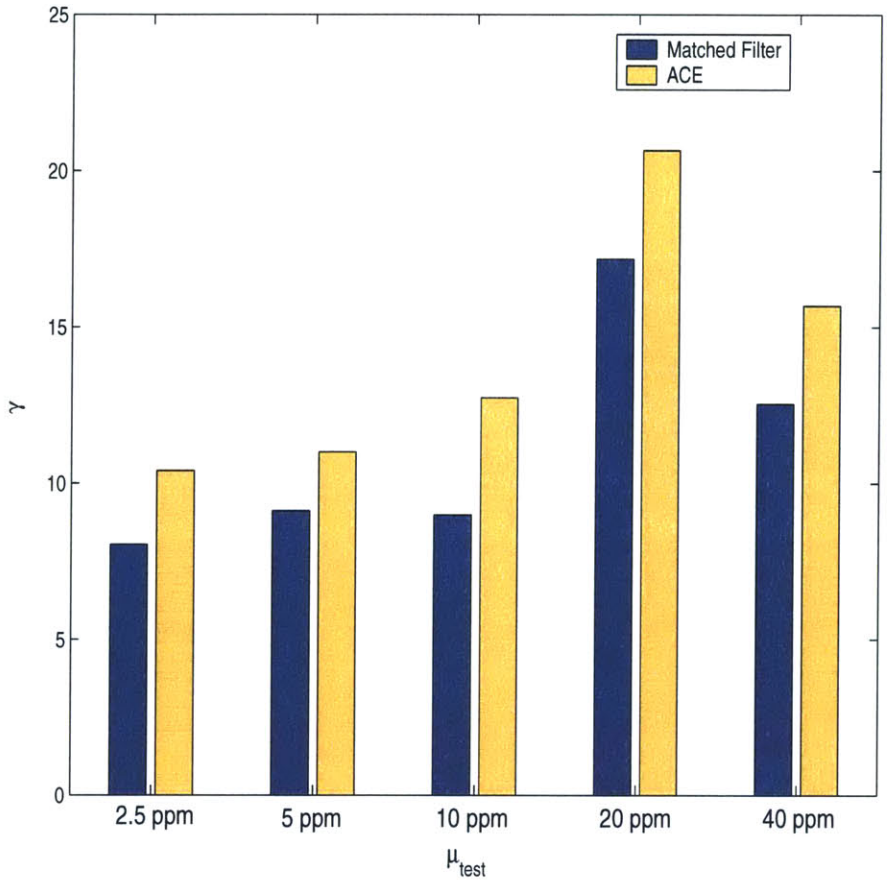


Figure 18: γ for matched filter and ACE multidimensional subspace detectors

concentration.

We would like to investigate the performance of the multidimensional detectors compared to that of the single dimensional detectors. In Figure 19, we compare the multidimensional results to the best and worst case results for that of the single dimensional detectors. In this case, the best case corresponds to $\mu_{train} = \mu_{test}$ and the worst case corresponds to the training concentration which yielded the lowest γ for each testing concentration. As can be seen, the best case single dimension detectors perform far better than the multidimensional detectors, with an average γ 1.78 times that of the multidimensional detectors. However, the multidimensional detectors perform generally better than the worst case single dimension detectors, with an average γ 1.38 times that

of the single dimensional detectors.

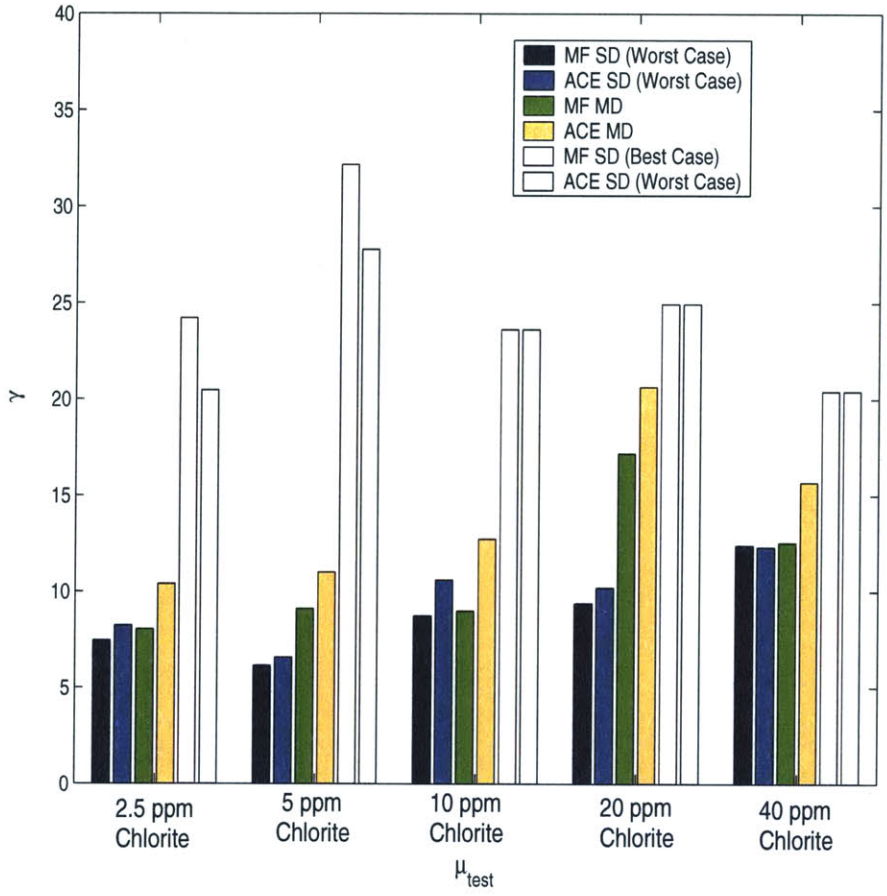


Figure 19: γ for multidimensional subspace detectors as compared to best and worst case single dimensional detectors

The results are not unexpected. When training on the subspace spanned by all of the means, we expect to get a better performance than when we test on a concentration different from the one we trained on. However, the best performance can still be obtained by training on the same concentration that we are testing on.

Table 4 below shows γ for both the matched filter and ACE multidimensional detectors. In column 2, for μ_{test}=2.5 ppm, we can see that the highest γ occurs when the training subspace only consists of the 2.5 ppm concentration, and uniformly decreases

as more concentrations of chlorite are added to the training subspace. In general, the performance measure, γ, decreases as more concentrations are added to the training subspace. By examining Figures 14 & 15, we can see that the training value of 2.5 ppm yields very good γ results (as compared to the other training values) across all testing concentrations, and thus it is expected that the γ value would decrease as more training concentrations are added. However, for a different training value, the addition of extra concentrations in the subspace can vastly improve the attainable γ as shown in Figure 19.

$\mu_{train} \backslash \mu_{test}$	2.5 ppm	5 ppm	10 ppm	20 ppm	40 ppm
[2.5 ppm]	28/34	14/16	21/27	23/25	16/18
[2.5 ppm 5 ppm]	14/19	21/26	13/19	23/31	17/21
[2.5 ppm 5 ppm 10 ppm]	11/14	13/15	14/20	19/24	15/19
[2.5 ppm 5 ppm 10 ppm 20 ppm]	9/12	10/13	10/14	21/28	15/19
[2.5 ppm 5 ppm 10 ppm 20 ppm 40 ppm]	8/10	9/11	9/13	17/21	13/16

Table 4: γ for matched filter & ACE multidimensional detectors (MF/ACE)

8 Wavelet Transform & Associated Detectors

8.1 Overview & Motivation

The wavelet transform is implemented as a method of reducing the dimensionality of the data. In both the matched filter and ACE detectors used on the raw data, all 50,000 data points in the signal were used by the detectors to develop a likelihood ratio test. In general, this requires an adequate estimate of the covariance, which in this case would be a 50,000 x 50,000 matrix. However, for the matched filter and ACE detectors, we made the key assumption that the covariance matrix was white and avoided the need to actually estimate the covariance.

An alternate route is to avoid making any assumptions about the covariance matrix, but instead perform detection by using a limited number of the coefficients in the signal. The covariance for a limited number of coefficients can be adequately estimated with the number of data samples available. However, the signal difference between pure water and chlorite-contaminated water is spread out over a vast number of coefficients, and thus it is be difficult to keep enough of those important coefficients to perform a reasonable detection.

This motivates the need to compress the relevant chlorite features into a smaller number of coefficients, and then use those coefficients, along with a reasonable estimate of the covariance of those coefficients in a corresponding matched filter and ACE detector.

A wavelet transform is suitable for the given data because of the shape of the signatures that are indicative of chlorite-contaminated water. Since each coefficient in the wavelet transformed data corresponds to the energy in the overlap of a wavelet and the raw data for a particular location and scale, the energy contained in the raw data over a range of points is now concentrated in a few wavelet coefficients. Thus, instead of spreading out the relevant information in a number of coefficients in the raw data, the wavelet transform will serve to consolidate the relevant information into a smaller

number of coefficients as represented in the wavelet transform.

8.2 Implementation

A 1-dimensional wavelet transform is performed on each time scan in the data. This is done instead of a 2-dimensional wavelet transform since the rolloff of the peaks occur only in neighboring compensation voltage points, and not in time points; thus, the inherent shape of the chlorite signature is matched only by wavelets in compensation voltage axis.

The wavelet coefficients resulting from the wavelet transform are given by:

$$c_k^{j-1} = \sum_n h_{n-2k} c_n^j \tag{27}$$

$$d_k^{j-1} = \sum_n g_{n-2k} c_n^j \tag{28}$$

where h and g are finite impulse response (FIR) quadrature-mirror filters corresponding to the Daubechies 6 wavelet [11], and the highest level coefficients, c_n^5, correspond to the original signal. A level 5 decomposition (equations (27) and (28)are iterated 5 times to produce c_n^0 & d_n^0) is performed on the 100 point signal, yielding 6 approximation coefficients and 94 detail coefficients of 5 scales for each time scan. A new 3-dimensional signal is produced as a concatenation of all the 1-dimension wavelet transform coefficients as a function of time.

In the resulting wavelet transformed data, the means for each of the two classes (pure water and a particular chlorite concentration) is computed, along with the difference of the means. A limited number of coefficients (1-6 in our tests) is kept according to the coefficients that demonstrate the largest difference between the means of the two classes are; detection is performed on those coefficients only.

8.3 Associated Matched Filter & ACE Detectors

The corresponding detectors for the new wavelet-transformed data are a direct extension of the matched filter and ACE detectors for the raw data. However, since we are no longer assuming a white covariance matrix, it becomes necessary to whiten both the received signal and the corresponding target mean signatures.

As mentioned previously, the coefficients that demonstrate the largest difference between the means of the two classes are kept. Let us denote that vector of coefficients as $\mathbf{w}$, where $\mathbf{w} \in \Re^n$, where n is between 1 and 6 and denotes the number of coefficients retained.

We are able to estimate a covariance for these n points, yielding a covariance matrix denoted here as Σ_{WT}, which can decomposed into its eigenvector decomposition as shown below [16]:

$$\Sigma_{WT} = U\Lambda U^T \tag{29}$$

where U denotes the matrix of eigenvectors and Λ denotes the diagonal matrix of the corresponding eigenvalues.

The associated whitening matrix, $R^{-\frac{1}{2}}$, is given by [8]:

$$R^{-\frac{1}{2}} = (U\Lambda^{\frac{1}{2}})^{-1} \tag{30}$$

By substituting whitened versions of our received signal and target signatures into the earlier derivation of the matched filter and ACE detectors, we obtain the following detectors:

$$\Lambda_{MFWT}(\mathbf{w}) \triangleq (R^{-\frac{1}{2}}\mathbf{w})^T(R^{-\frac{1}{2}}(m_0 - m_1)) \tag{31}$$

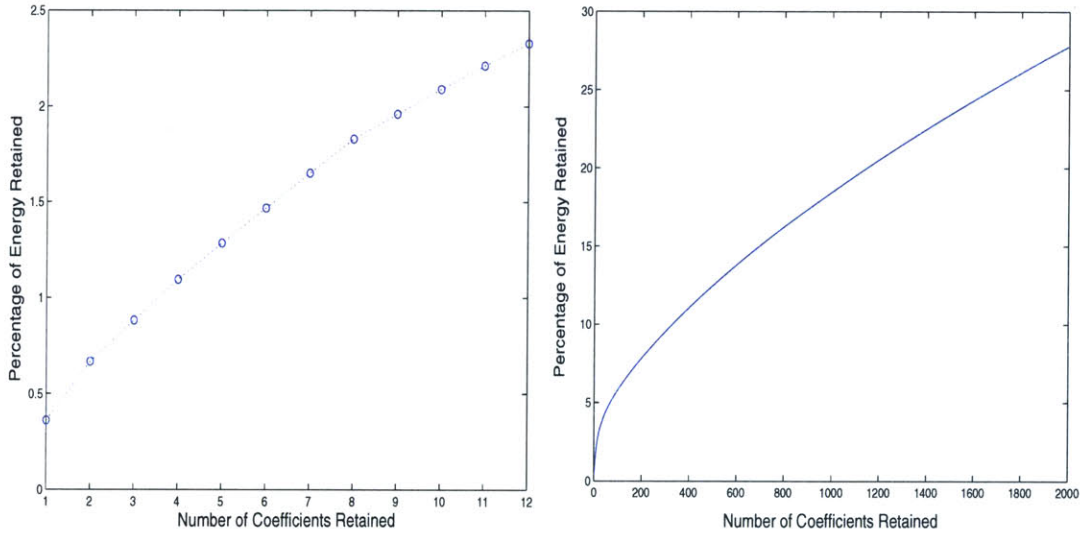
$$\Lambda_{ACEWT}(\mathbf{w}) \triangleq \frac{(R^{-\frac{1}{2}}\mathbf{w})^T(R^{-\frac{1}{2}}(m_0 - m_1))}{||R^{-\frac{1}{2}}\mathbf{w}|| * ||R^{-\frac{1}{2}}(m_0 - m_1)||} \tag{32}$$

Thus, the signals are whitened using the inverse of a reasonable estimate of the

covariance, and then implemented in the matched filter and ACE detectors as previously derived.

8.4 Results

We wish to investigate the performance of the wavelet transform detectors relative to that of the raw data detectors. To this end, we will first examine the amount of energy retained when the largest wavelet coefficients are kept and the remaining portion of the signal discarded. As can be seen in Figure 20, a relatively small portion of the overall signal energy is retained when only n coefficients are kept. This is due either to the inability to compress relevant information in such a small number of coefficients, or due to a large amount of noise energy on the signal that distorts the total energy relative to the relevant signal energy.



Range of 12 coefficients Range of 2000 coefficients

Figure 20: Percentage of energy retained versus number of coefficients retained after wavelet transform

In Figures 21 and 22, we can see the detector performance for the wavelet transform matched filter and ACE detectors. The first thing to notice is that the ACE wavelet

transform detector performance is not impressive. There is little, if any, level of differentiation possible between the different classes of chlorite and water. In contrast, the statistics for the matched filter wavelet transform detector show a clear level of differentiation between the classes.

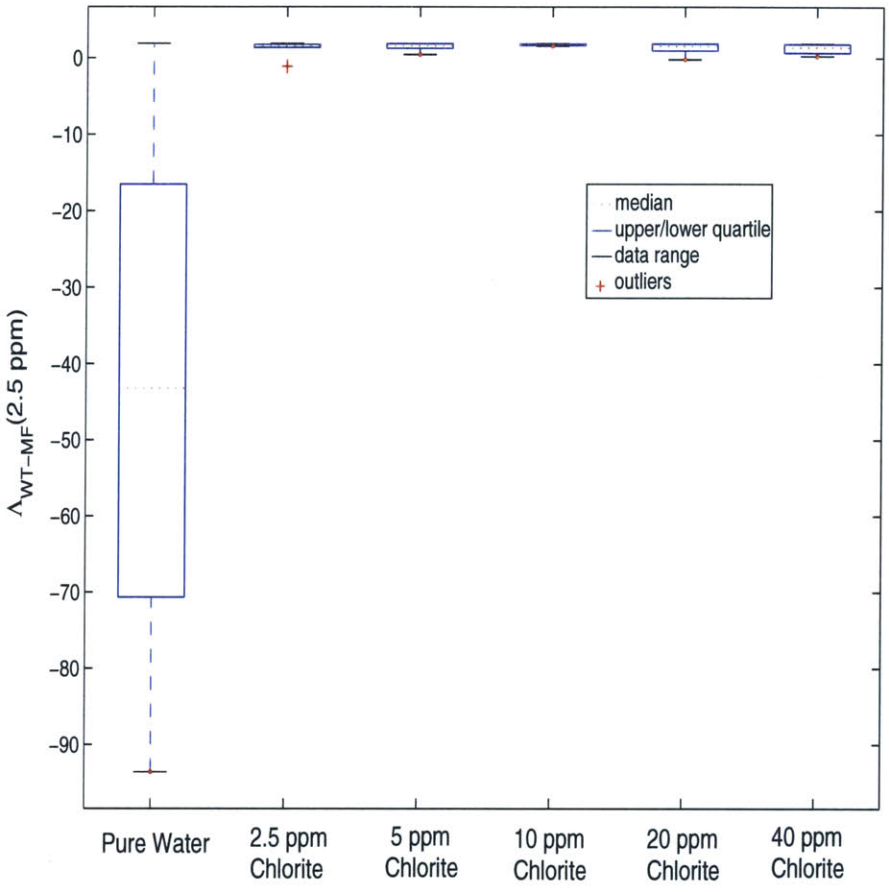


Figure 21: Wavelet transform matched filter detector statistics (μ_{train}=2.5ppm, 6 coefficients retained)

This is not particularly surprising. After applying a wavelet transform to the data, we chose to retain the coefficients that showed the greatest signal differentiation energy. As a result, we should be looking to the energy in those coefficients to determine to which class a received signal belongs. By normalizing the received signal by its energy, the ACE detector effectively removes much of the signal differentiation power, and the

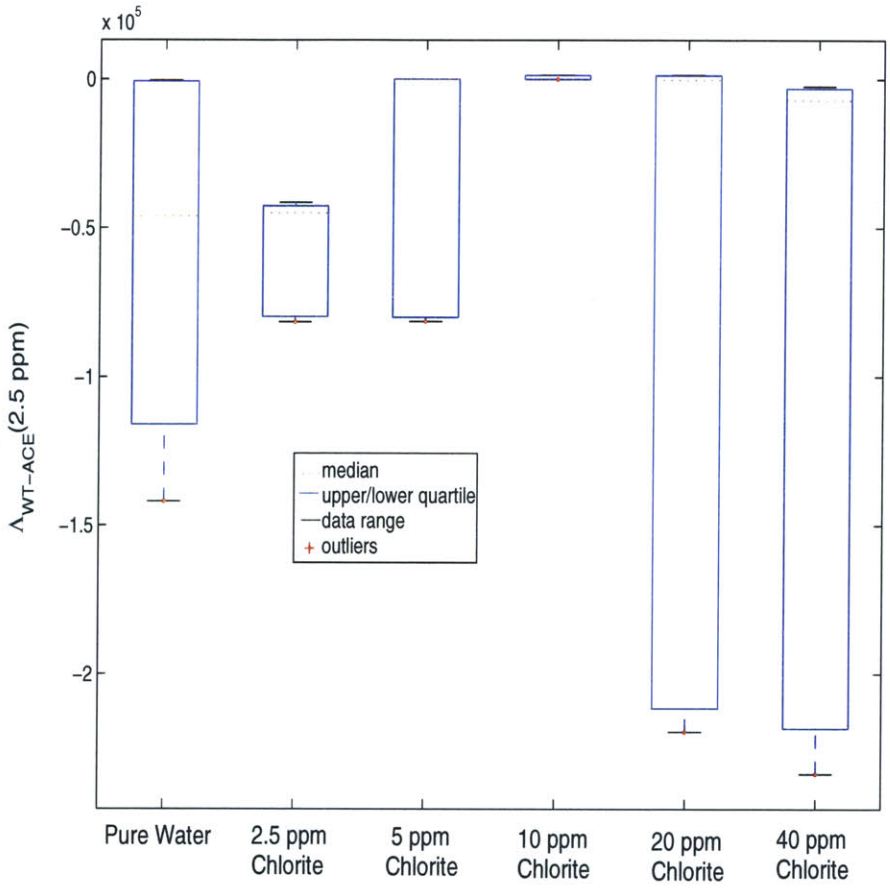


Figure 22: Wavelet transform ACE detector statistics (μ_{train}=2.5 ppm, 6 coefficients retained)

detector becomes very weak.

We will consider only the matched filter wavelet transform detector in subsequent sections. Consider the performance evaluation coefficient, γ, for the matched filter wavelet transform detector in Figure 23. In many cases, the matched filter wavelet transform detector does not perform as well as the raw data detectors (compare Figures 23 and 15), as the raw data detectors have a combined average γ of 16.7, while the matched filter wavelet transform detector has an average γ of 9.3.

However, as can be seen in Figure 24, for the testing and training concentration of 2.5 ppm, the wavelet transform detectors with 6 coefficients retained perform far better than

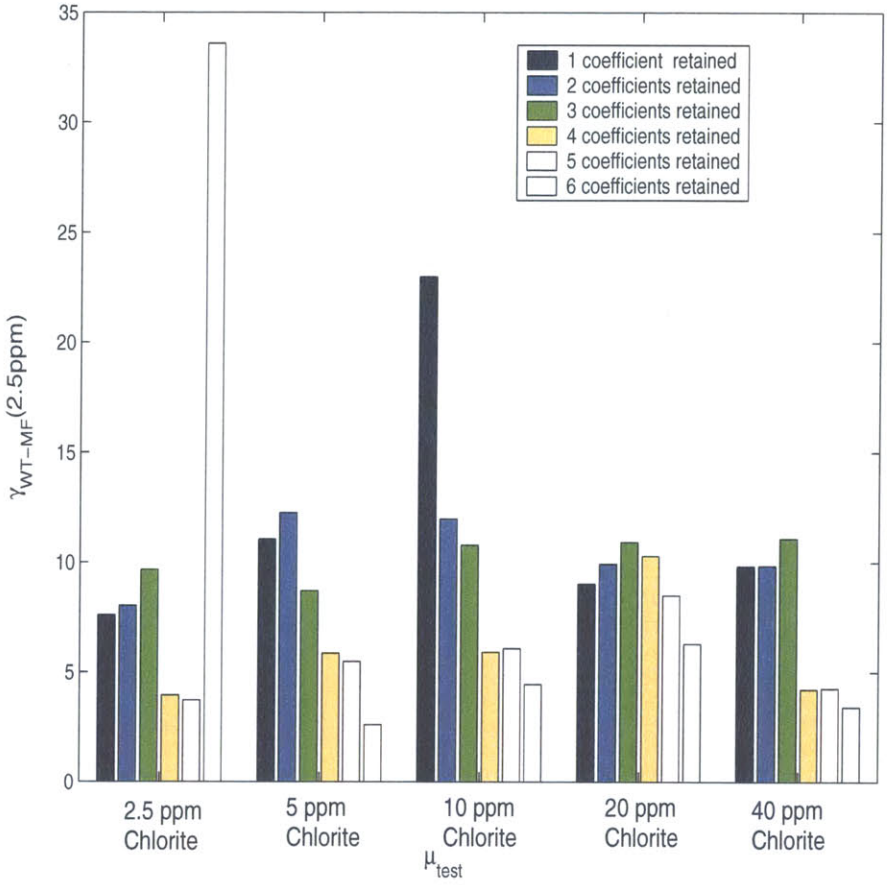


Figure 23: γ for matched filter wavelet transform detector at 6 different coefficient levels and all testing concentrations

the either of the raw data detectors. Thus, under certain circumstances, it is possible for the wavelet transform detector to perform better. This is likely due to the fact that there are fewer differences in the signal between the 2.5 ppm chlorite signal and the pure water signal, meaning that the relevant separation can be compressed into a smaller number of coefficients. In addition, the more accurate estimation of the covariance could play a role in the increased detection power. However, in all cases except for μ_{test} =2.5 ppm, the wavelet transform matched filter detector with 6 coefficients retained yields the worst performance (as compared to the detectors with fewer coefficients retained). This could indicate that although the number of coefficients is increasing, the relative quality of the

covariance estimate is decreasing and is thus giving a lower performance.

While the wavelet detector performs well in this case, in general we can see that the wavelet transform detectors perform at a lower level than the raw data (matched filter and ACE) detectors. However, this difference may not be as relevant as it seems. While

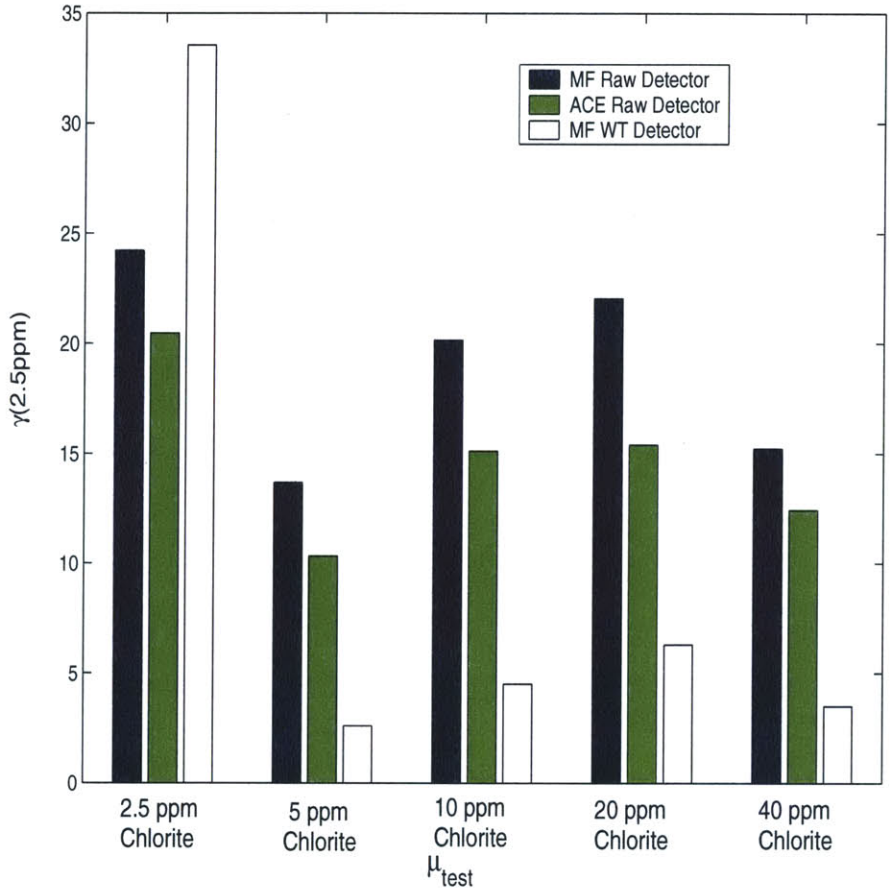


Figure 24: γ for raw and wavelet transform detectors with 6 coefficients retained (μ_{test}=2.5 ppm)

the wavelet transform detectors do not generally perform as well, by looking at the raw statistics on the detector output in Figur **??**, we can see the signal differentiation is still possible, and with the appropriate threshold, the thresholded binary output would frequently be the same for both types of detectors. Thus, a large amount of the relevant signal differentiation energy and relative detector performance is still retained while

allowing for a factor of 10,000 in reduction of data size.

9 Conclusion

9.1 Summary

The goal of this thesis was to develop and evaluate several methods for the detection of contaminants in water. We employed techniques well-established for their use on other types of data and examined their application to the FAIMS data.

With a goal of minimizing the error rate of the detector, we began by supposing an optimal strategy assuming an infinite quantity of data for use in analysis. One significant challenge to our work was the limited number of data samples available, and we examined the various assumptions that allowed us to proceed to a series of different detectors.

We began by analyzing and evaluating two complimentary subspace detectors: a matched filter detector and an adaptive cosine estimate (ACE) detector. Both of these detectors performed very well; in all cases the detector produced the correct output and yielded a high level of separation. The success which we attained with these detectors validated our earlier assumption that the noise was uncorrelated in time and compensation voltage.

A general trend that arose was that the ACE detector on average performed better than the matched filter, with a mean γ of approximately 1.2 times that of the matched filter γ. Since the ACE detector is scale-invariant to the received data while the matched filter detector relies on the overall energy in the received signal, we can conclude that an unknown and uncharacterized scaling factor may play a role in the acquisition of the raw data.

We also considered the case of a multi-dimensional target subspace for both the matched filter and ACE detectors. We showed that the energy in the means of various classes did not increase with concentration, thus disproving the simple model of the signal being scaled with concentration. Instead, we demonstrated that the relative angle

of the each class was changing with concentration, and this prompted the need to look in a subspace in which the received signal might lie.

We found that the resulting multidimensional detector trained on all concentrations was more robust but performed worse than a detector trained on a single concentration. Specifically, the multidimensional detectors performed better than the worst-case single dimensional detectors, but worse than the best-case single dimensional detectors. Thus, the multidimensional detectors traded in peak performance for a better worst-case performance, and served to make γ consistent across all testing concentrations and less prone to changes in performance with the introduction of a new class.

We also employed a wavelet transform to reduce the dimensionality of the original data and improve the covariance estimates. We implemented corresponding matched filter and ACE detectors on a limited number of those wavelet coefficients. The results were mixed. The ACE detector did not perform particularly well, but this was not surprising given that the wavelet coefficients had been picked based on their relative energy and not on their relative angle. Effectively, the wavelet coefficients were picked to maximize performance with the matched filter detector.

The matched filter detector, however, performed relatively well given that we were keeping such a small fraction of the energy in the signal. The detector was able to differentiate between pure water and chlorite, though it did so with a much smaller performance factor, γ, than the raw data detectors. Given the compression factor of 10,000, the detector performed remarkably well, and it is reasonable to assume that for many applications, this tradeoff is reasonable.

9.2 Future Work

From the work done on the detectors discussed above, it is evident that there is a plethora of future work to be investigated in the pursuit of detecting the lowest possible concentration of contaminant with the lowest rate of error.

In the implementation and evaluation of the wavelet transform and associated detectors, we mentioned that the ACE detector performed badly because we had made our initial choice of the coefficients to be retained based on the energy they contributed. Instead, however, it would be possible to choose the retaining coefficients based on the overall angle they produced between the two classes. In this case, the matched filter detector would probably not work very well, but the ACE detector would work much better. This is a detector situation that has yet to be pursued.

In the work on multidimensional subspaces, we defined the subspace based on varying chlorite concentrations. However, it would be equally possible to define a subspace based on the mean of a particular concentration and the direction of highest variance for that same concentration. This would serve to create a detector that was trained on only one concentration but could possibly serve to improve detector performance for that concentration over the single dimension detectors.

In many cases, we were limited by the amount of data available. It would be ideal to be able to take more data in a consistent fashion to yield better results. This would mean being able to have separate training and testing data, and having much better estimates of the covariance in the case of the wavelet transform results. This would also allow us to do experiments with retaining more and more wavelet coefficients, which would allow us to determine exactly how many wavelet coefficients needed to be kept to develop equal or better performance than that of the ACE and matched filter detectors.

References

[1] R. A. Miller, G. A. Eiceman, E. G. Nazarov, and T. A. King. A MEMS Radio-Frequency Ion Mobility Spectrometer for Chemical Agent Detection.
2000 Solid-State Sensor and Actuator Workshop.

[2] R. A. Miller, E. G. Nazarov, G. A. Eiceman, and T. A. King. A MEMS Radio-Frequency Ion Mobility Spectrometer for Chemical Vapor Detection.
Sensors and Actuators., 91, 307-318 (2001).

[3] C. S. Reese, S. J. Taraszewsky. Near-Real-Time Analysis of Toxicologically Important Compounds Using the Volatile Organic Analyzer for the International Space Station.
Proceedings of the 8th International Conference on IMS., (1999) Buxton, U.K.

[4] U. Kh Rasulev, E. G. Nazarov, V. V. Palitsin. Surface Ionization Gas-Analyzing Devices with Separation of Ions by Mobility.
4th International Workshop on IMS., (1995) Cambridge, U.K.

[5] E. A. Mason and E. W. McDaniel. Transport Properties of Ions in Gases.
Wiley, New York (1998).

[6] I.A. Buryakov, E.V. Krylov, E.G. Nazarov, and U.K. Rasulev. A New Method of Separation of Multi-atomic Ions by Mobility at Atmospheric Pressure using a High-frequency Amplitude Asymmetric Strong Electric Field.
Int. J Mass Spectrometry, v128, 143-148 (1993).

[7] I.S. Reed, J. D. Mallett, and L.E. Brennan. Rapid Convergence Rate in Adaptive Arrays.
IEEE Transactions on Aerospace and Electronic Systems, v6, 853-863 (1974).

[8] A.S. Willsky, G.W. Wornell, and J.H.S Shapiro. Stochastic Processes, Detection and Estimation: 6.432 Course Notes. *MIT, Cambridge, MA (2002).*

[9] S. Kraut, L. L. Scharf, and L. T. McWhorter. Adaptive Subspace Detectors.
IEEE Transactions on Signal Processing, v49, 1-16 (2001).

[10] S. Kraut, L. L. Scharf, and L. T. McWhorter. Adaptive Subspace Detectors.
Proceedings of the IEEE: Special Issue on Wavelets, v84 (1996).

[11] E. Aboufadel, S. Schlicker. Discovering Wavelets.
Wiley, New York (1999).

[12] R. Purves and R. Guevremont. Electrospray Ionization - High-Field Asymmetric Waveform Ion Mobility Spectrometry - Mass Spectrometry.
Anal. Chem., 71, 2346-2357 (1999).

[13] B. Ells, D. A. Barnett, K. Froese, R. W. Purves, S. E. Hrudey, R. Guevremont. Detection of Chlorinated and Brominated Byproducts of Drinking Water Disinfection Using Electrospray Ionization - High-Field Asymmetric Waveform Ion Mobility Spectrometry - Mass Spectrometry.
Anal. Chem., 71, 4747-4752 (1999).

[14] I.A. Buryakov, E.V. Krylov, E.G. Nazarov, and U.K. Rasulev. A New Method of Separation of Multi-atomic Ions by Mobility at Atmospheric Pressure using a High-frequency Amplitude Asymmetric Strong Electric Field.
Int. J Mass Spectrometry, v128, 143-148 (1993).

[15] T. Younos. Advances in Water Monitoring Research.
Water Resources Publications, LLC (2001).

[16] G. Strang. Introduction to Linear Algebra.
Wellesley Cambridge Press, Cambridge (2003).

[17] R. Miller, G. Eiceman, E. Nazarov and A. King. A novel micromachined high-field asymmetric waveform-ion mobility spectrometer.
Sensors and Actuators 67, 300-306 (2000).

[18] R. Miller, G. Eiceman, E. Nazarov and A. King. A MEMS radio-frequency ion mobility spectrometer for chemical vapor detection.
Sensors and Actuators 91, 307-318 (2001).

[19] C. Davis, J. Kang, C. Dube, J. Borenstein, E. Nazarov, R. Miller, and A. Zapata. Spore biomarker detection using a mems differential mobility spectrometer.
Transducers, Solid-State Sensors, Actuators and Microsystems, 12th International Conference, 1233-1237 (2003).

[20] M. Krebs, A. Zapata, E. Nazarov, R. Miller, I. Costa, A. Sonenshein, and C. Davis. Detection of Biological and Chemical Agents using Differential Mobility Spectrometry (DMS) Technology.
IEEE Sensors Journal (2005) In Press.

[21] G. Eiceman, B. Tadjikov, E. Krylov, E. Nazarov, R. Miller, J. Westbrook, and P. Funk. Miniature radio-frequency mobility analyzer as a gas chromatographic detector for oxygen-containing volatile organic compounds, pheromones, and other insect attractants.
Int. J Chromatography 917, 205-217 (2001).

[22] N. Krylova, E. Krylov, G. Eiceman, and J. Stone. Effect of Moisture on the Field Dependence of Mobility for Gas-Phase Ions of Organophosphorus Compounds at Atmospheric Pressure with Field Asymmetric Ion Mobility Spectrometry.
J Phys Chem A 107, 3648-3654 (2003).

[23] R. Miller, G. Eiceman, E. Nazarov, and A. King. A MEMS Radio-Frequency Ion Mobility Spectrometer for Chemical Agent Detection.

Solid-State Sensor and Actuator Workshop June 4-8. (2000) Hilton Head Island, SC.

[24] E. Krylov, E. Nazarov, R. Miller, B. Tadjikov, and G. Eiceman. Field Dependence of Mobilities for Gas-Phase-Protonated Monomers and Proton-Bound Dimers of Ketones by Planar Field Asymmetric Waveform Ion Mobility Spectrometer (PFAIMS).
it J Phys Chem 106, 5437-5444 (2002).

[25] G. Eiceman, E. Nazarov, R. Miller, E. Krylov, and A. Zapata. Micro-machined planar field asymmetric ion mobility spectrometer as a gas chromatographic detector.
The Analyst 127, 466-471 (2002).

[26] G. Eiceman, E. Nazarov, B. Tadjikov, and R. Miller. Monitoring Volatile Organic Compounds in Ambient Air Inside and Outside Buildings with the Use of a Radio-Frequency-Based Ion-Mobility Analyzer with a Micromachined Drift Tube.

[27] B. Ells, D. Barnett, R. Purves, and R. Guevremont. Trace level determination of perchlorate in water matrices and human urine using ESI-FAIMS-MS.
J. Environ. Monit. 2, 393-397 (2000).

[28] B. Ells, D. Barnett, K. Froese, R. Purves, S. Hrudey, and R. Guevremont. Detection of Chlorinated and Brominated Byproducts of Drinking Water Disinfection Using Electrospray Ionization-High-Field Asymmetric Waveform Ion Mobility Spectrometry-Mass Spectrometry.
Anal. Chem. v71, 4747-4752 (1999).

[29] W. Gabryelski, F. Wu, and K. Froese. Comparison of High-Field Asymmetric Waveform Ion Mobility Spectrometry with GC Methods in Analysis of Haloacetic Acids in Drinking Water.
Anal Chem v75, 2478-2846 (2003).

[30] R. Handy, D. Barnett, R. Purves, G. Horlick, and R. Guevremont. Determination of nanomolar levels of perchlorate in water by ESI-FAIMS-MS.
J. Anal. At. Spectrom. v15, 907-911 (2000).

[31] M. Unser and A. Aldroubi. A Review of Wavelets in Biomedical Applications.
Proceedings of the IEEE, v84, 4 (1996).

[32] M. Unser. Wavelets, Statistics and Biomedical Application.
8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing (1996) Corfu, Greece.

[33] A. Cohen and J. Kovacevic. Wavelets: The Mathematical Background.
Proceedings of the IEEE, v84, 4 (1996).

[34] M. Misiti, Y. Misiti, G. Oppenheim, J. Poggi. Wavelet Toolbox User's Guide for Use with MATLAB.
Mathworks, Natick (2000).

[35] L. Sharf and B. Friedlander. Matched subspace detectors.
IEEE Transactions on Signal Processing, v42, 8, 2146-2157 (1994).

[36] L. Scharf. Statistical Signal Processing.
Addison-Wesley, Reading (1991).

[37] W. Merrick, J. Zeskind, M. Krebs, and C. Davis. Monitoring Drinking Water Quality Using Pyrolysis-Differential Mobility Spectrometry.
Rapid Comm Mass Spec., In Review. (2005).

[38] Statistics Toolbox User's Guide for Use with MATLAB.
Mathworks, Natick (1997).