Computational Statistical Methods in Chemical Engineering

by

Mark Christopher Molaro

Submitted to the Department of Chemical Engineering
on January 29, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Chemical Engineering

Abstract

Recent advances in theory and practice, have introduced a wide variety of tools from machine learning that can be applied to data intensive chemical engineering problems. This thesis covers applications of statistical learning spanning a range of relative importance of data versus existing detailed theory. In each application, the quantity and quality of data available from experimental systems are used in conjunction with an understanding of the theoretical physical laws governing system behavior to the extent they are available.

A detailed generative parametric model for optical spectra of multicomponent mixtures is introduced. The application of interest is the quantification of uncertainty associated with estimating the relative abundance of mixtures of carbon nanotubes in solution. This work describes a detailed analysis of sources of uncertainty in estimation of relative abundance of chemical species in solution from optical spectroscopy. In particular, the quantification of uncertainty in mixtures with parametric uncertainty in pure component spectra is addressed. Markov Chain Monte Carlo methods are utilized to quantify uncertainty in these situations and the inaccuracy and potential for error in simpler methods is demonstrated. Strategies to improve estimation accuracy and reduce uncertainty in practical experimental situations are developed including when multiple measurements are available and with sequential data. The utilization of computational Bayesian inference in chemometric problems shows great promise in a wide variety of practical experimental applications.

A related deconvolution problem is addressed in which a detailed physical model is not available, but the objective of analysis is to map from a measured vector valued signal to a sum of an unknown number of discrete contributions. The data analyzed in this application is electrical signals generated from a free surface electro-spinning apparatus. In this information poor system, MAP estimation is used to reduce the variance in estimates of the physical parameters of interest. The formulation of the estimation problem in a probabilistic context allows for the introduction of prior knowledge to compensate for a high dimensional ill-conditioned inverse problem. The estimates from this work are used to develop a productivity model expanding on previous work and showing how the uncertainty from estimation impacts system
understanding.

A new machine learning based method for monitoring for anomalous behavior in production oil wells is reported. The method entails a transformation of the available time series of measurements into a high-dimensional feature space representation. This transformation yields results which can be treated as static independent measurements. A new method for feature selection in one-class classification problems is developed based on approximate knowledge of the state of the system. An extension of features space transformation methods on time series data is introduced to handle multivariate data in large computationally burdensome domains by using sparse feature extraction methods.

As a whole these projects demonstrate the application of modern statistical modeling methods, to achieve superior results in data driven chemical engineering challenges.

Thesis Supervisor: Richard D. Braatz
Title: Edward R. Gilliland Professor of Chemical Engineering
Acknowledgments

It is with great pleasure that I wish to thank some of the many people who have helped me in reaching this point in my academic and professional career. I would like to first thank my thesis committee members Prof. Michael S. Strano, Prof. William A. Tisdale, and especially my thesis advisor Prof. Richard D. Braatz. Without Richard’s support and supervision the work contained in this thesis would not have been possible. I’d like to thank Richard especially for allowing me the leeway to pursue my research objectives with broad freedom.

The work in my thesis has been motivated by problems experienced in industry and other scientific endeavors beyond my own. I’d like to highly thank Dr. Indrani Bhattacharyya and Prof. Gregory C. Rutledge first for bringing the electrospinning signal deconvolution problem to my attention and being helpful collaborators and co-authors. The work on anomaly detection in production oil wells has been funded by, and motivated in constant communication with folks at BP especially Richard Bailey. The carbon nanotube spectroscopy work would not have been possible without guidance and comments from Prof. Strano and numerous students and researchers within his research lab throughout my time at MIT.

The Braatz research group has been a constant source of advice and collegiality. Thank you to Joel, Kristen, Wit, Zack, and all of the other lab members and visitors who have kept my spirits high and expressed interest in my work.

The Chemical Engineering Department at MIT is full of wonderful students and staff members. Go ChemE Hockey!

To my friends, family and fiancée I could not have done this without you. Thank you.
Contents

1 Introduction

1.1 Historical Perspectives and Motivations for Computational Statistics and Data Science ....................................................... 25
1.1.1 Industrial Data Science ...................................................... 26
1.2 Probabilistic Modeling and Machine Learning .......................... 28
1.2.1 Constructing a Model ...................................................... 29
1.2.2 Regression Models ......................................................... 30
1.3 Uncertainty Quantification and Propagation .......................... 33
1.4 Anomaly Detection and Diagnosis ..................................... 34

2 Uncertainty Analysis in Spectroscopic Chemometrics: Regularization, Sequential Analysis, and Bayesian Estimation with a Detailed Model

2.1 Introduction ................................................................. 37
2.2 Notation ........................................................................... 38
2.3 Background ...................................................................... 39
2.3.1 Carbon Nanotubes ........................................................... 39
2.3.2 Electronic Structure and Optical Response of SWNTs ........ 41
2.3.3 Sources of Uncertainty in Analysis .................................. 42
2.3.4 Mathematical Formulation of Pure Component Signal ........ 48
2.4 Standard Linear Formulation ............................................ 51
2.4.1 Regularization and Prior Knowledge ................................. 58
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.2 Performance of Regularization Strategies for the Correctly Spec-</td>
<td>61</td>
</tr>
<tr>
<td>ified Problem</td>
<td></td>
</tr>
<tr>
<td>2.4.3 Uncertainty in Model Formulation</td>
<td>63</td>
</tr>
<tr>
<td>2.4.4 Estimation with Underspecified Models</td>
<td>65</td>
</tr>
<tr>
<td>2.4.5 Regularization and Model Underspecification</td>
<td>66</td>
</tr>
<tr>
<td>2.5 Nonlinear Estimation and Optimization</td>
<td>67</td>
</tr>
<tr>
<td>2.5.1 Bounded Parametric Uncertainty</td>
<td>69</td>
</tr>
<tr>
<td>2.5.2 Assuming a Prior Distribution for Parameters</td>
<td>70</td>
</tr>
<tr>
<td>2.5.3 Evaluating Uncertainty in Parameter Estimates</td>
<td>70</td>
</tr>
<tr>
<td>2.6 Sequential Analysis and Simultaneous Analysis</td>
<td>78</td>
</tr>
<tr>
<td>2.7 Conclusion</td>
<td>81</td>
</tr>
</tbody>
</table>

3 Deconvolution of Electroprocessing Data From Pharmaceutical Manufac-
   turing

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>83</td>
</tr>
<tr>
<td>3.2 Methods</td>
<td>86</td>
</tr>
<tr>
<td>3.2.1 Experimental Apparatus</td>
<td>86</td>
</tr>
<tr>
<td>3.2.2 Solution Preparation</td>
<td>87</td>
</tr>
<tr>
<td>3.2.3 Measurement of Entrained Liquid Layer</td>
<td>89</td>
</tr>
<tr>
<td>3.2.4 Electrospinning and Current Measurements</td>
<td>89</td>
</tr>
<tr>
<td>3.2.5 Electric Current Analysis</td>
<td>91</td>
</tr>
<tr>
<td>3.2.6 Bayesian Modeling of Electric Signals</td>
<td>93</td>
</tr>
<tr>
<td>3.2.7 Estimating the Electric Field Around the Wire Electrode</td>
<td>99</td>
</tr>
<tr>
<td>3.3 Results and Discussion</td>
<td>100</td>
</tr>
<tr>
<td>3.3.1 Liquid Entrainment</td>
<td>100</td>
</tr>
<tr>
<td>3.3.2 Influence of Electric Field on the Wavelength Parameter</td>
<td>102</td>
</tr>
<tr>
<td>3.3.3 Jet Initiation and Termination</td>
<td>103</td>
</tr>
<tr>
<td>3.3.4 Single Jet Lifetime</td>
<td>105</td>
</tr>
<tr>
<td>3.3.5 Jet Lifetime and Frequency</td>
<td>106</td>
</tr>
<tr>
<td>3.3.6 Linear Density of Jets</td>
<td>107</td>
</tr>
</tbody>
</table>
List of Figures


2-3 (a) Temperature dependence of the PL spectra (single excitation) of a single (9,8) SWNT from 40 to 297 K (b) Spectral fwhm as a function of temperature. Reprinted with permission from: Mechanism of exciton dephasing in a single carbon nanotube studied by photoluminescence spectroscopy Yoshikawa, Kohei and Matsunaga, Ryusuke and Matsuda, Kazunari and Kanemitsu, Yoshihiko, Applied Physics Letters, 94, 093109 (2009), Copyright 2009 American Institute of Physics.

2-4 Simulated absorbance spectra of a (6,4) SWNT.
2-5 Covariance matrix of estimated concentrations in linear system with iid normally distributed noise. ........................................ 56

2-6 Optical spectral of closely overlapping species. The pure component spectra of (15,2) and (16,0) SWNTs are very similar such that these species are difficult to identify as separate species from optical spectroscopic measurements in the presence of noise. ........................................ 56

2-7 Multiple closely overlapping pure component spectra. .................. 57

2-8 The best fit to a simulated spectra from a single species under the strategy of Nair et al., [67] reveals the large bias possible in their proposed regularization strategy. ........................................ 59

2-9 (6,4) and (9,1) pure component spectra. .................................. 73

2-10 Simulated measured data for an even mixture of (6,4) and (9,1) SWNTs with normally distributed noise. ......................... 73

2-11 Posterior distribution of estimated SWNT concentrations in the two component example ........................................ 74

2-12 Prior distribution assumed for MCMC sampling. ....................... 75

2-13 Posterior distribution of 6 parameters. ........................................ 75

2-14 The posterior distribution (histogram) is much narrower than the prior distribution for the center of the \( E_{11} \) transition for the (6,4) species. ................ 76

2-15 \( E_{11} \) center (6,4) nanotube marginal posterior distribution from MCMC (bars) and linearized estimate from optimization (line). ............ 76

2-16 Simulated measured data for an even mixture of (6,4) and (9,1) SWNTs when nearly no (9,1) nanotubes are present in the mixture .... 77

2-17 Posterior distribution of concentration estimates for two species with uncertain pure component spectral parameters when one species is near 0. ........................................ 78
2-18 Independent vs simultaneous analysis with five samples. Circles show the median and 95% range of the posterior. The line and + symbol show the true value. a) concentration of (6,4) species when estimated independently. b) concentration of (6,4) tubes when estimated simultaneously c) independent estimation of $E_{11}$ peak center d) simultaneous estimation $E_{11}$

2-19 Upper Right Panel: Concentration estimates shared between 7 samples. Lower Right Panel: Concentration estimates for 7 samples each taken independently. Upper Left Panel: The simulated UV-vis-NIR spectra for each of 7 samples. Lower Left Panel: The true concentration of each species in the simulated spectra.

3-1 (a) Schematic of the apparatus for free surface electrospinning from wire electrode on star-shaped spindles. The components are: (a) wire electrode, (b) solution bath, (c) solution, (d) DC motor, (e) high voltage power supply, (f) collector plate, (g) 1 MΩ resistor, (h) multimeter, (j) enclosure box. Refer to text for description of the apparatus. (b) Schematic of the wire electrode with spoked wheels (k) and threaded rod (l) containing only two wires; (i) side view, perpendicular to spindle axis, (ii) end-on view, parallel to spindle axis.

3-2 Formation of satellite droplets of aqueous solution of PVA (7 wt%) on the wire, midway between mother droplets.

3-3 a) The average single jet behavior observed by Forward et al., under different conditions of the PVP system b) The whole wire current observed as a sum of multiple overlapping individual jetting events c) The density of jets per length of wire estimated by Forward et al.'s deconvolution procedure. [27].
3-4 Example of current signal collected when multiple droplets are jetting (at 2.5 rpm and 45 kV). The current vs time contributions from individual jets at the onset of jetting clearly have a different shape than those of jets occurring later. The inset shows the average signal from a single jet under the same processing conditions.

3-5 The piecewise linear parameterization for the current contribution of a single droplet jetting from the wire electrode to the collection plate. This parameterization can provide an approximate description of the true single jet profiles shown in inset of Figure 3-4, with the exception of the final “tail”. It is sufficiently flexible to model both the apparent jet contributions of “early” and “late” jets observed in Figure 3-4.

3-6 Electric field strength near the surface of the wire electrode during rotation.

3-7 Liquid entrainment of 7 wt% aqueous PVA (146-186 kDa) solution. Black dotted line is the power law fit to the data.

3-8 The correlation between wavelength parameter, \(2\pi a_0/\lambda\), and applied electric potential at a constant spindle rotation rate of 2.5 rpm.

3-9 Current drawn from 7wt% PVA/water solution at 45 kV under different spindle rotation rates of 2.5 rpm (black), 3.2 rpm (red), 5.6 rpm (green), 7.2 rpm (blue) and 8.7 rpm (magenta) as a function of angular wire position. The differences in electric current profiles are indicatively of the variation in jetting with spindle rotation rate. (b) Angular range of jetting vs. rotation rate of spindle for an applied voltage of 45 kV. The dotted line is a guide to the eye.
3-10 Average current drawn from a single droplet at 45 kV at rotation rates of 3.2 rpm (dashed), 5.6 rpm (dash dotted), and 8.7 rpm (solid). Error bars correspond to one standard deviation determined from averaging over at least 10 current measurements for single jets. In the inset, the current profile is shown for a single jet that was allowed to jet only after the wire electrode crosses the apex of the arc of rotation (i.e., in between 90° and 203°).

3-11 Optimal model fit to the wire current generated by the Bayesian model, based on piecewise linearity of single jet profiles. The black solid line is the wire current data collected during a free surface electrospinning experiment by rotating the wire spindle at 5.6 rpm with $V_{\text{appl}} = 45$ kV. The heavy solid line that overlaps the data is the best fit obtained from model. The thin solid lines in the figure represent different single jets as they appear at different angular positions in order to generate the total current from the wire.

3-12 (a) The linear density of jets as a function of angular position and (b) electric field at different rotation rates: 3.2 rpm (dashed), 5.6 rpm (dash-dotted), and 8.7 rpm (solid line).

3-13 (a) mean and standard deviation number of jets in contact at 2.5V motor speed (b) mean and standard deviation number of jets in contact at 3V motor speed.

3-14 (a) sample total charge transferred for single jetting phenomena at 1.2V motor voltage, (b) empirical cumulative distribution of total charge transferred by single jets at the 1.2V motor voltage.
3-15 Optimal model fit to the wire current generated by the piece wise linear model with 4, 5, 6, and 7 total jets to the same total current. The blue solid line is the wire current data collected during a free surface electrospinning experiment by rotating the wire spindle at with 3V to the motor with $V_{app} = 45$ kV. The green and red lines are best fits. The green fit is unconstrained piece wise profiles. The red line shows constrained profiles by the prior knowledge of jet durations.

3-16 Model fitting to the experimental productivity data. Experimentally observed productivity (filled symbols) at different rpms at applied potentials of 40 kV (square), 42.5 kV (circle), 45 kV (triangle), and 50kV (diamond). The entrainment-limited productivity as calculated from Equation 3.20 is shown by the solid line and the field-limited productivities at different applied potentials as functions of rotation rate are shown by dashed lines. The transition from the entrainment-limited regime to the field-limited regime is most apparent in the data at 45 kV.

4-1 A diagram of sensor locations on a dual completion oil well. BHP: Bottom hole pressure. BHT: Bottom Hole Temperature. WHP: Well Head Pressure. WHT: Well Head Temperature. ACC1, ACC2: Acoustic Sensors.

4-2 Empirical cumulative distribution for bottom hole pressures and well head pressure during bean ups or main production modes.

4-3 Well routing history for the well under study. The well state changes many times over the course of a well's operation between the pre-identified labeled states associated with valve routings.

4-4 Parzen density estimator showing decision boundary tuned to reject 3% of data as outlying. The feature space axis are the first 2 principal components from 6888 time series features.
4-5 Left: Average weekly power demand in Dutch power demand dataset.
Right: Outlying weeks found by feature space transformation (offset).
Time is in 15 minute increments. ........................................... 136

4-6 Sample data from the Ford classification challenge dataset showing
time series measurements from an engine subsystem. a) nominal mea-
surements b) anomalous behavior ........................................ 138

4-7 Two classes of engine data shown projected down to two principal
components using 6093 features. ........................................... 139

4-8 Distribution of classification performance of individual features with a
linear decision boundary. .................................................... 139

4-9 Two classes of engine data shown in two highly separating feature di-
dimensions. ................................................................. 140

4-10 Scree plot showing the variance explained by each principal component
of the ford data with 6093 features. The red var shows the sum of the
rest of the principal components after the 30th component. ........ 141

4-11 Scree plot showing the variance explained by each sparse principal com-
ponent of the ford data with 6093 features. The SPCA was constrained
to have a specified number of features in each principal component. . 142

4-12 Two classes of engine data shown in two sparse principal component
with 6 features per sparse principal component. ...................... 143

4-13 The linear classification performance when trained on all 806 examples
considered in the Ford dataset using SLDA with a specified number
of features mapped as a linear combination to form a discriminating
direction. SLDA was trained with l2 penalty of 1e-6. ............... 144

4-14 Distribution of measurement responses in dual completion low pressure
configuration. ............................................................... 145

4-15 Correlation of 6 main measurements in dual completion low pressure
state. ........................................................................ 146

4-16 Autocorrelation of measurements in Well 1. ............... 147
4-17 Bottom hole pressure portions of operating history colored time in the well’s life from blue to green to yellow to red to dark red. The diamond points are shortly after restarts, the + symbols are near failure data points and the open circles are putatively normal conditions.  

4-18 Acoustic detector 1 portions of operating history colored time in the well’s life from blue to green to yellow to red to dark red. The diamond points are shortly after restarts, the + symbols are near failure data points and the open circles are putatively normal conditions. 

4-19 Three class classification performance trained with sparse linear discriminant analysis: a) acoustic detector 1, b) acoustic detector 2, c) bottom hole pressure, d) bottom hole temperature, e) well head pressure, f) well head temperature.  

4-20 Three class classification performance trained with Sparse Linear Discriminant Analysis with merged HCTSA features in acoustic 1 and 2, bottom hole temperature and pressure, and well head pressure. 

4-21 Scatter plot of the well 1 data projected into SLDA directions with a sparsity level of 40 features per direction. 

4-22 Boundaries for anomaly detection trained to reject 3% of putatively nominal data. Data was the first two principal components projected from HCTSA features. a) acoustic detector 1, b) acoustic detector 2, c) bottom hole pressure, d) bottom hole temperature, e) well head pressure, f) well head temperature. 

4-23 Number of dimensions each in which each day long portion of operating history was found to be anomalous at a 3% outlier selection rate by a Parzen density estimator in two principal component reduction of a HCTSA feature space. 

4-24 Anomalous behavior identified on March 4th. 

4-25 Anomalous behavior identified on March 6th. 

4-26 Anomalous behavior identified on March 6th.
4-27 Boundaries for anomaly detection trained to reject 3% of putatively nominal data with a Parzen density estimator. a) Data was the first two principal components projected from HCTSA features for 5 measurements (not using WHT). b) Data was the two discriminatory directions found via SLDA with a sparsity level of 40 components projected from HCTSA features for 5 measurements (not using WHT).

A1 SVD spectra with discretization .................................................. 163

B1 Empirical cumulative distribution function of the wht showing gaps in the data. ................................................................. 166

B2 Evaluation of the optimal number of clusters using the Calinski-Harabasz clustering evaluation criterion. Higher is better. ......................... 166

B3 Evaluation of the optimal number of clusters using the Davies-Bouldin clustering evaluation criterion. Lower is better. ......................... 167

B4 Distribution of data points in cluster 1 of 4 found by kmeans for LPLP data. ................................................................. 167

B5 Distribution of data points in cluster 2 of 4 found by kmeans for LPLP data. ................................................................. 168

B6 Distribution of data points in cluster 3 of 4 found by kmeans for LPLP data. ................................................................. 169

B7 Distribution of data points in cluster 4 of 4 found by kmeans for LPLP data. ................................................................. 170

B8 Four clusters identified by color in the LPLP data ......................... 171

B9 Static data colored by time point ............................................... 172

B10 Cluster identity of points through time ..................................... 173
List of Tables

2.1 Transition Energies for (15,2) and (16,0) SWNTs .......................... 55
2.2 Comparison of regularization strategies at recovering true concentrations of mixtures of nanotubes in the correctly specified pure component spectra setting with various regularization strategies, noise levels and sparsity of underlying signals. ................................. 62
2.3 Minimum Diagonal Element of Model Resolution Matrix ................. 66
2.4 Error from using overly simplified pure component models .............. 66
2.5 Parameters in a spectral model including peak centers, side peaks, peak widths, peak shape and relative oscillator strength. ......................... 68
3.1 Comparison of solution properties of 7 wt% PVA/water (this work) and 30 wt% PVP/ethanol [40] .......................................................... 88
3.2 Bayesian Optimization Prior Parameters and Model Parameter .......... 99
4.1 Multivariate Data Organization ..................................................... 132
4.2 Number of statistical significant bilinear and quadratic second order terms at the $p \leq 0.001$ level ......................................................... 143
A.1 The properties of 43 semiconducting nanotube species that were used to simulate spectra. From Rocha et al., 2011 [76]. .......................... 162
Chapter 1

Introduction

Rapid increases in instrumentation and computation have presented an increase in data collection capabilities in many industrial and scientific systems. Traditional paradigms of engineering have often been centered around first principles or semi-empirical mathematical modeling techniques. In this thesis an alternative scientific paradigm of working first from data, rather than predominantly relying on established theory is explored. Of course, most theories are developed from extended observations and controlled experimentation which naturally involve the collecting and understanding of data. However, in complex systems, the data can often be exploited without relying upon building detailed models from established theory for all parts of the system.

The applications of statistical modeling and learning contained in this thesis encompass three applications to chemical engineering problems which span a range of relative importance of data versus existing detailed theory. These projects introduce a broad conceptual tool which is understanding the gradient of available theory and existing models versus the quality of data available to understand the system. The first project involves the utilization of a detailed physical model, and small data. The use of the model for parameter estimation is evaluated with computational Bayesian methods allowing for the quantification of uncertainty in estimates. The second project involves as system where the complexities of the dynamic physical process are such that the direct application of a detailed physical model was not viable. In this project
an empirical model was derived to be sufficiently flexible to approximately model the dynamics of the system, and the results of the analysis made subsequent analysis of the experimental system possible. The final problem tackled relied on using data directly via the transformation of dynamic time series data to a high dimensional static feature space to build a probabilistic model for anomaly identification so that deviations from nominal behavior could be identified.

The common link between the projects in this thesis is the role of tools from computational statistics and machine learning in problems from chemical engineering. This chapter provides a brief background on computational Bayesian inference, probabilistic modeling, uncertainty quantification and anomaly detection and diagnosis necessary for the applications described in subsequent chapters.

Chapter two describes work on statistical estimation with a detailed physical model combined with softer assumptions. The application motivating the work is the quantification of uncertainty associated with estimating the relative abundance of mixtures of carbon nanotubes in aqueous solution.

Chapter three describes work on deconvolution of electrical signals generated from a free surface electro-spinning apparatus. In this application a detailed parametric model of the physical system with all uncertain parameters was not available so only an approximate model of the system was utilized. This chapter describes the experimental data collected, approximate model construction and utilization of the results of the approximate model in subsequent analysis.

Chapter four describes work utilizing a data driven approach for oil well monitoring. This work introduces a transformation of time series data into a feature space description. Discriminative techniques were utilized to analyze the data collected from the well.

Appendixes provide additional data and figures to supplement the main text.
1.1 Historical Perspectives and Motivations for Computational Statistics and Data Science

In conducting scientific and engineering work, a term has recently been introduced to describe a trending epoch towards data intensive science, this is called the fourth paradigm \[82\]. In this breakdown of the history of science, the first paradigm is empirical, where it consisted of describing natural phenomena. The second paradigm has been a more recent development of theoretical methods, using mathematical models and generalizations. With the advent of computing machines, a third paradigm of science related to simulation has emerged. The fourth paradigm is again a computational revolution, but seeks to unify theory, experiment and simulation. Data captured by instruments or generated by simulations has to be processed by computational work-flows before information, knowledge and insights can be generated. This scientific inquiry through data processing is at the heart of many recent revolutions in genetics, and the other -ohmics of bioinformatics.

At the same time as this revolution in scientific process has been possible, a massive increase in computational capacity and data acquisition has evolved. For the last several decades the computational performance of CPU chips has rapidly increased following the miniaturization of transistors (Moore’s Law). Problems with waste heat and fundamental physical limits have presented issues with per CPU performance gains with time. However, multiple strategies for parallelization have gained mainstream acceptance allowing for greater computational efficiency per watt or per dollar. The first is multi-core CPUs. The second is general purpose computing on graphical processor units which allows for very efficient parallelized operations of some computations. Additional parallelism exists in cluster computing infrastructures which exist in both high performance computing infrastructures and commodity cloud based systems. Most recently strongly enhanced performance for distributed computation on large data sets has become possible with increased use of in memory data storage.

By leveraging these foundational increases in computation and data storage capacity, machine learning has become a ubiquitous technology for much of the planet.
Cloud based Internet service providers such as Amazon Web Services, or Microsoft Azure offer machine learning pipelines as flexible and scalable services.

1.1.1 Industrial Data Science

With the rise of intense instrumentation and data recording capacities in industrial and computational systems, there has become a very large number of variables for system operators or process operators to monitor during chemicals manufacturing or other industrial tasks. Effectively monitoring these large complex systems is expected to increase in difficulty because the volume of data is so large and growing. Industrial systems such as chemical plants may have tens of thousands of variables which are measured continuously. Large server farms and cloud systems can similarly have a large number of hardware and software defined variables being tracked continuously.

Critically, the data are multivariate and have correlations or are otherwise interrelated and the maximum value of analyzing the data can only be accessed by simultaneous evaluation. To assist human operators in assimilating the volume of data available, the development of automatic data analysis systems has been an ongoing area of interest.

An example of the cross over of domains between industrial and “tech” systems is readily apparent in anomaly detection and diagnosis. The consumer web companies Twitter and Netflix released anomaly detection software packages for monitoring their server hardware and software processes.\(^1\) Both of these packages are designed for univariate analysis for a one dimensional time series of floating point values at a time. These packages are developed to account for critical factors of seasonality and periodicity in measurements due to changing human behavior with time of year, day of week and hour of day. In contrast in the engineering and industrial statistical process control literature, methods for handling multivariate methods are quite mature and most processes are designed with hierarchical control systems such that the process can be maintained in a steady state damping daily or weekly variability. In the engineering and industrial statistics literature multivariate control charts are a well

\(^1\)https://github.com/twitter/AnomalyDetection and https://github.com/Netflix/Surus
studied example of using multiple measurements simultaneously and gaining value from the correlation in multiple measurements.

Practitioners in a multiplicity of different domains are recognizing the value of data collected during normal business of scientific processes and the opportunities present, and organizing to work in a data driven fashion. One emerging challenge in generating value from data, is surfacing the data from databases or other systems designed to store historical data and supporting real time actions and decision making. Moving computational systems into real time applications requires engineering of not only the algorithms and statistical methods being applied but also the computational work flow and data handling processes. This promotes a synergistic relationship between expertise in data management and storage, software, machine learning methods and engineers and scientists.

With a sufficient degree of abstraction a lot of the challenges faced by a network operations engineer, or a site reliability engineering are very similar to those faced by someone monitoring and administering an industrial process. However, there exists a gap in familiarity with the specifics of the problems faced in each domain. This presents an ample opportunity to cross fertilize ideas from multiple previously isolated communities.

It is unlikely the industrial, manufacturing, or chemical engineers will know about columnar data stores, and distributed in memory computation, but traditional big data engineers will have a hard time selling to the chemical manufacturing industry because they don’t understand the problems, and don’t understand the kinds of data acquisition and control systems technologies that are already embedded.

In the final chapter of this thesis the understanding of an oil well system is generated predominantly from a data mining exercise. This bottom up approach of understanding engineered systems based on the data is made more powerful from the availability of low-cost computation and storage, the advancement in computational statistical methodologies, and the improvement in engineering understanding of how to handle data among the workforce.
1.2 Probabilistic Modeling and Machine Learning

The field of machine learning can be considered as a large and evolving set of methods for detecting relationships in datasets and predicting future relationships. The notions of probability provide a natural framework for understanding the relationships in data, uncertainty in data, and making predictions. Probabilistic modeling is used throughout this thesis to ensure robust and sensible procedures to map from data to actions or insights.

The work in this thesis involves both halves of the classic paradigm split of supervised and unsupervised learning. Supervised learning involves learning a mapping from input data to output labels, whereas unsupervised learning is focused on learning interesting patterns or relationships in data and not a specific predictive relationship. The work in Chapters 2 and 3 is formulated a regression problem, while Chapter 4 is focused on unsupervised learning and a one-class classification problem.

Often when tasked with the problem of interpreting some data the most conventional practice is to draw a line through the data, report the slope and intercept and move on. The most common form of fitting this line to the data is done via minimization of the least squares error of the line and the data. The least squares strategy represents a selection of a particular loss function to be minimized to select some parameters for the model. This selection of an objective function for minimization introduces assumptions about the underlying statistical behavior of the dataset that are often incorrect in practice. The selection of a least squares objective function is often done out of force of habit and lack of sophistication on the part of the analyst. There are many possible alternative choices for an objective function.

By introducing a generative model for the data, interpretable probabilistic inference is possible with, the assumptions of the model and analysis strategy being clearly specified and interpretable. A generative model is a mathematical expression or computation procedure that is parameterized to quantitatively describe the procedure that could have reasonably generated the data [46]. By first specifying the generative model for the system, the analyst's perspective on the problem is directly recorded.
Subsequently, the particular form of the objective function to be minimized to find a point estimate of the parameters can be often found explicitly and analytically as function of the data in simple problems, or as a relatively simple computational expression in cases where an analytic solution is not possible.

By working in the framework of probabilities, computational strategies or application logic such as rules for anomaly detection can be found from a firmer foundation than simply proscribing them based on the analysts whim and some validation that they work in some cases.

1.2.1 Constructing a Model

When constructing a probabilistic model for a system or dataset a valuable tool is consider the representation of a graphical model. See [66] and [65] for introductory discussion.

The benefits of constructing a generative model which relates quantities of interest to data are numerous. Upon constructing a probabilistic interpretation of data even if only approximate, the subsequent analysis procedures have arbitrariness removed and assumptions more directly composed.

One way to build a probabilistic model is to create a joint model of the data and parameters of interest \( p(y, x) \) where \( y \) is the data and \( x \) the parameters of interest and then to condition on \( x \), thereby deriving \( p(y|x) \). This is called the generative approach [66].

The generative model framework is contrary to discriminative modeling. The difference between these approaches is perhaps best understood in the context of classification. In classification discriminative approaches seek to maximize an objective by specifying some form of boundary between classes whereas in generative approaches the joint distribution of class labels and input values is learned from data. Generative models have strengths and weaknesses relative to discriminative approaches. In particular the discriminative approach may be less sensitive to model data mismatch because it doesn’t need to learn an entire distribution but rather in classification tends to be concerned with only the class boundary. Generative models
have the advantage of being symmetric in inputs and outputs and can be utilized for a variety of analytical tasks such as predicting the distribution of outputs given inputs. This is key for the working in Chapter on parametric modeling of optical spectra of mixture of carbon nanotubes. In Chapter 4 working on the feature space transformation of data the statistical assumptions about different features are harder to encode correctly when formulating a model.

1.2.2 Regression Models

The formulation of regression models with additive noise models application to simulated and experimental data is the central subject of Chapter 2 and Chapter 3. The formulation of regression models here follows the probabilistic interpretation of linear regression.

It is well known that the conventional weighted least squares is the maximum likelihood estimator and minimum variance unbiased estimator when a set of N points \((x_i, y_i)\) is available with perfect knowledge in the x coordinate and Gaussian uncertainties with known variance \(\sigma^2_{y_i}\) for each measurement. Positing a linear relationship between the x and y data gives a model for the y values as a function of x as in Equation 1.1.

\[
f(x) = mx + b
\]  

The objective of the analyst is to then find the quantities \(m\) and \(b\) such that a best fit is observed. The linear algebra solution to the such a circumstance is to stack up the data, indexed by \(i\), into the form \(y = A\theta\), \(\theta = [b, m]^T\), \(A_{i:} = [1, x_i]\), \(y_i = y_i\). The uncertainties in y then get introduced into a covariance matrix \(C_{i:j} = \sigma^2_{y_i}\), \(C_{i:j} = 0, i \neq j\).

\[
\begin{bmatrix}
  b \\
  m
\end{bmatrix} = \theta = [A^T C^{-1} A]^{-1} [A^T C^{-1} y]
\]  

This procedure minimizes a scalar objective function \(\chi^2\) which is the uncertainty
scaled total squared error between the right and left hand sides of $y = A\theta$. This makes this particular selection of parameters the best fit under a context where this scaled error is the error an analyst desires to minimize. The justification as a minimum variance estimator and maximum likelihood estimator requires a more detailed probabilistic description. The choice of the model and choice of objective function are the problems that must be fully justified in analysis. The choice of how to find optimal value of a scalar objective function can then be addressed via engineering expediency considerations.

Production of a generative model for the data and evaluating the posterior distribution of the unknown parameters is the most sensible and justifiable approach for analysis of regression problems. The generative model introduced in Equation 1.1 is that the process which generates the data is described by the conditional probability distribution for observations $y_i$, given everything else in Equation 1.3.

$$p(y_i|x_i, \sigma_{y_i}^2, m, b) = \frac{1}{\sqrt{2\pi \sigma_{y_i}^2}} \exp\left(-\frac{(y_i - mx_i - b)^2}{2\sigma_{y_i}^2}\right)$$ (1.3)

This introduces a view of the world where observations are generated, either through inherent stochasticity in observation of unavoidable random measurement error, such that in a hypothetical set of repeated experiments there is probability of observing different $y_i$ everything else being constant.

The transformation of this generative model to a scalar objective to be optimized, found explicitly deciding to determine the point estimate of the parameters that maximize the probability of the data we observed. This is the maximum likelihood estimate of the parameters. Because in this simple formulation the data is independent when conditioned on the particular value of the parameters the total data observation probability is a product of the condition probabilities of each observation. This joint probability is called the total likelihood function $L$ in Equation 1.4.

$$L = \prod_{i=1}^{N} p(y_i|x_i, \sigma_{y_i}^2, m, b)$$ (1.4)

The likelihood function is a scalar function of the parameters so this provides an
objective function which can be maximized using blackbox optimization procedures. However, taking the log transformation of Equation 1.4 gives the log-likelihood which is a more convenient function to optimize which is within constant factors of $\chi^2$ which gives that minimization of the $\chi^2$ error is the maximum likelihood estimator because the log transformation and dropping of scalar factors do not change the optimal parameter values found in optimization.

An alternative selection of estimators is to use Bayes’ theorem and derive the posterior distribution of the unknown parameters given the data and uncertainty. The posterior distribution for this two parameter linear model is given in Equation 1.5.

$$p(\theta|y_i, x_i, \sigma^2_y \forall i) = \frac{p(y_i|\theta, x_i, \sigma^2_y \forall i)p(\theta|x_i, \sigma^2_y \forall i)}{p(y_i|x_i, \sigma^2_y \forall i)}$$

In Equation 1.5 the posterior distribution is found via Bayes’ theorem first creating the joint distribution of the uncertain observations $y_i$ and unknown parameters $\theta$ as the product of the likelihood and the prior distribution of the parameters in the numerator and then conditioning on the data $y_i$, where everything is conditioned on the data considered to be known exactly ($\sigma$ and $x$).

Finding a point estimate from a posterior distribution can be done in a number of ways the simplest of which is to maximize this posterior distribution and find the MAP or maximum a posteriori value of the uncertain parameters. For uninformative selections of prior distributions this corresponds precisely with the maximum likelihood estimates. In some cases the posterior distribution may be analytically tractable to calculate and the entire distribution can be found often as a parametric expression of a few parameters, or the first several moments of the distribution calculated explicitly. In Chapter 2 for sufficiently complicated problems we find this to not be possible and resort to Markov Chain Monte Carlo methods to approximately sample from the posterior distribution and then construct the estimate of the mean from these samples.

In Chapter 2 and Chapter 3 the introduction of prior information is included
for purposes of regularizing point estimates. In this discussion and estimator is a procedure which maps from data to a point estimate of parameters of interest. The bias of an estimator (Equation 1.6) is the expected value of the deviation from the true value of parameters observed by an estimator. The point estimate of parameters is called $\hat{\theta}$ while the true value is $\theta^*$. This expected value is taken over the true distribution of parameters $p(\theta^*)$ which is the frequency distribution of true values we could ever expect to observe in our experiment.

$$\text{Bias}(\hat{\theta}) = E_{p(\theta^*)} (\hat{\theta} - \theta^*)$$

(1.6)

The variance of an estimator is an idea of the precision of an estimator or its sensitivity to the noise or inherent stochasticity of an estimation problem.

$$\text{Variance}(\hat{\theta}) = E_{p(\theta^*)} (\hat{\theta} - E(\hat{\theta}))^2$$

(1.7)

1.3 Uncertainty Quantification and Propagation

In the analysis of standard linear least squares problems as discussed in Chapter 2, the uncertainty in estimated parameters $\theta$ can be evaluated analytically provided the assumptions of the model are assumed to hold exactly. The resulting uncertainty around the point estimate in parameters is multivariate Gaussian with 0 mean and covariance matrix:

$$\text{cov}(\hat{\theta}) = \begin{bmatrix} \sigma_b^2 & \sigma_{m,b} \\ \sigma_{m,b} & \sigma_m^2 \end{bmatrix} = [A^T C^{-1} A]^{-1}$$

(1.8)

This description of the uncertainty in the estimated parameters is of course dependent on all of the assumptions holding exactly. When the posterior distribution of parameters is constructed then the uncertainty description is naturally described not in terms of an expected distribution of a point estimate, but rather as a “credible” interval for parameters combinations of parameters. Such intervals can be constructed analytically under some circumstances, but are more generally evaluated from drawing samples from a posterior distribution. This is the approach taken in Chapters 2
and 3 for evaluating uncertainty in complex models. Bayesian methods importantly offer the ability to evaluate the marginal uncertainty in some parameters accounting for the fact that their may be unknown but uninteresting parameters in many applications.

Uncertainty in estimates can be evaluated can also be evaluated non-parametrically by re-sampling methods such as the bootstrap or jackknife.

1.4 Anomaly Detection and Diagnosis

From a probabilistic perspective in this thesis anomaly detection and diagnosis is viewed as examining the posterior probability that a given input is anomalously different from nominal behavior. The input may be a scalar or vector input or as will be discussed in Chapter 5 sequence of observations. Anomaly detection has often been considered in the context of classification problems between two or more classes. Often this work is introduced under the paradigm of fault detection because the types of faults are assumed to be known in advance and in this context sufficient data is available about the distribution of data under faulty conditions. In this case the classification performance on training and test data is the core metric for evaluating the success of the methods under study or development.

An alternative related problem is that of novelty detection where data is available exclusively under known nominal conditions where the data is all labeled with one class in the training context with none of the training examples expected to be anomalous.

In the outlier detection problem context the available data is assumed to be largely in one class (nominal) but some fraction of the data which may be known or unknown should be properly identified as an anomaly. This context is especially difficult if the fraction of training data which contains outliers is unknown. This is principally the context of the oil well ADD problem. In this case without labeled data the success of the method is reliant of *ex post facto* investigation of the points or portions of operating history identified as outliers and inliers and the distributional or statistically
properties of the decision boundaries and datasets produced.

In the classification context, two perspectives on the problem exist a geometric or discriminative perspective and a generative or probabilistic perspective. In the generative perspective the classes belong to separate probability distributions, and the determination of which class a new example belongs to is given by the Bayes decision rule given in Equation 1.9 assuming an equal preference for misclassification. For two classes \( C_1 \) and \( C_2 \) a new observation \( x \) is assigned such that it is in the class for which it has maximum a posteriori probability:

\[
x \in C_i, \quad C_i = \arg \max_{C_i \in \{C_1, C_2\}} (P(C_i|x))
\]  

(1.9)

The posterior probability \( P(C_i|x) \) is found via Bayes rule which is given in Equation 1.10. The posterior probability joint probability of class assignment and observation conditioned on the marginal probability of the observation.

\[
P(C_i|X) = \frac{P(x|C_i)P(C_i)}{P(x)}
\]  

(1.10)

The joint probability of class assignment and observations is factored into two terms \( P(C_i) \) called the prior probability of a class and \( P(x|C_i) \) which is the likelihood for observations conditioned on being in a particular class. This is a generative probabilistic model which describes the probability of observations in fault and non-fault states. The specification of this model is at the heart of many generative classification problems. The identification, assumption or learning of a particular distributional relationship for the data is what drives the differences between generative classification methods. For example assuming that observations are multivariate normally distributed with different means and covariance structure in each class, and identifying these distributional parameters with maximum likelihood estimation is referred to as quadratic discriminant analysis (QDA) as the decision boundary is found as a quadratic function between the two classes. If shared covariances are assumed in
such a mixture of Gaussians model results in linear discriminant analysis (LDA) with a linear decision boundary between each class [66].
Chapter 2

Uncertainty Analysis in Spectroscopic Chemometrics: Regularization, Sequential Analysis, and Bayesian Estimation with a Detailed Model

2.1 Introduction

This chapter addresses the problem of estimating the concentration or relative abundance of a mixture of species of single-walled carbon nanotubes in solution. Analogous problems regularly occur in quantitative chemical analysis. The details of the problem as applied to carbon nanotubes allow for comparison of existing classical methods and more advanced optimization based analyses. To address the question of evaluating the uncertainty in analysis we resort to computational Bayesian inference with a hierarchical probabilistic model including prior distributions for uncertain spectral lineshape parameters. Strategies for rejecting noise and improving the quality of inference are reported for classical methods, optimization methods and in cases with sequential measurement data.

This chapter is organized first with a discussion of the physics of spectroscopic
measurement of single walled carbon nanotubes and why that gives rise to particular forms of uncertainty. Next estimation of relative abundance with noisy data under various modeling assumptions in linear models is reviewed. The impact of regularization strategies such as Tikhonov regularization and the Lasso method, for improving concentration estimates from chemometric models for concentration estimation from absorbance or photoluminescence excitation spectroscopy is demonstrated. The quality of performance of these methods is demonstrated in numerical simulation studies. The uncertainty and confidence intervals on from analysis using these procedures is introduced. Improvements in analysis in this simpler context for sequential data is presented. An increase in complexity of analysis for systems with incorrect model specification is subsequently introduced. This necessitates the introduction of a non-linear parametric model for the uncertainty in lineshape. The impact on the analysis with linear estimation strategies with model misspecification is examined theoretically and with computational examples. Regularization and sequential data analysis strategies are shown to reduce the impact of model misspecification for certain error metrics. Finally the complete set of parametric uncertainty in the model is evaluated with a hierarchical probabilistic model and compared to simpler analysis strategies.

Reduction in uncertainty in estimates is improved in contexts where multiple measurements of the same or related populations of species in solution are available. Improvements for sequential data analysis are demonstrated as a reduction in error from true concentration profiles in hypothetical separation experiments.

2.2 Notation

In this chapter vector quantities are represented as bold lower case symbols such as \( \theta \) or \( \mathbf{d} \). Matrix quantities are given as bold uppercase characters such as \( \mathbf{A} \). Normal weighted symbols will be used for scalar quantities. Subscripts will be used to indicate elements of vectors, or matrices. The notation \( \mathbf{A}_{i,:} \) will indicate the entire i-th row of the matrix \( \mathbf{A} \). Quantities estimated from data will be denoted with a hat such as \( \hat{\theta} \).
2.3 Background

2.3.1 Carbon Nanotubes

Single-walled carbon nanotubes (SWNTs) are a class nanomaterials that consist of a cylindrical arrangement of pure carbon atoms which are covalently linked by $sp^2$ bonds. The arrangement of these atoms can be considered as that of a sheet of graphene, which has been rolled up with a particular angle relative to the hexagonal structural unit cell of 2-d carbon. These materials are interesting to researchers in a wide variety of areas. In particular these materials have interesting mechanical, electronic, and thermal properties. SWNTs like many nanomaterials, exhibit heterogeneity in atomic structure. SWNTs can vary in a number of partially independent attributes such as the relative orientation of the tube axis to the carbon unit cell, the diameter of the tube, and the length of the tube. For the purposes of this work, SWNTs are considered with small diameters from 0-10nm and short lengths from hundreds of nanometers to a few micrometers. Though greater variation in length and diameter is possible, especially if the subject of interest is broadened to include multi-walled nanotubes rather than only single-walled carbon nanotubes. The diameters and roll up angles uniquely correspond to set of chiral indices $(n,m)$ which index all possible roll-up angles and diameters. The chirality of SWNTs strongly impacts the electronic structure. The chiral indices can be used to approximately understand the electronic behavior. SWNTs with $n-m \mod 3 = 0$ are metallic or semi-metallic and the rest are semiconducting. The electronic structure of SWNTs is complex. The best available data on the primary electronic transitions of semiconducting nanotubes is available in Liu et al., 2012 [58]. The variation of properties is of significant scientific and commercial interest as people endeavor to build sensitive chemical sensors or chemical components out of these materials. A major obstacle in the study of SWNTs and in the deployment of SWNTs in scientific and industrial applications has been that the conventional methods of synthesis such as arc discharge, laser ablation, and chemical vapor deposition, produce populations of CNTs that are not uniform [45]. As-synthesized nanotubes exhibit a distribution of lengths and chiralities as well as
impurities. Because of this variation, the purification and characterization of SWNTs is an important step in device development.

Great progress has occurred in recent years in characterizing populations of single walled carbon nanotubes (SWNTs) in solution through optical spectroscopy [76, 85, 4]. This progress has been developed in parallel to techniques focused on the separation of carbon nanotubes and isolating them in solution. These techniques have resulted in the ability to approximately uniquely identify different species of SWNTs by electronic bandgap by photoluminescence (PL) (excitation) spectroscopy [4]. The lack of uniformity in the properties of as-prepared SWNTs motivated significant efforts in the development of techniques to sort and separate populations of SWNTs. Separation methods have been simultaneously advanced with methods designed to support the selective growth and synthesis of SWNT populations. Current techniques in selective growth have not been fruitful enough to eliminate the need for sorting and separation. Techniques have been developed to sort and separate SWNTs based on diameter/chirality, length, and metallic/semiconductor character.

One of the earliest developed approaches for the separation of SWNTs by chirality and therefore electrical properties is by density gradient ultracentrifugation (DGU). DGU has been developed as a suitable technique for sorting by diameter, electronic properties, and of most interest to this work: specific chirality and helicity [45, 2, 94, 10, 33, 42, 96]. DGU operates by separating different types of SWNTs that have been wrapped in surfactant molecules on the basis of differing buoyant densities. The technique utilizes a density medium to establish a gradient of density with position in a centrifuge tube. Various combinations of surfactants can be used to establish the differences in buoyant density among nanotube species, based upon the electronic interactions between the surfactants and the nanotube. Common surfactants include short segments of DNA (ssDNA), sodium deoxycholate (SDC) and sodium dodecyl sulphate (SDS) among others.

In addition to DGU another promising technique for separations is gel chromatography [57]. This method exploits the structure-dependent strength of interaction of the SWNTs with an allyl dextran-based gel matrix in a series of vertically connected
columns along with nanotubes dispersed in SDS. The technique achieved the separation into individual chiralities through overloading the separation column with an excess amount of SWNTs so that SWNTs would selectively bind to the column in order of interaction strength. It was observed that small diameter semiconducting SWNTs preferentially bound to the gel matrix over larger diameter and metallic nanotube species.

The resulting populations from dispersion and separation techniques are frequently analyzed by means of spectroscopic analysis. This analysis can commonly involve photoluminescence, UV-vis-NIR absorption, and/or Raman spectroscopic techniques. These spectroscopies have been used for the evaluation of the relative purity [5], [22], chirality distribution [67, 85, 76], as well as length distribution [25, 80] and other properties of nanotube mixtures. This work is concerned with the problem of determining the chirality distribution and relative abundance of each chirality of semiconducting nanotubes in separation processes.

When spectroscopic measurements are taken in concert with separation processes, multiple measurements are made in sequence along a separation axis. This separation axis could be elution sample number in chromatography, volume fraction, or vertical position along an ultracentrifuge tube. Measurements made directly in the separation system are termed in situ as in directly measuring through a centrifuge tube after density gradient centrifugation as in [33]. In this work advancements of current methods of determining the relative abundance of each chirality in from absorbance or photoluminescence spectra taken in sequence with separation processes are reported and demonstrated in simulation studies.

2.3.2 Electronic Structure and Optical Response of SWNTs

The electronic density of states of single walled carbon nanotubes is dominated by van Hove singularities. This leads to a discrete number of transitions observable in optical processes. An example schematic of these transitions superimposed on the density of states is shown in Figure 2-1 for a semiconducting SWNT. The $E_{11}$ transition is a single electron transition between the first singularity of the valence band and first
singularity of the conduction band. The $E_{22}$ transition is from the corresponding second van Hove singularities. In detail, the electronic structure is more complicated due to the presence of strong excitonic effects in SWNTs. In PLE spectroscopy the excited state formed after absorption in resonance with an $E_{22}$ or higher transition is observed through an emission of a photon at the lower $E_{11}$ transition because the exciton produced quickly relaxes in energy level at a rate much faster than emission processes.

2.3.3 Sources of Uncertainty in Analysis

The spectral deconvolution of SWNT spectra into chirality contributions is a challenging procedure due to uncertainties about background in absorption experiments, potentially high levels of spectral congestion and uncertainties about the spectral line shape, quantum yield, and optical cross section. The spectra of carbon nanotubes
is dominated by optical transitions between van Hove singularities in the density of electronic states, however these specific electronic transitions are broadened by both homogeneous and inhomogeneous broadening effects. Local changes in the dielectric environment of carbon nanotubes are known to impact the observed transition energies [16], when combined with the natural inhomogeneity in local environment inhomogeneous broadening can be observed.

The impact of the local environment on the observed absorption photoluminescence spectra of SWNTs is a widely observed phenomenon. The change in the peak position of the $E_{11}$ optical transition as a function of local dielectric environment has been proposed to be described by a variety of empirical formulas. In [71] the shift in the observed first transition energy was modeled as a polynomial in the dielectric constant of two mediums $\epsilon_1$, $\epsilon_2$,

$$
\delta E_{\text{abs}}(\epsilon_1, \epsilon_2) \approx 43(\epsilon_1^{-1.6} - \epsilon_2^{-1.6}) \text{meV.} \quad (2.1)
$$

Alternatively in [16] the change in transition energy $\Delta E_{ii}$ is given as a function of the local dielectric constant $\varepsilon$ and the transition energy squared in vacuum $(E_{ii})^2$ in terms of empirical parameters $(L, k, R)$

$$
(E_{ii})^2 \Delta E_{ii} = -Lk \left[ \frac{2(\varepsilon - 1)}{2\varepsilon + 1} - \frac{2(\eta^2 - 1)}{2\eta^2 + 1} \right] \left( \frac{1}{R^4} \right) = \frac{c}{R^4}. \quad (2.2)
$$

These transitions in the local environment contribute to the red or blue shifting of peak centers in spectroscopic measurements. The high sensitivity of the nanotube photoluminescence spectra to the local environment allows for the differentiation between water filled and empty carbon nanotubes in measured spectra [11] and for chemical sensors based on modulation of the fluorescence properties in response to the presence or absence of a target molecule [44].

Isolated SWNTs experience extrinsic effects from the environment that impact the electronic spectra as well as intrinsic effects from the nanotube phonon modes and excitonic effects. The excitonic nature of excited states in SWNTs leads to changes in the photoluminescence excitation spectra observed in addition to the pri-
mary peaks [56, 38]. Two-dimensional Fourier-transform spectroscopy has resolved sidebands above the $E_{22}$ transition corresponding with phonon modes commonly observed in the Raman spectroscopy of SWNTs (RBM, G, G’ modes) [38]. In recent work published on the evaluation of the chiral (n,m) distribution of a sample of nanotubes through PL spectroscopy [76] these phonon shifted modes were represented as three Lorentzian sidebands above the main transition peak. The motion of the atoms due to thermal energy is observed to broaden the transitions as observed optically. At finite temperature the distribution of velocities among atoms in the CNT contribute to Doppler broadening. Doppler broadening gives the fwhm of the transition in wavelength units as

$$\frac{\Delta \lambda}{\lambda_0} = 2 \sqrt{2 \ln 2} \frac{kT}{m_0 c^2},$$

where $m_0$ is the atomic mass $\lambda_0$, and $\Delta \lambda$ are the wavelength of an atom at rest and the fwhm of the Doppler shift resulting Gaussian distribution.

The Doppler shift and the presence of phonon driven side bands are results of the overall interaction of nuclear motion and electronic transitions. There are two major delimiting regimes and an intermediate level for the relative coupling of these two phenomena. The coupling strength can be viewed as a relationship of timescales. The two delimiting regimes can be considered as fast nuclear motion and slow nuclear motion relative to the rate of fluctuations in the electronic energy gap [63]. The timescale for vibrational motion can be given by the reciprocal of the Debye frequency $T_{\text{vib}} \approx 1/\omega_D$. The second time scale, the timescale for fluctuations in energy gap, is related to the coupling strength of electronic and nuclear motions which is denoted as $S$ which is the Stokes shift. This time scale can be given as $T_{\text{fluc}} = h/\sqrt{k_B T S}$, where the denominator term is the mean energy of thermal motion and nuclear replacement upon excitation [63]. In the case of slow nuclear motion we observe the limit of $T_{\text{vib}} >> T_{\text{fluc}}$ the absorption line shape is given by a Gaussian line shape,

$$D_{\text{abs}}(\omega - \omega_{eg}) = \frac{T_{\text{fluc}}}{\sqrt{2\pi}h} \exp \left\{ -\frac{1}{2} \left(T_{\text{fluc}}(\omega - \omega_{eg} - S/2h)^2 \right) \right\}.$$  \hspace{1cm} (2.4)

This limit is inhomogeneous broadening and the nuclei can be considered to be frozen
in place and the transition is a vertical Franck-Condon transition with $\omega = \omega_{eg} + S/2$ [63].

If the relative size of the time scales is opposite the condition above, that is, $T_{vib} << T_{fluc}$ the line shape is given by the Lorentzian form expression

$$D_{abs}(\omega - \omega_{eg}) = \frac{1}{\pi\hbar(\omega - \omega_{eg})^2 - \gamma^2}, \quad \gamma = \frac{T_{vib}}{T_{fluc}}. \quad (2.5)$$

This is the limit of homogeneous broadening, the absorption center is now around the electronic transition frequency without a Stokes shift. This is the limit of very fast nuclear dynamics and motion of the atoms completely averaged out with respect to spectroscopic measurements [63].

In addition to the simplified results presented above, calculations have been performed using semi-empirical, as well as ab initio electronic structural computational methods to determine in silico the absorbance spectra of individual nanotubes [97]. The optical transitions of SWNTs as studied by theory and probed by optical spectroscopy have revealed results, convincingly different from the one electron picture, which can be explained by excitonic effects in the material. These intrinsic effects increase the complexity of simulating the excitation and emission profiles of SWNTs. These effects stem from the fact that the system is not perfectly described by the single electron model of the density of states (Figure 2-1) of the system. The strong binding energy between the excited electron and the hole it creates upon excitation result in a multiparticle picture of the manifold of exciton states that are occupied. The excitonic effects have been observed and compared against expected simpler resonance from the van Hove singularities associated with the one dimensional bond structure of the material in two photon PLE experiments [91].

In addition to theoretical analysis, single molecule studies, and computational studies which have shed light on the optical response of SWNTs, these molecules have been subject to an extensive amount of many particle empirical experimental analysis. In particular the a relationship between the fwhm of the second optical transition $\Delta\nu_{22}$ has been observed to be follow as a function of peak frequency $\nu_{22}$

with empirical constants in an exponential form based on observations from a large number of semiconducting nanotubes [76], as shown in Figure 2-2.

Variation of line width between different CNT species has been studied in detail, and the mechanism for spectral dephasing of excitons by phonon interactions has been studied spectroscopically [95, 24, 39]. The diameter and chirality dependence of the line width in single molecule photoluminescence studies has been used to investigate the mechanism of exciton dephasing through the temperature dependence. This lead to the observation that the PL spectra of single SWNTs varies with temperature as shown in Figure 2-3.

Carbon nanotubes have a large absorption cross-section. This results the creation of multiple excitons per SWNT [39]. The cross section is approximately $1 \times 10^{-17} \text{cm}^2$ per carbon atom [39]. The creation of multiple excitons per nanotube, results in significant exciton-exciton scattering events along the nanotube axis which can contribute to the measured optical dephasing [1]. The effective homogeneous line width,
Figure 2-3: (a) Temperature dependence of the PL spectra (single excitation) of a single (9,8) SWNT from 40 to 297 K (b) Spectral fwhm as a function of temperature. Reprinted with permission from: Mechanism of exciton dephasing in a single carbon nanotube studied by photoluminescence spectroscopy Yoshikawa, Kohei and Matsunaga, Ryusuke and Matsuda, Kazunari and Kanemitsu, Yoshihiko, Applied Physics Letters, 94, 093109 (2009), Copyright 2009 American Institute of Physics.

The relationship between dephasing time $T_2$ and the homogeneous line width is

$$\Gamma_h = \frac{2\hbar}{T_2}$$

[39]. The dephasing time is comprised of two terms the pure dephasing time and the population lifetime $T_1$,

$$\frac{1}{T_2} = \frac{1}{2T_1} + \frac{1}{T_2^*}.$$
optical dephasing times, if the population relaxation is considered significantly long. In SWNTs this was observed to not be a valid condition and the account of population decay time $T_1$ is necessary to properly account for the dephasing time and homogeneous line width. Using pump-probe spectroscopy of a (6,5) nanotubes Graham and coworkers found fast initial decay of pump-probe spectra with lifetime from 290-460fs for temperatures between 292-80 K [39]. To model the population dynamics of the system Graham et al., used a three state model consisting of a ground state, the optically active exciton state and a lower in energy dark state.

\[
\frac{dN_{E11}}{dt} = k_{E22}N_{E22} - (k_{E11} + k_-)N_{E11} + k_+N_D - \gamma N_{E11}^2
\]

\[
\frac{dN_D}{dt} = -(k_D + k_+)N_D + k_-N_{E11}
\]

In the model, the rates of the various processes in the system are included, such as transfer of population between excitonic states ($k_+$, $k_-$), radiative decay $k_{E11}$ and non-radiative decay from the dark excitonic state $k_D$. Also included is a second order recombination term. The connection between the D state and $E_{11}$ state is driven by thermal interactions with phonons and at sufficiently high temperatures.

### 2.3.4 Mathematical Formulation of Pure Component Signal

The process of spectral deconvolution of the contribution from individual nanotube chiralities relies upon a mathematical model for the relationship between the observed data and the physical parameters of interest. Several approximate models have been introduced in the literature previously. Models in the literature will be examined first from the prospective of interpretation of UV-vis-NIR optical absorbance spectra. This reduces a dimension in the data set when compared to PLE, and includes the effect of metallic tubes as well as a nonlinear broad background in as prepared samples attributed to amorphous carbon and non-SWNT impurities. The assertions and processes followed in references [67], [85] will be followed to demonstrate spectral deconvolution processes in 1-dimension with special attention to treatment line shapes for absorption at the first and second van Hove singularities. In [67] the line
shapes were chosen to be Voigt in shape, after comparing results found with utilizing Gaussian, Lorentzian and Voigt. In the later work in ref. [85], the absorbance were modeled as Gaussians, in addressing PLE data [76] used both Voigt and Lorentzian lineshapes. The measured optical spectra were modeled as a contribution from the band transitions in the sampled nanotubes as well as a background component.

In dilute samples the Beer-Lambert law indicates that for absorbance spectroscopy the measured signal is additively the contributions from the species in the sample. For absorbance spectroscopy ignoring side bands Nair et al., [67] modeled the background subtracted spectra $a_{\text{sub}}(\lambda)$ given in terms of concentration \((c)\) weighted contributions of pure spectra \((p)\) for two optical transitions $E_{11}$ and $E_{22}$ in semiconducting species and the first optical transition in metallic species:

$$a_{\text{sub}}(\lambda) = \sum_{(n,m)} c_{(n,m)} P_{(n,m)}^{11}(\lambda) + c_{(n,m)} P_{(n,m)}^{22}(\lambda) + c_{(n,m)} P_{(n,m)}^{M}(\lambda).$$  \hspace{1cm} (2.10)

In both [67] and [85] the fwhms of optical transitions were not independently modeled for each nanotube species. The fwhms were kept constant for each of the 3 spectral ranges in the former while all transitions were treated as having an fwhm of 50 meV in the later. The spectral deconvolution problem then takes the form of an optimization to vary the spectral weights so that the simulated spectra can best match the data. An objective function is defined to minimize some measure of the disagreement e.g the two norm of the spectral difference:

$$\min_{C_{n,m}} \|d_{\text{measured}}(\lambda) - a_{\text{simulated}}(\lambda)\|^2.$$  \hspace{1cm} (2.11)

The models for the spectra (background subtracted) are linear in the terms corresponding to relative abundance of each nanotube species. Therefore the system can be represented as the solution of a system of linear equations in matrix algebra as $A\theta = d$ where $A$ is a matrix containing all the spectra of pure species, $\theta$ is a vector containing the weights for each species and $d$ is a vector containing the measurement data. This system is likely ill-posed and can not be solved directly with the expec-
tation of a reasonable result. In order to address this issue the previously mentioned works solve a different optimization problem. In reference [67] the number of free parameters in the optimization is reduced to the number of distinct peaks in the bulk spectra by an empirical detailed weighting scheme that serves to strongly weight the contribution of nanotube species with transition energy close to observed distinct peak center for each peak. Tian et al., [85] addresses this problem by proposing a regularization scheme in the optimization giving an expression of the form:

$$
\min_{c_{n,m}} \|a_{\text{simulated}}(\lambda) - d_{\text{measured}}(\lambda)\|^2_2 + \alpha c_{n,m}^T \bar{\Theta} \bar{c}_{n,m}.
$$

(2.12)

The authors of [85] suggest that the matrix $\bar{\Theta}$ be a positive definite correlation matrix. The authors used the identity matrix which is otherwise well known as being 0th order Tikhonov regularization which biases the elements of the vector of relative abundances towards 0.

In the evaluation of relative abundances from PLE spectroscopy, Rocha and coworkers [76] developed a procedure for fitting emission profiles taken over a range of excitation energies. This work utilized refined line shapes for both processes including sidebands and made an effort to quantitatively measure concentration rather than the simpler relative abundance. The measured intensity is a function of two frequencies: excitation and emission. The expression used for the measured value is given as:

$$
S_{n,m}(v_{em}) = F_{\text{instr}}(v_{em}) \cdot [C]_{n,m} \cdot P(v_{exc}) \cdot \sigma_{rel}(v_{exc})_{n,m} \cdot f_{n,m}(v_{em} - v_{n,m}) \cdot (\sigma_{22} \Phi_{FI})_{n,m}.
$$

(2.13)

In this equation the first term, $F_{\text{instr}}$ is the efficiency of the detector which must be calibrated. The second term is the number of carbon atoms of each species $(n,m)$ of SWNT. The third term is the incident power at the excitation level. The forth term $\sigma_{rel}(v_{exc})_{n,m}$ is the excitation line shape. The fifth term is the emission line shape, and the final term is an adjustment for the extinction coefficient of each species multiplied by the quantum yield. The objective function is a double sum over excitation levels
i and measured emission levels k

\[
\min_{\vec{C}_{n,m}} \sum_i \sum_k \left[ S_{\text{sim}}^{i}(v_{em,k}) - S_{\text{exper}}^{i}(v_{em,k}) \right]^2.
\tag{2.14}
\]

As discussed previously, the excitation (absorption) profile near the \(E_{22}\) transition has been observed to contain a series of sidebands. Rocha et al., modeled these with Lorentzian functions for the main peak and phonon assisted sidebands as well as frequency independent term for uniform background [76]. In the model, the offset of the sidebands was given as fixed across all species at 211, 422 and 632 meV above the main peak each with equal widths given in relation to the width of the main peak. The widths of the main peak are not treated as constant for different nanotube types in this model, but instead exhibit variation based upon chirality. In addition to the phonon sideband peaks above the \(E_{22}\) transition, a subband above the \(E_{11}\) transition is included when excitation energy is low enough to be near the \(E_{11}\) transition for large-bandgap small-diameter tubes.

\subsection{2.4 Standard Linear Formulation}

The standard formulation of the Beer-Lambert law for estimation of concentration from excitonic optical spectra is a linear estimation problem. The model equation for the data is given as:

\[
A\theta = d,
\tag{2.15}
\]

where \(A\) is an \(n \times m\) matrix, \(\theta\) is a \(m \times 1\) vector, and \(d\) is a \(n \times 1\) vector. The data vector \(d\), is the measured absorbance spectra of mixture of SWNTs in solution sampled at \(n\) points over a range in frequencies, or equivalently wavelengths or energies. The matrix consists of the pure component spectra for each of \(m\) different SWNT chiralities being considered as possibly present in the mixture. The quantity to be estimated from the data, \(\theta\), is the relative abundance or concentration of each nanotube species present (depending on the normalization of \(A\)). If the pure component spectra have the appropriate oscillator strength or absolute absorption coefficient information.
correctly encoded, concentrations may be estimated directly, and otherwise the estimated quantities are subject to normalization for the differing oscillator strength or extinction coefficient of each species. In the context of photoluminescence excitation spectroscopy, the data output is usually the measured optical response as a matrix dependent on excitation and emission frequency. Such PLE matrix structured data can easily be reorganized into a larger vector and fit the same formalism as Equation 2.15.

Initially, we posit that the model Equation 2.15 is of the correct form for the true process that generated observed data, and any deviation from the model is due only to random noise in the signal recorded by the spectrometer. Critically this means that the pure component spectra are correctly encoded in the matrix $A$. Assuming that the noise is independently and identically distributed as a Gaussian random variable with shared standard deviation at each measured wavelength, the probability of observed particular measured data, $d$, can be written as function of the parameter $\theta$. The likelihood model describes the probability of observing the data conditioned on a particular value of the parameter to be estimated and is given in this equation:

$$P(d|\theta) = \prod_{i=1}^{n} N(A_{i,:} \theta - d_{i}; 0, \sigma). \quad (2.16)$$

Equation 2.16 describes the probability of observing the measurements $d$ conditioned on the observations $\theta$. The function $N(y; \mu, \sigma)$ indicates a normal distribution function with mean $\mu$ and standard deviation $\sigma$ evaluated at $y$. For the simulation studies below the data reported in [58], [76] and [17] for electronic and optical properties was used. The line shape was simulated for an absorbance spectra of an individual carbon nanotube in the spectral range of 500 to 1400 nm to cover the bulk of the observed electronic transitions of semiconducting nanotubes. The absorption profiles were modeled using a Voigt peak shape centered on the appropriate electronic transitions. The Voigt lineshape is a convolution of Lorentzian and Gaussian profiles with
characteristic function given by:

\[ f_V(t; \sigma, \gamma) = e^{-\gamma |t| - \sigma^2 t^2 / 2}. \]  (2.17)

The lineshape function is not analytically available and is efficiently computed via Fourier transform of the characteristic function or in terms of the real part of the Faddeeva function, for which high performance numerical libraries are available and are utilized for the calculations discussed in this work [50].

For semiconducting tubes the \( E_{11} \), and \( E_{22} \), transitions are considered and for metallic tubes the \( M_{11} \) transitions are considered only. The width of spectral broadening and the Voigt lineshape is attributable to the convolution of inhomogeneous and homogeneous broadening phenomena. Based on the data reported in [76] the appropriate Voigt fwhms were selected. A detailed table of the parameters used in different spectral models is included in the Appendix for 43 small diameter SWNTs. The nominal features of the spectral model used in this work is shown in Figure 2-4.
The $\mathbf{A}$ matrix was constructed with $n = 91$, and $m = 43$ for spectral range of 500 to 1400 nm with 10 nm increments (2.48 to 0.886 eV with variable energy spacing). From this construction we can examine the bias and variance of estimation of $\theta$ via the classical normal equations solutions solution to least squares problems. Under these conditions, the matrix is of full column rank, with a condition number of 63.15.

Given the likelihood model, a point estimate for the concentration vector maybe found by selecting the most likely value of given the data or optimizing the likelihood given in Equation 2.16, this is the principle of maximum likelihood estimation (MLE). Optimizing the data model Equation 2.16, the result is a standard linear least squares problem for the MLE estimate $\hat{\theta}$. This is found taking the monotonic transformation of the probability with a logarithm and subsequently solving the optimization problem for the KKT conditions. Important properties of the estimate $\hat{\theta}$ include the bias and variance of the estimator from this procedure. The maximum likelihood estimator for linear least squares problems is the minimum variance unbiased estimator of $\theta$. This analysis is extended for additional factors in the problem formulation in subsequent sections of this work. Bias is defined as the expectation of the difference of the estimated value $\hat{\theta}$, from the true value of the parameters $\theta_{\text{true}}$ that generated the data. The expectation is taken over the distributions of realizations of the data giving:

$$\text{Bias}_{\theta} = \mathbb{E}_d(\hat{\theta} - \theta_{\text{true}}).$$ (2.18)

In linear least squares problems the extent of bias in estimates can be found through construction of the model resolution matrix, $\mathbf{R}$ which is given by $\mathbf{R} = \mathbf{A}^\dagger\mathbf{A}$, when the solution is found by the generalized inverse and by $\mathbf{R} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{A}$ when the solution is found via the normal equations. When the model resolution matrix for the normal equations solution is calculated it is found to be the identity matrix. This indicates that the normal equations solution is an unbiased estimator for this problem [3].

Intuitively, in the absence of measurement noise the observed spectra would be per-
fectly resolvable so long as the pure component species were not completely collinear. The functional form of the pure component spectra used in this analysis satisfy this condition. In the presence of noise, pairs of SWNT spectra that differ the most from each other are more likely to be correctly identified by an estimation procedure.

In addition to the bias of estimation procedures and estimators, the variance of estimated quantities is also of interest. There is a natural bias and variance trade-off inherent in estimation problems. A desirable property of an estimation procedure is to be the unbiased estimator with minimum variance [3]. The normal equations solution of linear least squares problems in addition to being the MLE estimator is also the minimum variance unbiased estimator (MVUE). The covariance of estimates of concentration via the normal concentration when subject to independent and identically distributed Gaussian noise is given as $\sigma(A^T A)^{-1}$ [3]. This covariance matrix is shown graphically in Figure 2-5. From the diagonal entries in this matrix it is clear that under independent and identically distributed noise assumption, there are some species which will have drastically higher variance than others. In particular we can see some species are strongly inversely correlated. Species 41, and 42 in the Figure 2-5 which correspond to (15,2) and (16,0) chiralities respectively show strong inverse correlation. The pure component spectra of these two species are shown in Figure 2-6.

The parameters for the closest overlapping species are shown in Table 2.1. The primary features that peaks are associated with are dominated by the excitonic transition energies which are shown to be extremely similar in the data reported by Rocha et al., Table S2 [76]. Other reference data [58] also supports a very close overlap in these features.

<table>
<thead>
<tr>
<th>n</th>
<th>m</th>
<th>$E_{11}$ FWHM</th>
<th>$E_{11}$</th>
<th>$E_{22}$ FWHM</th>
<th>$E_{22}$</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>2</td>
<td>6164 cm$^{-1}$</td>
<td>189 cm$^{-1}$</td>
<td>12131 cm$^{-1}$</td>
<td>378 cm$^{-1}$</td>
<td>[76]</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>6159 cm$^{-1}$</td>
<td>191 cm$^{-1}$</td>
<td>12236 cm$^{-1}$</td>
<td>409 cm$^{-1}$</td>
<td>[76]</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>0.7829 eV</td>
<td>1.5058 eV</td>
<td></td>
<td></td>
<td>[58]</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0.7826 eV</td>
<td>1.5181 eV</td>
<td></td>
<td></td>
<td>[58]</td>
</tr>
</tbody>
</table>
Figure 2-5: Covariance matrix of estimated concentrations in linear system with iid normally distributed noise.

Figure 2-6: Optical spectral of closely overlapping species. The pure component spectra of (15,2) and (16,0) SWNTs are very similar such that these species are difficult to identify as separate species from optical spectroscopic measurements in the presence of noise.
The conclusion to this analysis is that while the model resolution matrix is very near the identity matrix, so elements will not be confused in the absence of noise, in a practical measurement scenario our models have components which may be so closely overlapping that it may be difficult to distinguish these species. Given the predictive peak width data from Rocha et al., and the oscillator strength information from Choi et al. also are very close for these two species and will provide a particular challenge for uniquely identifying mixtures of these species [76, 17]. Other relationships between estimates for combinations of species are striking in the covariance matrix. In particular species 36, 35, 34, and 26, which correspond to (14,1), (13,3), (12,5), (10,6) chiralities respectively, together have structured correlation. This is explained partially in that (14,1) nanotubes dominant peaks can separately be closely overlapped by contributions from the $E_{11}$ peak of (13,3) or (12,5) tubes and the $E_{22}$ peak of (10,6) species as shown in Figure 2-7.

The data resolution matrix gives insight into the overall spectral crowding of different regions of our measurement domain. From this data it is possible to identify the portions of the measurement spectra that tend to have the greatest influence on
uncertainty in the prediction of composition. The data resolution matrix is $A A^\dagger$. Portions of the spectral range at high and low energies don’t show up in the data model, and don’t have any informative power and the generalized inverse solution estimated from data won’t necessarily fit the data exactly if there is noise in this portion of the observed spectra. Equivalently, the null space of $A^T$ is non-trivial. If sampling at a wide range of wavelengths was not practical due to cost or other experimental considerations, such as in vivo or real time measurements, the data resolution matrix can be used to optimally design an experiment to maximize the information content of the measurement on the estimated parameters.

2.4.1 Regularization and Prior Knowledge

An estimation procedure with higher variance shows greater susceptibility to noise. At very noisy conditions the noise or error in the measurement data limits the ability to effectively resolve the input signal. Regularization is a general description for techniques that introduces bias in the estimate with the aim of reducing variance in estimates [3]. As is typical in problems where a large number of parameters maybe estimated it is often possible to fit the data with a high degree of fidelity even beyond the quality of the underlying true parameters. This phenomenon of overfitting is well known in high dimensional regression and estimation problems. This has been partially explored in estimation concentration from SWNT via optical spectroscopy [67, 85, 87]. In the work of Nair et al., an explicit weighting scheme was utilized to lessen the number of parameters which were to be estimated [67]. In Tian el al., and in our previous work Tikhonov regularization, also known as ridge regression was suggested as a technique to reduce the variance of estimates [85, 87].

The techniques of the above papers are briefly analyzed for the data model described previously. The key assumption in [67], is that the relative abundance of species with similar electronic transitions is conserved. The authors provide an empirical formula to assign weight to individual nanotube species to contribute to a particular observed peak. The critical motivation for this is to reduce the number of parameters to be estimated from the data. This can result in misleading estimates in
some simple examples, for example the fitting of nearly pure samples results in large errors. This procedure can be interpreted as asserting a strong prior distribution on the ratio of particular species. The use of Bayesian priors associated with composition can be effective in a different context as discussed in the portion of this work dedicated to analyzing sequential data. In Figure 2-8 a pure signal without noise is generated from a unit concentration of (9,8) SWNTs and no other species and deconvolved using the method of Nair et al., in this case it is clear that the regularization procedure introduces large errors. (Here a Gaussian line shape for a single transition is used for simpler demonstration.)

The 0th order Tikhonov regularized least squares problem is:

$$\min_{\theta} \| A\theta - d \|^2_2 + \lambda \| \theta \|^2_2.$$  \hspace{1cm} (2.19)

The 0th order Tikhonov regularized estimate can be found by using modified normal equations $\theta_{T0} = (A^T A + \lambda I)^{-1} A^T d$.

Tikhonov regularization can be viewed as an inclusion of additional prior knowledge into the fitting procedure. The maximum likelihood estimation problem is can now be viewed in a Bayesian context as a maximum a posteriori (MAP) estimate of the parameters with a uninformative prior distribution on the parameters. The regularized estimation problem is equivalent to a MAP estimation problem with a prior
distribution consisting of a normal distribution centered around 0 on each component of $\theta$. The MAP estimation problem is now to find the parameter values that maximize the posterior belief in the parameters conditioned on the observed data which is given by Bayes theorem,

$$P(\theta|d) = \frac{L(d|\theta)P(\theta)}{\int L(d|\theta)P(\theta)d\theta}. \quad (2.20)$$

The 0th order Tikhonov regularization with parameter $\lambda$ as defined previously, is equivalent to the MAP estimate with prior distribution $P(\theta) = N(0, \frac{\sigma^2}{\lambda})$, where $\sigma$ is the standard deviation of the noise in the generative model for the data, (Equation 2.16), the model is assumed to be correct, and the noise has 0 mean. This prior distribution is not reflective of the true expectation of these parameters, but clearly can be effective at reducing the sensitivity of estimates to noise in measurements.

One way this prior distribution is not consistent with the true prior expectation on concentration distribution is that it suggests that it is equally likely before measurements to have negative concentration measurements as positive. In the data model given by Equation 2.15, the concentration of each species in $\theta$ should be strictly non-negative. Benefits of regularization can be explored in reduction of sensitivity to noise by utilizing alternative forms of regularization which correspond to the introduction of other prior knowledge.

A key form of prior knowledge that can be employed to improve accuracy in estimation is that only a small number of the many possible species are likely to be present in large quantities. The prior belief that the support of $\theta$ is relatively small compared to the dimension $m$, can be implemented as a penalty term on the 0th norm of in a regularized optimization problem. More commonly, this form of prior knowledge can be implemented by relaxing this penalty term to be L1 norm penalty term in the least squares optimization problem. This technique is also known as total variation regularization or lasso and has been explored in some spectroscopic deconvolution problems [86]. Unlike in L2 regularization the solution cannot be analytically
formulated but is instead must be solved as an optimization problem as shown below:

\[
\minimize_{\theta} \| A\theta - d \|^2 + \lambda |\theta|. \tag{2.21}
\]

The optimal choice of regularization parameter in lasso regression is challenging, but clearly the inclusion of prior knowledge can improve the estimate. The overly flexible and under regularized models tend to fit the noise via inclusion of different species.

An additional key piece of prior information to improve estimates is to use constraints on the optimization problem. The concentration of each species is constrained to be greater than or equal to 0. This can be combined with regularization strategies or used independently.

The non-negative constraint and the L1 regularization can also be considered as prior distributions on the parameters in a Bayesian context, and the resulting penalized least squares optimization problem as a search for a MAP estimate. When non-negative least squares is applied as constraint the prior distribution assumed is an improper distribution that is uniform over \( \theta_i \geq 0 \). The L1 regularization penalty term corresponds to inclusion of a prior distribution which is independent and identically Laplace (i.e double exponential) [41].

2.4.2 Performance of Regularization Strategies for the Correctly Specified Problem

A numerical simulation study was performed using with additive noise data model to compare the relative impact of the inclusion of differing forms and combinations of prior knowledge to the estimation problem. A set of true concentration vectors were generated at sparsity levels of 43, 40, 20, 10, 6, 4 species with uniform random sampling between 0 and 100. For each sparsity level 10 differing true concentration vectors were generated. The true spectra was simulated by using the parameters in the Appendix, between 500 and 25000 cm\(^{-1}\) with 500 cm\(^{-1}\) increments, and random iid normally distributed noise was added at 3 different levels with 10 different realizations.
with various regularization strategies, noise levels and sparsity of underlying signals. The inclusion of prior knowledge can increase the quality of fit even in a correctly specified model. For example, as shown in Table 2.2 identifying the correct subset of species to include in advance with the highest level of noise with 10 species reduced the error from the full model average error ratio of 9.381% to only 0.253% error. However prior knowledge applied as regularization that is too strong can increase the error in estimated concentrations relative to the results in the absence of regularization terms. The best λ value for L2 regularization at 3% noise, and 10 species was the middle value (4.47 × 10⁻¹⁰), in fact when the regularization weight is too high the error is larger (9.902% versus 9.381%) than having no regularization term at all. As expected the regularization weight trends towards 0, the estimates match the regularized solutions.

Taking the Bayesian perspective for the parameter estimation problems studied
above with correctly specified models, the MAP estimate found in optimization procedures for the generalized linear models is the peak of a posterior distribution of the parameters conditioned on the data. It is possible to evaluate the uncertainty in parameter estimates and correlations between differing parameter estimates by exploring the shape of the multidimensional posterior distribution beyond such a point estimate. For unconstrained full normal solutions, correct subset solutions and L2 regularized estimation problems the posterior distribution of parameters is available analytically as they were generated via the product of normal distributions. The posterior distribution shape for Lasso and constrained optimization can be found via sampling procedures.

The empirical studies here introduced the concept of incorporation of additional prior knowledge into the estimation procedure when the pure component spectra are assumed to be known exactly and demonstrated that it is possible to improve performance in the relative abundance or concentration estimation procedure by utilizing carefully tuned prior knowledge. These topics are explored further with uncertainty in model formulation in the subsequent sections below.

2.4.3 Uncertainty in Model Formulation

In contrast to the analysis in the preceding sections typically in practice the exact form of the $A$ matrix of pure component spectra is unknown as the numerous physical factors can influence the effective pure component lineshape as discussed in Section 2.3.3. These factors are typically not completely known for all SWNT species or are subject to experimental variability in a given analysis. In a given experimental measurement SWNTs have a large optical response to environmental conditions which can result in changes due to spectral broadening or shifting of optical response. SWNTs are not naturally efficiently solvated in water and the application of surfactants is a typical processing procedure. Fundamental constants that appear as linear factors in $A$ for both UV-vis-NIR and PLE spectroscopies remain unknown or uncertain to relatively large degree. These linear factors include absorbance cross section and oscillator strength. Recently these quantities have been estimated through detailed
experimental studies, however they have not been determined accurately through experiment for all species of interest [17].

The absorbance cross section (equivalently oscillator strength) varies between given species of carbon nanotubes and is not known with same fidelity as the electronic transition information. This gives a multiplicative uncertain factor in the estimation of concentrations that may eclipse the scope of other uncertainties in the analysis. In conducting analysis with the aim of estimating concentration of species, several assumptions are necessary. Examples of other unknown linear uncertainty factors include: the as synthesized or processed length distribution and defect density.

These linear uncertainty factors on the pure component spectra can directly be shown to impact the estimate of concentrations. When a constant scaling factor is uncertain in pure component signals this linear uncertainty shows up in denominator of the uncertainty of each type. These linear factors may be the predominant uncertainty in estimation of concentration of tubes, but are increasingly being better understood. For the purposes of the rest of this analysis these factors are going to be assumed to be known exactly. In evaluating the quality of analysis it is assumed that there are no uncertain linear factors in the pure component spectral parameters. This follows the analysis of Rocha et al., [76] where the relative abundance is estimated on a pure carbon atom basis rather than on a molecule basis.

One class of uncertainty in the $A$ matrix of pure component optical spectra which is uncertain for both PLE data and absorbance data is side peaks or side bands driven from phonon coupling or additional electronic transitions in SWNTs beyond $E_{11}$, and/or $E_{22}$ transitions. To evaluate the impact of these components of the signal this variation was modeled with additive additional signals corresponding to side peaks in the pure component spectra. Under this model there are uncertain parameters $c_i$ indicating the relative magnitude of the side peaks modeled by $A_i$ for several side peak terms as

$$A = A_{main} + c_1A_1 + c_2A_2 + \ldots \quad (2.22)$$
In addition to the observability of certain higher order electronic transitions, there are other variations in the pure component signal, especially important is that the pure component signal may be changed due to the effect of the local chemical environment. Carbon nanotubes are noted as being particular sensitive to the local chemical environment and as such have been deployed as chemical sensors. The major phenomena at play are the shift in observed peak locations, or quenching of fluorescence depending on chemical environment. The extent of spectral broadening as given by the fwhm of the pure component signals can be modulated by the local chemical environment. These effects could be observed in terms of a bulk shift in pure component signals that effects all species of large number of nanotube species uniformly or as random variation. Both systemic variation of broad groups or random variation of individual species signal are explored. In PLE data Rocha et al., [76] suggested a mechanism for dealing with this variation in PLE estimation procedures by having different pure component signals selected for different chemical environments of the SWNTs where the estimation is being performed. This is not trivial to accomplish in advance of the analysis procedure.

2.4.4 Estimation with Underspecified Models

One of the perils of uncertainty in the construction of the $A$ matrix of pure component spectra is the possibility of mismatch between measured data and the model of the physics used to estimate the parameters. The model resolution can be explored analytically by constructing the model resolution matrix with a slightly different formula. When the model used to estimate parameters is different than the true model we adopt the notation $A$ for the matrix which is used in the parameter estimation, and $A_{\text{true}}$ for the true matrix which is used to generate the data. In this case the model resolution matrix is defined as $R = (A^T A)^{-1} A^T A_{\text{true}}$. The results are shown below when the true data includes more side peaks than the model used for estimation. In each case the true model that generated the data has $a_1, a_2, a_3, a_4$ defined based on the table in Appendix A, and these factors are scaled to 0 if not included when fitting with a more simplified model. By examining the trace of the model resolution
Table 2.3: Minimum Diagonal Element of Model Resolution Matrix

<table>
<thead>
<tr>
<th></th>
<th>Main Only</th>
<th>Main + $a_2$</th>
<th>Main + $a_2, a_3$</th>
<th>Main + $a_2, a_3, a_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Element</td>
<td>0.871</td>
<td>0.9437</td>
<td>0.9808</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 2.4: Error from using overly simplified pure component models

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Model</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Main Only</td>
<td>1.8389</td>
<td>0.1550</td>
</tr>
<tr>
<td>Main + $a_2$</td>
<td>0.5626</td>
<td>0.0525</td>
</tr>
<tr>
<td>Main + $a_2, a_3$</td>
<td>0.1028</td>
<td>0.0297</td>
</tr>
<tr>
<td>Main + $a_2, a_3, a_4$</td>
<td>0.0535</td>
<td>0.0322</td>
</tr>
</tbody>
</table>

matrices the minimum element informs the worst error expected by application of the wrong model. We find that the minimum element of the diagonal of the model resolution matrix is given in Table 2.3.

The minimum occurs for Species 30 which is a (15,1) chirality nanotube. If a pure spectra from a unit concentration sample of this species was analyzed for purity using only main peak information because of the presence of side peaks in the true signal it would be estimated to be only 87.1% pure due to lack of agreement between the model and the true process generating the data.

The issue of bias in estimates with model misspecification can also be addressed with numerical experiments. True parameter vectors were randomly generated and then data was generated with all side peaks included. The least squares estimate of parameters were found according to the normal equations with overly simplified models. The error $\frac{\|\hat{\theta}_{\text{true}} - \hat{\theta}\|_2^2}{\|\theta_{\text{true}}\|_2^2}$, where $\hat{\theta}$ is the least squares estimate, is given in Table 2.4 showing the results below. These results show the error level that can be expected even in noise free analysis.

2.4.5 Regularization and Model Underspecification

Regularization was introduced previously in the context of correctly specified models as a method to control the bias variance tradeoff and reduce the impact of noise on estimates. The model misspecification, or underspecification can be considered a
very structured class of noise in the data. In general when estimating parameters using a simpler model than the true model, the estimate tends to be overly broad including a broader mix of species in the estimate than in the true parameter vector. Regularization can be used to bias the estimate towards a parsimonious description of the data by Tikhonov regularization and the Lasso procedure.

2.5 Nonlinear Estimation and Optimization

To compensate for model misspecification, especially of the sort that can arise due to environmental effects, more advanced estimation procedures have been developed to simultaneously estimate parameters associated with the pure component lineshape and the concentration or abundance parameters. This results in a nonlinear optimization problem in the spectral parameters coupled with a linear least squares problem in the concentration vector, or a regularized linear least squares problem in the concentration vector. In order to accomplish this it is necessary to introduce a parameterization of the true spectra.

The parameterization of the pure component optical response in this work utilized for this analysis is the same as previously, the model introduced by Rocha et al., [76]. The parameters listed in 2.5 are included in the model, QY is obviously irrelevant for simulating absorbance spectra. Following Rocha et al., \(a_2, a_3, a_4, a_5\) are consistent across all species but \(a_6\) is allowed to vary independently for different species but is most often 0.

As introduced previously (Equation 2.13) the equation that describes the pure component spectra can be constructed as sum of Voigt components for main peaks with Lorentzian side peaks. Each semiconducting nanotube species has four characteristic parameters the peak width and peak centers for the \(E_{11}\) and \(E_{22}\) transitions. There are also several parameters that may vary but are shared over the dataset. In realistic scenarios we expect that the temperature effect and bathochromic shift from local dielectric environment to move most species of nanotube in a shared way as seen in Equation 2-3 and Equation 2.2.
Table 2.5: Parameters in a spectral model including peak centers, side peaks, peak widths, peak shape and relative oscillator strength.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\nu_{11}$</td>
<td>Center of $E_{11}$ Resonance (cm$^{-1}$)</td>
</tr>
<tr>
<td>$\nu_{22}$</td>
<td>Center of $E_{22}$ Resonance (cm$^{-1}$)</td>
</tr>
<tr>
<td>$d\nu_{11}$</td>
<td>fwhm $E_{11}$ Resonance (cm$^{-1}$)</td>
</tr>
<tr>
<td>$d\nu_{22}$</td>
<td>fwhm $E_{22}$ Resonance (cm$^{-1}$)</td>
</tr>
<tr>
<td>QY</td>
<td>Quantum Yield</td>
</tr>
<tr>
<td>$\phi_{11}$</td>
<td>Oscillator Strength Per Carbon Atom (a. u) for $E_{11}$ Transition</td>
</tr>
<tr>
<td>$\phi_{22}$</td>
<td>Oscillator Strength Per Carbon Atom (a. u) for $E_{22}$ Transition</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Relative Lorentzian character, of Voigt lineshape for all transitions (0 is Gaussian, 1 is Lorentzian)</td>
</tr>
<tr>
<td>$a_2$</td>
<td>Relative height of $E_{22}$ side peak at $+1700$ cm$^{-1}$</td>
</tr>
<tr>
<td>$a_3$</td>
<td>Relative height of $E_{22}$ side peak at $+3400$ cm$^{-1}$</td>
</tr>
<tr>
<td>$a_4$</td>
<td>Relative height of $E_{22}$ side peak at $+5100$ cm$^{-1}$</td>
</tr>
<tr>
<td>$a_5$</td>
<td>Relative broad spectrum flat, absorbance relative to 1 (a. u)</td>
</tr>
<tr>
<td>$a_6$</td>
<td>Relative height of $E_{11}$ side peak at $+1600$ cm$^{-1}$</td>
</tr>
</tbody>
</table>

The parameters to be estimated that define the pure component spectra can be aggregated into a parameter vector $\theta_{\text{spec}}$. The data model is then given as:

$$A \cong A(\theta_{\text{spec}})$$

$$A(\theta_{\text{spec}}) \theta + \epsilon = \mathbf{d}. \quad (2.23)$$

$$\epsilon \sim N(0, \sigma)$$

Here it is indicated that the pure component spectra has a functional dependence on the spectral parameters, and that the measured signal $\mathbf{d}$ is the result of a linear combination of these pure component signals and additive iid Gaussian noise $\epsilon$. The likelihood model for the parameters in the system is then defined as

$$P(\mathbf{d}|\theta) = \prod_{i=1}^{n} N(A_{i;}(\theta_{\text{spec}}) \theta - d_i; 0, \sigma). \quad (2.24)$$

To find a single point estimate the maximum likelihood estimator can be derived as

$$\min_{\theta} \frac{1}{2\sigma^2} \sum_{\lambda=\lambda_0}^{\lambda_N} (\mathbf{d}(\lambda) - A(\theta_{\text{spec}}, \lambda) \theta_{\text{conc}})^2. \quad (2.25)$$
This is a separable nonlinear least squares problem. The concentration elements of $\theta$ have a very different functional dependence than the spectral parameters $\theta_{\text{spec}}$. This specific structure allows for efficient solution of a non-linear least squares problem via the variable projection method [34].

### 2.5.1 Bounded Parametric Uncertainty

The optimization problem to fit the data (Equation 2.25), is a non-linear optimization problem. This non-linear optimization problem has at least 8 parameters for each nanotube allowed to be considered in the mixture if all parameters are allowed to vary freely. If all of these parameters are allowed to vary freely then the problem is not uniquely identifiable. In the non-linear parameter estimation problem non-identifiability occurs when there is not a unique value for each of the parameters that maximizes the likelihood of observing the data. One way that the full estimation problem is not identifiable is that when the peak centers and widths can vary freely then the identity of each species can be trivially relabeled and an equivalent likelihood found. Furthermore, the identity of each species is tied to its optical resonances. The observed values of these quantities can vary slightly due to effects of surfactants or solution, but would not be expected to vary without bounds. The introduction of bounds for these parameters can solve the identifiability problem and constrain the result of the optimization to physical solutions. As seen in the analysis of the standard linear formulation, certain closely overlapping species are present in the default formulation so great care should be taken to ensure that these species remain identifiable if peak centers and widths are flexible in optimization.

Solving a non-linear optimization problem with this large number of parameters can be computationally difficult. In [76] this is addressed with a sequential relaxation of constraints to achieve a particular local minima solution starting initially with most of the spectral parameters being fixed and then allowing freedom only within specified bounds.
2.5.2 Assuming a Prior Distribution for Parameters

An alternative to using upper and lower bounds for each non-linear parameter is to modify the likelihood function to encode the likely range of values for each parameter to observe. This is accomplished by introducing penalty functions into the least squares curve fitting objective. This can be viewed as inclusion of a prior distribution of the parameters and performing MAP estimation. If each parameter is normally distributed around a nominal value then a least squares penalty term in the objective results from the log transformation of the posterior likelihood function for the parameters given the data. One advantage of using penalty terms rather than constraints is that it can be difficult to define a strict bounds on the parameters. If the true value of one parameter lies just outside the bounds the increased flexibility a penalty function or Bayesian prior allows can improve estimates. Finally it may be more natural over a series of experiments in various conditions to identify a normal distribution for a range of peak centers than bounds on their range. The introduction of peaked probability distributions for spectral parameters of closely overlapped species can retain the identifiability characteristics of the original problem if the peaks are sharp enough, if prior distributions are broad and flat, for closely overlapping species the problem may become unidentifiable.

2.5.3 Evaluating Uncertainty in Parameter Estimates

The form of the likelihood function for non-linear parameter estimation problem is analytic and comes as a sum of several terms. The uncertainty in parameter estimates found from maximum likelihood estimation can be determined by using a linearized estimate. The multivariate parameters estimated by the data are approximately distributed, in the limit of large data, as a normal distribution with the covariance of the distribution given by the inverse of the observed Fisher information matrix.
Linearized Uncertainty Estimates

The observed Fisher information matrix is given as the matrix of second derivatives of the negative log likelihood function as

\[ J_n(\hat{\theta}_n) = -\nabla^2 l_n(\hat{\theta}_n). \]  

(2.26)

The matrix of second derivatives is calculated using finite differences rather than analytic differentiation. The uncertainty of estimates around the point estimate \( \hat{\theta} \) is approximated as given as a multivariate normal distribution with center at the point estimate and scaled by the inverse of the Fisher information matrix as

\[ \hat{\theta}_n \sim \mathcal{N} \left( \hat{\theta}, J_n \left( \hat{\theta}_n \right)^{-1} \right). \]  

(2.27)

Calculating the confidence intervals in analogy to scalar parameters can be done for any linear combination of the parameters \( t^T \theta \) which is estimated by \( t^T \hat{\theta}_n \), then the confidence intervals are found with extreme value statistics as

\[ t^T \hat{\theta}_n \pm c \sqrt{t^T J_n \left( \hat{\theta}_n \right)^{-1} t}, \]  

(2.28)

where \( c \) is the appropriate critical \( z \) value (e.g. 1.96 for 95% confidence). This allows for the calculation of one parameter at a time confidence intervals as well as estimates of joint quantities like the sum of all nanotube concentrations.

Computational Bayesian Inference

Bayesian inference, inferring the posterior distribution of parameters conditioned upon the observation of data, can be efficiently approximately performed under a variety of computational schemes [32]. The methods of Bayesian inference are suited for situations where not only a MLE or MAP point estimate is desired but the entire posterior distribution of one or more parameters is of interest. The Markov Chain Monte Carlo (MCMC) method allows for samples to be drawn from this posterior.
distribution even in parameters spaces with a vary large number of dimensions. By sampling from the posterior distribution of the parameters rather than approximating the posterior distributions, the uncertainty in estimates can be more appropriately described in terms of univariate or multivariate credible intervals. This gives more sensible results than approximation when parameters are strongly correlated.

By approximating the entire posterior distribution if only a few parameters are of interest experimentally the remaining parameters which are uncertain, can be efficiently marginalized over or integrated out of the joint uncertainty description. To estimate the posterior distribution of parameters of interest a cascaded affine invariant ensemble MCMC sampler based on [35] and [26] was utilized in MATAB.1

If the pure component spectral components are not uncertain, than confidence intervals can be constructed analytically. Assuming the magnitude of the noise is known then the posterior distribution of the parameters is a multivariate normal distribution centered around the MLE estimate. This is illustrated in a simple example.

The pure component absorbance spectra for (6,4) and (9,1) nanotubes were simulated over a range of wavenumbers. The pure spectra are shown in Figure 2-9. An equal mixture \( \theta = [10000, 10000] \) of these two species with additive 0 mean Gaussian noise (\( \sigma = 0.1 \)) was simulated as shown in Figure 2-10.

The MLE estimate based on simulated data is: [9991.2, 10083.7] (as found by least squares estimation). The covariance matrix is given by \( \sigma(A^TA)^{-1} \) and is equal to
\[
\begin{bmatrix}
655729 & -176700 \\
-176700 & 434542
\end{bmatrix}
\]. From this the 95% confidence intervals constructed by Equation 2.28 is \( \pm [1587, 1292] \) around the MLE estimate. Drawing MCMC samples with fixed spectral parameters confirms that the full posterior distribution is multivariate normally distributed. The MCMC chain rapidly mixes and a large number of samples are able to drawn extremely efficiently for this simple example. The joint posterior distribution and marginal posterior distribution of the estimated concentrations is shown in Figure 2-11. The median of the posterior distribution is found to be [10044.8, 10044.8] and the 95% credible interval for each is identically [9740.9−10345].

1 https://github.com/grinsted/gwmcmc
Figure 2-9: (6,4) and (9,1) pure component spectra.

Figure 2-10: Simulated measured data for an even mixture of (6,4) and (9,1) SWNTs with normally distributed noise.
Figure 2-11: Posterior distribution of estimated SWNT concentrations in the two component example

The results differ much more greatly between linearized estimates and the full uncertainty description as revealed by MCMC when the spectral parameters are included. In the same example if the $E_{11}$ and $E_{22}$ peak centers are uncertain the nonlinear uncertainty description is most readily produced by MCMC. A prior distribution over the peak centers is that the are all independently normally distributed around the nominal value with $\sigma = 200\text{cm}^{-1}$ as shown in Figure 2-12.

The posterior distribution found via MCMC for the now 6 uncertain parameters is shown in Figure 2-13. The reduction in uncertainty from the prior distribution for the center of the $E_{11}$ transition for the (6,4) nanotube is shown in Figure 2-14.

When the maximum likelihood estimates are found by non-linear least squares using constrained optimization with sequential quadratic programing the Hessian at the maximum likelihood estimate was evaluated and used to construct linearized estimates of the uncertainty in the parameters. These estimates closely matched the results from MCMC because the MCMC estimates are approximately normally distributed for all parameters.
Figure 2-12: Prior distribution assumed for MCMC sampling.

Figure 2-13: Posterior distribution of 6 parameters.
Figure 2-14: The posterior distribution (histogram) is much narrower than the prior distribution for the center of the $E_{11}$ transition for the (6,4) species.

Figure 2-15: $E_{11}$ center (6,4) nanotube marginal posterior distribution from MCMC (bars) and linearized estimate from optimization (line).
Figure 2-16: Simulated measured data for an even mixture of (6,4) and (9,1) SWNTs when nearly no (9,1) nanotubes are present in the mixture

An example of the error that will result in estimates using optimization based methods and linearized statistics is shown below. The two parameter example is considered again this time with true concentrations 10000 for (6,4) and 100 for (9,1). The simulated data is shown in Figure 2-16. The prior distribution on concentrations is assumed to uniform and positive, while the optimization method is solved with constraints for positivity. The posterior distributions for concentrations found from MCMC are shown in Figure 2-17. The results estimated concentrations from the MLE estimate are: 10110, and 155. However the standard deviation from linearized estimates of these parameters found from the Hessian of the negative log-likelihood around the MLE are 262 and 268. A one standard deviation confidence internal of $155 \pm 268$ for (9,1) tubes is misleading when estimating the error in purity of mixtures of nanotube. In particular it is expected that many species will have near 0 concentration so appropriately constructing the uncertainty in estimates to take this into account is a strong motivation for estimation of full posterior distributions with MCMC or other methods for computational Bayesian inference.
2.6 Sequential Analysis and Simultaneous Analysis

Strategies for improving the estimation of parameters in mixtures of SWNTs are introduced. These techniques are especially valuable when considering sequential data such as might be produced during separations experiments. In general the key principle applied is that multiple measurements may exactly or approximately share some parameters between the measurements, while other parameters of interest may be completely independent.

One application of this is that measurements of differing populations but with shared spectral parameters. For instance a shared population of SWNTs may have been produced from an individual synthesis and then separated into mixtures in such a way that the expected pure component spectral parameters are identical but the relative composition of each sample is different. This is illustrated in the following example showing the relative reduction in error and uncertainty from evaluation of multiple measurements with shared spectral parameters but freely varying concentrations.

Figure 2-17: Posterior distribution of concentration estimates for two species with uncertain pure component spectral parameters when one species is near 0.
Five simulated spectra were generated for varying compositions of a two component mixture of (6,4) and (9,1) SWNTs. The true spectral parameters for peak centers were fixed but assumed unknown. Error and uncertainty for concentrations and peak centers considering each sample independently and considering them all jointly are shown in Figure 2-18.

An additional situation which frequently occurs in separations data is that when data is collected in a sequence the variation in the parameters between samples is expected, but is expected to occur gradually over the course of the sequence. In density gradient ultracentrifugation if \textit{in situ} measurement is performed or if sequential samples of chromatography based procedures are analyzed the relative concentration profile is observed to be smooth \cite{57}. The prior knowledge that sequential samples are near each other can be implemented in a prior distribution on the difference between samples or a regularization penalty term. A first order Tikhonov regularization penalty term introduces a penalty on the first derivative of concentration estimates in sequential samples \cite{3}. The equivalent Bayesian prior interpretation is that the difference in concentration estimates in sequential samples is 0.

An alternative situation is to introduce prior knowledge that a collection of samples should approximately match the results from a joint analysis of the components in the mixture. This could occur when a pre-separation process mixture is compared to individual samples during separations. In such a case the natural prior distribution a 0 mean prior on the deviation between the pre-separation sample and the sum of concentrations from post-separation samples.

A frequent simultaneous analysis problem can occur when multiple identical samples are generated. This can occur for example in a multiple dilution of a single original sample. In these cases the concentrations are assumed to be constant in each of multiple samples but the noise is independent. By simultaneously estimating the concentration in each sample rather than solving independently noise rejection is achieved with superior performance. As shown in the Figure 2-19 the estimated purity in each sample is much closer to the true value in a simultaneous analysis rather than an independent analysis. Even if independent analyses were subsequently merged to
Figure 2-18: Independent vs simultaneous analysis with five samples. Circles show the median and 95% range of the posterior. The line and + symbol show the true value. 
a) concentration of (6,4) species when estimated independently. b) concentration of (6,4) tubes when estimated simultaneously c) independent estimation of $E_{11}$ peak center d) simultaneous estimation $E_{11}$
provide confidence in the correct estimated concentration of the dominant species, the impurity contribution would not cancel out between samples as it does in the simultaneous estimation problem. This can be seen in the black circles and lines representing the total impurity concentration estimated (nanotube species fit by the estimation procedure) that were not in the true species concentration) in the right upper and lower panels of Figure 2-19.

2.7 Conclusion

In this work the problem of estimating the concentration or relative abundance of a mixture of species of carbon nanotubes in solution was addressed. The methods introduced can have application in a broader context of spectroscopic data analysis. In particular the improvements to estimates from sequential analysis are likely to be useful in a wide range of experimental applications. The simplest curve fitting methods from a linear least squares problem through to advanced computational Bayesian
inference with Markov Chain Monte Carlo methods were reviewed. Strategies for re-
jecting noise and improving the quality of inference are reported for classical methods,
optimization methods and in cases with sequential measurement data. In particular
the improvements in estimates that are addressed by the utilization of prior knowl-
edge of various kinds are shown to help ameliorate the challenges placed by having a
high dimensional system with many free parameters.

The challenge of addressing uncertainty in estimated abundances of mixtures of
SWNTs evaluated from solution is found to be well addressed by computational
Bayesian methods. The frequent practical use case of estimating purity or sepa-
ations products is particularly highlighted as a desirable case to utilize advanced
methods to understand the uncertainty in estimates.
Chapter 3

Deconvolution of Electroprocessing Data From Pharmaceutical Manufacturing

3.1 Introduction

Free surface electrospinning has evolved as a method to have a high productivity method for generating electrospun fibers from solution. Electrospun fibrous materials have drawn attention for their potential applications in such areas as energy storage, drug delivery, tissue engineering, filtration, and defense and security [74, 75]. The main advantage of electrospinning technology over other fiber-forming techniques is its ability to generate non-woven mats with fibers less than 1 micrometer in diameter and with high surface-to-volume ratio. Traditionally, the electrospinning technique uses an electrically charged metal needle or spinneret to feed a viscoelastic liquid into a continuously operating, electrically-charged jet. A grounded metal plate at a certain distance from the needle tip is used as a collector for the fibers, which usually dry before they reach the collector due to the high surface area of the jet.

1 Portions of this chapter have been published as Free surface electrospinning of aqueous polymer solutions from a wire electrode by I. Bhattacharyya, M. C. Molaro, R. D. Braatz, and G. C. Rutledge in Chemical Engineering Journal
made possible by the electrical stretching forces. However, a major drawback of traditional electrospinning with a single needle spinneret is low throughput, usually 0.1-1 g/hr/needle, depending on solution and process parameters [23], which has limited the use of the traditional technique in industrial settings. Several different ways of increasing the throughput of electrospun fibers have been studied over the past few years, from the use of arrays of single-needle spinnerets to a wide variety of other configurations that may collectively be called “needle-less” or “free surface” electrospinning methods [23, 52, 84, 88].

Free surface electrospinning is a technique that takes advantage of the remarkable capacity to launch multiple jets from a charged liquid surface, in principle, if the surface charge density is high enough and curvature can be introduced to the air-liquid interface [60]. One of the earliest configurations of free surface electrospinning employed a magnetic liquid to create liquid “spikes” that perturbed the charged liquid surface. Since then, numerous configurations have been reported to induce free surface electrospinning, e.g., liquid-filled trenches, slits, wetted spheres, rotating wires and fixed wires, cylinders, disks, conical wire coils, and gas bubbles rising through the liquid surface [23, 88, 13, 27, 40, 49, 53, 64, 89, 93]. Many of these approaches have shown promising increase in throughput, and a number of reviews are available [61, 98, 92]. However, for implementation of a process in an industrial setting, a thorough and quantitative understanding of the design parameters is required. Choosing a relatively simple design for high throughput electrospinning allows one to analyze the process for important factors that affect the throughput and product quality.

In recent work [27], free surface electrospinning from a wire electrode was analyzed to understand the effects of different process parameters and solution properties on productivity of fiber, and an empirical model was developed using solutions of poly(vinyl pyrrolidone) (PVP) in ethanol as the fluid. In that work, a relatively simple experimental design was used, in which a wire electrode was connected to high voltage power supply and mounted on a rotating spindle such that electrode both charged the fluid and delivered it into the high electric field [27]. Details of the equipment design were described there [27]. The entire electrospinning process
was decomposed into a sequence of three steps: (i) the entrainment of liquid on the wire as it passes through a liquid-air interface, (ii) the breakup of an annular layer of liquid into droplets on the cylindrical wire, and (iii) the formation of jets from the droplets by electrostatic forces. A review of the understanding of each of these steps was provided previously by Forward and Rutledge [27]. Briefly, the amount of liquid entrained on the wire was found to be predominantly a function of the capillary number, \( Ca = u \eta / \gamma \), where \( u \) is the velocity of the wire, \( \eta \) is the fluid viscosity and \( \gamma \) is the fluid surface tension. Once entrained, the cylindrical layer of liquid on the wire would break up into droplets by a Plateau-Rayleigh instability. The separation distance between droplets is dominated by the most rapidly growing disturbance, whose wavelength \( \lambda \) is a function of the radius \( a_0 \) of the annular liquid surface and the Ohnesorge number, \( Oh = \eta / (\rho \gamma r)^{1/2} \), where \( \rho \) is the density of the liquid and \( r \) is the radius of the wire [36]. For low levels of entrainment, \( (a_0/r < 2) \) the wavelength should be approximately constant \( (2 \pi a_0/\lambda \sim 0.69) \), but Forward and Rutledge [27] found it to depend on applied voltage as well. Once formed, the droplets on the charged wire emit jets above a critical threshold of the local electric field. Since the wire was mounted on a rotating spindle, the local electric field at the wire was a function of angular position, with a maximum at the apex of rotation. It was also observed that the droplets did not all jet simultaneously. This observation was explained to be a consequence of the electrostatic fields created by nearby jets, which suppress simultaneous jetting from neighboring droplets. The linear jet density distribution was characterized by measurements of the electrical current flow at the collector for single jets and for multiple jetting of droplets on the wire, and reported to be a function of local electric field at the wire. These observations were used to construct a model of productivity for the PVP/ethanol system as a function of solution properties (i.e. surface tension, viscosity, density and concentration) and process parameters (rotation rate and electric field). Details of this model are described elsewhere [27].

In this chapter, a regression parameter estimation problem is introduced to solve the problem of analyzing data from a related system. This data analysis procedure allows for the extension the applicability of the previously developed productivity
model ([27]) to aqueous solutions of poly(vinyl alcohol) or PVA. The applicability of this model to aqueous solution systems is not obvious because of the significant differences in solution properties and viscoelastic behavior of aqueous PVA solutions from ethanolic PVP solution. A comparison of solution properties between PVP/ethanol and PVA/water is given in Table 3.1. This chapter discusses the experimental data collection, the data modeling and interpretation, and the application of the model to aqueous polymeric solution, new experimental observations and modifications to the model to incorporate this understanding. The data analysis procedure naturally returns a distribution of results that can be understood with varying confidence based on the probabilistic model formulated. The distribution of results based on the totality of the data collected is reported here.

3.2 Methods\(^2\)

3.2.1 Experimental Apparatus

The apparatus is shown in Figure 3-1a; it consisted of several different components as described below. A wire electrode spindle (a) was placed in a bath (b) filled with polymer solution (c). The spindle was rotated by a DC motor (Zheng gear box motor ZGA25RP216) that was driven by a power supply (d). The rotation speeds ranged from 1.6 to 8.7 rpm; this corresponds to wire velocities \( u \) of 0.5-2.7 cm/s. The shear rate at the wire was \( \sim 100-700 \, s^{-1} \), where \( z = a_0 - r \) is the thickness of the liquid coating on the wire. The liquid in the bath was electrified by a high voltage power supply (e) that was connected to the metal pin at the bottom of the bath. A metal plate (f) collected the non-woven fibers. The metal plate was connected to a 1 M\( \Omega \) resistor (g) across which the voltage drop was measured by a multi-meter (h) in order to quantify the current flow to the collector plate. The whole set-up was placed in a Plexiglas\textregistered chamber (j) inside which the relative humidity was maintained at 20% by a humidity controller (ETS, Model: 5100) for most of the experiments.

\(^2\)Laboratory experiments were designed and conducted by I. Bhattacharyya
In order to successfully electrospin a solution using a wire electrode, the events of liquid entrainment, break-up of droplets and jetting should occur promptly and sequentially. However, due to the highly viscoelastic nature of the aqueous solutions of PVA used in this work, a stable film of liquid frequently formed between the consecutive wires on the spindle. Hence, a cylindrical layer of liquid could not form around the wire within the time that the electrode resided in the regime of electric field in excess of the critical value. In order to promote consistent annular liquid entrainment and thereby facilitate the following droplet break-up, the spindle was modified as follows to prevent film formation and allow successful electrospinning of aqueous PVA solutions. The round Teflon wheels of Ref 10 were replaced by two spoked wheels (k) made of Ultem® to hold the wires in place. The wheels were held in place on a threaded rod (l) by nuts. Grooves centered on the end of each spoke held the stainless steel wires in place. A wire of diameter 200 μm was strung between spokes on the two wheels, as shown in Figure 3-1b, to create the spindle. Up to six wires could be accommodated on the spindle. However, in this work, only two wires were used, 180° apart from each other.

3.2.2 Solution Preparation

Poly(vinyl alcohol) (PVA) with Mw = 146-186 kDa (87% hydrolyzed) was obtained from Sigma-Aldrich. A 7 wt% polymer solution in de-ionized (DI, laboratory supply) water was used in these experiments. Viscosities were determined using an AGR2 Rheometer (TA Instruments). The liquids showed Newtonian behavior for shear rates below 1000 s⁻¹. Conductivities were measured with a VWR Traceable digital conductivity meter. Surface tensions were measured using a Krüss K100MK2 Processor-Tensiometer. Solution properties are listed below in Table 3.1. It is important to notice here that the aqueous PVA solution is not only five times more viscous than PVP/Ethanol solution, but also has a significantly higher surface tension and electrical conductivity as well.
Figure 3-1: (a) Schematic of the apparatus for free surface electrospinning from wire electrode on star-shaped spindles. The components are: (a) wire electrode, (b) solution bath, (c) solution, (d) DC motor, (e) high voltage power supply, (f) collector plate, (g) 1 MΩ resistor, (h) multimeter, (j) enclosure box. Refer to text for description of the apparatus. (b) Schematic of the wire electrode with spoked wheels (k) and threaded rod (l) containing only two wires; (i) side view, perpendicular to spindle axis, (ii) end-on view, parallel to spindle axis.

Table 3.1: Comparison of solution properties of 7 wt% PVA/water (this work) and 30 wt% PVP/ethanol [40]

<table>
<thead>
<tr>
<th>Material</th>
<th>146-186 kDa PVA</th>
<th>55 kDa PVP</th>
</tr>
</thead>
<tbody>
<tr>
<td>concentration (wt%)</td>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>density, ρ (kg/m³)</td>
<td>1075</td>
<td>0.898</td>
</tr>
<tr>
<td>conductivity (S/m)</td>
<td>0.0633</td>
<td>8.89x10⁻⁴</td>
</tr>
<tr>
<td>viscosity, η (Pa·s)</td>
<td>0.520</td>
<td>0.105</td>
</tr>
<tr>
<td>surface tension γ (N/m)</td>
<td>0.0495</td>
<td>0.023</td>
</tr>
<tr>
<td>Ohnesorge number, Oh</td>
<td>7.22</td>
<td>2.3</td>
</tr>
</tbody>
</table>
3.2.3 Measurement of Entrained Liquid Layer

In order to determine the total volume of solution entrained on a wire, the droplets formed on the wire after de-wetting of the entrained liquid were imaged, and their volumes were determined using Carroll’s equations for barrel-shaped drops [12]. The total volume of the liquid was then determined by summing the liquid droplet volumes. Images of the droplets on the wire were recorded photographically with 1st Vision MC 433 Firewire Camera and Quantaray AF 70-300 mm f/4.0-5.6 LD Tele Micro Lens. Image J software (National Institute of Health) was used to measure the equatorial height, length, contact angle between the liquid layer and the wire, and the center-to-center distance \( \lambda \) between the droplets. The contact angle was found to be relatively constant, with a value of 5 ± 3 degrees. The diameter of the wire was kept constant (200 μm) throughout all the experiments. Satellite droplets were observed in between the major droplets, as shown in Figure 3-2. Their volumes were obtained by the same method and added to the volumes of major droplets in order to obtain the total volume of entrained liquid. A MATLAB routine was used to calculate the shape and volume \( V_d \) of a droplet, to a tolerance of 0.01% [27]. For detailed description of the algorithm for determining \( V_d \), the reader is referred to [27]. To obtain an average value for volume and center-to-center-distance, measurements are taken over at least 100 droplets. Using the volumes of the droplets and the number of droplets on a wire, an equivalent thickness \( z \) of the entrained annular liquid layer on a wire was calculated, using the equation:

\[
z = \left( \frac{V_d}{\pi \lambda} + r^2 \right)^{1/2} - r.
\]

3.2.4 Electrospinning and Current Measurements

In order to initiate electrospinning, a high potential was applied to the solution bath. The wire electrodes were also connected to the same applied voltage, to ensure that the surface of the bath and the wires were at the same potential as the external power supply. The applied potential was increased slowly until at least a few of the droplets
on the wire formed Taylor-like cones and finally emitted jets. The droplets on a wire did not all jet simultaneously. The number of droplets jetting at any given time varied depending on the applied potential. The volume of each droplet varied during the lifetime of its jet as liquid was removed by the jet itself. Also the local electric field around a droplet varied due to motion of the wire through the arc of rotation.

As the jets travelled through the chamber, water evaporated and dry solid PVA fibers were deposited on a copper plate collector at a distance of 26 cm from the central rod of the electrode spindle. In order to monitor the electrical charge carried by the liquid to the collector in the form of jets, the voltage drop was measured across a 1 MΩ resistor between the collector and true ground (Figure 3-1a). This current flow was assumed to be that due to free charge remaining on the fibers at time of deposition; it is possible that some of charges were carried away by the evaporated solvent [28]. The relative humidity within the chamber was maintained at 20%. It was observed that if the relative humidity was higher than 20%, the critical field required to initiate jetting increased; this effect was attributed to dissipation of surface charge at the droplets by excessive moisture. The potentials applied in these experiments were below 55 kV, which is significantly lower than the threshold of 126 kV/cm required for corona discharge in air at standard temperature and pressure. A measurement of the background current, performed without any solution in the bath and under the application of same potential as used during spinning experiment, confirmed that corona discharge was not significant. To avoid any changes in solution properties over the duration of electrospinning due to evaporation of solvent from the bath itself, the experiments were limited to a maximum duration of 15 minutes. In

Figure 3-2: Formation of satellite droplets of aqueous solution of PVA (7 wt%) on the wire, midway between mother droplets.
any case, “solution aging” was much less prominent for aqueous solutions of PVA, than for the ethanolic solutions of PVP used in Ref 10. The masses of the deposited electrospun mats were measured after each experiment using a micro-balance.

3.2.5 Electric Current Analysis

Under the conditions for these experiments, the measurement of a finite electric current (or voltage) across the 1 $M\Omega$ resistor was considered to be a reliable indication of jetting behavior. In order to estimate the total volumetric flow of liquid in the form of jets, the total current was measured continuously throughout the arc of rotation of the wire electrode. In order to deconvolute subsequently the total electric current into contributions from several individual jets, the current from a single jet was measured as well. Multiple experiments were conducted in which the effective length of the wire was reduced by covering the wire with insulating tape, leaving a small portion exposed such that only three to four droplets could form after the wire was coated with liquid at each rotational speed and applied potential. Of those three to four droplets, only one droplet could jet within one rotation of the wire. By analyzing the data for electric current versus time, for both a wire supporting multiple jets and one supporting a single jet, a quantitative measure of the jet density along the wire could be obtained.

In this work, a statistical model was developed to deconvolute the measured current signal into contributions from individual droplets. In their previous work, Forward and Rutledge used a deconvolution strategy based on the assumption that each individual jet produced the same current profile, and that these may be combined additively to produce the overall measured current [27].

Forward et al., modeled the total electrical signal measured $i_w(t)$ as the convolution of multiple jetting events with identical profiles $i_s(t - \tau)$ with differing offset times as shown in Equation 3.2. Forward et al., obtained the total current $i_w(t)$ and an averaged single jet signal $i_s(t)$ from measurements comparable to the measurements of PVA solutions in this work. The single jet, total wire current and estimated density of jets from Forward et al., is shown in Figure 3-3.
Figure 3-3: a) The average single jet behavior observed by Forward et al., under different conditions of the PVP system b) The whole wire current observed as a sum of multiple overlapping individual jetting events c) The density of jets per length of wire estimated by Forward et al.’s deconvolution procedure. [27].

\[ i_w(t) = (j * i_s)(t) = \int_0^\infty j(\tau)i_s(t-\tau)d\tau \quad (3.2) \]

The deconvolution procedure of Forward et al., is somewhat reasonable given the data observed. In particular in the data in Figure 3-3b the observed measurement current is continuous and smooth indicating no breaks in contact between jetting events during a rotation. The observed data at peak current is many times peak of a single jet. The best fit found by Forward et al., was very noisy and the results in Figure 3-3c show the results after smoothing of this signal with a five point moving average.

This assumption of the same current profile for each jet was found not to hold for the solutions studied in this work. Examining Figure 3-11 shows that the current vs time contributions of individual jets which additively combine to the total measured signal are not constant. Similarly the data of whole wire currents often shows period of 0 or little current being transferred before current is re-established. Thus, a new approach was taken to analyze the present data. The model used to fit the data in this work made several simplifying assumptions. Firstly, the single jet current profile was approximated as piecewise linear. The data was modeled assuming independent and identically distributed Gaussian noise. The model allowed for multiple jets to occur simultaneously, or with overlapping periods of contact with the collection plate. Jet currents were assumed to be additive. The time domain was discretized in accordance
with the sampling rate of the electrical signal. The details of the model used to describe multiple jetting and its solution as an optimization problem are described in the following section.

### 3.2.6 Bayesian Modeling of Electric Signals

An example of the current measured from a single wire with multiple droplets forming and subsequently jetting is shown in Figure 3-4. In the inset, an example of the variations associated with the current jetting from individual droplets under identical conditions is shown. The data produced from multiple droplets jetting during the rotation of the wire electrode clearly demonstrate that the electric current associated with all the individual droplets could not have had the same profile with time.

The complex nature of the jetting behavior of the system is illustrated by considering the following example. In the data plotted in Figure 3-4, the current decreases to 0 at approximately 3.5 seconds. The increase in current immediately after this drop is associated with at least one droplet beginning to jet. The shape of the current vs time curve at the end of the data set in Figure 3-4 is inconsistent with the single jet behavior shown in the inset of Figure 3-4. This behavior motivated a reformulation of the analysis to extract jetting information from electric current measurements, described next.

The total current $J_{total}$ at each time $t$ is modeled as the sum of the contributions of $n_{jets}$ terms, $J_i$, each of which is the current associated with an individual jet as shown in Equation 3.3.

$$J_{total}(t) = \sum_{i=1}^{n_{jets}} J_i(t) \quad (3.3)$$

Each of the individual jet profiles is modeled as being piecewise linear, and is parameterized by five scalars $(x_0, x_1, y_1, x_2, y_2)$, as illustrated in Figure 3-5. A single jet
Figure 3-4: Example of current signal collected when multiple droplets are jetting (at 2.5 rpm and 45 kV). The current vs time contributions from individual jets at the onset of jetting clearly have a different shape than those of jets occurring later. The inset shows the average signal from a single jet under the same processing conditions.

The profile is defined by the piecewise linear equation:

\[
J_i(t) = \begin{cases} 
0 & t \leq x_0 \\
\frac{y_1}{x_1-x_0} (t-x_0) & x_0 \leq t \leq x_1 \\
\frac{y_2-y_1}{x_2-x_1} (t-x_1) + y_1 & x_1 \leq t < x_2 \\
-\frac{y_2}{c} (t-x_2) + y_2 & x_2 \leq t 
\end{cases}
\]  \hspace{1cm} (3.4)

As seen in Figure 3-5; the \(x_0\) parameter controls the start time of the jet, and represents physically the time origin of the jetting phenomena. \((x_1, y_1)\) represent the point in current-time space at which the sharp initial rise in current stops. The last two parameters \((x_2, y_2)\) represent the point in current-time space at which the sharp drop off in current begins. The time between the final time-point and zero end-current is fixed in the model based on observed single jet behavior to be 0.1 seconds.
Figure 3-5: The piecewise linear parameterization for the current contribution of a single droplet jetting from the wire electrode to the collection plate. This parameterization can provide an approximate description of the true single jet profiles shown in inset of Figure 3-4, with the exception of the final “tail”. It is sufficiently flexible to model both the apparent jet contributions of “early” and “late” jets observed in Figure 3-4.

This parameterization was chosen to balance the need for a flexible description of individual jet behavior consistent with observed data, while introducing a minimum of parameters for computational efficiency. An alternative formulation, considering a direct scaling of a nominal single jet profile in the vertical and horizontal axis, was found to be insufficiently flexible to model the observed data with high accuracy.

The final estimation problem was solved is formulated as a nonlinear least squares problem:

$\min_{\theta} \frac{1}{2\sigma^2} \sum_{i=1}^{T} (Y(t) - J_{\text{total}}(t|\theta))^2 - \log(P(\theta))$

subject to $f_i(\theta) \leq b_i, i = 1, \ldots, m.$

where $Y(t)$ is the data to be fitted, $\sigma$ is the standard deviation of noise in the measurement signal, $\theta$ is the vector of parameters describing the piecewise linear single jet profiles and $f_i(\theta) < b_i$ are the constraints on these parameters. The full mathematical model derivation and pertinent numerical constants are given below.

The optimization problem (Equation 3.5) is a least squares fit of the model, plus a penalty term associated with the expectation of droplet volume (and total charge) being normally distributed, $P(\theta)$, as defined later. To fit the model given by Equation
3.5 to the data for current versus time, the `fmincon` function in MATLAB was used, with the sequential quadratic programming algorithm. The data was fit over a range of values for an assumed number of droplets jetting in each rotation. For each case, the model parameters defining the jetting behavior were fit in the maximum a posteriori sense. The number of droplets jetting in each rotation was then varied, and the quality of fit was compared between different numbers of jets using the Bayesian Information Criteria (BIC) [78]. The BIC score was used to select the overall best fit model for subsequent analysis. This procedure compares models using different numbers of parameters fairly, so that no bias is given to more highly parameterized models that may fit the data more closely. Details of the BIC calculation are given below.

The probabilistic data model used in this work was defined by Equation 3.6. The model assumes independent and identically distributed Gaussian noise with zero mean and variance $\sigma^2$ in the electrical signal measurements. Equation 3.6 describes the probability of observing a particular data set $\mathbf{Y}$ with a given set of parameters. The parameterization comprises five values to be determined for each of $n_{\text{jets}}$ jets; these parameters are represented by the vector $\theta$. The function $N$ indicates a univariate normal probability distribution function parameterized by a mean and standard deviation:

$$ P(\mathbf{Y}|\theta) = \prod_{t=1}^{T} N(\mathbf{Y}(t) - J_{\text{total}}(t|\theta); 0, \sigma) $$

(3.6)

The maximum likelihood estimate of the vector of parameters $\theta$ is found by solving the optimization problem:

$$ \begin{aligned} & \text{maximize } & & P(\mathbf{Y}|\theta) \\ & \text{subject to } & & f_i(\theta) \leq b_i, \ i = 1, \ldots, m. \end{aligned} $$

(3.7)

This optimization, after the monotonic transformation via the natural logarithm,
is a standard nonlinear least squares problem [8]:

$$\min_{\theta} \frac{1}{2\sigma^2} \sum_{t=1}^{T} (Y(t) - J_{\text{total}}(t|\theta))^2$$

subject to $f_i(\theta) \leq b_i, i = 1, \ldots, m.$

(3.8)

In Equation 3.7 and Equation 3.8, the constraints are represented as arbitrary functions; their exact forms for the current problem are specified later. These functions were used to restrict the parameters describing the current contributed by each jet to be within certain allowed physical ranges. The problem as formulated in Equation 3.8 can be solved to local optimality via black-box optimization algorithms such as sequential quadratic programming.

The posterior distribution of parameters $P(\theta|Y)$ is found using Bayes’ theorem:

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

(3.9)

where the denominator is calculated by marginalizing the joint distribution of the data and parameters in the numerator. The denominator normalizes the output value to the a proper probability distribution but this is not necessary for optimization as it does not change the optimal values found.

A Bayesian maximum a posteriori (MAP) estimation procedure was utilized to find the optimal value of the posterior distribution of the parameters given the observations: $P(\theta|Y)$. The data model (Equation 3.7) and resulting maximum likelihood nonlinear least squares problem in Equation 3.8 was modified to include the prior knowledge that droplet size, and therefore total charge transfer, had predictable variation. This procedure, includes the prior distribution $P(\theta)$ resulting in Equation 3.10 and Equation 3.11:

$$\max_{\theta} \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

subject to $f_i(\theta) \leq b_i, i = 1, \ldots, m.$

(3.10)
\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2\sigma^2} \sum_{t=1}^{T} (Y(t) - J_{total}(t|\theta))^2 - \log(P(\theta)) \\
\text{subject to} & \quad f_i(\theta) \leq b_i, i = 1, \ldots, m.
\end{align*}
\] (3.11)

Including this term improves the accuracy of results relative to those produced from optimizing Equation 3.8, so that the match to the experimental data \(Y(t)\) has better agreement with experimental experience. This experimental and observational experience is included in the prior probability distribution:

\[
P(\theta) = \prod_{i=1}^{njets} P_i(\theta_i)
\] (3.12)

where the symbol, \(\theta_i\), indicates the vector of parameters associated with the \(i\)-th jet. The prior knowledge associated with the parameter \(\theta\) is the approximation that the total charge transferred by a droplet is normally distributed, and that the total charge on each droplet is independent of other droplets. Therefore the prior distribution \(P(\theta)\) can be written as the product of normal distributions on the total charge transferred:

\[
P_i(\theta_i) = N(\int J_i(t; \theta_i) dt; \mu_{\text{charge}}, \sigma_{\text{charge}})
\] (3.13)

The constraints were defined so that the electric current associated with each of the jets on the wire conforms physically to the form observed for the single jet, as exemplified by the inset of Figure 3-4, as shown in Figure 3-5 and modeled by Equation 3.4, so that a left-to-right ordering of the single jet profiles is enforced. This eliminates degenerate solutions that entail only a relabeling of the jet profiles. The constraints used were:

\[
\begin{align*}
x_2 & > x_1 > x_0 > 0 \\
y_1 & > 0 \\
y_2 & > 0 \\
x_{0,i} & > x_{0,(i-1)}
\end{align*}
\] (3.14)
### Table 3.2: Bayesian Optimization Prior Parameters and Model Parameter

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>Jet current time for decrease</td>
<td>0.1 (s)</td>
</tr>
<tr>
<td>$\mu_{\text{charge}}$</td>
<td>Mean integrated current for single jets</td>
<td></td>
</tr>
<tr>
<td>$\sigma_{\text{charge}}$</td>
<td>Standard deviation for integrated current for single jet</td>
<td></td>
</tr>
<tr>
<td>Motor Voltage</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>1.2 V</td>
<td>31.2</td>
<td></td>
</tr>
<tr>
<td>2 V</td>
<td>38.2</td>
<td></td>
</tr>
<tr>
<td>2.5 V</td>
<td>27.8</td>
<td></td>
</tr>
<tr>
<td>3 V</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>Motor Voltage</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>1.2 V</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>2 V</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>2.5 V</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>3 V</td>
<td>2.6</td>
<td></td>
</tr>
</tbody>
</table>

The formulation of the optimization procedure has been expressed explicitly in terms of the number of jets. The data for electrical current from a wire was modeled independently assuming different total numbers of jets (typically 3-8 in this study), based on experimental observations.

The BIC score is calculated as a function of the MAP estimate with posterior probability $\hat{L}$ found by optimizing Equation 3.11 the number of parameters in the model, and the number of observations:

$$
BIC = -2\log(\hat{L}) + n_{\text{param}} \log(n_{\text{obs}})
$$

(3.15)

The lowest BIC score gives the model parameterization for which the associated estimate of jetting behaviors is utilized for subsequent analysis.

#### 3.2.7 Estimating the Electric Field Around the Wire Electrode

A 2D finite element method (FEM) electrostatics model implemented in COMSOL Multiphysics @ 4.0 Modeling Software was used to calculate static electric field around the wire electrode placed between two charged finite-sized planes (the solution bath with applied potential and the grounded collector plate). The simulation
did not take into account the liquid entrainment, the droplets and the presence of liquid jets from the droplets. The electrostatic equations were solved over a rectangular geometry consisting of the air between solution bath (dash-dot line showing the upper level of solution (c) in Figure 3-1(i)) and collector plate (bold line showing the lower end of the metal plate (f) and the electrode wire. The electric field at the surface of a 200 \( \mu \text{m} \) wire was modeled under the application of 45 kV potential. The model was meshed and solved for the electric field at the wire at 19 angular positions along the arc of rotation of the wire about the spindle axis. At each wire location, the model was solved using zero charge boundary conditions on the sides of the domain, the collector plate set as ground potential, and the solution bath and wire held at the applied potential of 45 kV. The results for electric field at the wire are well-described by a second degree polynomial in \( \phi \), the angular position of the wire electrode:

\[
\frac{E_w(\phi)}{V_{\text{appl}}} = A_1 \phi^2 + A_2 \phi + A_3.
\]

(3.16)

with \( A_1 \), \( A_2 \), and \( A_3 \) equal to 4.5 \( \times \) 10\(^{-5} \) deg\(^{-2} \) cm\(^{-1} \), 0.0083 deg\(^{-1} \) cm\(^{-1} \), and 0.685 cm\(^{-1} \), respectively; these values are dependent upon the geometry of the apparatus. The results from the FEM simulation of electric field are shown in Figure 3-6 for the electric field near the surface of the wire as a function of the angular rotation of the wire electrode with the center of rotation 2 cm from the bath.

### 3.3 Results and Discussion

#### 3.3.1 Liquid Entrainment

The entrainment of liquid on the wire was examined for aqueous PVA solution. With dilute solutions of PVA in water, only barrel-shaped droplets were observed, as was the case with PVP/ethanol solutions [27]. The formation of small satellite droplets in between the larger "mother" droplets was observed in case of these aqueous PVA solutions. However, the volume of the satellite droplets was only 3% of the total volume of the mother droplets. The satellite droplets were barrel-shaped as well
The normalized film thickness is expected to scale with capillary number $Ca$ according to a power law relation, as observed previously by Goucher, Quéré, and Forward [27, 37, 73]:

$$\frac{z}{r} = a(Ca)^b$$

(3.17)

For $Ca < 0.20$, the normalized film thickness is well-described by $a = 2.06 \pm 0.08$ and $b = 0.21 \pm 0.02$ with an $R^2 = 0.95$ as shown in Figure 3-7. The film thickness appeared to reach a plateau as the capillary number increased beyond 0.20. The relationship of normalized film thickness to capillary number beyond 0.20 was found to be approximately linear with a slope of 0.21. Beyond $Ca \sim 0.40$, the desired annular liquid film was observed not to form on the wires, due to the relatively high spindle rotation rate compared to the time required for breakup of the film that formed between the wires, and the short residence time of the wire within the electric field for jet initiation. Thus, due to the highly viscoelastic nature of the PVA solution, there exists an upper limit in rotation rate beyond which jetting, and fiber formation, was not observed. Nevertheless, for a given rotation rate, the entrainment of liquid on the wire was observed to be much higher for the aqueous PVA solutions studied here, compared to the solutions of PVP in ethanol reported previously [27].
increased entrainment can be attributed to the higher viscosity of the aqueous PVA solution, resulting in slower drainage of liquid back to the bath.

### 3.3.2 Influence of Electric Field on the Wavelength Parameter

The wavelength parameter serves as a measure of the distance between the droplets in relation to the thickness of the annular liquid layer on the wire. With increase in rotation rate of the spindle, the amount of liquid entrained increased. The center-to-center distance between the droplets was observed to change systematically with rotation rate. In the absence of any applied potential, the average value of wavelength parameter was found out to be $0.44 \pm 0.02$. In contrast to ethanolic solutions of PVP, the PVA/water system exhibited almost no effect of applied potential on $\lambda$, according to the relationship between $2\pi a_0$ and $\lambda$ given by

$$\frac{2\pi a_0}{\lambda} = 0.0013V_{\text{apply}} + 0.44$$

(3.18)
where $V_{\text{app}}$ is given in units of kV see Figure 3-8. This relationship is attributed to the dominance of viscoelastic forces of the liquid over electrostatic forces on the droplet breakup.

![Graph showing correlation between wavelength parameter and applied voltage](image)

Figure 3-8: The correlation between wavelength parameter, $2\pi a_0/\lambda$, and applied electric potential at a constant spindle rotation rate of 2.5 rpm.

### 3.3.3 Jet Initiation and Termination

As the spindle rotates and the electrode wires travel up through the liquid-air interface, the distance between the wire and the grounded collector plate continuously changes. Thus, the electric field strength on the wire changes along the arc of rotation as is detailed in calculations and approximated as quadratic function in Equation 3.16. With increasing voltage applied to the solution bath and the wires, the droplets on the wires were observed to deform from the symmetric barrel-shaped droplets, exhibiting an increasingly prominent conical shape, culminating in emission of a jet. As was the case observed previously for PVP/ethanol solutions, jetting was found to occur when conditions at the wire exceeded a critical field strength. From observations of the height of the wire above the surface of the solution bath at the moment of initiation of the first jet (refer to Figure 3-1 for geometry, with the reference angle of...
90° corresponding to the apex of rotation), the angular position of the wire was determined, and the electric field on the wire at that position was obtained from Equation 3.16. For five different applied potentials in the range of 37-50 kV, the critical field strength for jet initiation was found to occur at 38 ± 2.1 kV/cm, about 10% higher than the critical electric field is observed for PVP/ethanol systems (34 ± 1.3kV/cm).

This difference is attributed to the higher viscosity and surface tension of the aqueous PVA solutions. The implication of the critical electric field condition for jetting is that a larger applied potential on the wire results in a larger angular range where jetting is possible, while a faster rotation speed reduces the time spent by the wire within this angular range. However, in contrast to the PVP/ethanol solutions, aqueous PVA solutions were observed to continue jetting even after the wire exited the angular range in which the critical field strength was exceeded, as confirmed by electric current measurements. This indicates that there exists some hysteresis in the electric field dependence of jet initiation and termination. Such hysteresis was also reported by Shin et al. [79] for a needle-based apparatus. Figure 3-9a presents some typical results for electric current measured as a function of angular position of the wire. It is immediately apparent that jetting generally initiates at the same point (~23.5° for VAppl = 45 kV), but is generally not symmetric about the apex of the arc of rotation (90°). The angular range of jetting versus rotation rate of the spindle is plotted in Figure 3-9b. It can be seen that the angular range of jetting from initiation to termination increases slowly with rotation rate and apparently plateaus at high rotation rate. For rotation rates below 5 rpm, the entrained liquid on the wire electrode is completely depleted by jetting before the wire travels the full angular range (i.e., 23.5° - 156.5°) in which the electric field is above the critical value of ~38kV; this behavior has been called the “entrainment-limited regime” [27]. For rotation rates above 5 rpm, jetting continues throughout the angular range in which the electric field is above critical, and even beyond this, to angles as large as 203°. Under these conditions, jetting is limited not by the amount of liquid entrained on the wire, but by the amount of liquid that can be jetted from the wire electrode in the time required for the electrode to travel from the position at which the electric
field is sufficient to initiate jetting, to the position at which the electric field falls below the value required to sustain jetting. Thus, this behavior has been called the “field-limited regime”.

3.3.4 Single Jet Lifetime

The current measurements such as the one shown in Figure 3-9a are constituted of contributions from many individual jets. In order to estimate the total volume of liquid jetted from each wire in the field-limited regime, it was necessary to deconvolute the wire current into a summation of several individual jet currents. For this purpose, the current originating from a single jet was measured at different rotation rates. An average over at least 10 jets was taken for an applied voltage of 45 kV, and the single jet current versus time is plotted in Figure 3-10. The average lifetime of a single jet of PVA/water solution was calculated to be $t_J = 1.6$ sec and was considered to be independent of applied voltage for the purposes of predicting productivity [27].

In another set of experiments, drops on the wire were allowed to jet only after the wire crosses its apex of rotation (by keeping the voltage off until the wire reached the apex). It is interesting to note that the current profiles for individual jets initiating
at different angular positions were not similar. In the first half of rotation, i.e., before reaching the apex of the arc of rotation, the current profile was observed to have an increasing slope in the middle, as shown in Figure 3-10. However, if a droplet was allowed to jet in the second half of the rotation, i.e., after reaching the apex of rotation, the current profile had a decreasing slope in the middle. These observations contrast with those of Ref [27].

Figure 3-10: Average current drawn from a single droplet at 45 kV at rotation rates of 3.2 rpm (dashed), 5.6 rpm (dash dotted), and 8.7 rpm (solid). Error bars correspond to one standard deviation determined from averaging over at least 10 current measurements for single jets. In the inset, the current profile is shown for a single jet that was allowed to jet only after the wire electrode crosses the apex of the arc of rotation (i.e., in between 90° and 203°).

### 3.3.5 Jet Lifetime and Frequency

The contribution of individual jetting droplets to the total measured current versus time was estimated by fitting a Bayesian model to the data. As described in Subsection 3.2.6 a simplified model for a single jet profile introduces 5 parameters per jets to be estimated by fitting to the measured current versus time profile for a wire
having many jetting droplets. In general, this problem may be unidentifiable. This modeling method requires the determination of a large number of parameters from a limited amount of available data. To improve the tractability of the analysis, the solution space was narrowed by inclusion of constraints on the allowable ranges of parameters, and by the introduction of penalty terms in the final fitting procedure to include the prior knowledge for expected values of these parameters in the physical scenario being mathematically modeled.

For a given set of experimental conditions, the measured volume of a droplet was observed to vary less than 3%. Similarly, the total charge associated with a droplet, as determined by integration of the current versus time profiles, was observed to vary with coefficient of variation <0.1. From 40-60 single jet measurements at each experimental condition, the integral of the current versus time was calculated. These experimental observations were used to introduce a Bayesian component to the data model. The prior distribution included was a normal distribution for the total charge transferred by a jet, with mean and variance determined by this observational analysis. The size of each droplet was assumed to be independent of every other droplet, with the same mean value.

Using this method of analysis, the distribution of single jets as they appear during the rotation of the wire electrode could be obtained. An illustrative example is presented in Figure 3-11 for the measured current from a wire rotating at 5.6 rpm under an applied potential of $V_{appi} = 45$ kV. The figure also shows the best fit set of single jets that reproduce the current profile of the wire.

### 3.3.6 Linear Density of Jets

The linear density of droplets jetting concurrently was obtained by dividing the total number of jets at a particular time point by the length of the wire electrode. The linear density of jets as a function of angular position is shown in Figure 3-12a, and is not symmetrical about the apex of the arc of rotation of electrode (i.e., around 90°) as might have been expected. Rather, the linear density of jets is symmetric about an angular position of ~123°. For lower rotation rates, the linear density of jets drops
Figure 3-11: Optimal model fit to the wire current generated by the Bayesian model, based on piecewise linearity of single jet profiles. The black solid line is the wire current data collected during a free surface electrospinning experiment by rotating the wire spindle at 5.6 rpm with $V_{app} = 45$ kV. The heavy solid line that overlaps the data is the best fit obtained from model. The thin solid lines in the figure represent different single jets as they appear at different angular positions in order to generate the total current from the wire.

The linear density of jets is shown as a function of electric field in Figure 3-12b. Shortly after jet initiation at $\sim 38$ kV/cm, the linear density of jets increased rapidly, and similarly for the three different rotations rates shown here. Beyond the apex of rotation, however, the linear density of jets decreased in a manner that depended on the rotation rate. The lower rotation rate exhibited a more rapid decline in density of jets than the higher rotation rate. For modeling purposes, the dependence of linear
Figure 3-12: (a) The linear density of jets as a function of angular position and (b) electric field at different rotation rates: 3.2 rpm (dashed), 5.6 rpm (dash-dotted), and 8.7 rpm (solid line).

density of jets on electric field was expressed by two second degree polynomials in two different ranges of angular positions:

\[ n_c(\text{jets/cm}) = \begin{cases} 
0.0026E_w^2 - 0.18E_w + 3.21 & \phi < 123^\circ \\
-0.00002E_w^2 + 0.03E_w - 0.10 & \phi \geq 123^\circ 
\end{cases} \quad (3.19) \]

where \( \phi \) is the angular position of the wire.

### 3.3.7 Fitting Results

Fitting the raw current profile under the experimental conditions shows a wide range of behavior. The data under two different rotation rates is shown in Figure 3-13. The standard deviation in the number of jets that are in contact with the electrode as a function of time is very large relative to the mean. This is partially an artifact of the alignment of different time alignment of the jetting behavior.

### 3.3.8 Uncertainty in Electric Current Estimation

The downstream analysis of the models rely on an averaged behavior of for the linear density of jets through time along the wire. This analysis is subject to multiple forms of uncertainty. The first unavoidable error is due to the approximate nature of the model. The model is not flexible enough to account for the true behavior of the
system. Additional error is present due to the non-convex nature of the optimization problem and the fact that the results of analysis only find a local minimum. The local minimum problem is addressed by solving for model parameters with multi-start optimization.

The productivity is ultimately considered by evaluating the number of jetting events per length of wire as a function of electric field. The variation in jetting behavior shown in Figure 3-13 shows that this is a highly variable behavior over multiple estimated jetting profiles. The distribution of jetting behavior found under optimal behavior is more severe due to the presence of a limiting regime where not all droplet jetting phenomena occur at high rotational speed.

### 3.3.9 Regression Model Quality Evaluation

The model prior distributions were evaluated based on the data collected empirical from single jetting studies. The total charge transferred (integral of current over time) was numerically evaluated and the mean and standard deviation were fit to approximate the data as normally distributed. A scatter plot of observed charge transfers at a 1.2V motor rotational speed are shown in Figure 3-14 along with the empirical CDF showing a mean of 31 μC and a standard deviation of 0.7 μC.

The tight distribution of observed total charge transferred in single jet droplets suggests that the regularizing effect of including this prior distribution is very strong.
Figure 3-14: (a) sample total charge transferred for single jetting phenomena at 1.2V motor voltage, (b) empirical cumulative distribution of total charge transferred by single jets at the 1.2V motor voltage.

A demonstration of how this regularizing effect strongly improves model fitting results is seen in Figure 3-15. In this case it is clear that the constraints of the prior knowledge on the jet behavior impact the model understanding strongly. This data set is properly interpreted as being generated with 7 jetting events as the constrained and unconstrained estimates show very similar model residuals.

Figure 3-15: Optimal model fit to the wire current generated by the piece wise linear model with 4,5,6, and 7 total jets to the same total current. The blue solid line is the wire current data collected during a free surface electrospinning experiment by rotating the wire spindle at with 3V to the motor with $V_{appl} = 45$ kV. The green and red lines are best fits. The green fit is unconstrained piece wise profiles. The red line shows constrained profiles by the prior knowledge of jet durations.
3.3.10 Productivity model

For free surface electrospinning from a wire electrode, productivity is defined as the total mass of product deposited on the collector per unit of time and length of the wire electrode; this is the same convention as employed earlier by Forward and Rutledge. Productivity was determined from the mass of fibers deposited on the collector as a function of rotation rate of the spindle, for four different applied electric potentials: 40kV, 42.5kV, 45kV, and 50kV. The productivity model that was developed earlier was modified to fit the current experimental data. As was the case for the PVP/ethanol system, two regimes were observed for free surface electrospinning of the PVA/water system. In the entrainment-limited regime, productivity was determined by the amount of liquid entrained on the electrode and can be estimated using:

\[ P_E = \pi \Omega C r^2 [(aC a^b + 1)^2 - 1] \]  

where \( C \) is the concentration of polymer, in mass per unit volume, assuming that the amount of liquid entrained does not vary with applied voltage [27].

In the field-limited regime, the productivity is determined by the volumetric flow of liquid from the wire electrode to the collector plate, and can be obtained using:

\[ P_F = \frac{V_d \Omega C n_j}{2t_j} \left[ (aC a^b + 1)^2 - 1 \right] \int_{\phi(E_{\text{init}})}^{\phi(E_{\text{term}})} n_c(E_w(\phi')) d\phi' \]  

where \( n_J \) is the number of jets emitted from the wire electrode during one single rotation of the spindle; \( E_{\text{init}} \) is the critical voltage at which jetting is first observed ("initiation"); \( \phi(E_{\text{init}}) \) is the angle at which such jet initiation is observed for a given applied voltage, as determined by Equation 3.18 using \( E_w = E_{\text{init}} = 38 \) kV/cm; and \( \phi(E_{\text{term}}) \) is the angle at which jet termination is observed for a given applied voltage and rotation rate. For purposes of this work, the upper limit for jetting was set to be 203°, as observed from experiments. The liquid entrained was considered to be independent of applied voltage and only a function of capillary number. Both
experimental and predicted productivities were considered to be valid for one or more wires, as long as the wires were not close enough to influence the electric field of the other.

The productivities measured experimentally at the four different applied voltages are plotted as functions of rotation rate in Figure 3-16. Also shown are the productivities predicted by Eq. 6 and Eq. 7 for the entrainment-limited and field-limited regimes, respectively. The model for the field-limited regime, $P_F$, in particular, shows good agreement with the experimental data. Although the linear density of concurrent jets was only measured at 45 kV, the model predicts the productivity observed at other applied voltages reasonably well.

By applying the model to PVA/water, a significantly different system than PVP/ethanol, the general form of the model and the basic physics behind the model are shown to be robust. Keeping the general form of the model the same as earlier, with only a few modifications, the model can predict productivities for two completely different systems. Also, based on the observations for both systems, it remains desirable to operate the free surface electrospinning experiment from the wire electrode at high potentials and at high rotation rates. However, for the PVA/water system, the field-limited regime is reached at a much lower rotation rate than was the case for the PVP/ethanol system, and a higher applied potential is required to initiate jetting. These observations are believed to be due to the higher viscosity and higher surface tension of the aqueous polymer system relative to the ethanol-based system. Also, at a particular desired electric field for electrospinning, the rotation rate at which the transition occurs from entrainment-limited to field-limited regime should be picked for operation with optimal performance.

### 3.4 Conclusion

The analytical model that was previously developed for polymeric solution in organic solvent has been applied to an aqueous polymeric solution. This model, with only a few empirically determined parameters, proved to be sufficiently robust to describe
Figure 3-16: Model fitting to the experimental productivity data. Experimentally observed productivity (filled symbols) at different rpms at applied potentials of 40 kV (square), 42.5 kV (circle), 45 kV (triangle), and 50kV (diamond). The entrainment-limited productivity as calculated from Equation 3.20 is shown by the solid line and the field-limited productivities at different applied potentials as functions of rotation rate are shown by dashed lines. The transition from the entrainment-limited regime to the field-limited regime is most apparent in the data at 45 kV.

two very different types of systems. The relationship for entrainment of the aqueous PVA solutions on the wire was found to be only slightly different from that reported for the PVP/ethanol solutions, which we have attributed to the higher viscosity of the PVA solutions. By contrast, in the present case, the thickness of the entrained liquid layer was comparable to the wire diameter, which is quite different from the earlier observations for PVP/ethanol. Satellite droplets formed, comprising ~3% of the volume of annular film of liquid, but they did not lead to jet formation under the potential applied in these experiments So, although the volumes of these small droplets were considered for purposes of correlating the thickness of the liquid layer with capillary number, they were subtracted when estimating “entrainment-limited” productivity. This treatment proved to be appropriate, as the model could predict the experimental data quite well. Once again, there exists a critical field for aqueous
systems, just like organic systems. For aqueous systems, the magnitude of the critical field was larger; we attribute this to a higher viscosity and higher surface tension of the aqueous solutions.

As seen earlier, two limiting regimes for productivity were observed in the case of aqueous polymeric solution electrospinning. When the rotation rate of the electrode spindle is relatively low, the productivity is limited by the liquid entrainment on the wire. Under this situation, all the liquid entrained on the wire can be converted into jets during the rotation of the electrode as long as the electric field is sufficiently high. In the second regime, when the rotation rate of the electrode spindle is high, there are two limiting factors that come into play in particular for viscous aqueous polymeric solutions. First, the time required for Plateau-Rayleigh breakup of the annular film needs to be sufficiently small such that the highly curved surfaces of the droplets are formed before the wire reaches the critical field. Second, the wire rotation speed needs to be slow enough such that all the droplets have sufficient time to jet before the wire again dives into the solution bath. Thus, this second regime is limited by the angular range of jetting as well as the Plateau-Rayleigh breakup time.

The analysis presented here for aqueous polymeric systems, along with the previous one for ethanol-based polymeric systems, provides a platform to understand the productivity behavior from a free surface electrospinning setup. The key parameters by which the productivity of such a process can be characterized and optimized are not only the applied electric potential, but also the rotation rate of the spindle, spindle geometry (wire diameter, length, number of wires and distance between the wires) and fluid properties.

The role of statistical modeling in developing the best overall estimate for the jetting behavior of the system is paramount in developing the system understanding. The distribution of jetting behavior with time is a crucial input to understanding the productivity of this system. The use of a statistical modeling and then a model selection procedure was necessary to understand the angular range of jetting behavior and variation of the system with rotation speed revealing field limited and entrainment limited productivity regimes.
Chapter 4

Anomaly Detection and Diagnosis in Production Oil Wells

4.1 Introduction

The high stakes nature of oil and gas production has lead to the industry being motivated to expend great effort in data acquisition and process monitoring technology. The production of oil and gas has been increasingly instrumented to gather data during a well operation. The data collected can be used to develop systems to improve the operation and management of the oil field.

Oil and gas well and subsurface systems can develop into one of many undesirable states that are believed to precede system deterioration and potential failure. Most significantly, failure to anticipate problems in producing wells can lead to deferred production, costly interventions, and risk to the integrity of the well. Given the large cost to develop a new well, such losses can cost hundreds of millions of dollars.

Oil and gas production wells have been increasingly deployed with more advanced instrumentation and measurement technology as operators strive to achieve maximum awareness of the well and reservoir conditions to avert losses. In the systems studied here, continuous sensor measurements are available providing insight into the conditions at the well head, at the bottom of the well bore and are integrated into surface equipment. In other industrial scenarios supervised learning techniques have
been successfully deployed via commercial software solutions (e.g. rotating equipment). The practical challenges of well operations has motivated the development of a different technique: using a feature based representation of portions of operating history.

Oil wells exhibit batch process-like behavior. Wells have a wide range of operating conditions. The state of the system and associated dynamic response to operator actions or disturbances changes during the long-running history of production. Wells are subject to external perturbations and disturbances from the reservoir, as well as interventions in the form of automatic and manual control inputs due to surface activities. The fluid flow experiences both discrete and continuous changes over the course of operations. For example, the flow can be routed through a different manifold, a discrete change, or it can be altered using an artificial lifting strategy, a continuous change. The data collected has informed the development of methodologies and tools to support a wide variety of operational situations. Some of the first areas for applications are the diagnosis of abnormalities in flow rates, unusual operating conditions, and the identification and prediction of faults or integrity risks.

The task of monitoring oil wells has several challenges, a few which have been identified in collaboration with industrial experts. To have practical value, methods must support adaptation as wells and reservoirs mature. Each well is instrumented with only a relatively small number of sensors, but many time points collected in sequence. This leads to a dataset that is very large in the sample direction relative to the dimension of each sample.

Monitoring, analytics, fault detection, and anomaly detection methods have been widely deployed in industrial process systems [15, 77]. The adaptation of these methods to production of oil and gas has limited discussion in the academic literature. Smart monitoring and control of oil production equipment is not novel. However, the specific problems addressed in the literature are not the same as those investigated in this work. Investigations in the literature have been have studied problems in the drilling phase of oil wells, and modeling of reservoirs and other tasks [69, 70]. Nybø et al.'s work coupled machine learning with physical models from process simulation,
and used a predetermined set of failure modes and physical signatures associated with each. In other work, fault signatures associated with particular trending behaviors were predetermined in advance manually, and used to develop decision criteria to identify faulty behavior [47]. Work also exists on sensor failure using support vector regression and a soft sensor to identify sensor drift in oil wells as the only fault class [9]. Support vector regression has also been used to predict time between failure for wells in an oil field based on well characteristics but with no dynamics considered and given a labeled history of failures. In work on failure prediction, a large human factor ad-hoc system was considered with human labeled of failure and a 2-class RBF kernel SVM used for fault identification [59]. Other efforts exist in using machine learning methods such as using neural networks to statically identify sensor failure in down-hole sensors [20].

This work describes the characteristics of process measurement data in production oil and gas wells in real datasets provided by BP. Based on the analysis of the available data, a system for identifying anomalous periods of process operating history has been developed. This system relies on a transformation of multivariate time series measurements into static feature descriptions capturing the salient information contained in the measurements. This allows for identification of anomalous events which are interesting in expert analysis and correspond with observed events in the well’s operation. Three attributes of the data collected from production oil wells are highlighted as they impact the applicability of several conventional anomaly detection and diagnosis methods. In comparison with previous efforts, in this work we have identified the importance of having multiple operating modes, and dynamics in the available data. Labeled failure examples to build a two class classifier are not available, neither are preexisting physical signatures for specific classes of faults or anomalous behavior.

Feature based classification has been recently used to identify anomalous whole time series for further analysis in data from monitoring computers and software processes [55]. In this chapter, a large number of potentially salient features is utilized to transform subsegments of a single large time series. Additionally, this chapter ex-
tends the application of univariate feature descriptions of time series data to a highly practical multivariate scenario. The method introduced in this work uses approximate system information to bootstrap a selection of features for time series anomaly detection when faulty information is available. This is performed with feature based classification of non-overlapping time series consisting of portions of the well operating history.

4.1.1 Production Oil Well Measurement Data

BP has provided anonymized data from production oil wells from both on-shore and off-shore operations. There is heterogeneity in the data due to a non-uniform number of possible flow routings and varying sensor installations. This is a practical concern for enterprises developing monitoring technologies to be distributed at scale. Data is available as continuously sampled time series for the sensor data. In addition to the sensor data, non-continuous sampling measurements or other data is available but not used in this work. The inclusion of infrequent well test, sand sampling and other data is being investigated for future work. This work shows results from one highly characterized well.

The largest set of measurements is available for wells which have dual completions on the surface and artificial gas lift. These consist of bottom hole pressure, bottom hole temperature, well head pressure, master valve position, wing valve position, choke valve position, acoustic detector, gas lift flow and choke, gas lift wing valve, two annulus pressures and three diverters valve positions on each flow line for high pressure, low pressure and test flow paths, as well as the pressure for each of the possible receiving manifolds. This continuous measurement data is available at a fixed 2 minute interval representing an average for each sensor over a 120 second period. For wells which are single completion or don't feature artificial lift corresponding measurements are not available. In addition for some wells, well head temperature measurements are not available. The diagram of well sensor locations and manipulatable valves is shown in Figure 4-1.

The signal from the acoustic detector is of high interest. Acoustic detectors have
Figure 4-1: A diagram of sensor locations on a dual completion oil well. BHP: Bottom hole pressure. BHT: Bottom Hole Temperature. WHP: Well Head Pressure. WHT: Well Head Temperature. ACC1, ACC2: Acoustic Sensors
been used to monitor flow regimes of multiphase flow [21]. Acoustic detectors have also been used to monitor sand production and other flow assurance concerns such as hydrate formation [19].

**Well Restarts**

At various points during a well’s operating history it may be have been shut in and production of fluid stopped. The transition from a shut-in state to a full open production state is performed gradually by opening the choke valve(s). This process is called “beaning-up” the well. This is known to be high risk operating period for failures of the well’s systems. It is also known that the changes in pressures experienced by the well bore and reservoir during production can result in different sand production and fluid flow compared to nominal operations. The distribution of observed bottom hole pressure and well head pressures shown in Figure 4-2 illustrates that when comparing bean-up versus nominal production that the well is experiencing quite different fluid flow and pressure stresses.

Figure 4-2: Empirical cumulative distribution for bottom hole pressures and well head pressure during bean ups or main production modes.
Deferrals and Losses of Production

Another critical portion of wells operating history occurs when the well is forced to have lower than desired or predicted production. This occurs when the well is forced to operate in a reduced capacity. These periods of low production are called deferrals. The well conditions that lead to lost or delayed production are of maximum economic value for identification. In some cases the production of sand or particulates can be associated as a cause of these deferrals. Wells naturally produce solid particles such as chalks and sand. These produced particulates can cause issues with surface equipment or could be indicative of more serious problems in the integrity of the reservoir. Sand production alone is not necessarily something to be optimized against but unexpected changes in sanding behavior or abnormal behavior is worthy of being identified. The sanding behavior is expected to be identified by changes in the signal on the acoustic detector. There is an unrecoverable loss in production which was associated with very high levels of sand production for the well that is studied in this report. The portion of the well's operating history ahead of this loss was treated separately from the remained of the well's production history. In other wells some deferrals are temporary and the well resumes normal production. The portion of operating history immediately before these losses of production was treated separately from the rest of the well's operating history.

4.1.2 Outlier Detection, Novelty Detection, and Classification

Anomaly detection has often been considered in the context of classification problems between two or more classes. Often this work is introduced under the paradigm of fault detection because the types of faults are assumed to be known in advance and in this context sufficient data is available about the distribution of data under faulty conditions. In this case the classification performance on training and test data is the core metric for evaluating the success of the methods under study or development.

An alternative related problem is that of novelty detection where data is available exclusively under known nominal conditions where the data is all labeled with
one class in the training context with none of the training examples expected to be anomalous.

In the outlier detection problem context the available data is assumed to be largely in one class (nominal) but some fraction of the data which may be known or unknown should be properly identified as an anomaly. This context is especially difficult if the fraction of training data which contains outliers is unknown. This is principally the context of the oil well ADD problem. In this case without labeled data the success of the method is reliant of *ex post facto* investigation of the points or portions of operating history identified as outliers and inliers and the distributional or statistically properties of the decision boundaries and datasets produced.

In the classification context, two perspectives on the problem exist a geometric or discriminative perspective and a generative or probabilistic perspective. In the generative perspective the classes belong to separate probability distributions, and the determination of which class a new example belongs to is given by the Bayes decision rule given in Equation 4.1 assuming an equal preference for misclassification. For two classes $C_1$, and $C_2$ a new observation $x$ is assigned such that it is in the class for which it has maximum a posteriori probability:

$$x \in C_i, C_i = \arg \max_{C_i \in \{C_1, C_2\}} (P(C_i|x)).$$

(4.1)

The posterior probability $P(C_i|x)$ is found via Bayes rule which is given in Equation 4.2. The posterior probability of the class label is the joint probability of class assignment and observation conditioned on the marginal probability of the observation.

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)}$$

(4.2)

The joint probability of class assignment and observations is factored into two terms $P(C_i)$ called the prior probability of a class and $P(x|C_i)$ which is the likelihood for observations conditioned on being in a particular class. This is a generative probabilistic model which describes the probability of observations in fault and non-fault states. The specification of this model is at the heart of many generative classification
problems. The identification, assumption or learning of a particular distributional relationship for the data is what drives the differences between generative classification methods. For example assuming that observations are multivariate normally distributed with different means and covariance structure in each class, and identifying these distributional parameters with maximum likelihood estimation is referred to as quadratic discriminant analysis (QDA) as the decision boundary is found as a quadratic function between the two classes. If shared covariances are assumed in such a mixture of Gaussians model results in linear discriminant analysis (LDA) with a linear decision boundary between each class [66, 62].

QDA and LDA can have reduced performance in the presence of non-informative variables [90]. Non-informative variables for the classification task do not assist in separating the classes but can be viewed as adding noise which increases the classification error. Verron proposed using a mutual information criterion to identify those variables which contribute to the classification of faults in QDA or LDA classification fault diagnosis settings [90].

4.1.3 Data Driven Dynamic Models

In some cases it may be possible to learn and validate models from the data to which are representative of the process by which the data was generated. In well understood problems, such as sufficiently well characterized physical systems, a mathematical model may be able to be constructed which maps from knowledge of the system state to the observed measurements. Inverse problem methods can be used then to map from the measurements to estimates of the system state. Measurements can be identified as abnormal to the extent that the model residuals, or estimates of the underlying state of the system differ. In these reconstruction methods, if the model is well chosen to be representative of the system, the expected generalization capacity of the model is to be strong. However, the extent to which the model is an inappropriate approximation of reality can introduce a large bias. These models introduce additional complexity beyond the distributional assumptions of mixture of Gaussian type models. Dynamics and nonlinear relationship between measurement
variables can be effectively incorporated.

One class of models that has been widely applicable in developing fault models for chemical processes has been to infer a linear dynamic system representation through canonical variate analysis. This description can then be used for anomaly detection and diagnosis such as in combination with Fisher discriminant analysis [48].

4.1.4 Multivariate Extreme Value Statistics and Probability Density Estimation

One approach for identifying anomalies is to model the probability distribution, \( f(x) \), of nominal observed data, \( x \). In this case, a cut off probability is introduced to classify examples as inliers or outliers: \( f(x) \geq \text{threshold} \rightarrow \text{inlier} \). Making various distributional assumptions about the values of nominal data leads to the construction of differing extreme value statistics. The performance of one-class classifiers can be evaluated by the difference between the results on training data and data withheld from the classifier. An example of this is constructing a Gaussian density around the dataset and then optimizing the spread parameters and a cut off level to determine a specified number of outliers. A more flexible approach is to model the data as a mixture of Gaussians where the data density is described by.

\[
    f(x) = \sum_{i=1}^{K} P_i \mathcal{N}(x; \mu_i, \Sigma_i)
\]

(4.3)

where \( K \) is the number of Gaussians, \( P_i \) a probability weight for each Gaussian and \( \mathcal{N}(x; \mu_i, \Sigma_i) \) is a multivariate normal distribution with mean \( \mu_i \) and covariance \( \Sigma_i \), evaluated at the observed values of a sample \( x \). In high dimensional data the number of parameters to be estimated by a mixture of Gaussians model is very large so the covariance matrices are often assumed to be diagonal or diagonal with constant variance for all features.

An even more flexible density model for the probability density is to construct a Gaussian distribution around each training sample, this is called Parzen density

126
estimation. In this case the density model used contains a sum over all of the \(N\) training examples which are given equal weight and equal covariance matrices,

\[
f(x) = \sum_{i=1}^{N} \exp(-1(x - x_i)^T \Sigma^{-1}(x - x_i))
\]  

(4.4)

where the covariance is given as \(hI\) where \(I\) is the identity matrix and \(h\) is a scalar parameterizing the spread. The \(h\) scale parameter and associated classification threshold are then estimated from data to achieve a desired outlier fraction in the training set. The mixture of Gaussians and Parzen models assume an underlying density distribution of the observations that is static. In time varying system of an oil well application of these methods requires a transformation of the dataset into a static feature space.

### 4.1.5 Feature Modeling of Time Series

A library of 9613 time series algorithms and analysis procedures was identified and collected from the scientific literature in [31] and subsequently used for classification of time series in [30]. In these related works, the authors produced a library of features as computational procedures to map from an input of univariate equally sampled time series to a scalar value. These summary statistical features include information theoretic quantities, model fit information, and others such as wavelet methods. A full detailing of features is available in [31, 30].

The implementation used for this analysis was the Beta release of the software excluding the TISEAN features [43] which weren’t able to be run on our computing platform.\(^1\) Using these “HCTSA” feature based representation of the time series the authors were able to perform multi-class classification of EEG signals into healthy and abnormal conditions. Using a large library of features many dimensions used individually were found to be useful in constructing linear classifiers with high accuracy [31]. When examining the 20 multi-class labeled examples in the UCR Time Series Classification Archive [14], a small number of features and linear classifiers were competitive with or yielded superior performance than 1-NN instance based classifiers.

\(^1\)https://github.com/SystemsAndSignalsGroup/hctsa
using dynamic time warping distance metrics [30].

4.1.6 Feature Selection and Classification

When constructing classifiers utilizing a large feature space dimension [30] used greedy forward feature selection to identify which features to include. Greedy forward feature selection grows the set of features utilized by incrementally adding a feature and calculating the classification rate on the training data. The feature with the highest resulting classification rate is selected. The procedure of adding features ends when the improvement in classification performance was below a target level or a 0 training error was found. Over 20 example classification problems from the UCR database the mean number of features selected was less than 4 at 3% improvement level stopping criteria [30]. Reducing the stopping improvement threshold resulted in an increase of features included in the model and an improvement in mean misclassification accuracy indicating that the linear classifiers examined did not tend to over-fit the data when considering 6 or more dimensions.

A related task in classification to embedded feature selection methods such as greedy forward feature selection is ensembling of various classifiers. Bagnall et al., showed that ensembling of multiple feature space classifiers can lead to state of the art performance on many test bench problems in time series classification [6].

The ranking of features can also be conducted by evaluating the mutual information between class labels and the scalar feature. The mutual information between the labels and the feature value has been utilized to select feature for fault detection and identification [90].

4.1.7 Feature Transformation and Anomaly Detection

Recently, feature based identification of entire anomalous time series was introduced in the context of monitoring time series dynamics seen in web-traffic and server monitoring metrics [55]. Laptev et al., mapped time series to a collection of features, used PCA to reduce the system to two dimensions from linear combinations of these
features, and identified anomalies in the two-dimensional PCA projection with an $\alpha$-convex hull technique.

They utilized 15 features were selected manually and all features were used in the anomaly detection procedure. In [55], the evaluation of anomaly detection performance was considered by using labeled examples in a synthetic dataset. Performance was compared against time series distance metrics, and clustering based techniques in feature space. Their technique is reported to be deployed at scale at Yahoo. The details on the source of ground truth labels for anomalies for real process data examined by [55] is not detailed and presumably comes from manual labeling.

4.2 Methods

4.2.1 Data Preprocessing

Data preprocessing was necessary because the data may be corrupted by data handling errors. Some of the errors are recorded in the primary data storage but, other failures were not identified in the original data storage. Measurements where a value was indicated as being suspect were replaced with an internal NaN indication. Extreme values, recorded with a value so large as to be physically unreasonable were also present in the data, a small number of measurements with such non-physical values were also replaced with NaN. These two preprocessing data cleaning steps impacted less than 1% of the available data. Further preprocessing was necessary to increase the continuity of the data. The proposed time series analysis methods require continuous portions of process operating history. If 1-3 sequential measurements were recorded as NaN values these were replaced with linearly interpolated values of the nearest neighbors. If a larger portion of process history was marked by a continuous set of NaN recordings these were considered a break in the recorded data for subsequent analysis.
4.2.2 Segmentation by Well Operational State and Routing

The valves which control the flow of fluid from the reservoir into the receiving manifold can be open or closed in a combination of possible states. Some valves are digital and represent an open or closed switch, while the chokes can be closed, partially or fully open. Based on the recorded valve position data it is possible to preidentify and code the operating modes associated with these valve positions. Changes to the flow path impact the system dynamics, for example some measurements may be correlated under some conditions but then otherwise irrelevant. An example of this phenomena is that the behavior of the acoustic detector is a function of the pressure drop over the detector. However, selection of the appropriate downstream pressure is a function of valve positions associated with production manifold routing.

Wells are ideally operated to be flowing to production manifolds, with fully open chokes to maximize the flow of fluid and profit for the operator. However, in practice, operational changes are made to react to surface equipment conditions or undesirable well behavior. These changes in how the well is operated were identified by inspection and the categorization of the system’s behavior into 15 categories was done by constructing rules based on valve positions. A trace of the process history of this well over its observed life ~800,000 data points and what state the system is in based on valve routings is shown in Figure 4-3.

4.2.3 Feature Transformation and Anomaly Detection

Expanding upon the work of [55] and [30], we introduce a paradigm for feature based identification of anomalous subsequences of large multivariate time series. Our approach directly adapts the feature based approach to a multivariate context producing features for each dimension of the data independently. We also note several features from the HCTSA database [31] are naturally adaptable to a multivariate context to use the mutual information content of measurements jointly changing in time, but do not use any new features in this work.
Figure 4-3: Well routing history for the well under study. The well state changes many times over the course of a well's operation between the pre-identified labeled states associated with valve routings.

**Multivariate Information**

Two approaches to using multivariate data are utilized. In the first, multivariate time series are converted into feature representations by concatenating the features mapping of each dimension. The feature transformation functions are several maps,

\[ [f_a(x_k), f_b(x_k), \ldots] \rightarrow [f_{a,k}, f_{b,k}, \ldots] \]  

where is a \( f_a(x_k) \) function from a time series of measurement values \( x_k \) to scalar values \( f_{b,k} \). The inclusion of multiple measurements in a feature vector can be constructed by using the same features when calculated on new measurement data and appending these results to the feature vector. This means none of the features explicitly encompass the multivariate nature of the measurements of the system. The data collected is organized as in Table 4.1 and consists of \( m \) measurements (different sensors) \( M_1 \) through \( M_m \) collected at \( N \) time points. Each portion of operating history is constructed as a univariate time series of \( n \) measurements.

Using this data organization, the resulting feature vector can be calculated at each
The multivariate feature vector is given as

$$\begin{bmatrix} f_a(x_k^{(1)}), f_b(x_k^{(1)}) \ldots f_a(x_k^{(2)}), f_b(x_k^{(2)}) \ldots \end{bmatrix} \rightarrow \begin{bmatrix} f_a^{(1)}(x_k), f_b^{(1)}, \ldots f_a^{(2)}(x_k), f_b^{(2)}(x_k), \ldots \end{bmatrix}.$$  

The same feature transformations are conducted on each univariate time series as separated from the original multivariate data. This large dimensional feature vector is reduced to a lower dimension and a one-class classification method is used to identify anomalous data.

The other strategy for multivariate data inclusion is to construct a one-class classifier for each measurement variable separately. The feature transformation and dimensionality reduction are also considered separately. Then to aggregate the results a voting procedure is utilized to score time series segments by the number of dimensions for which it was determined to be anomalous.

**Dimensionality Reduction and Feature Selection**

After feature transformation, dimensionality reduction was performed on the resulting feature vector in multiple ways. When multiple classes of data are available, features with strong classification performance are identified using the labeled classes in the data. The set of features identified by this process is used to determine a decision boundary for anomaly detection by using a one-class classification strategy. In many situations multiple classes of data are not available.

Dimensionality reduction methods based upon principal component analysis were
utilized as would be applied having only one class of data.

When using classification to evaluate which sets of features are important in the experimental data from production oil wells, the dataset was split into three portions of process history by manual labeling. These portions of process history are near restarts, far from restarts and near deferrals. The varying stresses on the well and formation during well restarts result in different sanding behavior than typical long term operations. This variation in qualitative sanding behavior from normal operational conditions is one behavior we wish to identify during otherwise nominal operating conditions.

Feature selection was done based on two principles: 1) recovering a small number of features to aid human interpretation, and 2) achieving a high degree of classification capability. An embedded feature selection was evaluated through sparse discriminant analysis SLDA [18]. The sparse discriminant analysis method performs linear discriminant analysis with a sparseness criteria imposed. Using SLDA is desirable because we expect to use a large number of features that may exceed the number of training examples and many features may contain redundant information. Using SLDA allows selection of an upper bound on a desired number of non-zero component features in each discriminating direction [81].

Feature ranking was conducted by independent classification accuracy on labeled data. The linear separation between the two classes was utilized. In this case a loss function such as the 0/1 loss is minimized by optimal selection of a single parameter $\lambda$ with the associated decision rule that $x > \lambda \rightarrow \text{Class 1}, \ x \leq \lambda \rightarrow \text{Class 2}$ where $x$ is the scalar feature value of interest.

Feature selection in classification problems for one-class classification problems here deserves special attention. The previously discussed methods were largely considered in the context of two-class classification. In the anomaly detection and diagnosis context without labeled anomalies, it is impossible to know a priori the effectiveness of a particular feature. In order to accomplish features selection in such a context we can examine the cross validation performance of various features and the relative agreement they have.
In [30], for multi-class classification problems with linear classifiers different features were found to be discriminatory between the multiple classes in a 1 vs 1 coding. Novelty detection or outlier detection frameworks are inherently a 1 vs all coding. The outlying behaviors may exist in multiple classes whose characteristics are largely unknown.

This motivates the selection of features in a context of dimensionality reduction. There is value to having highly varying features in the training set as these features are then known to be sensitive to the kinds of variation observed in the data. Dimensionality reduction for visualization can be conducted by principal component analysis (PCA) which constructs linear combinations of features that are orthogonal and additively capture the maximum variance of the original feature space description. PCA as well as Robust PCA was used as the dimensionality reduction scheme in [55] so that boundary based method could encompass the data and it could be easily visualized. When studying the dataset of [55] we found superior performance using PCA (non-robust) and utilized it for dimensionality reduction ignoring classification in this work.

To construct a decision boundary around the data with Parzen density estimators the implementations in the software package DDtools were utilized [83].

4.3 Results

4.3.1 Benchmark On Other Data

A range of existing data sets exists for benchmarking time series classification problems. The University of California at Riverside has a repository of multiple real world and synthetic datasets consisting mostly of short univariate time series consisting of data belonging to multiple classes [14]. Yahoo released a dataset of server monitoring metrics that contain nominal and abnormal operating data [54]. Fault and fault free datasets are also available from various databases produced NASA Glenn Research Center and NASA Ames Research Center for prognostics research [68]. Long run-
ning univariate data sets were made available for finding discordant subsequences by Keogh et al., [51].

The dutch power demand data set is a multivariate dataset showing the power demand of a dutch city over the course of 1 year. The data is available publicly.² The data is univariate and is sampled at 15 minute intervals. For analysis the data was split into 52 segments each of which contained 672 data points and represents a week of demand data. The splitting of the data into one week periods is a natural transformation of the data for comparison because the data shows strongly weekly periodicity. The data was transformed into a feature space description using the HCTSA features.

The dutch power demand data does not include any labels that could be used for feature selection for the identification of salient features. The data was projected down to two dimensions for visualization using principal component analysis. A parzen density estimator (Equation 4.4) was evaluated targeting 3% of the data as outlying this projection of the data is shown in Figure 4-4.

![Figure 4-4: Parzen density estimator showing decision boundary tuned to reject 3% of data as outlying.](image)

The feature space axis are the first 2 principal components from 6888 time series features.

Examining the data found to be outlying with this simple transformation of the data yields qualitatively interesting results that agree with human judgment. Three weeks of data each found to be outlying in the feature space representation are shown in Figure 4-5 along with the mean weekly demand data.

![Graph showing average weekly power demand and outlying weeks found by feature space transformation](image)

Figure 4-5: Left: Average weekly power demand in dutch power demand dataset, Right: outlying weeks found by feature space transformation (offset). Time is in 15 minute increments.

In the average data the five peaks correspond to the five work days with the first peak Wednesday. (Jan 1 was a Wednesday in this dataset). The most outlying datasets in the right pane of Figure 4-5 show weeks that correspond to holidays. The unusual sequences are the 8th (black), 18th (red) and 52nd (green) weeks of the year. The 52nd week of the year is clearly different than the mean behavior because the presence of holidays on the dutch calendar shown low demand for Thursday and Friday corresponding to Christmas and the second day of Christmas which clearly have lower than normal demand. The 18th week of the year shows a holiday on Wednesday in late April corresponding to the King/Queen’s birthday in the Netherlands and Liberation day on the following Monday. The outlying 8th week was identified not by an obvious drop in demand but by unusual variability on the Saturday of that week. This demonstrates the power of the feature space representation to find sequences that are humanly interesting and justifiable in as differing from each other in the absence of labeled ground truth. Keogh et al., similarly found weeks containing holidays as
the most outlying subsequences when using SAX representation to estimate a DTW distance between sequences [51].

A feature space transformation was successfully deployed across many datasets by both [30] and [6]. The transformation to a feature based representation and either using a large library of features or ensembling many classifiers in feature space was beneficial across many classification problems. The power of the large feature space representation is shown here in an example on the Ford data set from the UCR archive. The Ford dataset was originally constructed as a challenge problem “Ford Classification Challenge” for the World Congress on Computational Intelligence, 2008. This dataset was analyzed with ensemble methods in [7] and best performing test error of 0.1447 was achieved by using a rotation forest classifier with 30 base classifiers. The performance on the preallocated test set using the training set as a 1-NN euclidean distance classifier is has an error rate of 0.341.

The Ford set is split into two classes and consists of 500 sample long time series that either contain or do not contain an anomalous condition in an automotive sub-system. This represents labeled examples of failure in an industrial system that does not have clearly human interpretable features. Some example traces from the Ford dataset (FordA on UCR) are shown in Figure 4-6. Nominal data is plotted on the top and anomalous data is shown on the bottom.

After calculating the HCTSA features on 806 example time series from the ford data set and normalizing the features to remove time series features that produced errors or were otherwise uninformative 6093 features remained. After projecting the data into two dimensions via PCA the two classes (nominal and anomalous) are not well separated as shown in Figure 4-7.

Even though when projecting the data into two principal components does not clearly show class separation several of the features calculated shown good classification performance. The distribution of linear discriminant performance on the training set of 806 examples is shown for 6093 features in Figure 4-8.

Many individual features show strong classification performance even through the data set is not linearly separable after using principal component dimensionality
Figure 4-6: Sample data from the Ford classification challenge dataset showing time series measurements from an engine subsystem. a) nominal measurements b) anomalous behavior

reduction. The best performing feature has 98.01% classification accuracy on the training set and the top 10 features all have > 95% classification accuracy on an individual basis training an LDA model. If instead of visualizing the data through principal component analysis two highly separating features are used data is now clearly well separated as shown in Figure 4-9.

The feature transformation approach here has been highlighted as effective in identifying human interesting or labeled anomalous portions of operating history in multiple contexts. The large feature space was used directly for the Dutch Power Demand dataset where good performance was achieved in identifying human relevant subsequences. On the Ford engine classification test problem the highly comparative feature selection procedure identifies several salient metrics which achieve world best performance at the classification task even with extremely simple decision boundaries.
Figure 4-7: Two classes of engine data shown projected down to two principal components using 6093 features.

Figure 4-8: Distribution of classification performance of individual features with a linear decision boundary.

The engine data classification problem highlights the importance of feature selection when utilizing a large number of features as many features which may be calculated are likely irrelevant to the classification or anomaly detection problem.

**Sparse Methods**

Knowing that many of the features are potentially irrelevant to the challenges of anomaly detection or classification in practical scenarios selection of a small number of features is of high importance. The variance explained by each principal component of the Ford data after feature transformation is shown in Figure 4-10. Using sparse
Figure 4-10: Scree plot showing the variance explained by each principal component of the forod data with 6093 features the red var shows the sum of the rest of the principal components after the 30th component.

(ADD) procedures. The properties of the system of highest interest were the measurements of the acoustic detectors, the bottom hole and well head pressures and temperatures. After segmentation based on the valve positions of the system the measurements associated with one well are visualized by plotting pairwise density plots and smoothed kernel density estimates of the univariate distribution of each of these measurements. When the well is operating in its dominant operating mode, routed to dual completion "low pressure" manifolds with fully open chokes, the distribution of measurement responses is shown in Figure 4-14. The acoustic signal is pre-transformed with by being raised to the (1/2) power.

The distribution of the well head temperature as shown in Figure 4-14 seems to be quite unusual with gaps in the density. These are an actual feature of the data. The well head temperature empirical cumulative distribution function for all the data available, and the dual completion low pressure manifold flow restricted data is shown in Figure B1.

A framework for data driven modeling of systems was developed with a focus on methods for estimation and model building. Evaluation of three critical characteristics of a system based on the available data can be used to inform model building decisions. These three factors are correlation, non-linearity and dynamics. These factors have
principal component analysis a scree plot was created showing the variance explained with 6, 12, 24 and 48 feature per principal component in Figure 4-11.

The sparse linear combination of features in the SPCA calculations show much less explained variance for the original data compared to dense PCA. However as seen in Figure 4-9 a small number of features is all that is necessary to separate the data. The sparse representation does not do a good job at recovering the majority of the features in the dataset but does offer improved separation between the classes compared to the dense PCA as shown in Figure 4-12.

Using SLDA to select features for classification with quickly classifiers with 100% accuracy on the training data can be found as shown in Figure 4-13.

4.3.2 Oil Well Production Data

Exploratory Data Analysis and Summary Statistics

Exploratory data analysis was conducted to evaluate the properties of the data that could have an impact on the selection or design of anomaly detection and diagnosis
been found to successfully guide the selection of modeling techniques based on a match between the assumptions of the method and characteristics of the data.

The correlation of the well data is evaluated between the six measurements discussed in Figure 4-15. The two acoustic signals show strong correlation, but the remainder of the measurements have only weak correlation, with the next strongest between the bottom hole pressure and the well head pressure which is also quite sensible. The lack of strong correlation structure in the primary measurement signals suggests that besides the two acoustic signals the information content of these measurements is likely additive. This will be examined further in the context of multivariate feature space representations.

A heuristic to diagnose the presence of strong nonlinearity in the relationship between the variables is found by constructing multiple regression problems between polynomial transformations of the data. Each of the six measurement variables previously analyzed for correlation were regressed against the remaining 5 measurements, bilinear combination of these measurements and the squared value transformation of these measurements. To the extent that the second order terms were estimated
to be statistically significant provides some validation of the presence of a nonlinear relationship between these variables. The results of this analysis is shown in Table 4.2.

<table>
<thead>
<tr>
<th>Output</th>
<th>Num Sig Bilinear</th>
<th>Num Sig Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ac 1</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Ac 2</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>BHP</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>BHT</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>WHP</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>WHT</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4.2: Number of statistical significant bilinear and quadratic second order terms at the $p \leq 0.001$ level

Dynamics was evaluated by examining the autocorrelation and the cross-autocorrelation of these primary measurements. The autocorrelation of these measurements are shown in Figure 4-16. From the autocorrelation analysis a strong dynamic character of the system was established.

### 4.3.3 Operational Mode Identification By Static Clustering

Wells are expected to naturally have changes in behavior over time. It is believed that some of the non-linear character of the system is best understood as fundamental switches in behavior beyond that of the well routing. The multivariate measurements of the system were statically clustered based on euclidean distance in the dual com-
Figure 4-13: The linear classification performance when trained on all 806 examples considered in the Ford dataset using SLDA with a specified number of features mapped as a linear combination to form a discriminating direction. SLDA was trained with l2 penalty of 1e-6.

pletion to low pressure manifold primary production state to evaluate the natural clusters associated with operating modes in a static context. Because there are naturally occurring extreme values that a clustering procedure searching for dominant operating modes would be overly sensitive too, only the 99.5% of the data most central in each dimension was used. If an observation was in the extreme distributional tails in one or more of the 6 dimensions it was not considered in the static clustering. The clustering is static because the time stamp, or observation number associated with each measurement was ignored. The data was centered, and converted to have unit norm in each dimension before clustering to avoid skewing the clustering results based on larger magnitude measurements. The kmeans clustering algorithm was used with multiple random initializations using squared euclidean distance metric. The clusters generated by kmeans clustering were evaluated using the Calinski-Harabasz and Davies-Bouldin metrics to determine the optimal number of clusters see Figure B2 and Figure B3. In both these metrics the optimal number of clusters was determined to be 4. The corner plot of these four clusters can be seen in the Appendix in Figures
Figure 4-14: Distribution of measurement responses in dual completion low pressure configuration.
Visually examining the clusters found, the most similar static measurements were grouped into clusters that agreed with temporal occurrence of the data. Examining Figure B9 and Figure B10 shows that as the wells behavior changes over time, the wide range of values observed over the lifetime of the well is not informative of anomalies we may wish to identify. This strongly motivates learning decision boundaries with local temporal and measurement character that measurements that are absolutely similar to one portion of process history but locally dissimilar are likely anomalous. In summary developing anomaly detection and diagnosis methods based on extremal observations of the raw data is expected to have poor performance due to the wide range in observed values over the lifetime of the well, but substantially more limited variation in reduced time frames.

**Feature Transformation**

In the provided data for the well reported in this study an unrecoverable loss in production occurred that is attributed to excessive sand production. The portion of the well’s operating history ahead of this loss was treated separately from the
Figure 4-16: Autocorrelation of measurements in Well 1.

remainder of the well’s production history.

Data was segmented into 720 sample (1 day) periods of operating history with three classes of behavior. The near failure period, immediately after bean ups that subsequently entered status 10 (dual completion low pressure production) and putatively nominal dual completion low pressure production. The data was required to be continuous in each mode to be included in this analysis.

Initially each of the six principal measurements (BHP, BHT, AC1, AC2, WHP, WHT) were treated separately and the HCTSA feature set was evaluated on these portions of well operating history. After projecting the data to two principal components these differing operating modes appeared well separated. This can be seen for BHP in Figure 4-17.

From the color coding in Figure 4-17 the separation between near failure and putatively nominal data is not driven by the age of the well. The open circles move further away (in distance on the figure) from the red + symbols as the well matures in this representation, the opposite of the trend of the points in time. There are relatively few data points that are available for near after bean up conditions where the well stably returns to status 10 production.
Figure 4-17: Bottom hole pressure portions of operating history colored time in the well's life from blue to green to yellow to red to dark red. The diamond points are shortly after restarts, the + symbols are near failure data points and the open circles are putatively normal conditions.

The separation of the data from near failure conditions and otherwise nominal conditions is seen in feature space through a projection to two principal components not only in BHP but in most of the measurement variables. The results for acoustic detector 1 are shown in Figure 4-18.

**Classification and Feature Selection**

The classification performance between the post bean up, near failure and otherwise nominal dual completion low pressure production data was evaluated for each of the features present for each of the 6 dimensions treated independently. The best features in the measurement variables have linear classification performance of approximately 80%. The best feature for BHP has classification performance of 94% while the best feature for WHT is only 70% accurate. Because the classes are imbalanced a naive classifier that assigned all points to the largest class (nominal) would achieve an
Figure 4-18: Acoustic detector 1 portions of operating history colored time in the well’s life from blue to green to yellow to red to dark red. The diamond points are shortly after restarts, the + symbols are near failure data points and the open circles are putatively normal conditions.

accuracy of 77.6%. This indicates that many features do not linearly separate the data. Sparse linear discriminant analysis was conducted for multiclass classification for each variable to find two discriminating directions containing a linear combination of a small number of features. The classification error is shown in Figure 4-19.

The data was merged to concatenate the feature vectors for each measurement dimension excluding WHT (the reasoning for this is described in the Use of Multiple Signals Section). The three class linear classification performance when using SLDA with concatenated feature vectors was evaluated for sparsity levels of 1 to 40 features. The SLDA classification performance is shown in Figure 4-20.

Based on the SLDA analysis, anomaly detection algorithms were conducted with the full set of features available using PCA for dimensionality reduction to two dimensions and with SLDA with a sparsity level of 40 and two discriminatory directions. When the data is projected into these sparse discriminator directions the separation...
Figure 4-19: Three class classification performance trained with sparse linear discriminant analysis: a) acoustic detector 1, b) acoustic detector 2, c) bottom hole pressure, d) bottom hole temperature, e) well head pressure, f) well head temperature.
Figure 4-20: Three class classification performance trained with Sparse Linear Discriminant Analysis with merged HCTSA features in acoustic 1 and 2, bottom hole temperature and pressure, and well head pressure.

between the classes becomes very clear. This is shown in Figure 4-21.

4.3.4 Parzen Density Estimation

Parzen density estimation on the putatively nominal data was conducted utilizing all six of the measurements after projection to two principal components. The outlier fraction was selected to be 3%. The resulting decision boundaries are shown in Figure 4-22.

Anomalous Portions of Operating History

Anomalous portions of operating history were identified as being outside the decision boundary from the Parzen density estimator. In each of 6 measurement signals approximately 10 days of operating history were identified as being outliers. The outliers selected by each feature were aggregated to evaluate the shared selection as an outlier by multiple measurement features. A histogram showing the number of
dimensions in which a day long portion of operating history was found to be outlying is shown in Figure 4-23.

Several days of the well's operating history were identified as anomalous independently by multiple measurements. The most frequently voted as anomalous portions of operating history were March 4, March 5 and March 6th. These measurements are shown for March 4 in Figure 4-24, and March 6 in Figure 4-25 respectively. The March 4th and March 6th data show a sharp transition in the acoustic measurements and pressure data. This is not justified by the valve movements used to partition the data. This change in measurement is perhaps justified as a step change in another characteristic of the system. The March 6th corresponds to a period of operating history when the gas lift was engaged. This disturbance to the system correlates strongly with the change in bottom hole pressure as shown in Figure 4-26. The March 5th data doesn't show this step change behavior however it shows rapid short duration fluctuations in the acoustic signal. The March 4th data also correlates with sharp changes in gas lift application.
Use of Multiple Signals

The results for Parzen density estimation shown above show a voting procedure to aggregate the information of multiple features.

An alternative approach is to combine the data from multiple measurements into a single large feature vector as described in the methods section. The results of a 3% Parzen density estimator using the five (non WHT) feature vectors and applying PCA for dimensionality reduction are shown in Figure 4-27. WHT was excluded because a smaller number of features were effectively calculated on this variable after applying the recommended post processing steps in the HCTSA software. The data handling process would be overly restrictive reducing the number of available data points.

The 9 days of process operating history identified as anomalous by this method are: '19-Feb-2010', '03-Mar-2010', '04-Mar-2010', '05-Mar-2010', '17-Mar-2010', '05-Apr-2010', '18-Apr-2010', '03-Jun-2010', and '28-Jul-2010'. In comparison with the outlier voting histogram the early March data points are represented in both but not March 6th. The 5th of April data point had a voting score of 2, the 18th or April had a voting score of 1, the other points were not identified as outlying in any of the dimensions when considered independently. Both strategies for incorporation of multivariate data identified that the early March data points were quantitatively different in the high dimensional features space.

Parzen density estimation was also conducted in the SLDA discriminator dimensions from Figure 4-21. The boundary at 3% outlier rejection is shown in Figure 4-27.

The 9 days of process operating history identified as anomalous by this method are: '18-Dec-2011', '19-Feb-2010', '04-Mar-2010', '18-Apr-2010', '28-Apr-2010', '14-May-2010', '03-Jun-2010', '17-Jun-2010', and '20-Jun-2010'. In comparison with the previously identified outlying points early March is again identified however there is increased heterogeneity compared to the voting or PCA based methods.

The SLDA selection of dimensions for density based boundaries here show some qualitatively desirable features. The majority of points are tightly clustered near the
center of the data. The outlying points are clearly separated from the majority of the data. Some of the outlying points have local neighbors which prevent them from being identified as outlying or anomalous. These points can be visually identified because of the projection down to a sufficiently low number of dimensions. In a higher number of dimensions this would not have been possible.

4.4 Discussion

The methodology introduced in this work to transform the data from multivariate time series in to a feature space representation is well motivated by problems in time series classification. Here this technique has been utilized in an anomaly detection context. In this work we apply density based outlier detection methods to find anomalous portions of an oil well’s operating history. The selection of a boundary for one class classification was introduced based on an expected fraction of outliers in the data. This outlier fraction was introduced manually and could not be tested in the oil well dataset. In other situations the classification performance with labeled outliers could be utilized.

There are a few difficulties with using the classification approach for feature selection being directly applied to the one-class modeling context. Among these are the problem that the anomaly class is un-sampled or severely undersampled. Working to conduct feature selection on the basis of classification is the methodology proposed in this work but runs a risk of the method not clearly being able to distinguish between groups that may be present in a future test set but not available at training time.

In the novelty detection or outlier detection contexts for anomaly detection and diagnosis, it is especially difficult to apply this sort of method. Without existing labels, in the multi-class context, the mutual information strategy is not effective. An example of the challenge of feature selection for one class modeling methods is the case where nominal operating conditions might be modeled with a constant or nearly constant feature which might be assumed to be not useful for describing the variation in the nominal data but would have a step change difference in a faulty or
anomalous condition.

Working in a large dimensional feature space several features may address very similar fundamental characteristics of the data. For instance if in a classification problem the autocorrelation is a distinguishing attribute of the data features from time series modeling fits may be equally informative of the same underlying fundamental difference. In the HCTSA feature list autoregressive models of various orders are utilized and the resulting best fit model parameters are likely to be similarly informative.

The intelligent selection of features is critical to the success of having a high performing decision boundary with the minimum number of dimensions. In a filtering approach that doesn’t consider combinations of features, two features which are nearly identical might be simultaneously selected without offering any increase in classification efficiency.

It remains a challenge to assess the performance of the anomaly detection strategy introduced in this work and its sensitivity to design factors such as the length of operating history segments. In order to make effective comparisons in this work it was necessary to segment the data into continuous operating history. In reality the most complex portions of data represent situations when the system is dynamically receiving inputs. In this work only the largest input free portions of operating history are addressed. Extensions of this work to address situations when the well is being actively manipulated are under investigation.

Human interpretation of the data points identified as anomalous via the voting procedure show strong candidates as being truly representative of anomalous system behavior. Acquiring detailed human labels of an entire well life cycle is an important next step in bringing this strategy closer to practical deployment.

4.5 Conclusions

This work has introduced the characteristics of on-line measurement data from production oil and gas wells. The particular attributes of the data most important to
building anomaly detection, fault detection, and prognostic health monitoring applications have been principally discussed. A new method for identifying abnormal conditions as evident in the acoustic detectors and other measurements has been identified. In order to be able to identify portions of operating history when anomalous events are suspected of occurring and separate these events from nominal conditions a transformation of measured signals into a feature space was employed. The particular features found to be most strongly capable of separating labeled portions of the wells operating history were utilized to build a one-class classifier. The one-class classifier performance was evaluated by examining the data points selected as anomalous under varying modeling decisions.

The data collected from the oil well under study showed a complex range of behaviors including multimodality and nonlinear dynamics. By taking a data based approach we have demonstrated human acceptable performance at the industrially relevant task of identifying anomalous periods of oil well operation for the first time.
Figure 4-22: Boundaries for anomaly detection trained to reject 3% of putatively nominal data. Data was the first two principal components projected from HCTSA features. a) acoustic detector 1, b) acoustic detector 2, c) bottom hole pressure, d) bottom hole temperature, e) well head pressure, f) well head temperature.
Figure 4-23: Number of dimensions each in which each day long portion of operating history was found to be anomalous at a 3% outlier selection rate by a Parzen density estimator in two principal component reduction of a HCTSA feature space.

Figure 4-24: Anomalous behavior identified on March 4th.
Figure 4-25: Anomalous behavior identified on March 6th.

Figure 4-26: Anomalous behavior identified on March 6th.
Figure 4-27: Boundaries for anomaly detection trained to reject 3% of putatively nominal data with a Parzen density estimator. a) Data was the first two principal components projected from HCTSA features for 5 measurements (not using WHT). b) Data was the two discriminatory directions found via SLDA with a sparsity level of 40 components projected from HCTSA features for 5 measurements (not using WHT).
Appendix A

Appendix for Uncertainty Analysis in Spectroscopic Chemometrics: Regularization, Sequential Analysis, and Bayesian Estimation with a Detailed Model
Table A.1: The properties of 43 semiconducting nanotube species that were used to simulate spectra. From Rocha et al., 2011 [76].

<table>
<thead>
<tr>
<th>n</th>
<th>m</th>
<th>v11 (cm(^{-1}))</th>
<th>Δv11</th>
<th>v22 (cm(^{-1}))</th>
<th>Δv22</th>
<th>μ</th>
<th>(a_6^{785})</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>4</td>
<td>11465</td>
<td>197</td>
<td>17294</td>
<td>618</td>
<td>0.7</td>
<td>0.36</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>10966</td>
<td>140</td>
<td>13960</td>
<td>267</td>
<td>0.68</td>
<td>0.22</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>10498</td>
<td>198</td>
<td>15085</td>
<td>679</td>
<td>0.5</td>
<td>0.15</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>10232</td>
<td>203</td>
<td>17613</td>
<td>271</td>
<td>0.59</td>
<td>0.05</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>10075</td>
<td>262</td>
<td>20014</td>
<td>270</td>
<td>0.63</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>9762</td>
<td>202</td>
<td>15516</td>
<td>1029</td>
<td>0.85</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>9612</td>
<td>235</td>
<td>21045</td>
<td>1036</td>
<td>0.93</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>9476</td>
<td>217</td>
<td>13583</td>
<td>183</td>
<td>0.63</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>9070</td>
<td>159</td>
<td>13751</td>
<td>597</td>
<td>0.9</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>8963</td>
<td>143</td>
<td>17062</td>
<td>305</td>
<td>0.49</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>8897</td>
<td>200</td>
<td>15446</td>
<td>414</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>8758</td>
<td>200</td>
<td>17484</td>
<td>1782</td>
<td>0.66</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>8537</td>
<td>94</td>
<td>12622</td>
<td>263</td>
<td>0.32</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>8506</td>
<td>268</td>
<td>14054</td>
<td>532</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>8338</td>
<td>147</td>
<td>12598</td>
<td>797</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>8040</td>
<td>162</td>
<td>14809</td>
<td>598</td>
<td>0.96</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>8002</td>
<td>112</td>
<td>15629</td>
<td>636</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>7988</td>
<td>149</td>
<td>12741</td>
<td>217</td>
<td>0.86</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>7886</td>
<td>185</td>
<td>16241</td>
<td>704</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>7858</td>
<td>162</td>
<td>13708</td>
<td>337</td>
<td>0.96</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>7739</td>
<td>193</td>
<td>11667</td>
<td>186</td>
<td>0.59</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>7626</td>
<td>189</td>
<td>11647</td>
<td>216</td>
<td>0.56</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>7551</td>
<td>124</td>
<td>12678</td>
<td>520</td>
<td>0.81</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>7452</td>
<td>230</td>
<td>11628</td>
<td>290</td>
<td>0.38</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>7292</td>
<td>195</td>
<td>13916</td>
<td>529</td>
<td>0.32</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>7266</td>
<td>137</td>
<td>14677</td>
<td>629</td>
<td>0.47</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>7251</td>
<td>190</td>
<td>13262</td>
<td>356</td>
<td>0.48</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>7166</td>
<td>155</td>
<td>11726</td>
<td>257</td>
<td>0.54</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>7074</td>
<td>174</td>
<td>12409</td>
<td>429</td>
<td>0.54</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>7026</td>
<td>206</td>
<td>10862</td>
<td>341</td>
<td>0.76</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>6909</td>
<td>215</td>
<td>10796</td>
<td>222</td>
<td>0.78</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>6780</td>
<td>197</td>
<td>11478</td>
<td>320</td>
<td>0.77</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>6728</td>
<td>200</td>
<td>10841</td>
<td>271</td>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>6672</td>
<td>144</td>
<td>12652</td>
<td>343</td>
<td>0.46</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>6676</td>
<td>187</td>
<td>13101</td>
<td>477</td>
<td>0.64</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>6658</td>
<td>197</td>
<td>13330</td>
<td>534</td>
<td>0.67</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>6585</td>
<td>158</td>
<td>11947</td>
<td>432</td>
<td>0.46</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
<td>6479</td>
<td>207</td>
<td>10821</td>
<td>243</td>
<td>0.75</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>6425</td>
<td>228</td>
<td>11319</td>
<td>371</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>6191</td>
<td>208</td>
<td>10465</td>
<td>179</td>
<td>0.54</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>6164</td>
<td>189</td>
<td>12131</td>
<td>378</td>
<td>0.65</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>6159</td>
<td>191</td>
<td>12236</td>
<td>409</td>
<td>0.69</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>6163</td>
<td>205</td>
<td>11845</td>
<td>295</td>
<td>0.7</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure A1: SVD spectra with discretization
Appendix B

Oil Well ADD Appendix
Figure B1: Empirical cumulative distribution function of the wht showing gaps in the data.

Figure B2: Evaluation of the optimal number of clusters using the Calinski-Harabasz clustering evaluation criterion. Higher is better.
Figure B3: Evaluation of the optimal number of clusters using the Davies-Bouldin clustering evaluation criterion. Lower is better.

Figure B4: Distribution of data points in cluster 1 of 4 found by kmeans for LPLP data.
Figure B5: Distribution of data points in cluster 2 of 4 found by kmeans for LPLP data.
Figure B6: Distribution of data points in cluster 3 of 4 found by kmeans for LPLP data.
Figure B7: Distribution of data points in cluster 4 of 4 found by kmeans for LPLP data.
Figure B8: Four clusters identified by color in the LPLP data
Figure B9: Static data colored by time point
Figure B10: Cluster identity of points through time
Bibliography


[40] T B Green, S L King, and L Li. Apparatus and method for reducing solvent loss for electro-spinning of fine fibers.


