CHEMICAL IDENTIFICATION UNDER A POISSON MODEL FOR RAMAN SPECTROSCOPY

A Thesis
Presented to
The Academic Faculty

by

Ryan D. Palkki

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2011

CHEMICAL IDENTIFICATION UNDER A POISSON MODEL FOR RAMAN SPECTROSCOPY

Approved by:

Dr. Aaron D. Lanterman,
Committee Chair
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. David S. Citrin
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. W. Dale Blair
Georgia Tech Research Institute
Georgia Institute of Technology

Dr. Brani Vidakovic
Department of Biomedical Engineering
Georgia Institute of Technology

Dr. Steven W. McLaughlin
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: November 2, 2011

This dissertation is dedicated to my wife, Brenna, and to my parents, Dennis and Marilyn Palkki. Thank you for all you have done to encourage my continuing education.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my wife, Brenna. She moved to Atlanta with me so that I could enjoy a Georgia Tech education, and along the way she has made many sacrifices that have made this thesis possible. I thank her for her endless support, love, and encouragement.

I am indebted to my advisor, Dr. Aaron Lanterman, for the guidance he has given me throughout my studies. I thank him for making sure that I was always funded, and for all of his advice regarding my coursework, research, and career planning. He contributed many ideas to this thesis, and I would not have made it this far without his help.

I would also like to thank my colleagues and supervisors at GTRI and on the Chem/Bio Benchmark Team for their support and guidance. Dale Blair, Phil West, Mahendra Mallick, Andy Register, Barry Drake, Darren Emge, and Randy Paffenroth offered invaluable feedback that helped shape this research. The work in this thesis was funded in part by the ONR grant N00014-07-1-0378 and in part by the Air Force HC1047-05-D-4000-0102 DTRA Test Demonstration.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Raman spectroscopy provides a powerful means of chemical identification in a variety of fields, partly because of its non-contact nature and the speed at which measurements can be taken. The development of powerful, inexpensive lasers and sensitive charge-coupled device (CCD) detectors has led to widespread use of commercial and scientific Raman systems. However, relatively little work has been done developing physics-based probabilistic models for Raman measurement systems and crafting inference algorithms within the framework of statistical estimation and detection theory.

The objective of this thesis is to develop algorithms and performance bounds for the identification of chemicals from their Raman spectra. This involves the following thrusts:

- *Measurement Model:* A Poisson measurement model based on the physics of a dispersive Raman device is presented. Placing a statistical model on parameters and data allows one to draw on information theory to quantify how much information needed for the required task is available in the data provided by the sensor.

- *Parameter Estimation:* The problem is expressed as one of deterministic parameter estimation, and several methods are analyzed for computing the maximum-likelihood (ML) estimates of the mixing coefficients under our data model. The performance of these algorithms is compared against the Cramér-Rao lower bound (CRLB). The non-negative iteratively reweighted least squares (NNIRLS) algorithm is seen to give performance that is nearly identical to the more computationally demanding expectation-maximization approach.

- *Target Detection:* The Raman detection problem is formulated as one of multiple hypothesis detection (MHD), and an approximation to the optimal decision rule is presented. The resulting approximations are related to the minimum description length

(MDL) approach to inference. In our simulations, this method is seen to outperform two common general detection approaches, the spectral unmixing approach and the generalized likelihood ratio test (GLRT). The MHD framework is applied naturally to both the detection of individual target chemicals and to the detection of chemicals from a given class.

- *Accounting for Unknown Chemicals:* The common, yet vexing, scenario is considered in which chemicals are present that are not in the known reference library. A novel variation of nonnegative matrix factorization (NMF) is developed to address this problem. Our simulations indicate that this algorithm gives better estimation performance than the standard two-stage NMF approach and the fully supervised approach when there are chemicals present that are not in the library.

- *Dealing with Library Error:* Estimation algorithms are developed that take into account errors that may be present in the reference library. In particular, an algorithm is presented for ML estimation under a Poisson errors-in-variables (EIV) model. It is shown that this same basic approach can also be applied to the nonnegative total least squares (NNTLS) problem.

Most of the techniques developed in this thesis are applicable to other problems in which an object is to be identified by comparing some measurement of it to a library of known constituent signatures.

CHAPTER I

INTRODUCTION

1.1 The Raman Effect

Raman spectroscopy is the study of the inelastic scattering of photons by molecules. Quantum mechanics requires that only certain vibrational modes are allowed in a molecule. When a photon hits a molecule, it can interact with the bonds in the molecule, causing it to change vibrational modes. Consider the simplified example of Figure 1. The frequency of the incident light is ν_0 and its energy is $h\nu_0$ (where h is Planck's constant). Upon striking the chemical sample, some of the photons induce changes in the vibrations of the molecules, causing a jump to a higher, excited energy state. The light loses energy in the process, and the scattered light thus has a lower frequency than the incident light. This is known as Stokes scattering.[1] The intensity of the scattered light is plotted vs. Raman (Stokes) shift in Figure 2. The measured Raman spectrum can be used as a "fingerprint" by which to identify the chemical. The spectrum of andradite, obtained from [14], is shown as an example in Figure 3.

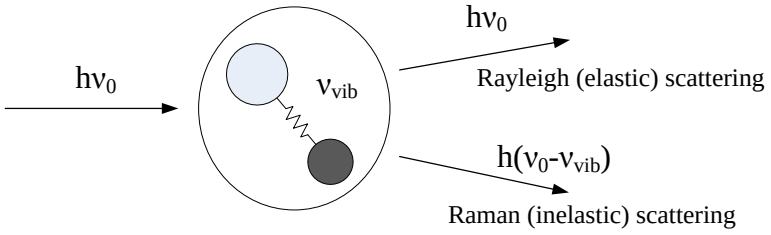


Figure 1: The Raman effect.

[1] Some photons cause the molecules to fall to a *lower* vibrational state, in which case the scattered light has a *higher* frequency than the incident light. This is known as Anti-Stokes scattering. Since there are initially many more molecules in the ground state than in the excited state, the intensity of the Anti-Stokes scattering is much lower than the intensity of the Stokes scattering. For this reason, the Anti-Stokes bands are rarely used in the interpretation of Raman spectra, and we will ignore the Anti-Stokes effect in this thesis.

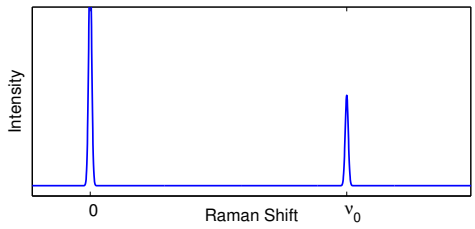


Figure 2: Raman spectrum for the diatomic molecule of Figure 1.

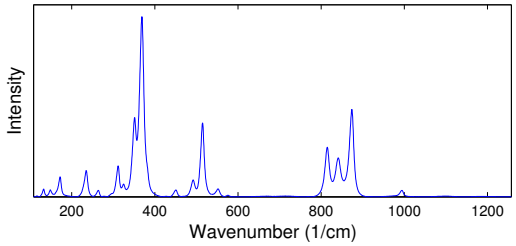


Figure 3: Raman spectrum of andradite.

1.2 History and Development of Raman Instrumentation

The first experimental verification of the effect was performed by Sir C. V. Raman [72] in 1928.[2] Raman let "white" sunlight pass through a narrowband photographic filter, and the resulting near-monochromatic light then passed through a liquid sample. The inelastic (Raman) scattering is extremely weak compared to the elastic (Rayleigh) scattering (see Table 1, from [57]), making it difficult to detect. Thus, to reduce the Rayleigh light, a second optical filter was placed before the detector. In his first experiments, the "detector" was his eyes; this was soon replaced with photographic plates. While this initial setup was quite simple, it demonstrates several key components of any Raman measurement system:

1. Monochromatic light source.

2. Filter to reduce strong Rayleigh-scattered light.

3. Photodetector.

Over the years, these three components have undergone many stages of advancement in an effort to overcome the serious problem of detecting the Raman photons in the presence of the

[2] In 1930, Raman was awarded the Nobel Prize in Physics for this discovery.

strong Rayleigh light (see [18, 69]). In particular, the development of powerful, inexpensive lasers and sensitive charge-coupled device (CCD) detectors has led to widespread use of commercial and scientific Raman systems. In addition, advances in computing technology and the availability of inexpensive, powerful personal computers have enabled automatic data processing of the Raman spectra.

Table 1: Relative intensities (approximate).

Incident light	Rayleigh Scattering	Raman Scattering
1	10^{-6}	10^{-10}

While the excitation source is typically in the visible or UV range, the Raman *shifts* are in the IR region and correspond to the same vibrational shifts that are studied in IR spectroscopy. The spectra generated from these two methods arise from different physical phenomena, however: IR spectroscopy relies on an absorption process while Raman spectroscopy uses a scattering process, and the two effects are governed by different selection rules. Because of this, a noticeable peak in one type of spectrum may be weak or missing in the other. Thus, Raman and IR spectra are complementary and often used together in chemical classification.

1.3 Application: Detecting Surface Contamination with the Joint Contaminated Surface Detector

Current US military reconnaissance systems use a double-wheel sampling system and a mass spectrometer to detect surface contamination. This system is rather sensitive and can detect small quantities of chemical and biological agents, but it is also slow (walking speed) and disrupts operational tempo. It also requires substantial user interaction. The Joint Contaminated Surface Detector (JCSD) has recently been developed in an effort to make the reconnaissance systems faster, safer, and more automated [85].

The JCSD is a small Raman sensor that is mounted in a traditional reconnaissance vehicle.[3] In the typical configuration, it sits about a meter above the ground and points

[3]Unmanned robots are also being developed to bring the JCSD into rough terrain such as buildings, tunnels, or caves.

downward, shooting laser pulses 25 times per second as the vehicle moves along. It has been shown to be successful at speeds of up to 45 miles per hour. An analysis computer performs the detection/classification by comparing the measured Raman spectra with a "reference library" of known signatures [28, 70].

There are several challenges inherent in this detection problem. First, each measured spectrum contains noise introduced by the sensor. The signal-to-noise ratio (SNR) is typically low, due partly to the fact that the laser can destroy the target chemical if it is too powerful. Also, multiple chemicals may be present, in varying quantities, in a single shot of data. The spectra in the reference library are often highly correlated, and there may be substances present that are not in the library. These are some of the fundamental problems that we have addressed in our research.

1.4 Thesis Organization

This thesis is organized as follows. In Chapter 2, a physics-based probabilistic model for a dispersive Raman device is presented. Several approaches to estimating the mixing coefficients of the target spectra are discussed, and the results of the various algorithms are compared with the Cramér-Rao Lower Bound (CRLB). Chapter 3 addresses the *detection* problem and outlines several of the challenges and drawbacks of two of the common general detection approaches. The Raman detection problem is formulated as one of multiple hypothesis detection, and an approximation to the optimal decision rule is presented. Chapter 4 considers the scenario in which chemicals are present that are not in the known reference library, and presents a novel variation of nonnegative matrix factorization to address this problem. Chapter 5 expands on the work of Chapter 2 by taking into account the error present in the reference library. The contributions and conclusions of our research, as well as several potential avenues for future work, are summarized in Chapter 6.

To understand individual chapters of this thesis, it is not essential to read them in order. Figure 4 shows the conceptual dependencies of Chapters 2 through 5 and the appendices.

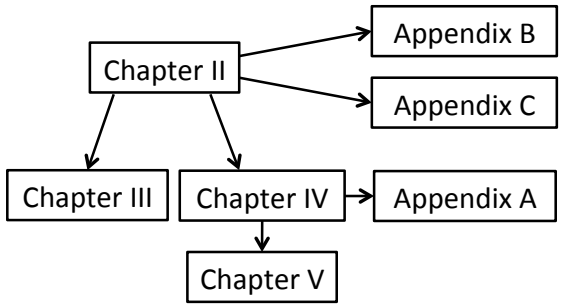


Figure 4: Conceptual dependencies of the chapters and appendices.

CHAPTER II

CHEMICAL MIXTURE ESTIMATION UNDER A POISSON RAMAN SPECTROSCOPY MODEL

2.1 Introduction

Given an $N \times M$ reference library $\mathbf{A}$ of known spectra $\{\mathbf{a}_1, \cdots, \mathbf{a}_M\}$ and a measured spectrum $\mathbf{y} \in \Re^N$, perhaps the simplest way to estimate the relative quantities of each of the M components is to use the least-squares (LS) approach:

$$\hat{\mathbf{x}}^{LS} = \arg\min_{\mathbf{x}} ||\mathbf{A}\mathbf{x} - \mathbf{y}||^2 , \tag{1}$$

where $\mathbf{x} = (x_1, \cdots, x_M)$ is the unknown vector of mixing coefficients. The solution to (1) is well known [89]:

$$\hat{\mathbf{x}}^{LS} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} . \tag{2}$$

However, the LS approach does not take into account the fact that none of the mixing coefficients can be negative. Therefore, a nonnegativity constraint is usually applied to the minimization problem, yielding

$$\hat{\mathbf{x}}^{NNLS} = \arg\min_{\mathbf{x}} ||\mathbf{A}\mathbf{x} - \mathbf{y}||^2 \text{ subject to } x_j \geq 0 \text{ for } 1 \leq j \leq M . \tag{3}$$

Note that the solution to (3) is typically not the same as the ad hoc approach of simply solving (1) and setting the negative results to zero. The standard nonnegative least squares (NNLS) algorithm is due to Lawson and Hanson [44], but other algorithms have also been employed on Raman data [54, 85] for faster computational performance.

The estimated mixing coefficients are typically each compared to an ad hoc threshold to decide if the corresponding chemical is present [85]. More details about this and other detection approaches are the subject of Chapter 3.

Note that these least-squares approaches assume linear mixing of the spectra. The assumption of linearity is made in most of the literature, and in this thesis.

2.1.1 Noise Model

If a statistical estimation viewpoint is taken, the least-squares methods of Section 2.1 implicitly assume an additive white Gaussian noise (AWGN) model. If the data is generated by

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}, \tag{4}$$

where $\mathbf{v}$ is white Gaussian noise, then the least-squares approach is justified because it yields the maximum-likelihood estimate for this particular model [35]. Many popular processing algorithms, such as the subspace version of the Generalized Likelihood Ratio Test [84, 55], make this Gaussian noise assumption.

Visual inspection of real Raman spectra indicates that the noise, if thought of as additive, tends to be more positive than negative and that the Gaussian distribution, being symmetric, is not appropriate. Based on this observation, asymmetric distributions such as the Gumbel and Chi-square distributions have been proposed [82, 93, 34]. However, relatively little work has been done on developing probabilistic models based on the underlying physics of the instrumentation.

2.1.2 Organization of this Chapter

This chapter is organized as follows. Section 2.2 discusses several key components of a dispersive Raman device and presents a new mathematical model. The modified Richardson-Lucy algorithm, which computes maximum-likelihood estimates under the model, is presented in Section 2.3. Section 2.4 derives the Cramér-Rao lower bound (CRLB) for our model. The weighted least-squares (WLS) approach is presented in Section 2.5, and its properties (under the assumption of known variances) are explored. Section 2.6 discusses several methods to estimate the weights from the data and presents simulation results for the algorithms.

2.2 Measurement Model

The two basic classes of Raman measurement systems are interferometer-based Fourier transform systems and traditional dispersive devices. This section models a basic dispersive instrument. In particular, as explained in Section 1.2, this work is motivated by the Joint Contaminated Surface Detector (JCSD), a portable Raman instrument used for standoff detection and identification of surface-deposited chemical agents [84].

2.2.1 Key Components of a Dispersive Raman System

Several key components of the measurement system are shown in Figure 5. The Raman-scattered light is collected with a telescope and directed to a diffraction grating spectrograph. The spectrograph is similar to a prism, separating the light by wavelength. The different frequencies of light land on different columns of the charge-coupled device (CCD) detector.

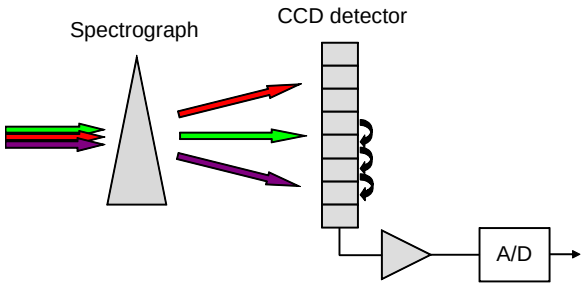


Figure 5: Key components of a dispersive Raman measurement system.

The CCD [33, 90] is an array of photodetectors. During the collection time interval of the device, the photons hitting each element of the array are converted to electrons. After the collection time is over, the charge is transferred down to a shift register, which in turn transfers the charge to an on-chip output amplifier. The resulting output is then digitized by an analog-to-digital converter. In this thesis, the A/D converter quantization effects are ignored, and the characteristics of the A/D converter are assumed to be known such that the photoelectron count may be extracted from the digitized data.

2.2.2 Snyder's Model

Much work has previously been done by the astronomical image restoration community to model the data obtained by CCD cameras. Similar to the dispersive Raman instrument of Figure 5, these imaging systems feature an optical system preceding a CCD detector. We adapt the mathematical model of Snyder et al. [88] to describe the data collected by each frequency bin in our Raman device:

$$y_i = n_i^s + n_i^b + g_i, \qquad i = 1, \ldots, N, \tag{5}$$

where n_i^s is the number of photoelectrons generated on the i^{th} bin due to the signal, n_i^b is the number of photoelectrons generated on the i^{th} bin due to the background counts, and g_i is Gaussian readout noise from the amplifier. These three random variables are assumed to be statistically independent of one another and of their values in other bins. The mean m and standard deviation σ of the readout noise are assumed to be known. Note that g_i is characterized by a continuous random variable, whereas n_i^s and n_i^b are described by discrete random variables. The number of photoconversions that occur during the collection time interval is Poisson-distributed:

$$n_i^s \sim Pois(\mu_i^s) \tag{6a}$$

$$n_i^b \sim Pois(\mu_i^b), \tag{6b}$$

where μ_i^s and μ_i^b are the mean values corresponding to the "ideal noiseless signals." The background counts n_i^b are primarily thermoelectrons generated because of heat ("dark noise"), but also arise from other noise sources such as internal background radiation and bias. Each of these individual components of the background noise is Poisson-distributed and each is independent of each other. Thus, their sum, the "total background counts," is also Poisson-distributed as indicated by (6b). The values of μ_i^b are assumed to be known through a calibration measurement.

If the spectrograph were perfect and the CCD had 100% efficiency, then (6a) would become

$$n_i^s \sim Pois(s_i), \tag{7}$$

where $\mathbf{s} = (s_1, \ldots, s_N)$ is the "true" underlying spectrum. Even if no extra noise is introduced by the instrumentation, the measurements will vary because of the Poisson nature of the photoconversions. This "shot noise" is one of the most fundamental noise sources of the measurement system.

If we still assume that the CCD is perfectly efficient but now allow the "true" signal to be distorted by the diffraction grating, then μ_i^s in (6a) is given by

$$\mu_i^s = \sum_{\tau=1}^{N} p_{i|\tau} s_\tau, \tag{8}$$

where p is the point-spread function (PSF) of the diffraction grating, which we assume is normalized such that it sums to one. The PSF can be thought of as a conditional probability density, or as a "blurring" function. For the special case in which it is shift invariant, (8) becomes a convolution sum, and p is then the impulse response of the spectrograph. In matrix form, (8) is given by

$$\boldsymbol{\mu}^s = \mathbf{P}\mathbf{s}. \tag{9}$$

The PSF is assumed to be known, either through theoretical prediction or through measurements of it [86].

In reality, the CCD is not perfectly efficient; not all of the photons that hit the CCD are converted to electrons. The detector efficiency, commonly known as the "quantum efficiency" or "flat-field response," is represented by the term β_i:

$$\mu_i^s = \beta_i [\mathbf{P}\mathbf{s}]_i. \tag{10}$$

The flat-field response generally varies along the array. It is also assumed to be known through calibration measurements.

The fact that β_i and $\mathbf{P}$ appear in the mean of the Poisson distribution is a consequence of the "binomial selection theorem" as described in [41]. To illustrate this, we here consider only the quantum efficiency; the derivation for the PSF is quite similar and is omitted here for brevity. Once the PSF has been accounted for via (9), the number of photons incident on the ith bin of the detector is distributed by

$$n_i^p \sim Pois([\mathbf{P}\mathbf{s}]_i). \tag{11}$$

Suppose that for a given detector, each photon incident on the i^{th} bin has a probability β_i of producing an electron. Then the conditional distribution of n_i^s (the number of photoelectrons) given n_i^p (the number of incident photons) is binomial. The marginal distribution of n_i^s is then given by

$$p(n_i^s) = \sum_{n_i^p=n_i^s}^{\infty} p(n_i^s|n_i^p)p(n_i^p) \tag{12a}$$

$$= \sum_{n_i^p=n_i^s}^{\infty} \binom{n_i^p}{n_i^s} \beta_i^{n_i^s}(1-\beta_i)^{n_i^p-n_i^s} \frac{([\mathbf{Ps}]_i)^{n_i^p} e^{-[\mathbf{Ps}]_i}}{n_i^p!}. \tag{12b}$$

This can be manipulated to obtain

$$p(n_i^s) = \frac{(\beta_i[\mathbf{Ps}]_i)^{n_i^s} e^{-\beta_i[\mathbf{Ps}]_i}}{n_i^s!}, \tag{13}$$

so that

$$n_i^s \sim Pois(\beta_i[\mathbf{Ps}]_i), \tag{14}$$

which verifies (10). A similar derivation can be performed for (9).

Finally, $\mathbf{s}$ is given by

$$\mathbf{s} = \mathbf{Ax}, \tag{15}$$

where, as in Section 2.1, $\mathbf{A}$ is the reference library of known spectra and $\mathbf{x}$ is the unknown vector of mixing coefficients.[1] Thus, we have

$$\mu_i^s = \beta_i(\mathbf{PAx})_i = \beta_i(\mathbf{Bx})_i, \tag{16}$$

where

$$\mathbf{B} \triangleq \mathbf{PA}. \tag{17}$$

Our objective is to estimate the mixing coefficients $\mathbf{x}$ given the measured spectrum $\mathbf{y}$ and given the parameters $\mathbf{B}$, $\boldsymbol{\beta}$, $\boldsymbol{\mu}^b$, m, and σ.

To illustrate our sensor model, Figure 6 shows the clean spectrum of andradite (obtained from [14]) and Figure 7 shows a simulated noisy spectrum that was generated by passing the clean spectrum through our model.

[1] In general, mapping values of the elements of $\mathbf{x}$ to physical units can be difficult. Although the normalization conventions of the library $\mathbf{A}$ can be easily addressed, issues such as the exposure time, inclination angle of the sensor relative to the tested material, atmospheric transmission losses, and pulse-to-pulse variations in laser strength—all of which combine to define a potentially time-dependent scalar mapping for each chemical—can become complicated. These issues lie outside the scope of this thesis.

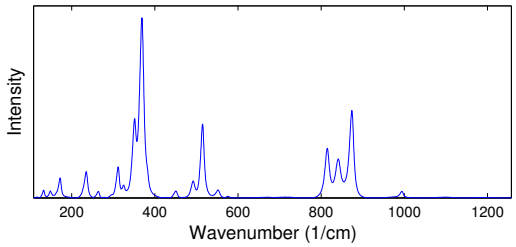


Figure 6: "True" spectrum of andradite, obtained from [14].

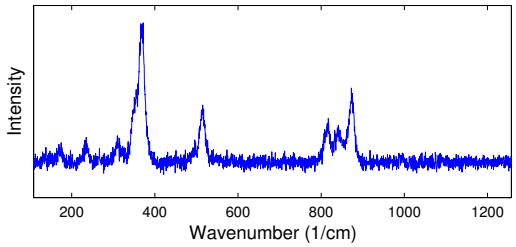


Figure 7: Simulated noisy spectrum of andradite for case of $m = 0$, $\sigma = 15$, $\beta_i = 1$ for all i, and $\mu_i^b = 300$ for all i.

We note that the randomness in our measurements is not all modeled as additive noise. A consequence of this is that the signal and noise are coupled—it is impossible to increase the signal without also increasing the noise. Consider the shot noise described in (6a): if the mean of n_i^s is increased, its variance is also increased, since the mean and variance are equal. However, the signal-to-noise ratio (SNR) of n_i^s still increases with μ_i^s:

$$SNR_{n_i^s} \triangleq \frac{\mathrm{E}[n_i^s]}{\sqrt{\mathrm{Var}(n_i^s)}} = \frac{\mu_i^s}{\sqrt{\mu_i^s}} = \sqrt{\mu_i^s}. \tag{18}$$

Similarly, the SNR for the data y_i is given by

$$SNR_{y_i} \triangleq \frac{\mathrm{E}[\text{signal component of } y_i]}{\sqrt{\mathrm{Var}(y_i)}} = \frac{\mu_i^s}{\sqrt{\mu_i^s + \mu_i^b + \sigma^2}}. \tag{19}$$

Figure 8 plots the SNR vs. the signal level (μ_i^s) for two different cases. The top red line represents the ideal case in which no noise is introduced by the detector. In this shot noise limited case, $\mu_i^b = 0$ and $\sigma^2 = 0$, and the SNR increases with the square root of the signal. The bottom blue curve represents the case in which $\mu_i^b = 50$ and $\sigma^2 = 4$. Because the signal shot noise increases with μ_i^b while the other noise sources remain constant, the shot noise is the dominant noise effect when the signal is powerful.

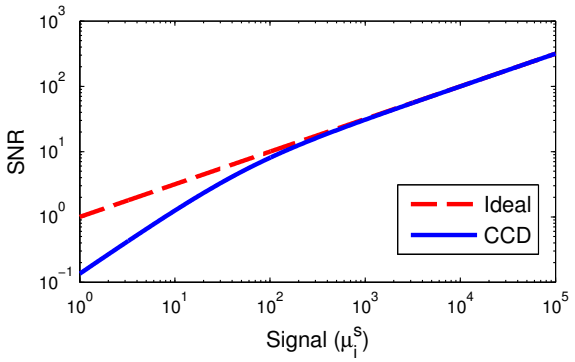


Figure 8: SNR for a CCD pixel.

We note that the SNR does not by itself characterize the difficulty in our estimation problem. Roughly speaking, there are two main sources of difficulty in estimating $\mathbf{x}$ under our model: there is noise, and there is confusion between library elements. Even if the SNR is high, deciding which chemicals are present may still be difficult if the library elements are highly correlated with each other. On the other hand, even if there is only one chemical in our library (and thus no confusion between library elements), detection may still be difficult if the SNR is sufficiently low. The Cramér-Rao lower bound (CRLB) derived in Section 2.4 is a preferable metric to gauge the difficulty of the problem, as it takes both of these sources of difficulty into account from within the information space formed by the sensor model.

To keep the notation simple, this discussion treated the CCD as a $1 \times N$ linear array. In reality, it is generally an $R \times N$ rectangular array. However, in spectroscopy applications, the R rows of the CCD are typically moved down simultaneously and summed together into the shift register. This "full vertical binning" results in a $1 \times N$ vector, so not much has changed with our model. The main difference is that while the signal, the shot noise, and the background counts all increase with R, the Gaussian readout noise remains the same as if no binning were performed. This is one of the main differences between our spectroscopy application and the image restoration problem: the binning makes the Gaussian readout noise smaller in comparison with the other noise sources, and the readout noise is thus often neglected. We will likewise neglect the readout noise for the rest of this thesis.

Since the signal n_i^s and the background n_i^b are assumed to be statistically independent,

the sum $n_i = n_i^s + n_i^b$ is Poisson-distributed: $n_i \sim Pois(\mu_i^s + \mu_i^b)$. Neglecting the readout noise, our model is then given by

$$y_i \sim Pois(\mu_i), \tag{20}$$

where

$$\mu_i \triangleq \mu_i^s + \mu_i^b = \beta_i \left(\mathbf{Bx}\right)_i + \mu_i^b. \tag{21}$$

Since the y_i are assumed to be independent, the distribution of $\mathbf{y}$ is given by

$$p(\mathbf{y}; \mathbf{x}) = \prod_i \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}. \tag{22}$$

2.2.3 Related Work

Similar Poisson models have been employed in other applications such as shifted excitation Raman spectroscopy (SERS) and quantum-limited hyperspectral imaging (HSI). In [99] and [58], a SERS observation model is given as

$$\mathbf{y} \sim \text{Pois}(\mathbf{Hs}), \tag{23}$$

where $\mathbf{s} = [\mathbf{s}_F^T \ \mathbf{s}_R^T]^T$ is a concatenation of the Raman spectrum $\mathbf{s}_R$ and the fluorescence background $\mathbf{s}_F$, and $\mathbf{H}$ describes the linear relationship between these and the observed data. This SERS model differs from our Poisson model of Section 2.2.2 in several ways. First, the SERS approach of [99] and [58] directly estimates the entire Raman spectrum, whereas in our approach we estimate the mixing coefficients corresponding to a known reference library. Second, the SERS model explicitly accounts for the fluorescence background, which is not currently included in our measurement model; this is left as a topic for future work. Third, our model includes several additional sources of noise and distortion that are present in a realistic dispersive measurement system.

In [21], a similar Poisson model is used to describe the interferogram data obtained from a quantum-limited HSI system:

$$\mathbf{y} \sim \text{Pois}(\mathbf{H}'\mathbf{s}'). \tag{24}$$

The main difference between (24) and the model of Section 2.2.2 is that $\mathbf{H}'$ encapsulates the linear model for the interferometer-based device, whereas our model is for a dispersive

device. Also, as in the SERS approach of [99] and [58], the entire spectrum $\mathbf{s}'$ is to be reconstructed; there is no reference library. (However, nothing inherent in the formulations in [99], [58], and [21] would prevent them from being adapted to include a reference library.)

2.3 The Modified Richardson-Lucy Algorithm

Given the measured spectrum $\mathbf{y}$, one approach to estimating the mixing coefficients $\mathbf{x}$ is to seek the maximum-likelihood (ML) estimate

$$\begin{aligned}\hat{\mathbf{x}}^{ML} &= \arg\max_{\mathbf{x}\geq\mathbf{0}} p(\mathbf{y};\mathbf{x}) \\ &= \arg\max_{\mathbf{x}\geq\mathbf{0}} \ln p(\mathbf{y};\mathbf{x}). \end{aligned} \tag{25}$$

For notational simplicity, we will drop the superscript and refer to the ML estimate simply as $\hat{\mathbf{x}}$. Plugging (22) into (25) and dropping the terms that do not depend on $\mathbf{x}$,

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}\geq\mathbf{0}} \sum_i y_i \ln(\mu_i) - \mu_i \tag{26a}$$

$$= \arg\max_{\mathbf{x}\geq\mathbf{0}} \sum_i y_i \ln\left[\beta_i (\mathbf{B}\mathbf{x})_i + \mu_i^b\right] - \beta_i (\mathbf{B}\mathbf{x})_i \,. \tag{26b}$$

It is straightforward to show that this objective function is concave in $\mathbf{x}$. This means that any local maximum of the loglikelihood function is also a global maximum.

Building on the approach developed by Shepp and Vardi [80], Snyder et al. [88] derived the following expectation-maximization (EM) iteration:

$$\hat{x}_j^{new} = \frac{1}{\sum_i \beta_i [\mathbf{B}]_{ij}} \sum_i \frac{\beta_i [\mathbf{B}]_{ij} \hat{x}_j^{old}}{\sum_{j'} \beta_i [\mathbf{B}]_{ij'} \hat{x}_{j'}^{old} + \mu_i^b} y_i \,. \tag{27}$$

This iterative formula turns out to be a variation of the Richardson-Lucy algorithm [73, 52], modified to include the various noise effects in the CCD. The sequence of estimates will converge toward a ML estimate. Also, the estimates at any iteration will clearly be nonnegative, provided that the initial estimate is nonnegative. Since a parameter initialized to zero will stay at zero, it is important to initialize with a positive estimate. Algorithm 1 sketches the main steps of the modified Richardson-Lucy (MRL) algorithm.

Algorithm 1 Modified Richardson-Lucy Algorithm

Input: $\mathbf{y}, \mathbf{B}, \boldsymbol{\beta}, \boldsymbol{\mu}^b$
Output: $\hat{\mathbf{x}}^{MRL}$

```
Initialize with some positive x̂^old
while not converged do
    Compute x̂^new using Equation (27)
    x̂^old ← x̂^new
end while
x̂^MRL ← x̂^new
```

2.3.1 Intuition behind the Richardson-Lucy Algorithm

The MRL algorithm becomes more intuitive if we first view the trivial case in which $\mathbf{B}$ is an $N \times 1$ matrix, and then consider how to expand to the $N \times 2$ problem. In this discussion, we let $\beta_i = 1$ and $\mu_i^b = 0$ for all i (i.e., we consider the original Richardson-Lucy algorithm). Let $\boldsymbol{\alpha}$ be the single vector in the reference library, so that $y_i \sim \text{Pois}(\alpha_i x)$. It is straightforward to show that the ML estimate of x is given by

$$\hat{x}^{ML} = \frac{\sum y_i}{\sum \alpha_i} \,. \tag{28}$$

When there are two vectors in the reference library, i.e., when $y_i \sim \text{Pois}(\alpha_i x_1 + \gamma_i x_2)$, it is impossible to solve for the ML estimates of x_1 and x_2 directly. The EM approach considers the "complete" data $\{n_i^1, n_i^2\}$, where $n_i^1 \sim \text{Pois}(\alpha_i x_1)$ and $n_i^2 \sim \text{Pois}(\gamma_i x_1)$. We have knowledge of only the sum $y_i = n_i^1 + n_i^2$, or the "incomplete data." However, if we somehow magically knew the values of n_i^1 and n_i^2, then the ML estimates would simply be given by (28):

$$\hat{x}_1^{ML} = \frac{\sum_i n_i^1}{\sum_i \alpha_i} \qquad\qquad \hat{x}_2^{ML} = \frac{\sum_i n_i^2}{\sum_i \gamma_i} \,. \tag{29}$$

The EM algorithm iterates as follows:

1. Make an initial guess $\hat{\mathbf{x}}^{old}$.

2. Since we do not know n_i^1, we find its expected value *given* $\mathbf{y}$ and $\hat{\mathbf{x}}^{old}$:

$$E_{n_i^1 | y_i; \hat{\mathbf{x}}^{old}} \left[n_i^1 \right] = \frac{\alpha_i \hat{x}_1^{old}}{\alpha_i \hat{x}_1^{old} + \gamma_i \hat{x}_2^{old}} \, y_i \,, \tag{30}$$

where we have used the fact (see [87],[80]) that if $A \sim \text{Pois}(\lambda_A)$ and $B \sim \text{Pois}(\lambda_B)$, then $E(A|(A+B)) = \frac{\lambda_A}{\lambda_A+\lambda_B}(A+B)$. We then update $\hat{x}_1^{new}$ using (29), but with the expectation:

$$\begin{aligned}\hat{x}_1^{new} &= \frac{\sum_i E_{n_i^1|y_i;\hat{\mathbf{x}}^{old}}\left[n_i^1\right]}{\sum_i \alpha_i} \\ &= \frac{\sum_i \frac{\alpha_i \hat{x}_1^{old}}{\alpha_i \hat{x}_1^{old} + \gamma_i \hat{x}_2^{old}} y_i}{\sum_i \alpha_i} \,. \end{aligned} \tag{31}$$

Similarly,

$$\hat{x}_2^{new} = \frac{\sum_i \frac{\gamma_i \hat{x}_2^{old}}{\alpha_i \hat{x}_1^{old} + \gamma_i \hat{x}_2^{old}} y_i}{\sum_i \gamma_i} \,. \tag{32}$$

3. Repeat (31) and (32) until convergence has been reached.

Algorithm 1 is a generalization of these three steps, and (27) is a generalization of (31) and (32), to allow for arbitrary values of $\boldsymbol{\beta}$, $\boldsymbol{\mu}^b$, and M.

2.3.2 Relationship to Minimum I-divergence Methods

Csiszár's I-divergence is a generalization of the Kullback-Leibler divergence to functions that do not sum to unity. Csiszár showed that if the functions involved are nonnegative, then the I-divergence is the only discrepancy measure that satisfies a desired set of intuitive postulates [10, 8]. The I-divergence from the vector $\mathbf{p}$ to the vector $\mathbf{q}$ is defined by

$$I(\mathbf{p}||\mathbf{q}) = \sum p_i \ln \frac{p_i}{q_i} - p_i + q_i \,. \tag{33}$$

Suppose we are given the data $\mathbf{y}$, and are seeking the nonnegative mixing vector $\mathbf{x}$ that minimizes the I-divergence from $\mathbf{y}$ to $\mathbf{Bx}$:

$$\hat{\mathbf{x}}^{I-div} = \arg\min_{\mathbf{x}\geq\mathbf{0}} I(\mathbf{y}||\mathbf{Bx}) \tag{34a}$$

$$= \arg\min_{\mathbf{x}\geq\mathbf{0}} \sum_i \left[y_i \ln y_i - y_i \ln \sum_j B_{ij}x_j - y_i + \sum_j B_{ij}x_j \right] \tag{34b}$$

$$= \arg\max_{\mathbf{x}\geq\mathbf{0}} \sum_i \left[y_i \ln \sum_j B_{ij}x_j - \sum_j B_{ij}x_j \right] \,. \tag{34c}$$

By comparing (34c) with (26a), we see that for the special case in which $\beta_i = 1$ and $\mu_i^b = 0$ for all i, the ML approach under the Poisson model is equivalent to the minimum I-divergence approach.

2.4 Cramér-Rao Lower Bound (CRLB)

The Cramér-Rao lower bound (CRLB) is a lower bound on the variance of any unbiased estimator. This performance bound gives algorithm developers an idea of the fundamental limits of the sensor. If acceptable performance is not obtained, it is tempting to assign fault to the algorithm and hope that a better algorithm is all that is needed to coax the desired result out of the same data. However, this may be a hopeless quest. If better performance is needed than the bounds indicate is possible, then the amount of needed information is not available in the data, regardless of how smart the algorithm is. In this case, a more informative sensor will be needed.

2.4.1 Derivation of the CRLB

The Fisher Information Matrix (FIM) for a vector parameter is defined as

$$[\mathbf{I}(\mathbf{x})]_{kl} = E_{\mathbf{y};\mathbf{x}}\left[-\frac{\partial^2}{\partial x_k \partial x_l}\ln p(\mathbf{y};\mathbf{x})\right]. \tag{35}$$

For any unbiased estimator $\hat{\mathbf{x}}$ with covariance matrix $\mathbf{C}_{\hat{\mathbf{x}}}$, the matrix $\left[\mathbf{C}_{\hat{\mathbf{x}}} - \mathbf{I}^{-1}(\mathbf{x})\right]$ is nonnegative definite. Since the diagonal elements of a nonnegative definite matrix are nonnegative, this implies that

$$\mathrm{Var}(\hat{x}_j) \geq [\mathbf{I}^{-1}(\mathbf{x})]_{jj}. \tag{36}$$

The diagonal elements of the inverse FIM are the Cramér-Rao lower bounds on the individual parameters [92].

Taking the logarithm of (22) and differentiating with respect to x_l, we have

$$\begin{aligned}\frac{\partial}{\partial x_l}\ln p(\mathbf{y};\mathbf{x}) &= \frac{\partial}{\partial x_l}\sum_i y_i \ln\mu_i - \mu_i - \ln(y_i!) \\ &= \sum_i \frac{y_i}{\mu_i}\frac{\partial}{\partial x_l}\mu_i - \frac{\partial}{\partial x_l}\mu_i \\ &= \sum_i \frac{y_i\beta_i[\mathbf{B}]_{il}}{\mu_i} - \beta_i[\mathbf{B}]_{il}. \end{aligned} \tag{37}$$

Differentiating with respect to x_k yields

$$\frac{\partial^2}{\partial x_k \partial x_l}\ln p(\mathbf{y};\mathbf{x}) = -\sum_i \frac{y_i\beta_i^2[\mathbf{B}]_{il}[\mathbf{B}]_{ik}}{\mu_i^2}\,. \tag{38}$$

The FIM is then given by

$$\begin{aligned}[\mathbf{I}(\mathbf{x})]_{kl} &= \sum_i \frac{\beta_i^2[\mathbf{B}]_{il}[\mathbf{B}]_{ik}}{\left(\beta_i\sum_j[\mathbf{B}]_{ij}x_j + \mu_i^b\right)^2} E_{y_i;\mathbf{x}}[y_i] \\ &= \sum_i \frac{\beta_i^2[\mathbf{B}]_{il}[\mathbf{B}]_{ik}}{\beta_i\sum_j[\mathbf{B}]_{ij}x_j + \mu_i^b}\,, \end{aligned} \tag{39}$$

where we have used the fact that $E_{\mathbf{y};\mathbf{x}}\left(\sum_i \alpha_i y_i\right) = \sum_i \alpha_i E_{y_i;\mathbf{x}}(y_i)$.

For the Cramér-Rao inequality to hold, the FIM must exist for all valid $\mathbf{x}$. This condition will be satisfied if the denominator in (39) is nonzero for all i. Since $\boldsymbol{\beta}$, $\mathbf{B}$, and $\mathbf{x}$ are all nonnegative, the expression $\beta_i\sum_j[\mathbf{B}]_{ij}x_j + \mu_i^b$ will be positive for all i if μ_i^b is positive for all i. Thus, if $\boldsymbol{\mu}^b > 0$, the FIM will exist and be finite. The FIM might also exist and be finite for $\boldsymbol{\mu}^b = 0$, if the $\boldsymbol{\beta}$, $\mathbf{B}$ and x_j are set appropriately.

A second "regularity condition" that must be satisfied for the CRLB to be valid [35] is that

$$E_{\mathbf{y};\mathbf{x}}\left[\frac{\partial}{\partial x_l}\ln p(\mathbf{y};\mathbf{x})\right] = 0. \tag{40}$$

We verify this by taking the expected value of the expression in (37):

$$\begin{aligned}E_{\mathbf{y};\mathbf{x}}\left[\frac{\partial}{\partial x_l}\ln p(\mathbf{y};\mathbf{x})\right] &= \sum_i \frac{E_{y_i;\mathbf{x}}[y_i]\beta_i[\mathbf{B}]_{il}}{\mu_i} - \beta_i[\mathbf{B}]_{il} \\ &= 0. \end{aligned} \tag{41}$$

2.4.2 CRLB Examples

2.4.2.1 Special Cases of One and Two Estimated Mixture Coefficients

Some intuition concerning this CRLB may be be gained by considering the degenerate case in which there is a single spectrum $\mathbf{b} = (b_1, \ldots b_N)$ in the reference library.[2] If we simplify further by letting $\beta_i = 1$ and $\mu_i^b = 0$ for all i, then $y_i \sim Pois(b_i x)$, and plugging (39) into (36) yields the scalar CRLB

$$\mathrm{Var}(\hat{x}) \geq \frac{x}{\sum_i b_i}. \tag{42}$$

In this case, the minimum achievable variance of any unbiased estimator $\hat{x}$ increases with the true parameter x. This is a reflection of the fundamental property of shot noise discussed in Section 2.2.2: the signal cannot be increased without the variance also being increased. We note that while the *absolute* variance of $\hat{x}$ increases with x, the *relative* accuracy of the estimate improves with stronger chemical concentration:

$$SNR_{\hat{x}} = \frac{\mathrm{E}[\hat{x}]}{\sqrt{\mathrm{Var}(\hat{x})}} \propto \frac{x}{\sqrt{x}} = \sqrt{x}. \tag{43}$$

In contrast with this single-vector Poisson example, consider the single-vector AWGN model $\mathbf{y} = \mathbf{b}x + \mathbf{v}$, where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The CRLB for x under this model is given by

$$\mathrm{Var}(\hat{x}) \geq \frac{\sigma^2}{||\mathbf{b}||^2}. \tag{44}$$

Here the variance decreases as the L^2 norm of $\mathbf{b}$ increases, whereas under the Poisson model, it decreased with the sum, or L^1 norm, of $\mathbf{b}$. This suggests that for the shot noise dominated case, if normalization of the spectra is to be performed as a preprocessing step, it may be more natural to use the L^1 norm rather than the L^2 norm. Another difference between (42) and (44) is that the CRLB for the AWGN model does not depend on the true parameter x. These two differences are illustrated in Table 2, which shows the CRLBs under each model for different values of $\mathbf{b}$ and x.

[2]Here we have followed the notation of (17): $\mathbf{b} = \mathbf{Pa}$, where $\mathbf{a}$ is the reference library vector for a single chemical.

Table 2: CRLB for the $M = 1$ case.

$\mathbf{b}$	x	AWGN model CRLB(x)	Poisson model CRLB(x)
$(0, 4, 0, 0, 0)^T$	1	$\sigma^2/16$	1/4
$(0, 4, 0, 0, 0)^T$	10	$\sigma^2/16$	10/4
$(1, 1, 1, 1, 0)^T$	1	$\sigma^2/4$	1/4
$(1, 1, 1, 1, 0)^T$	10	$\sigma^2/4$	10/4

If we again consider the single-parameter case but now relax the requirement that $\beta_i = 1$ and $\mu_i^b = 0$, then the CRLB is given by

$$\mathrm{Var}(\hat{x}) \geq \frac{1}{\sum_i \frac{\beta_i^2 b_i^2}{\beta_i b_i x + \mu_i^b}}. \tag{45}$$

It is clear that the CRLB increases as the detector efficiency $\boldsymbol{\beta}$ decreases and as the mean of the background counts $\boldsymbol{\mu}^b$ increases.

The bounds will typically be higher if the parameters are to be jointly estimated. This is illustrated in Table 3. The "uncoupled" bounds are the scalar CRLBs found in Table 2. The "coupled" CRLBs are given by (36) and (39). We used the tiny value of $\mu_i^b = 0.01$ for all i to be sure that the FIM could be computed. The CRLB (nonstrictly) increases as we estimate more parameters, because $[\mathbf{I}^{-1}(\mathbf{x})]_{jj} \geq 1/[\mathbf{I}(\mathbf{x})]_{jj}$. We see that the bounds are highly dependent on the specific combination of true mixing coefficients; having a lot of the first chemical present does not greatly undermine our ability to estimate the second (an increase from 0.25 to 0.33), but a large presence of the second chemical will have a more significant effect on the bound for the first (an increase from 0.25 to 1.08).

Table 3: CRLB for the $M = 2$ case.

$\mathbf{b}$	x	CRLB (uncoupled)	CRLB (coupled)
$(0, 4, 0, 0, 0)^T$	1	0.25	1.08
$(1, 1, 1, 1, 0)^T$	10	2.5	3.33
$(0, 4, 0, 0, 0)^T$	10	2.5	2.58
$(1, 1, 1, 1, 0)^T$	1	0.25	0.33

2.4.2.2 CRLB for a Real Reference Library

This section explores a reference library provided by Darren Emge of the Edgewood Chemical Biological Center. We consider the detection of a target chemical characterized by the Raman spectrum shown in Figure 9. We let this spectrum be the first column of our reference library (with corresponding mixing coefficient x_1), and we examine how our estimate $\hat{x}_1$ is degraded as the library is expanded to include more chemicals. For this simulation, we normalize each spectrum in the library to sum to 10,000, and we let $\mu_i^b = 5$ and $\beta_i = 1$ for all i. Each spectrum in the reference library has 1024 frequency samples ($N = 1024$).

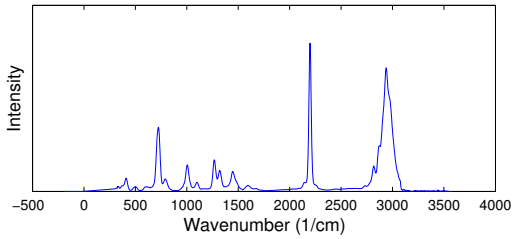


Figure 9: Raman spectrum of Chem. 1.

The standard deviation bound is plotted vs. M (the number of chemicals in the reference library) in Figure 10. The two curves are for two different true mixing vectors: for the top curve, $x_j = 1$ for all j, while for the bottom curve, $x_1 = 1$ and $x_j = 0$ for all $j \neq 1$. The sharp jump at $M = 13$ reflects the fact that the 13^{th} element of the reference library is one of the elements most correlated with the first. From the bottom curve we see that the additional spectra in the reference library make it more difficult to estimate x_1, even if no other chemicals are actually present in the measurement sample. Note that there was no particular logic to the ordering of the spectra—if the 50 spectra in the reference library were rearranged, the plot would look different, although the overall trends would stay the same.

2.5 Weighted Least Squares (WLS) with Known Weights

This section begins by revisiting the simple Gaussian model,

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}, \tag{46}$$

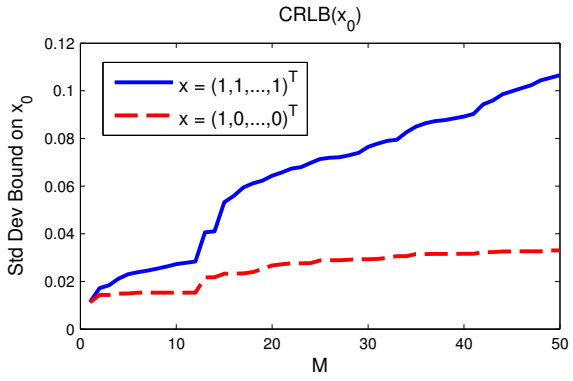


Figure 10: Standard deviation bound on x_1 as a function of M.

but we now consider the case in which the variance depends on the bin number:

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{Hx}, \mathbf{C}\right), \tag{47}$$

where $\mathbf{C}$ is a diagonal matrix defined by $[\mathbf{C}]_{ii} = \sigma_i^2$. It is well known that the maximum-likelihood estimate is given by

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} \ln p(\mathbf{y};\mathbf{x}) \\ &= \arg\min_{\mathbf{x}} (\mathbf{y}-\mathbf{Hx})^T \mathbf{C}^{-1} (\mathbf{y}-\mathbf{Hx}) \\ &= \arg\min_{\mathbf{x}} \|\mathbf{Wy}-\mathbf{WHx}\|^2, \end{aligned} \tag{48}$$

where $\mathbf{W}$ is the diagonal matrix defined by $[\mathbf{W}]_{ii} = w_i \triangleq 1/\sigma_i$, i.e., $\mathbf{W}^2 \triangleq \mathbf{C}^{-1}$. Equation (48) formalizes the intuitive notion that the measurements with lower variance should have more influence on our estimate than the measurements in which we have less confidence. Assuming the true weights (inverses of the variances) are known, the ML estimate is thus given by the weighted least-squares (WLS) estimator

$$\hat{\mathbf{x}} = \left(\mathbf{H}^T\mathbf{W}^2\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{W}^2\mathbf{y}. \tag{49}$$

This estimator is easily shown to be unbiased and is also shown to achieve the CRLB [35]. Since it is the minimum variance unbiased (MVU) estimator, no other unbiased estimator can achieve a lower mean-squared error.

2.5.1 WLS for Poisson Data

The WLS approach has also been applied to Poisson data. The next four subsections discuss four justifications for its use.

2.5.1.1 Large-Signal Approximation

For strong signals, the Poisson-distributed measurement y_i is approximated well by a normal distribution:

$$y_i \sim Pois\left((\mathbf{Hx})_i\right) \tag{50}$$

can be approximated by

$$y_i \sim \mathcal{N}\left((\mathbf{Hx})_i, (\mathbf{Hx})_i\right), \tag{51}$$

or, in vector form,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{Hx}, \mathbf{C}). \tag{52}$$

Thus, under this "large-signal approximation," the WLS approach is "approximately" MVU [54].

2.5.1.2 Best Linear Unbiased Estimator Property

If we relax the assumption of Gaussianity but still assume that $E[\mathbf{y}] = \mathbf{Hx}$ and $\mathrm{Cov}(\mathbf{y}) = \mathbf{C}$ (as is the case for the Poisson model), then by the Gauss-Markov theorem [35], the WLS estimator defined by (49) is the best *linear* unbiased estimator (BLUE). Therefore, if the signal intensity is too low for the large-signal approximation to be valid, the WLS estimator still has the lowest variance out of the restricted class of linear unbiased estimators.

2.5.1.3 Minimum Variance Unbiased Estimator Property

The above properties are common justifications for applying the WLS approach to Poisson data. We now present a more powerful, lesser-known property: *if the "true" weights corresponding to the "true" parameters are precisely known, the WLS method yields the MVU estimator for our Poisson data model.* We recognize that this is an artificial formulation

(since the true weights will not be known in practice), but following this path will provide insights into algorithm behavior.

We begin by defining

$$[\mathbf{H}]_{ij} \triangleq \beta_i [\mathbf{B}]_{ij} \tag{53}$$

and

$$\mathbf{y}' \triangleq \mathbf{y} - \boldsymbol{\mu}^b. \tag{54}$$

Then the WLS method, applied to the calibrated data $\mathbf{y}'$, is unbiased:

$$E(\hat{\mathbf{x}}^{WLS}) = \left(\mathbf{H}^T\mathbf{W}^2\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{W}^2 E(\mathbf{y}') \tag{55a}$$

$$= \mathbf{x}. \tag{55b}$$

Furthermore, if we let

$$\mathbf{P} \triangleq \left(\mathbf{H}^T\mathbf{W}^2\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{W}^2, \tag{56}$$

then the covariance matrix of $\hat{\mathbf{x}}^{WLS}$ is given by

$$\begin{aligned}
\mathrm{Cov}(\hat{\mathbf{x}}^{WLS}) &= \mathrm{Cov}(\mathbf{P}\mathbf{y}') \\
&= \mathbf{P}\mathrm{Cov}(\mathbf{y}')\mathbf{P}^T \\
&= \mathbf{P}\mathrm{Cov}(\mathbf{y})\mathbf{P}^T \\
&= \left(\mathbf{H}^T\mathbf{W}^2\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{W}^2(\mathbf{W}^2)^{-1}\mathbf{W}^2\mathbf{H}\left(\mathbf{H}^T\mathbf{W}^2\mathbf{H}\right)^{-1} \\
&= \left(\mathbf{H}^T\mathbf{W}^2\mathbf{H}\right)^{-1}.
\end{aligned} \tag{57}$$

The Fisher information matrix given by (39) can be rewritten as

$$[\mathbf{I}(\mathbf{x})]_{kl} = \sum_i \frac{[\mathbf{H}]_{il}[\mathbf{H}]_{ik}}{\sum_j [\mathbf{H}]_{ij}x_j + \mu_i^b}. \tag{58}$$

This is expressed in matrix form as

$$\mathbf{I}(\mathbf{x}) = \mathbf{H}^T\mathbf{W}^2\mathbf{H}, \tag{59}$$

where $\mathbf{W}^2$ is the diagonal matrix defined by $[\mathbf{W}^2]_{ii} \triangleq 1/(\sum_j [\mathbf{H}]_{ij}x_j + \mu_i^b)$, i.e., $\mathbf{W}^2 \triangleq \mathbf{C}^{-1}$. Since $\mathrm{Cov}(\hat{\mathbf{x}}^{WLS}) = \mathbf{I}^{-1}$, $\hat{\mathbf{x}}^{WLS}$ is an MVU estimator. This is stronger than the BLUE property guaranteed (for any distribution of y_i) by the Gauss-Markov theorem.

This subsection assumed the variances of the data points are known exactly. In reality, they will never be known exactly, since they depend on the unknown true mixing vector $\mathbf{x}$.[3] In practice, it is necessary to estimate the weights from the data, and the optimality properties of WLS are lost. However, this section provides an important piece of insight: the failings of WLS are not caused by the Gaussian distribution failing to approximate the Poisson data *per se*; rather, they are caused by our lack of knowledge of the true weights.

2.5.1.4 Relationship between ML and WLS

Lastly, we will show that the ML approach can lead to a WLS formulation. Section 2.3 showed that the ML estimate under our Poisson model is found by maximizing the concave objective function

$$\psi(\mathbf{x}) = \sum_i y_i \ln u_i - u_i, \tag{60}$$

where $\mu_i \triangleq \beta_i \left(\mathbf{B}\mathbf{x}\right)_i + \mu_i^b = \left(\mathbf{H}\mathbf{x}\right)_i + \mu_i^b$. Since $(\partial/\partial x_j)\mu_i = [\mathbf{H}]_{ij}$, we have

$$\frac{\partial}{\partial x_j}\psi(\mathbf{x}) = \sum_i \frac{y_i[\mathbf{H}]_{ij}}{\mu_i} - [\mathbf{H}]_{ij} \tag{61}$$

if $\mu_i > 0$ for all i. The objective function ψ is not defined for negative μ_i; the mean parameter for a Poisson distribution cannot be negative. Therefore, the point $\tilde{\mathbf{x}}$ satisfying

$$\left(\sum_i \frac{y_i[\mathbf{H}]_{ij}}{\mu_i} - [\mathbf{H}]_{ij}\right)\Bigg|_{\mathbf{x}=\tilde{\mathbf{x}}} = 0 \quad \text{for all } j \tag{62}$$

will be the ML estimate only if the resulting $\tilde{\boldsymbol{\mu}} = \mathbf{H}\tilde{\mathbf{x}} + \boldsymbol{\mu}^b$ happens to contain no negative values. In addition to the nonnegativity of $\boldsymbol{\mu}$ required by the Poisson model, we have a stricter requirement that the mixing coefficients themselves be nonnegative ("stricter" in the sense that $\tilde{\mathbf{x}} \geq 0$ implies that $\tilde{\boldsymbol{\mu}} \geq 0$). Thus, the $\tilde{\mathbf{x}}$ satisfying (62) will be the solution to the constrained ML problem (26a) if $\tilde{\mathbf{x}} \geq \mathbf{0}$.

Even if we momentarily ignore the nonnegativity constraint, (62) is a nonlinear system of M equations and M unknowns that cannot be solved directly (analytically). However, we may gain insight into the problem by following the approach of Wheaton et al. [95] and

[3] Furthermore, if the weights *were* somehow perfectly known, then $\mathbf{x}$ could be directly recovered from them exactly; $\mathbf{y}$ would not be needed.

Dixon et al. [13] and rewriting (62) as

$$\sum_i \frac{y_i[\mathbf{H}]_{ij}}{\tilde{\mu}_i} = \sum_i \frac{[\mathbf{H}]_{ij}\tilde{\mu}_i}{\tilde{\mu}_i} \tag{63a}$$

$$= \sum_i \frac{[\mathbf{H}]_{ij}\left(\sum_{j'}[\mathbf{H}]_{ij'}\tilde{x}_{j'} + \mu_i^b\right)}{\tilde{\mu}_i}. \tag{63b}$$

Equivalently,

$$\sum_i \frac{[\mathbf{H}]_{ij}(y_i - \mu_i^b)}{\tilde{\mu}_i} = \sum_i \frac{[\mathbf{H}]_{ij}\sum_{j'}[\mathbf{H}]_{ij'}\tilde{x}_{j'}}{\tilde{\mu}_i} \tag{64a}$$

$$= \sum_{j'}\sum_i \frac{[\mathbf{H}]_{ij}[\mathbf{H}]_{ij'}}{\tilde{\mu}_i}\tilde{x}_{j'}. \tag{64b}$$

Equation (64) can be expressed in matrix form as

$$\left[\mathbf{H}^T\tilde{\mathbf{W}}^2\left(\mathbf{y} - \boldsymbol{\mu}^b\right)\right]_j = \left[\mathbf{H}^T\tilde{\mathbf{W}}^2\mathbf{H}\tilde{\mathbf{x}}\right]_j, \tag{65}$$

where $\tilde{\mathbf{W}}^2$ is the diagonal matrix defined by

$$[\tilde{\mathbf{W}}^2]_{ii} \triangleq \frac{1}{\tilde{\mu}_i}. \tag{66}$$

If we again let $\mathbf{y}' \triangleq \mathbf{y} - \boldsymbol{\mu}^b$, then

$$\mathbf{H}^T\tilde{\mathbf{W}}^2\mathbf{y}' = \mathbf{H}^T\tilde{\mathbf{W}}^2\mathbf{H}\tilde{\mathbf{x}}. \tag{67}$$

Thus, $\tilde{\mathbf{x}}$ may be viewed as the solution to a WLS problem, but one in which the "optimal" weights are unknown:

$$\tilde{\mathbf{x}} = \left(\mathbf{H}^T\tilde{\mathbf{W}}^2\mathbf{H}\right)^{-1}\mathbf{H}^T\tilde{\mathbf{W}}^2\mathbf{y}'. \tag{68}$$

We again emphasize that the $\tilde{\mathbf{x}}$ given by (68) is the ML estimate only if it happens to satisfy $\tilde{\mathbf{x}} \geq \mathbf{0}$. We also note that the "optimal" weighting matrix $\tilde{\mathbf{W}}^2$, which is a function of the estimate $\tilde{\mathbf{x}}$, will generally be different from the "true" weighting matrix $\mathbf{W}^2$ (which is a function of the "true" $\mathbf{x}$) used in the previous section. The "true WLS" nonetheless provides a useful reference point in our simulations.

2.6 WLS with Estimated Weights

In Section 2.5, we saw that if the "true" weighting matrix, defined by

$$[\mathbf{W}]_{ii} = w_i \triangleq \frac{1}{\sqrt{\mu_i}} = \frac{1}{\sqrt{\beta_i\sum_{j'}[\mathbf{B}]_{ij'}x_{j'} + \mu_i^b}}, \tag{69}$$

is assumed to be known, then the resulting WLS estimator is the MVU estimator for our Poisson model. In practice, however, the true weights depend on $\mathbf{x}$ and are therefore unknown, making the "true WLS" method unrealizable. One approach is to estimate the weights from the data and then apply WLS with the estimated weights:

$$\hat{\mathbf{x}} = \left(\mathbf{H}^T \hat{\mathbf{W}}^2(\mathbf{y})\mathbf{H}\right)^{-1} \mathbf{H}^T \hat{\mathbf{W}}^2(\mathbf{y})\mathbf{y}'. \tag{70}$$

We will use the nomenclature of [5] and refer to this approach as generalized least squares (GLS). The estimated weights, being a function of $\mathbf{y}$, are now random and are generally correlated with the data. Thus, unlike the "true" WLS or the LS or any other method in which the weights are fixed, the GLS approach is generally biased (because (55a) no longer holds). Similarly, the covariance matrix of $\hat{\mathbf{x}}$ is no longer given by (57).

2.6.1 A Simple Generalized Least Squares Algorithm

Perhaps the simplest and most common GLS method is to estimate the variances μ_i directly from the data: $\hat{\mu}_i = y_i$. If we neglect the structure on $\boldsymbol{\mu}$ created by (21) and assume that each value is uncorrelated with the next, this method can be shown to give the ML estimate of μ_i given y_i:

$$\hat{\mu}_i^{ML} = \arg\max_{\mu_i} p(y_i; \mu_i) = y_i. \tag{71}$$

Furthermore, it is unbiased: $E(\hat{\mu}_i) = E(y_i) = \mu_i$. It follows from the invariance property of ML estimation that the resulting estimated weights are also ML estimates:

$$\hat{w}_i^{ML} = \frac{1}{\sqrt{\hat{\mu}_i^{ML}}} = \frac{1}{\sqrt{y_i}}. \tag{72}$$

These estimates of the weights, however, are biased: $E(\hat{w}_i) = E(1/\sqrt{y_i}) \neq 1/\sqrt{\mu_i}$ in general. They are also unstable; the variances of the estimated weights can be quite large. This is especially true for small measurement values; for example, if $y_i = 0$, this method will give infinite weight to the i^{th} measurement.

We now illustrate this simple GLS method by revisiting the $M = 1$, $\boldsymbol{\mu}^b = \mathbf{0}$ case of Section 2.4.2.1, in which $y_i \sim Pois(h_i x)$. For this case,

$$\hat{x}^{WLS} = \frac{\mathbf{h}^T \mathbf{W}^2 \mathbf{y}}{\mathbf{h}^T \mathbf{W}^2 \mathbf{h}} = \frac{\sum_i y_i}{\sum_i h_i} \tag{73}$$

and

$$\hat{x}^{GLS} = \frac{\mathbf{h}^T \hat{\mathbf{W}}^2 \mathbf{y}}{\mathbf{h}^T \hat{\mathbf{W}}^2 \mathbf{h}} = \frac{\sum_i h_i}{\sum_i h_i^2 / y_i}, \tag{74}$$

where we have used the fact that $[\mathbf{W}^2]_{ii} = 1/(h_i x)$ and $[\hat{\mathbf{W}}^2]_{ii} = 1/y_i$. It is straightforward to show that

$$\hat{x}^{ML} = \arg\max_x p(\mathbf{y}; x) = \frac{\sum_i y_i}{\sum_i h_i}. \tag{75}$$

In this *specific* case, the "true WLS" and ML are equivalent, but as discussed in Section 2.5.1.4, this is not true in general.

We let $\mathbf{h} = [4\ 6]^T$ and $x = 1$, and computed 1 million Monte Carlo runs of the above algorithms. We also computed the unweighted least-squares estimate $\hat{x}^{LS} = \mathbf{h^T y}/\mathbf{h^T h}$. The sample biases of the LS and WLS estimators were approximately zero, as expected, while the sample bias of GLS was approximately -0.1. The GLS method underestimates x because the weights are negatively correlated with the data; the small measurements are given more weight because they are thought to have smaller variance. Table 4 contains the mean square error of each method relative to the MSE of WLS with known weights. Because h_1 and h_2 are similar in magnitude, weighting does not have much influence, and WLS is not much better than unweighted least squares. This somewhat contrived example emphasizes that GLS can theoretically perform *worse* than LS; here, the benefit from weighting is not as great as the harm introduced by needing to estimate the weights. However, in practice, we have yet to see this effect with typical Raman spectra.

Table 4: Relative MSE for the simple $M = 1$, $N = 2$ case.

method	relative MSE
WLS	1
LS	1.035
GLS	1.318
GLS2	1.230

Part of the error in GLS comes from the infinite weight given to any zero measurements. One ad hoc way to address this problem is to change those values from zero to some constant for the calculation of the weights. We force the zero measurements to 1, which is the smallest nonzero value that the measurements can take under our Poisson model. The effect of this

tweak, and the choice of which constant to use, are topics of future research outside of the scope of this thesis. This method is called "GLS2" in Table 4, and is seen to offer some improvement. We will use this approach in all of the remaining GLS simulations in this thesis.

2.6.2 Iteratively Reweighted Least Squares

The simple GLS method of the previous section, given by (72), suffers because it does not take into account the known structure of the variance function. It is known that $\boldsymbol{\mu} = \mathbf{H}\mathbf{x} + \boldsymbol{\mu}^b$, yet (71) might yield a $\hat{\boldsymbol{\mu}}$ that cannot be formed by $\hat{\boldsymbol{\mu}} = \mathbf{H}\hat{\mathbf{x}} + \boldsymbol{\mu}^b$. A preferable approach, iteratively reweighted least squares (IRLS), views the estimation of the variances as a form of regression [5]. Algorithm 2 sketches the main steps of the IRLS algorithm.

Algorithm 2 Iteratively Reweighted Least Squares Algorithm

Input: $\mathbf{y}, \mathbf{H}, \boldsymbol{\mu}^b, K$
Output: $\hat{\mathbf{x}}^{IRLS}$

1: Initialize with LS estimate $\hat{\mathbf{x}}$
2: **for** K iterations **do**
3: $\quad \hat{\boldsymbol{\mu}} \leftarrow \mathbf{H}\hat{\mathbf{x}} + \boldsymbol{\mu}^b$
4: $\quad \hat{w}_i \leftarrow 1/\sqrt{\hat{\mu}_i}$ for all i
5: $\quad$ Compute WLS estimate $\hat{\mathbf{x}}$ using updated weights
6: **end for**
7: $\hat{\mathbf{x}}^{IRLS} \leftarrow \hat{\mathbf{x}}$

The IRLS algorithm is typically initialized with the unweighted least-squares estimate.[4] At each iteration, the algorithm estimates the weights based on the estimated mixing vector from the previous iteration. These estimated weights are then used to compute a new $\hat{\mathbf{x}}$ for the next iteration.

In our simulations, we found the algorithm to typically give good results (and more or less converge) after just two iterations.

[4] An alternative approach, not considered here, is to initialize with a generalized least-squares estimate such as that defined by (72).

2.6.2.1 Fisher's Method of Scoring

For our measurement model, the IRLS algorithm is equivalent to Fisher's method of scoring, which is a common modification [35] of the Newton-Raphson approach to maximizing likelihoods. In the method of scoring, the Hessian matrix in the Newton-Raphson update is replaced by its expected value, resulting in the iterative formula

$$\hat{\mathbf{x}}_{(k+1)} = \hat{\mathbf{x}}_{(k)} + \mathbf{I}^{-1}(\mathbf{x}) \frac{\partial}{\partial \mathbf{x}} \psi(\mathbf{x}) \Big|_{\mathbf{x}=\hat{\mathbf{x}}_{(k)}}. \tag{76}$$

Sections 2.4 and 2.5.1.3 showed that the Fisher information matrix for our measurement model is given by

$$\mathbf{I}(\mathbf{x}) = \mathbf{H}^T \mathbf{W}^2(\mathbf{x}) \mathbf{H} \tag{77}$$

and that the gradient of the loglikelihood objective function is given by

$$\frac{\partial}{\partial \mathbf{x}} \psi(\mathbf{x}) = \mathbf{H}^T \mathbf{W}^2(\mathbf{x}) \mathbf{y}' - \mathbf{H}^T \mathbf{W}^2(\mathbf{x}) \mathbf{H} \mathbf{x}, \tag{78}$$

where we have now made explicit the dependence of $\mathbf{W}^2$ on $\mathbf{x}$. The method of scoring then becomes

$$\hat{\mathbf{x}}_{(k+1)} = \hat{\mathbf{x}}_{(k)} + \left(\mathbf{H}^T \mathbf{W}^2(\mathbf{x}) \mathbf{H}\right)^{-1} \left(\mathbf{H}^T \mathbf{W}^2(\mathbf{x}) \mathbf{y}' - \mathbf{H}^T \mathbf{W}^2(\mathbf{x}) \mathbf{H} \mathbf{x}\right) \Big|_{\mathbf{x}=\hat{\mathbf{x}}_{(k)}} \tag{79a}$$

$$= \left(\mathbf{H}^T \hat{\mathbf{W}}^2_{(k)} \mathbf{H}\right)^{-1} \mathbf{H}^T \hat{\mathbf{W}}^2_{(k)} \mathbf{y}', \tag{79b}$$

where $\left[\hat{\mathbf{W}}^2_{(k)}\right]_{ii} \triangleq 1/(\mathbf{H}\hat{\mathbf{x}}_{(k)} + \boldsymbol{\mu}^b)_i$. The update equation (79b) is equivalent to the update equation for the IRLS algorithm.

Since the objective function is concave, the method of scoring, at first glance, seems appropriate for our problem. However, as in Section 2.5.1.4, the resulting estimate will be the solution to the constrained ML problem (26a) only if it happens to satisfy $\hat{\mathbf{x}}^{IRLS} \geq \mathbf{0}$. In fact, if any of the $\hat{\mu}_i$ are negative on a given iteration, then the derivatives required by the scoring method do not exist, and IRLS ceases to be equivalent to the scoring algorithm.

2.6.2.2 Ensuring nonnegativity

We would like to modify the IRLS approach to ensure that $\hat{\mathbf{x}} \geq \mathbf{0}$. The simplest way to force nonnegativity on each iteration is to simply truncate any negative values to zero.

This "clipping" method does not generally converge to the constrained maximizer [19]. A preferred approach is to replace the WLS estimate

$$\hat{\mathbf{x}}_{(k+1)} = \left(\mathbf{H}^T\hat{\mathbf{W}}_{(k)}^2\mathbf{H}\right)^{-1}\mathbf{H}^T\hat{\mathbf{W}}_{(k)}^2\mathbf{y}' \tag{80a}$$

$$= \arg\min_{\mathbf{x}} \|\hat{\mathbf{W}}_{(k)}\mathbf{y}' - \hat{\mathbf{W}}_{(k)}\mathbf{H}\mathbf{x}\|^2 \tag{80b}$$

with the nonnegative weighted least-squares estimate

$$\hat{\mathbf{x}}_{(k+1)} = \arg\min_{\mathbf{x}\geq\mathbf{0}} \|\hat{\mathbf{W}}_{(k)}\mathbf{y}' - \hat{\mathbf{W}}_{(k)}\mathbf{H}\mathbf{x}\|^2, \tag{81}$$

which can be solved for using any standard NNLS algorithm. This nonnegative iteratively reweighted least-squares (NNIRLS) algorithm appears to converge to the constrained ML estimate in our simulations, although we do not have an explicit mathematical proof of convergence.

Our physics-based model differs from the "standard" Poisson regression model described in [60, 59], in that the "link function" inherently built into it is the identity link, whereas the standard Poisson regression model features a log link function to ensure nonnegativity of the means of the Poisson distribution. The Poisson regression model with the log link function is a special case of the family of generalized linear models, for which the equivalence between IRLS and Fisher's method of scoring (summarized in the previous section) is well-known. However, in Raman spectroscopy, there is no physical justification for applying the mixture model in a logarithmic domain, so additional care must be taken to ensure nonnegativity. In our case, it does not appear that NNIRLS and Fisher's method of scoring are equivalent.

We note that while the nonnegativity constraint reduces the estimation error, it also introduces bias. This is seen, for example, in the common case in which $x_j = 0$: $\hat{x}_j$ will be either zero or positive but never negative; thus, on average, it will be positive and therefore biased.

2.6.3 Simulation Results

We now compare the experimental performance of all of these algorithms to each other and to the CRLB. We again use the reference library of Section 2.4.2.2, but now consider

the detection of the 27^{th} chemical (with corresponding mixing coefficient x_{27}). The Raman spectrum of Chem. 27 is shown in Figure 11. We consider the case in which this is the only substance present, i.e., $x_j = 0$ for all $j \neq 27$. We vary the signal energy by sweeping the mixing coefficient from $x_{27} = 0.1$ to $x_{27} = 10.0$, and examine the performance of the algorithms under the different energy levels. For this simulation, we again normalize each spectrum in the library to sum to 10,000, and we let $\mu_i^b = 5$ and $\beta_i = 1$ for all i. Each spectrum in the reference library has 1024 frequency samples ($N = 1024$).

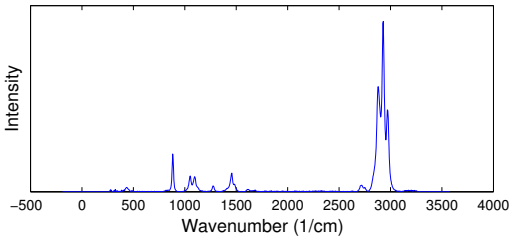


Figure 11: Raman spectrum of Chem. 27.

In this discussion, we will refer to the simple generalized least-squares method of Section 2.6.1 as "GLS." We will use "WLS" to refer to the unrealizable estimator that assumes knowledge of the "true" weights. We used two iterations (in addition to the initial unweighted least-squares estimate) in the IRLS and NNIRLS algorithms. We initialized the NNIRLS and modified Richardson-Lucy (MRL) algorithms with the NNLS estimate. Any zero values in the MRL initial estimate, however, were changed to a small constant; otherwise, they would remain at zero after each MRL iteration. The MRL algorithm was terminated when $\|\hat{\mathbf{x}}^{new} - \hat{\mathbf{x}}^{old}\|_2 <$ 1e-5. The algorithms are summarized in Table 5.

The sample root mean square error (RMSE) and sample bias of the algorithms are plotted in Figures 12 and 13, respectively, where

$$\begin{aligned} \mathrm{RMSE}(\hat{x}_{27}) &\triangleq \sqrt{E\left[(\hat{x}_{27} - x_{27})^2\right]} \\ &= \sqrt{\mathrm{Var}(\hat{x}_{27}) + [\mathrm{Bias}(\hat{x}_{27})]^2}. \end{aligned}$$

We used 1,000 Monte Carlo runs.

The WLS approach, which is unrealizable in practice but provides a useful reference point, is seen to have approximately zero bias, and its RMSE matches up almost exactly

Table 5: Summary of algorithms.

Acronym	Full name	Nonnegativity Constrained (Y/N)	Weighting (None/True/Estimated/Implicit)
LS	Least squares	N	None
NNLS	Nonnegative least squares	Y	None
WLS	Weighted least squares	N	True
GLS	Generalized least squares	N	Estimated
IRLS	Iteratively reweighted least squares	N	Estimated
NNIRLS	Nonnegative iteratively reweighted least squares	Y	Estimated
MRL	Modified Richardson-Lucy	Y	Implicit

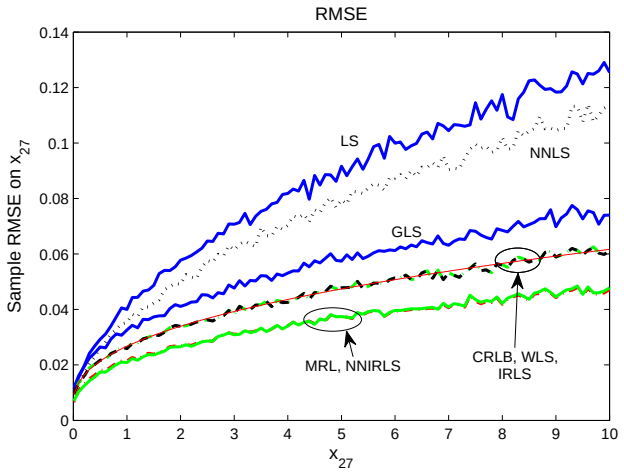


Figure 12: Sample RMSE of the algorithms.

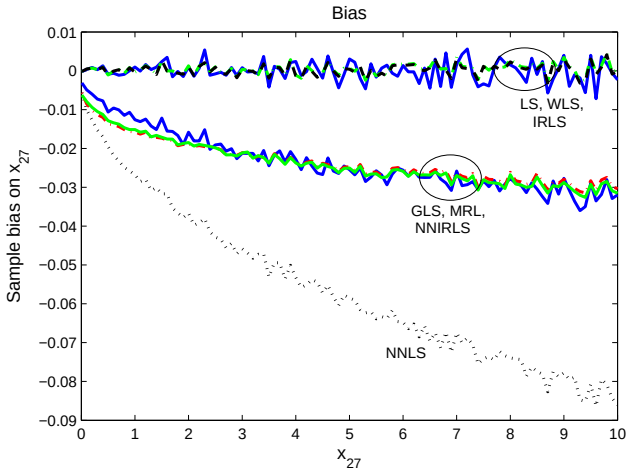


Figure 13: Sample bias of the algorithms.

with the CRLB. This is to be expected, since it was shown in Section 2.5 to be an efficient estimator for our model.

The GLS method has noticeably higher RMSE than WLS for all of the energy levels. This is not surprising; as we saw in Section 2.6.1, this simple GLS approach often yields estimates that are biased and that have increased variance. However, in this case, this error is not as large as that introduced by using uniform weights; LS has the largest RMSE of any of the algorithms. For all of the signal levels considered, weighting leads to improved performance over unweighted least squares; this is true even if the weights are simply estimated using (72). Furthermore, the IRLS algorithm matches WLS (and the CRLB) almost exactly, even after just two iterations; this shows that if the weights are estimated appropriately, little is lost by not knowing the true weights.

The NNIRLS results match up almost exactly with the MRL results; this suggests that NNIRLS is indeed converging to the constrained ML estimate. We note that the MRL and NNIRLS algorithms *beat* the standard deviation bound. This is possible because the CRLB is defined for unbiased estimators.[5] The unbiased estimators (LS, WLS, and IRLS) are less accurate because they sometimes estimate negative values. While the nonnegativity constraint introduces bias, this is more than made up for by the reduction in variance that the extra information brings.

2.7 Conclusions

This chapter presented a probabilistic model, based on previous work in astronomical image restoration [88], for a basic dispersive Raman measurement system. We considered the supervised unmixing framework for chemical identification, in which the problem is formulated as one of deterministic parameter estimation. The maximum-likelihood estimates for the mixing coefficients can be found using the modified Richardson-Lucy algorithm [88].

If the variances of the measurements are assumed to be known, the weighted least-squares method is the minimum variance unbiased estimator for our Poisson model. In

[5]There also exists a biased version of the CRLB, which we do not consider here, since it requires an analytic expression for the bias, which we do not have.

practice, however, the true weights are unknown and must be estimated from the data. Our simulations suggest that the resulting generalized least-squares techniques can give performance that is comparable to the Richardson-Lucy algorithm. In particular, in our simulations, the NNIRLS algorithm achieved nearly identical accuracy to the MRL algorithm but was computationally less demanding.

One possible avenue for future research is to apply this work more generally to the simple passive reflectance EO spectral detection of gasses and materials [78].

CHAPTER III

DETECTING CONSTITUENT CHEMICALS USING MINIMUM DESCRIPTION LENGTH

3.1 Introduction

Most detection algorithms for Raman spectroscopy compare the measured Raman spectrum to a given library of known signatures. One common general approach is to first estimate the relative amount of each chemical present, and then compare each of the estimated mixing coefficients to a threshold. We will call this technique the "spectral unmixing" approach. As discussed in Chapter 2, the method of estimating the mixing coefficients depends on the probabilistic measurement model for the Raman instrument. Chapter 2 considered two such models, Gaussian and Poisson, along with corresponding estimation algorithms. While *parameter estimation* under this approach may be optimal in some sense, the resulting *detection* algorithm will not generally possess any particular optimality.

An alternative method that has been employed on Raman data is the subspace version of the generalized likelihood ratio test (GLRT). Section 3.2.2 reviews the common subspace GLRT that assumes an additive Gaussian noise model. We will also derive the GLRT for the Poisson measurement model. It is generally challenging to set an appropriate threshold for the GLRT; because of the nonnegativity of the parameters, the GLRT does not generally attain its nominal asymptotic performance.

Section 3.3 will frame the problem as one of multiple hypothesis detection (MHD). We will apply Schwarz's approximation [79] to the maximum a posteriori (MAP) decision rule, which minimizes the probability of classification error. The MHD framework is also applied naturally to the detection of individual target chemicals. Our simulations indicate that this method gives better detection performance than the spectral unmixing and GLRT approaches. For large libraries, brute force enumeration of the hypotheses becomes infeasible,

and it becomes necessary to either apply prior knowledge or seek approximations to the multiple hypothesis detection approach [66, 64].

3.2 Two General Detection Approaches for Raman Spectroscopy

This section will explore two general detection methods for Raman spectroscopy, the "spectral unmixing" approach and the generalized likelihood ratio test.

3.2.1 Spectral Unmixing Approach

One common general approach is to first estimate the relative amount of each chemical present, and then compare the estimated mixing coefficients to a threshold. The parameter estimation step, which was the subject of Chapter 2, is briefly reviewed in Section 3.2.1.1; the thresholding step is addressed in Section 3.2.1.2.

3.2.1.1 Estimating the Mixing Coefficients

Given an $N \times M$ reference library $\mathbf{A}$ of known spectra $\{\mathbf{a}_1, \cdots, \mathbf{a}_M\}$ and a measured spectrum $\mathbf{y} \in \Re^N$, one approach to estimating the relative quantities of each of the M components is the maximum-likelihood (ML) technique:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{y}; \mathbf{x}), \tag{83}$$

where $\mathbf{x} = (x_1, \cdots, x_M)$ is the unknown vector of mixing coefficients. If the data are generated by

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}, \tag{84}$$

where $\mathbf{w}$ is Gaussian noise, then the ML approach coincides with the least-squares method. If the inherent nonnegativity of the mixing coefficients is taken into account, then the ML estimate under the Gaussian model is found by solving the nonnegative least squares (NNLS) problem

$$\hat{\mathbf{x}}^{NNLS} = \arg\min_{\mathbf{x} \geq \mathbf{0}} ||\mathbf{A}\mathbf{x} - \mathbf{y}||^2. \tag{85}$$

Many common processing algorithms either explicitly or implicitly assume an additive Gaussian noise (AGN) model.

Visual inspection of real Raman spectra indicates that the noise, if thought of as additive, tends to be more positive than negative and that the Gaussian distribution, being symmetric, is not appropriate. In Section 2.2, a model was presented for a dispersive Raman instrument, based on the physics of several key components of the measurement system [62, 63, 88]. A simplified form of the model is given by

$$y_i \sim Pois(\mu_i), \tag{86}$$

where $\mu_i \triangleq (\mathbf{A}\mathbf{x})_i + \mu_i^b$ and μ_i^b is the mean of the "background counts" (primarily due to dark current) on the i^{th} frequency bin of the charge-coupled device (CCD) detector.[1] The values of the μ_i^b are assumed to be known through a calibration measurement. The y_i are assumed to be statistically independent of each other, so the PDF of $\mathbf{y}$ is given by

$$p(\mathbf{y};\mathbf{x}) = \prod_i \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}. \tag{87}$$

Sections 2.3 and 2.6 discussed several iterative methods for computing the ML estimate under this Poisson model. The spectral unmixing simulations in this chapter will use the modified Richardson-Lucy (MRL) algorithm described in Section 2.3. We initialize the MRL algorithm with the NNLS estimate (any zero values in this initial estimate, however, are replaced with a small constant). The MRL algorithm is terminated when $\|\hat{\mathbf{x}}^{new} - \hat{\mathbf{x}}^{old}\|_2 / \|\hat{\mathbf{x}}^{new}\|_2 <$ 1e-5.

3.2.1.2 Determining the Threshold

After the estimates of the mixing coefficients are obtained, each may be compared to a threshold to decide if the corresponding chemical is present. Choosing an appropriate threshold is a vital, and often challenging, task. The usual tradeoff applies: lowering the threshold will increase the probability of detection (P_D), but also increase the probability of false alarm (P_{FA}).

[1] In this simplified model, $\boldsymbol{\mu}^b$ is the only parameter that varies from one device to another. The more detailed model of Section 2.2 includes other parameters that are specific to the sensor, such as the point-spread function of the spectrograph (which affects the measured linewidth) and the nonuniform quantum efficiency of the CCD detector (which affects the relative peak intensities). Those parameters are easily incorporated into the work presented in this chapter but are omitted here for clarity.

For a system to be operational in the field, it must run automatically (i.e., without user intervention), and a typical practical requirement is that it be able to maintain a reasonable false alarm rate. Thus, one common aim is to set the threshold to obtain a given permissable P_{FA}. To find the threshold that achieves a specified P_{FA}, the distribution of the test statistic under the null hypothesis must be known. Detectors that have this property are referred to as constant false alarm rate (CFAR) detectors.

As discussed in Section 3.2.1.1, the ML estimate under the AGN model is simply the LS estimate (assuming no nonnegativity constraint is applied). In this case, $\hat{\mathbf{x}}$ is a linear transformation of the Gaussian vector $\mathbf{y}$, so $\hat{\mathbf{x}}$ is itself normally distributed. Thus, under the AGN model without a nonnegativity constraint, it is trivial to find the threshold that corresponds to a desired P_{FA}. This threshold will generally be different for each mixing coefficient.

Under the Poisson model (86), on the other hand, $\hat{\mathbf{x}}$ is a nonlinear function of the Poisson data, and the distribution of $\hat{\mathbf{x}}$ under the null hypothesis is not known. Most non-Gaussian models have this same limitation, as do most algorithms that constrain $\hat{\mathbf{x}}$ to be nonnegative; while the spectral unmixing approach has the CFAR property for the special case of the AGN model, it does not have the CFAR property in general.

Detection algorithms having the CFAR property do not necessarily have detection performance that is optimal in any way. For example, consider the detector that generates a uniform random number $u \sim U(0,1)$ and then declares a target chemical to be present if $u < \alpha$. For this contrived detector, the probability of false alarm is equal to the threshold α, so this trivial approach has the CFAR property. However, this is clearly a poor detection method because u is not even a function of the data $\mathbf{y}$. In this case, the probability of detection (P_D) is equal to the probability of false alarm. While the CFAR property allows us to set the threshold to achieve a desired P_{FA}, it says nothing about the P_D achieved for that P_{FA}.

Thus, the spectral unmixing approach has two basic shortcomings. First, for most noise models (including the Poisson model), it does not have the CFAR property, so determining a suitable threshold becomes difficult. Second, even if the parameter estimation is optimal

in some sense, the resulting detection algorithm will not necessarily have any particular optimality.

3.2.2 Generalized Likelihood Ratio Test

Another detection method that has been successfully employed on Raman data [85, 84] is the subspace version of the generalized likelihood ratio test (GLRT). Applied to a single target chemical, the subspace GLRT computes how much *worse* you can fit the data if you remove the target from the reference library. For the case of additive Gaussian noise, it considers the following two competing hypotheses:

$$\mathcal{H}_0 : \mathbf{y} = \mathbf{A}^*\mathbf{x}^* + \mathbf{w} \qquad \text{(target absent)} \tag{88a}$$

$$\mathcal{H}_1 : \mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w} \qquad \text{(target present)}\,, \tag{88b}$$

where $\mathbf{A}$ is the full $N \times M$ reference library, $\mathbf{A}^*$ is the $N \times (M-1)$ "reduced" library, $\mathbf{x}$ $(M \times 1)$ and $\mathbf{x}^*$ $((M-1) \times 1)$ are the corresponding mixing vectors, and $\mathbf{w}$ is white Gaussian noise with (unknown) variance σ^2. The GLRT [36] test statistic is then given by

$$T(\mathbf{y}) = \frac{p(\mathbf{y}; \hat{\mathbf{x}}, \hat{\sigma_1^2}, \mathcal{H}_1)}{p(\mathbf{y}; \hat{\mathbf{x}}^*, \hat{\sigma_0^2}, \mathcal{H}_0)} \gtrless \eta\,, \tag{89}$$

where $\hat{\mathbf{x}}$ and $\hat{\sigma_1^2}$ are the ML estimates under $\mathcal{H}_1$, and $\hat{\mathbf{x}}^*$ and $\hat{\sigma_0^2}$ are the ML estimates under $\mathcal{H}_0$. Under the linear Gaussian model, the ML approach yields the LS estimate [35]. Thus, if a detection decision is being made for a single target chemical, the subspace GLRT (for Gaussian $\mathbf{w}$) essentially does the following:

1. Find the LS estimate of $\mathbf{x}$, and compute the resulting squared error e_1.
2. Remove the target of interest from the reference library.
3. Again find the LS estimate for the mixing coefficients (but now there are only $M-1$ of them), and the corresponding squared error e_0.
4. Compare the ratio e_0/e_1 to a threshold to decide if the target of interest is present.

Under the Gaussian model, the GLRT is the uniformly most powerful invariant (UMPI) detector for a somewhat general set of transformations [77].

GLRT for the Poisson Model The GLRT for the Poisson measurement model (86) compares the following two competing hypotheses:

$$\mathcal{H}_0 : \mathbf{y} = (y_1, \ldots, y_N) \qquad \text{(target subset absent)}$$

$$y_i \sim Pois((\mathbf{A}^*\mathbf{x}^*)_i + \mu_i^b) \tag{90a}$$

$$\mathcal{H}_1 : \mathbf{y} = (y_1, \ldots, y_N) \qquad \text{(target subset present)}$$

$$y_i \sim Pois((\mathbf{A}\mathbf{x})_i + \mu_i^b)\,. \tag{90b}$$

The GLRT test statistic is then given by

$$T(\mathbf{y}) = \frac{p(\mathbf{y}; \hat{\mathbf{x}}, \mathcal{H}_1)}{p(\mathbf{y}; \hat{\mathbf{x}}^*, \mathcal{H}_0)} \gtrless \eta\,, \tag{91}$$

where $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}^*$ are the ML estimates under $\mathcal{H}_1$ and $\mathcal{H}_0$, respectively. These estimates may be found using the MRL algorithm of Section 2.3. The likelihood under each hypothesis is given, as in (87), by

$$p(\mathbf{y}; \hat{\mathbf{x}}^*, \mathcal{H}_0) = \prod_i \frac{(\hat{\mu}_i^*)^{y_i} e^{-(\hat{\mu}_i^*)}}{y_i!} \tag{92a}$$

$$p(\mathbf{y}; \hat{\mathbf{x}}, \mathcal{H}_1) = \prod_i \frac{\hat{\mu}_i^{y_i} e^{-\hat{\mu}_i}}{y_i!}, \tag{92b}$$

where

$$\hat{\mu}_i \triangleq (\mathbf{A}\hat{\mathbf{x}})_i + \mu_i^b \tag{93a}$$

$$\hat{\mu}_i^* \triangleq (\mathbf{A}^*\hat{\mathbf{x}}^*)_i + \mu_i^b. \tag{93b}$$

The test statistic is then given as

$$T(\mathbf{y}) = \prod_i \left(\frac{\hat{\mu}_i}{\hat{\mu}_i^*}\right)^{y_i} e^{\hat{\mu}_i^* - \hat{\mu}_i} \tag{94}$$

or as

$$T'(\mathbf{y}) = \ln T(\mathbf{y}) = \sum_i y_i \left[\ln \hat{\mu}_i - \ln \hat{\mu}_i^*\right] + \hat{\mu}_i^* - \hat{\mu}_i. \tag{95}$$

The GLRT for the Poisson model, applied to a single target chemical,[2] is summarized in Algorithm 3.

[2]The GLRT can also be used to detect a *set* of K chemicals; in that case, $\mathbf{A}^*$ would be formed by removing the corresponding K columns of $\mathbf{A}$. In this thesis, we will apply the GLRT to a single chemical at a time.

Algorithm 3 Subspace GLRT for Poisson Model

1: $\hat{\mathbf{x}} = \text{MRL}(\mathbf{y}, \mathbf{A}, \boldsymbol{\mu}^{\boldsymbol{b}})$
2: Form reduced library $\mathbf{A}^*$ by removing the column of $\mathbf{A}$ corresponding to the spectrum of interest.
3: $\hat{\mathbf{x}}^* = \text{MRL}(\mathbf{y}, \mathbf{A}^*, \boldsymbol{\mu}^{\boldsymbol{b}})$
4: Compute test statistic using Equation (95). If greater than threshold, declare the target to be present.

One reason the GLRT is popular is that under the Gaussian model, a monotonic function of (89) yields a detection statistic whose distribution under $\mathcal{H}_0$ is known. This means that for Gaussian data, the threshold can be set to obtain the CFAR property [77, 55, 17]. Unfortunately, this property does not hold in general for other noise models, and the GLRT under the Poisson model is not CFAR for finite data records.

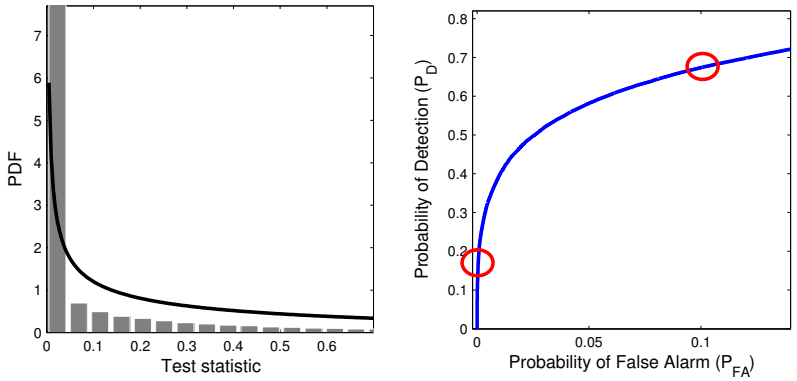
In theory, the GLRT is "asymptotically CFAR;" as $N \to \infty$, the distribution of the GLRT test statistic under $\mathcal{H}_0$ approaches a chi-squared PDF [97, 7]. However, this asymptotic property holds only to the extent that the ML estimate attains its asymptotic distribution $\hat{\mathbf{x}} \sim \mathcal{N}(\mathbf{x}, \mathbf{I}^{-1}(\mathbf{x}))$ [36]. This requirement is not met under the Poisson model or under any other model for which the parameters are constrained to be nonnegative [71]. To see why, note that under $\mathcal{H}_0$, some of the mixing coefficients are zero, while the corresponding estimates, which will be either zero or positive but never negative, will on average be positive and therefore biased. Because some of the parameters lie on the boundary of the parameter space, the distribution of the estimates will be asymmetric and non-Gaussian. As a result, the test statistic does not attain its asymptotic chi-squared distribution.

To illustrate this, we consider a simple scenario, described in more detail in Section 3.3.3, in which the target chemical is not present. We simulated the data according to the Poisson model, and computed the GLRT test statistic for 10,000 different Monte Carlo runs. A histogram of the test statistic values is shown in Figure 14(a) along with the nominal asymptotic chi-squared distribution. For this example, the distribution of the test statistic appears to have much lighter tails than the chi-squared distribution. If the detection threshold is set to obtain a particular P_{FA} *assuming the nominal asymptotic distribution*, then the actual P_{FA} will be much smaller. The resulting P_D will then be much smaller than

what is actually possible for the nominal P_{FA}.

To demonstrate this, we also ran 10,000 runs in which the target chemical *is* present, and generated the resulting receiver operating characteristic (ROC) curve. A section of the ROC curve is shown in Figure 14(b). If the user wishes to obtain a P_{FA} of 0.1, and assumes the chi-squared PDF in the calculation of the threshold, then the resulting threshold ($\eta = 2.71$) will yield an actual P_{FA} of 6e-4. Instead of obtaining the desired (P_{FA}, P_D) point of (0.1, 0.67), this threshold will correspond to the point (6e-4, 0.18). These two points are marked in Figure 14(b).

A more fundamental limitation of the GLRT is that it addresses the binary hypothesis testing problem, while in many applications—such as the detection of chemicals from a known reference library—the problem may be more naturally expressed as one of multiple hypothesis detection. The next section will discuss this in more detail.



(a) Distribution of the GLRT test statistic.

(b) P_D versus P_{FA} for the GLRT.

Figure 14: The distribution of the GLRT test statistic under $\mathcal{H}_0$ is quite different from the nominal asymptotic chi-squared distribution. If the threshold is chosen based on the chi-squared distribution, the resulting detection performance will be much different than anticipated.

3.3 Multiple Hypothesis Detection Framework

This section seeks a more rigorous detection scheme by formulating the problem as one of multiple hypothesis detection (MHD). For example, if there are only two chemicals in the reference library, then there are four possible hypotheses: neither is present, only the first is present, only the second is present, or both are present. In general, for a library

of M spectra, there are $\sum_{k=0}^{M} \binom{M}{k} = 2^M$ hypotheses. This is illustrated in Table 6, which lists the 8 hypotheses for the $M = 3$ case. Our problem is to choose between these 2^M hypotheses.

Table 6: Hypotheses for the $M = 3$ case.

$\mathcal{H}_1$:	0	0	0
$\mathcal{H}_2$:	0	0	+
$\mathcal{H}_3$:	0	+	0
$\mathcal{H}_4$:	0	+	+
$\mathcal{H}_5$:	+	0	0
$\mathcal{H}_6$:	+	0	+
$\mathcal{H}_7$:	+	+	0
$\mathcal{H}_8$:	+	+	+

3.3.1 MAP Decision Rule and Schwarz's Approximation

One approach to the multiple hypothesis testing problem is to use the Bayesian paradigm and seek to minimize the probability of classification error P_e, defined as

$$P_e = \sum_{k=1}^{N_h} p(\text{error}|\mathcal{H}_k)p(\mathcal{H}_k), \tag{96}$$

where $N_h \triangleq 2^M$ is the number of hypotheses. It is well known that P_e is minimized by using the maximum a posteriori (MAP) decision rule [36]: choose the hypothesis $\mathcal{H}_k$ for which

$$p(\mathcal{H}_k|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{H}_k)p(\mathcal{H}_k)}{\sum_{k=1}^{N_h} p(\mathbf{y}|\mathcal{H}_k)p(\mathcal{H}_k)} \tag{97}$$

is maximized (i.e., decide $\mathcal{H}_k$ if $p(\mathcal{H}_k|\mathbf{y}) > p(\mathcal{H}_i|\mathbf{y})$ for all $i \neq k$). If the priors on the hypotheses are equal, this becomes the ML decision rule. Since $\mathbf{x}$ is unknown, the likelihood is given by

$$p(\mathbf{y}|\mathcal{H}_k) = \int p(\mathbf{y}|\mathbf{x}, \mathcal{H}_k)p(\mathbf{x}|\mathcal{H}_k)d\mathbf{x}, \tag{98}$$

assuming a prior on $\mathbf{x}$ is available. This integral is difficult to evaluate in practice, and an approximation is needed. We will use Schwarz's general approach [79], which employs Laplace's method (the "saddle-point approximation") to approximate (98). While we do not have actual priors $p(\mathbf{x}|\mathcal{H}_k)$, we can reasonably assume the prior to be "sufficiently flat" in the regions where the likelihood function $p(\mathbf{y}|\mathbf{x}, \mathcal{H}_k)$ is significant [42]. Under this

assumption, the prior term $p(\mathbf{x}|\mathcal{H}_k) \approx p(\hat{\mathbf{x}}^{ML}|\mathcal{H}_k)$ can be moved outside of the integral, yielding

$$p(\mathbf{y}|\mathcal{H}_k) \approx p(\hat{\mathbf{x}}^{ML}|\mathcal{H}_k) \int p(\mathbf{y}|\mathbf{x}, \mathcal{H}_k) d\mathbf{x}. \tag{99}$$

With the prior removed from the integrand, the remaining integral may now be approximated using Laplace's method [53]:

$$\int p(\mathbf{y}|\mathbf{x}, \mathcal{H}_k) d\mathbf{x} \approx p(\mathbf{y}|\hat{\mathbf{x}}^{ML}, \mathcal{H}_k) \sqrt{\frac{(2\pi)^{n_k}}{\det \mathbf{G}}}\,, \tag{100}$$

where n_k is the number of chemicals (i.e., the number of nonzero mixing coefficients) in hypothesis $\mathcal{H}_k$, and

$$[\mathbf{G}]_{\gamma l} = -\frac{\partial^2}{\partial x_\gamma \partial x_l} \ln\, p(\mathbf{y}|\mathbf{x}, \mathcal{H}_k)\bigg|_{\mathbf{x}=\hat{\mathbf{x}}^{ML}}\,. \tag{101}$$

Combining (99) and (100) and taking the logarithm, we have

$$\ln p(\mathbf{y}|\mathcal{H}_k) \approx \ln p(\hat{\mathbf{x}}^{ML}|\mathcal{H}_k) + \ln p(\mathbf{y}|\hat{\mathbf{x}}^{ML}, \mathcal{H}_k) + \frac{n_k}{2} \ln 2\pi - \frac{1}{2} \ln \det \mathbf{G}\,. \tag{102}$$

The magnitudes of the second and fourth terms grow with N (the number of data points, not to be confused with n_k) and begin to dominate the expression for large N, yielding

$$\ln p(\mathbf{y}|\mathcal{H}_k) \approx \ln p(\mathbf{y}|\hat{\mathbf{x}}^{ML}, \mathcal{H}_k) - \frac{1}{2} \ln \det \mathbf{G}\,. \tag{103}$$

If the priors on the hypotheses are equal, the MAP decision rule selects the hypothesis that maximizes (103).

The first term in (103) is the likelihood assuming the maximum-likelihood estimate is plugged in for the unknown parameter. The second term penalizes higher-order models to avoid overfitting the data. In the spectral unmixing approaches of Section 3.2.1, only the first term was used; however, a threshold is then generally applied to the estimated parameters to eliminate the small ones. This is done because it is intuitively presumed that small mixing coefficients are most likely just fitting the noise; thus, this thresholding is essentially an ad hoc way of penalizing higher-order models. Our proposed technique hopes to provide a more rigorous way to achieve the same effect.

The matrix $\mathbf{G}$ is sometimes known as the "observed" or "empirical" Fisher information matrix. For the Poisson model, plugging (87) into (101) gives [62]

$$[\mathbf{G}]_{\gamma l} = -\sum_i \frac{y_i [\mathbf{A}]_{il} [\mathbf{A}]_{i\gamma}}{\left(\sum_j [\mathbf{A}]_{ij} \hat{x}_j^{ML} + \mu_i^b\right)^2}, \tag{104}$$

where $\hat{\mathbf{x}}^{ML}$ is the ML estimate of $\mathbf{x}$ under hypothesis k.

A major drawback to this method is its computational complexity. While the spectral unmixing approach requires the computation of a single ML estimate, and the GLRT computes two ML estimates per detection decision, the multiple hypothesis detection approach computes 2^M ML estimates. Thus, whereas the spectral unmixing and GLRT methods are suitable for real-time implementation [85], the full MHD method is clearly infeasible for most reference libraries of interest. We will address this in more detail in Section 3.3.5.

3.3.2 Minimum Description Length Interpretation

We must be careful about how we interpret the result of (97). Since we most likely do not have any reasonable priors $p(\mathcal{H}_k)$ (in particular, it is difficult to imagine a scenario in which each hypothesis is equally likely), the final result may not be properly thought of as a typical posterior "probability." An alternative approach is to interpret the result as an abstract score corresponding to a description length, as described below.

In the same year that Schwarz proposed his Bayesian model order criterion discussed above, Rissanen developed a similar formula using a completely different approach, the "Minimum Description Length" (MDL) philosophy [74]. The basic idea behind the MDL principle is that the more we are able to compress the data, the more we have learned about the data and its regularities [26]. Rissanen suggested that the best hypothesis for the data is the one that minimizes the number of bits (or equivalently, nats, the base-e equivalent of base-2 bits) needed to describe the data. He showed that under certain conditions, the minimum number of nats needed to encode the data in a two-part code is given by an equation similar to (103). While the formulas appear to be similar for these two methods, the mindset is completely different: rather than seek the "true" hypothesis that generated the data (an approach that requires specification of the priors), the MDL method merely

evaluates the candidate hypotheses and scores each one based on how efficiently it encodes the data. The resulting description length can then be interpreted as an abstract score for the hypothesis. If desired, a ranked list of the m best hypotheses can be returned as the output.

3.3.3 Example

To illustrate this approach, we consider a simple scenario in which there are five spectra in the reference library. The library spectra were provided by Darren Emge of the Edgewood Chemical Biological Center. Using such a small library (which is necessary to keep the MHD method computationally feasible), an unrealistically high noise level was required to "push" the algorithms. We simulated the data according to the Poisson model (86) and let $\mu_i^b = 5,000$ for all i. Each spectrum in the library was normalized to sum to 10,000.[3] We used 20,000 Monte Carlo runs; on half of the runs, only asphalt (the fourth spectrum of the library) was present, while on the other half, asphalt plus a hazardous chemical (the third column of the library) were both present.

In this test, we assumed uniform priors on the hypotheses and used (103) to approximate the likelihood term. We used the modified Richardson-Lucy (MRL) algorithm to compute the ML estimate required by (103) for each hypothesis. Table 7 lists some of the ranked hypotheses for a particular run in which both substances are present. The two terms of (103) are listed for each hypothesis, along with the total score. The "unpenalized likelihood" and total score are quite large and vary from one hypothesis to another only in the least few significant digits, which are displayed in Table 7. For this particular run, the correct hypothesis has the highest score. The hypothesis containing all five chemicals has the highest unpenalized likelihood, but is penalized enough that it has a lower total score. The MHD approach chose the correct hypothesis on 82% of the runs.

[3]The choice of 10,000 is somewhat arbitrary; making another choice would simply scale the mixing coefficients. In general, mapping the coefficient values to precise physical units is difficult, involving such properties as the exposure time, laser power, atmospheric attenuation, the incident angle of the beam, and other factors, many of which may require careful calibration measurements. Following the pattern of many papers addressing signal processing algorithms for spectral data analysis, we consider these issues to be outside the scope of this thesis.

Table 7: Ranked hypotheses for one particular run.

Ranking	Hypothesis	"Unpenalized Likelihood"	Penalty	Total Score
1	{3,4}	...97.20	4.52	...92.68
2	{2,4}	...96.42	4.78	...91.64
3	{2,3,4}	...97.21	5.70	...91.51
⋮				
12	{1,2,3,4,5}	...97.39	10.61	...86.79
⋮				
31	{ }	...58.43	0	...58.43
32	{5}	...59.17	2.95	...56.21

Table 8 contains the probability of correct classification (P_C) for the MHD method as well as for the exact MAP decision rule and the GLRT. The exact MAP rule compares the posterior probabilities of (97) using the true mixing vector under each hypothesis and the "true" priors $p(\mathcal{H}_k)$ that were used in the simulation. For this example, the "true" priors are $p(\{4\}) = 1/2$, $p(\{3,4\}) = 1/2$, and zero probability for each of the remaining hypotheses. Although the exact MAP rule clearly cannot be implemented in practice (since the true priors and mixing coefficients are unknown), it provides an instructive upper bound on P_C.

Table 8: Probabilities of correct classification.

Method	P_C
Exact MAP	0.95
MHD	0.82
GLRT	0.57

For the GLRT, we computed the Poisson test statistic (95) for each chemical in the library, one by one, and compared the resulting set of M values to a threshold to decide which chemicals are present. As we saw in Section 3.2.2, choosing an appropriate detection threshold for the GLRT is typically a difficult task. Table 8 shows the GLRT at its best; we swept the entire reasonable range of thresholds and used the threshold that resulted in the maximum P_C. Because this optimal threshold will not be known in practice, the actual P_C for the GLRT will generally be lower than the value shown in Table 8.

The MHD method gives worse performance than the exact MAP rule for two reasons: it must make some (generally incorrect) assumption about the hypothesis priors, and it

must estimate the mixing vector under each hypothesis. To analyze how much each of these factors contributes to the classification error, we ran the MHD algorithm using the true priors instead of uniform priors, but still made it deal with the unknown $\mathbf{x}$. This resulted in a P_C of 0.94, which is almost identical to that of the exact MAP rule. This suggests that the classification error is mostly a result of the hypothesis priors being unknown rather than the mixing coefficients being unknown.

This observation offers some insight into the poor performance of the GLRT. The GLRT makes a detection decision for each chemical by framing the problem as one of binary detection. For example, to decide whether or not the third chemical is present, the GLRT tests the hypothesis $\{1, 2, 3, 4, 5\}$ versus the hypothesis $\{1, 2, 4, 5\}$. If some complex combination of the other four spectra is highly correlated with the spectrum of the third chemical, then the GLRT will find that including the third spectrum does not help us to obtain a much better fit of the data than what could be obtained using the other four spectra alone. Unlike the MHD approach, the GLRT does not penalize the alternative that requires a large number of spectra to do what the single spectrum #3 could do on its own.

We note that our MHD method was derived using several approximations and assumptions that may or may not be accurate for a given scenario; the important basic point is that it views the problem as one of multiple hypothesis detection. Other model order estimation techniques may also be applicable here; for example, the multifamily likelihood ratio test (MFLRT) of Kay [37] yields a P_C of 0.76, which is comparable to our MHD method.

3.3.4 Detection of an Individual Target Chemical

Usually, one of two basic goals is sought after by a Raman detection algorithm:

1. Output the best global hypothesis (or a list of the m best hypotheses), where a hypothesis describes which chemicals are present and which are not.

2. Output the detection decision for a given chemical of interest.

As explained in Section 3.3.1, the spectral unmixing approach, when applied to the detection problem, can be viewed as an approximate method to find the best hypothesis. The MAP

decision rule is a statistical-theoretic optimal method to find the best hypothesis. These methods directly seek to solve #1, and use the result to indirectly solve #2 (by assuming the best hypothesis to be true and declaring a given chemical to be present if it is found in the best hypothesis). One drawback to this indirect strategy to solving #2 is that a given chemical may be present in the best hypothesis but absent in the next 10 best hypotheses; even if the target is found in the best hypothesis, it might be *improbable* that it is present. Likewise, even if a target is absent from the best hypothesis, it might be probable that it is present, if it is found in many of the other good hypotheses. Thus, rather than using only the best hypothesis to decide if a given target is present, we might instead take the Bayesian view and estimate the probability that it is present by summing the probabilities of all the hypotheses that contain the target of interest:

$$\Pr(\text{Chem. } j \text{ present}) = \sum_{k \in S_j} p(\mathcal{H}_k|\mathbf{y}), \tag{105}$$

where S_j is the set of indices corresponding to the hypotheses that contain Chem. j, and $p(\mathcal{H}_k|\mathbf{y})$ is given by (97). We can then compare the estimated probability to a threshold to make a hard detection decision on the chemical of interest. The detection rule would then be to decide that Chem. j is present if

$$\sum_{k \in S_j} p(\mathcal{H}_k|\mathbf{y}) > \eta. \tag{106}$$

This detection scheme may be developed more thoroughly as follows. First, we generalize the P_e metric (96) to include different costs for the different classification errors. The expected cost or Bayes risk $\mathcal{R}$ is defined as

$$\mathcal{R} = \sum_{l=1}^{N_h} \sum_{k=1}^{N_h} C_{lk} p(\mathcal{H}_l|\mathcal{H}_k) p(\mathcal{H}_k), \tag{107}$$

where C_{lk} is the cost assigned to the error in which we choose $\mathcal{H}_l$ but $\mathcal{H}_k$ is true. It is well known [36] that $\mathcal{R}$ is minimized by choosing the hypothesis that minimizes

$$C_l(\mathbf{y}) = \sum_{k=1}^{N_h} C_{lk} p(\mathcal{H}_k|\mathbf{y}) \tag{108}$$

over $l = 1, 2, \ldots, N_h$. For example, if we wished to assign an equal penalty to any misclassification, then the costs would be

$$C_{lk} = \begin{cases} 0 & l = k \\ 1 & l \neq k \,, \end{cases} \tag{109}$$

in which case $\mathcal{R} = P_e$. In this case, it is straightforward to show that minimizing (108) is accomplished by maximizing (97). However, if we are only concerned about the detection of a specific target chemical, and do not care about detection errors on other chemicals (i.e., if we wish to solve goal #2 above), then the costs are

$$C_{lk} = \begin{cases} 0 & l, k \in S_j \quad \text{or} \quad l, k \notin S_j \\ 1 & l \in S_j, k \notin S_j \quad \text{or} \quad l \notin S_j, k \in S_j \,. \end{cases} \tag{110}$$

In general, missed detections may be more or less severe than false alarms, so the two could be assigned different penalties. The costs are then

$$C_{lk} = \begin{cases} 0 & l, k \in S_j \quad \text{or} \quad l, k \notin S_j \\ C_{FA} & l \in S_j, k \notin S_j \\ C_{MD} & l \notin S_j, k \in S_j \,. \end{cases} \tag{111}$$

For any hypothesis $\mathcal{H}_l$ containing Chem. j (i.e., $l \in S_j$), (108) becomes

$$C_l(\mathbf{y}) = \sum_{k \notin S_j} C_{FA} p(\mathcal{H}_k|\mathbf{y}). \tag{112}$$

Likewise, for any hypothesis $\mathcal{H}_l$ *not* containing Chem. j (i.e., $l \notin S_j$), (108) becomes

$$C_l(\mathbf{y}) = \sum_{k \in S_j} C_{MD} p(\mathcal{H}_k|\mathbf{y}). \tag{113}$$

In this case, there is no unique hypothesis that minimizes (108); half of the hypotheses result in (112) while the other half result in (113). We will choose any of the hypotheses containing Chem. j if

$$\sum_{k \notin S_j} C_{FA} p(\mathcal{H}_k|\mathbf{y}) < \sum_{k \in S_j} C_{MD} p(\mathcal{H}_k|\mathbf{y}). \tag{114}$$

In other words, Chem. j is declared to be present if

$$C_{MD} \sum_{k \in S_j} p(\mathcal{H}_k|\mathbf{y}) > C_{FA} \left(1 - \sum_{k \in S_j} p(\mathcal{H}_k|\mathbf{y}) \right), \tag{115}$$

or equivalently, if

$$\sum_{k \in S_j} p(\mathcal{H}_k|\mathbf{y}) > \frac{C_{FA}}{C_{FA} + C_{MD}}. \tag{116}$$

This is the decision rule given by (106), with the threshold determined by the relative costs assigned to the Type I and Type II errors. Changing the hypothesis priors will similarly change the threshold.

Up to this point, our general approach has been to make a detection decision on each individual chemical. However, in many applications, the end user may only care if one of a given *class* of chemicals is present. For example, for the detection of hazardous chemical agents, the user might prefer there to be only two possible outputs from the detection algorithm: either one or more hazardous chemicals is present, or no hazardous chemicals are present.

Typically, this problem is solved by first making a detection decision on each individual target chemical, and then declaring a hazardous chemical to be present if any of the hazardous chemicals in the reference library were detected. However, the intermediate step of detection on each individual chemical is unnecessary. One of the advantages of the MHD approach is that it is easily extended to the problem of detecting a chemical from a given set. In this case, we decide that a hazardous chemical is present if

$$\sum_{k \in S_H} p(\mathcal{H}_k|\mathbf{y}) > \frac{C_{FA}}{C_{FA} + C_{MD}}\,, \tag{117}$$

where S_H is the set of indices corresponding to the hypotheses that contain any hazardous chemicals.

Unfortunately, as explained in Section 3.3, we will generally not have actual priors $p(\mathcal{H}_k)$. In this case, the results of the model order selection rule are not really hypothesis probabilities; they may be better interpreted as abstract scores corresponding to description lengths. This method is then perhaps more accurately described as an intuitive way to mix models, where the weight of each model is chosen by the description length score.

3.3.4.1 Detection Performance Evaluation

We now revisit the simple example of Section 3.3.3, in which there are five spectra in the reference library. In this test, we assume uniform priors on the hypotheses and compute the "posterior score" for each of the $2^M = 32$ hypotheses:

$$p(\mathcal{H}_k|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{H}_k)}{\sum_{k=1}^{N_h} p(\mathbf{y}|\mathcal{H}_k)} \tag{118}$$

$$= \frac{e^{\ln p(\mathbf{y}|\mathcal{H}_k)}}{\sum_{k=1}^{N_h} e^{\ln p(\mathbf{y}|\mathcal{H}_k)}} \tag{119}$$

where the loglikelihoods are found (approximately) by (103). To avoid numerical problems, we found it necessary to shift the loglikelihoods:

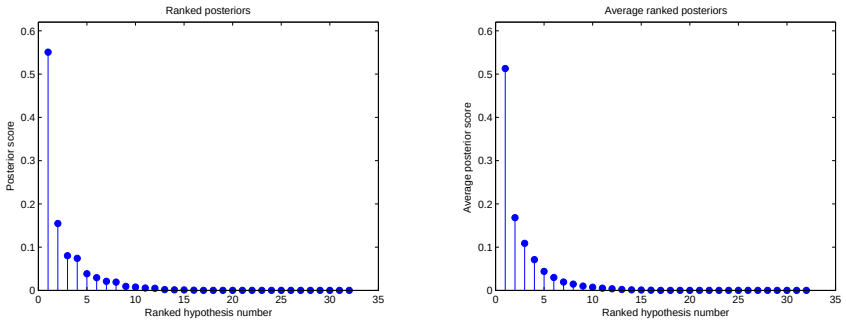
$$p(\mathcal{H}_k|\mathbf{y}) = \frac{e^{-L_0} e^{\ln p(\mathbf{y}|\mathcal{H}_k)}}{e^{-L_0} \sum_{k=1}^{N_h} e^{\ln p(\mathbf{y}|\mathcal{H}_k)}} \tag{120}$$

$$= \frac{e^{\ln p(\mathbf{y}|\mathcal{H}_k) - L_0}}{\sum_{k=1}^{N_h} e^{\ln p(\mathbf{y}|\mathcal{H}_k) - L_0}} \,. \tag{121}$$

We used the largest of the 32 loglikelihoods for L_0.

The ranked hypothesis posteriors for the single MC run of Table 7 are shown in Figure 15(a). The score of the best hypothesis is 55% of the total (summed) scores from all of the hypotheses. A similar plot can be created for each run, and the average ranked posterior scores are shown in Figure 15(b). Note that a particular library entry may be assigned different ranks on different MC runs, so points on the horizontal axis of Figure 15(b) should not be interpreted as corresponding with specific library members, as one could with Figure 15(a). On average, the best hypothesis accounts for over half of the total score, and the best five hypotheses account for over 90% of the total score.

We now analyze the detection performance of the different algorithms by evaluating how well they detect the hazardous chemical in the $M = 5$ scenario discussed above. Figure 16 contains the receiver operating characteristic (ROC) curves for the third chemical in the library, which is present in half of the 20,000 MC runs. The top curve of the ROC plot is for the Neyman-Pearson (NP) detector, which "cheats": we tell it that only one of two hypotheses are possible, and we also tell it the true mixing vector under each hypothesis, so it just has to decide which of the two is correct. The NP detector is an upper bound on the ROC curve for any algorithm.



(a) Ranked posterior scores for one particular MC run.

(b) Ranked posterior scores averaged over the runs.

Figure 15: Ranked posterior hypothesis scores, $p(\mathcal{H}_k|\mathbf{y})$.

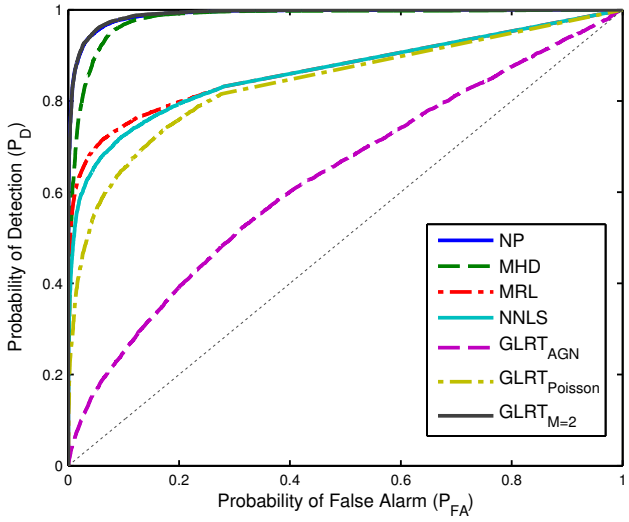


Figure 16: Probability of detection versus probability of false alarm for the five detectors and the Neyman-Pearson bound.

Spectral unmixing using the modified Richardson-Lucy (MRL) algorithm performs only slightly better in these tests than spectral unmixing using NNLS, but both outperform the GLRTs (both the AGN and Poisson versions). It appears that choosing an appropriate general detection philosophy (for example, spectral unmixing versus GLRT) makes a greater difference than using the correct noise model.

The "Multiple Hypothesis Detection" (MHD) method, whose detection rule is given by (106), shows the best performance of any of the algorithms and comes the closest to reaching the NP bound.

The bottom curve, representing the conventional subspace GLRT (89), shows by far the

worst performance in this test. This is largely because the data was simulated using the Poisson measurement model, while the conventional GLRT assumes an additive Gaussian noise model with no nonnegativity constraint on the mixing coefficients. This merely illustrates that the conventional GLRT is not appropriate when the model for which it was derived is not accurate. The GLRT for the Poisson model, represented by the second-to-bottom curve, shows significant improvement.

We note that, regardless of the performance shown in Figure 16, these methods may not be useful in practice without an appropriate strategy for choosing the detection threshold. As discussed in Section 3.1, threshold selection is generally a difficult task for the spectral unmixing (MRL and NNLS) and GLRT approaches.

The GLRT performs relatively poorly here because it is deciding between the wrong two hypotheses. The true hypothesis on any run is either $\{3, 4\}$ (if the target chemical is present) or $\{4\}$ (if the target chemical is absent), but the GLRT tests the hypothesis $\{1, 2, 3, 4, 5\}$ versus the hypothesis $\{1, 2, 4, 5\}$. If we knew a priori that only the third and fourth chemicals would ever be present, we could run the GLRT using the smaller library consisting of only these two elements. In this case, the GLRT would be deciding between the correct two hypotheses, so the error would be entirely caused by not knowing $\mathbf{x}$ under each hypothesis. The resulting ROC curve, labeled "GLRT$_{M=2}$" in Figure 16, matches up almost exactly with the NP bound, suggesting that most of the error from the GLRT is a result of the wrong hypotheses being tested rather than the mixing coefficients being unknown. Of course, reference libraries used in practice will typically include more than just these two spectra. Other chemicals are included in the library precisely because they have some nonzero prior probability of being present; if a chemical has no chance of being present, then there would be no reason to include it in the library as it would only increase the confusion. If we are detecting chemicals from a reference library, then we are inherently dealing with a multiple hypothesis detection problem.

3.3.5 Applying Prior Knowledge

Since the number of hypotheses grows exponentially with the number of spectra in the library, this multiple hypothesis detection method clearly becomes infeasible for most reference libraries of interest. However, the hypothesis space can be reduced by applying realistic prior information. For example, in the detection of surface contaminants, it might be reasonable to assume that two hazardous chemicals will not both be present in a given measurement. In the Bayesian framework, this prior knowledge will be taken into account by setting $p(\mathcal{H}_k) = 0$ for any hypothesis containing multiple hazardous chemicals. Alternatively, in the MDL framework, this prior knowledge limits the number of candidate models that we try. By reducing the hypothesis space, this prior information will lead to improvements in both detection performance and computational feasibility. In particular, if T is the number of target chemicals in the library and M is the total number of chemicals in the library, then this prior knowledge reduces the number of hypotheses from 2^M to $(T+1)2^{M-T} = 2^{M-(T-\log_2(T+1))}$. For example, if $M = 40$ and $T = 28$, this reduces the number of hypotheses from approximately 1e12 to approximately 1e6. We note that this is still optimal; no approximations have been made while incorporating this prior knowledge. Even with this reduction, however, the number of hypotheses still grows exponentially with $(M - T)$, and the problem will still become infeasible for larger libraries.

One way to address this issue is to further reduce the number of candidate hypotheses by assuming that all non-hazardous chemicals may be present. For example, if there are two hazardous chemicals in our 5-element library (i.e., $M = 5$ and $T = 2$), then this approach would evaluate only the $T + 1 = 3$ hypotheses illustrated in Table 9.

Table 9: Candidate hypotheses after prior knowledge has been applied.

$\mathcal{H}_1$:	+	+	+	0	0
$\mathcal{H}_2$:	+	+	+	+	0
$\mathcal{H}_3$:	+	+	+	0	+

An alternative general approach to address this issue is to apply *approximate* methods that do not rely on exhaustive enumeration of all the hypotheses. For example, if the ML

estimate in our $M = 5$ example is $\hat{\mathbf{x}} = (0.3, 0, 0.1, 0.6, 0)^T$, then we could begin searching the hypothesis space by "sweeping the threshold" and evaluating the resulting candidate hypotheses $\{\}$, $\{4\}$, $\{1, 4\}$, and $\{1, 3, 4\}$. The highest-scoring of these hypotheses could then be used as an initial point from which the hypothesis space is searched; other hypotheses that are "close" to it in some sense could be evaluated, and so on.

We leave a full analysis of these approaches to reduce the number of candidate hypotheses (by applying prior knowledge and/or using approximate methods) as a major topic for future research.

3.4 Conclusions

This chapter presented a multiple hypothesis detection framework for identifying chemicals present in a measured Raman spectrum. We applied Schwarz's approximation to the MAP decision rule for a Poisson Raman spectroscopy model. For a small reference library, this method gave better detection performance than the spectral unmixing and GLRT approaches.

A fundamental limitation of the GLRT is that it addresses the binary hypothesis testing problem, while in many applications—such as the detection of chemicals from a known reference library—the problem may be more naturally expressed as one of multiple hypothesis detection. Our simulation results suggest that most of the error in the GLRT arises from the wrong hypotheses being tested rather than the mixing coefficients being unknown. It is thus unsurprising that the simple spectral unmixing approach performed better in our simulations than the GLRT. Furthermore, because of the nonnegativity of the parameters, the GLRT does not generally attain its nominal asymptotic performance; it does not generally have the CFAR property. The GLRT is not recommended for the problem of detecting chemicals from a known reference library.

Because the number of hypotheses grows exponentially with the size of the library, the exhaustive enumeration of hypotheses is infeasible for the large libraries typically used in practice. However, we can significantly reduce the hypothesis space by applying realistic

prior knowledge. Furthermore, even when it cannot be employed directly as a real-time algorithm, the MHD framework provides basic insight about the Raman detection problem; for example, the spectral unmixing approaches can be viewed as an approximate method to seek the best hypothesis. Future work could focus on developing other approximations that do not rely on brute force enumeration of the hypotheses.

CHAPTER IV

ACCOUNTING FOR UNKNOWN CHEMICALS

The previous chapters examined the *supervised* detection framework, in which the measured Raman spectrum is compared to a reference library of known spectra. A well-known shortcoming of the supervised approach is that no comprehensive library exists, and when chemicals are present that are not in the library, the supervised algorithms may confuse those chemicals with library members.

One way to deal with this problem is to use an *unsupervised* method such as nonnegative matrix factorization (NMF) to estimate both the constituent spectra and their relative quantities directly from a block of measured spectra. Chemical identification may then be performed by associating the extracted spectra with the reference library spectra. This two-stage NMF approach often fails because knowledge of the reference library was not used in extracting the spectra.

To address these issues, this chapter presents an alternative *partially-supervised* framework. Because this framework is applicable to a variety of signal processing problems, we present it here as a general new approach, and consider Raman spectroscopy as a specific example application. We begin by briefly discussing the general field of unsupervised data analysis.

4.1 Introduction

For many signal processing applications, the amount of available data has steadily grown in recent years because of various technological advances. Data analysis methods have increasingly employed low-rank approximations to reduce the number of variables and detect structure in the data. Many of these methods are *unsupervised*, or *data-adaptive*, in that the representations are learned directly from the particular data being analyzed (as

opposed to, for example, Fourier analysis or linear regression, whose bases are not tailored to the observed data) [30]. Principal component analysis (PCA) [81], independent component analysis (ICA) [32], and nonnegative matrix factorization (NMF) [61, 45] are common techniques belonging to this class of unsupervised subspace approximations. Nonnegative matrix factorization differs from the other methods in that all of the elements in the factorization are constrained to be nonnegative; because it does not allow subtractive combinations, NMF is a natural approach for many problems involving physical quantities that mix only additively. We will briefly review NMF and its properties in Section 4.2; for a more thorough overview, see [1].

For many machine learning problems, the extracted basis vectors in these subspace approximations do not correspond to known physical quantities. For example, the principal components in PCA are chosen to best capture the variation in the data; while this makes PCA an effective tool for dimensionality reduction, the resulting principal components (e.g., eigenfaces for face images) do not necessarily correspond to physically meaningful constituent components of the data.

However, for other problems, we may seek basis vectors that do have physical significance. In spectral data analysis, for instance, the extracted basis vectors may correspond to the spectral signatures of the constituent materials [68, 75, 1]. For these cases, one often desires to detect whether any signatures from a known library are present, and/or estimate the relative quantities of each. This chapter addresses this problem. One straightforward approach to perform detection for a given target in the library is to compare its known signature with the extracted bases estimated by an NMF algorithm. Section 4.3 will show that this two-stage NMF approach often performs poorly because knowledge of the reference library was not used in extracting the basis vectors.

Section 4.4 presents a novel version of NMF in which a subset of the extracted spectra are constrained to be equal to the known reference library. This algorithm is applicable to any problem in which an object is identified by comparing a block of measured data to a library of known constituent signatures; the example simulations in this chapter consider the spectral unmixing of Raman spectroscopy data. In our simulations, this method performs

better than the two-stage NMF approach and the fully supervised approach when there are chemicals present that are not in the library.

4.2 Nonnegative Matrix Factorization (NMF)

The goal of nonnegative matrix factorization [61, 45] is to approximate an (elementwise) nonnegative data matrix $\mathbf{Y}$ with a lower-rank factorization

$$\mathbf{Y} \approx \mathbf{W}\mathbf{X}, \tag{122}$$

where all of the elements of $\mathbf{W}$ and $\mathbf{X}$ are constrained to be nonnegative. Each column of $\mathbf{Y}$ typically represents an object, and NMF approximates each as a nonnegative linear combination of the columns of $\mathbf{W}$. The columns of $\mathbf{W}$ can thus be interpreted as *basis vectors* for $\mathbf{Y}$, and each column of $\mathbf{X}$ is the *mixing vector*, or *encoding*, for the corresponding column in $\mathbf{Y}$. The number of basis vectors is specified by the user in advance. The general NMF problem is given by

$$(\hat{\mathbf{W}}, \hat{\mathbf{X}}) = \arg\min_{\mathbf{W},\mathbf{X}} \Psi(\mathbf{W}, \mathbf{X}) \qquad \text{s.t.} \;\; [\mathbf{W}]_{ij} \geq 0, [\mathbf{X}]_{j\gamma} \geq 0, \;\; \forall i, j, \gamma \,, \tag{123}$$

where the objective function $\Psi(\mathbf{W}, \mathbf{X})$ measures the discrepancy between the data $\mathbf{Y}$ and the product $\mathbf{W}\mathbf{X}$. The most common choice of objective function is the squared error[1]

$$\Psi(\mathbf{W}, \mathbf{X}) = \|\mathbf{Y} - \mathbf{W}\mathbf{X}\|_F^2, \tag{124}$$

although other cost functions such as Csiszár's I-divergence (a generalized form of the Kullback-Leibler divergence) are also common [45, 9]. Along with the data mismatch term, the objective function often contains regularization terms, usually to impose prior knowledge about the particular application. For example, auxiliary constraints such as sparseness [30] and smoothness [1] have been incorporated into the NMF cost function, with improved results when applied in appropriate cases.

[1] It is well known that the singular value decomposition finds the factorization of a given rank that minimizes the squared error; however, the resulting factorization will generally contain negative values, which in many applications contradicts physical realities.

There are a number of methods to minimize the objective function. One common approach is the gradient descent technique, for which each iteration is defined by

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} - \alpha_{ij}^{(n)} \frac{\partial}{\partial [\mathbf{W}]_{ij}} \Psi(\mathbf{W}, \mathbf{X}) \Big|_{\mathbf{W}=\hat{\mathbf{W}}^{(n)}, \mathbf{X}=\hat{\mathbf{X}}^{(n)}} \tag{125a}$$

$$[\hat{\mathbf{X}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{X}}^{(n)}]_{j\gamma} - \delta_{j\gamma}^{(n)} \frac{\partial}{\partial [\mathbf{X}]_{j\gamma}} \Psi(\mathbf{W}, \mathbf{X}) \Big|_{\mathbf{W}=\hat{\mathbf{W}}^{(n)}, \mathbf{X}=\hat{\mathbf{X}}^{(n)}} . \tag{125b}$$

The step sizes $\alpha_{ij}^{(n)}$ and $\delta_{j\gamma}^{(n)}$ are chosen differently for different gradient descent algorithms. In general, little can be said about convergence [1]. Also, special care must be taken to ensure that the updated matrices remain nonnegative. However, for certain objective functions (such as the squared error and the I-divergence), $\alpha_{ij}^{(n)}$ and $\delta_{j\gamma}^{(n)}$ can be chosen such that (125) becomes a multiplicative update rule [46, 45, 75]:

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} [f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)})]_{ij} \tag{126a}$$

$$[\hat{\mathbf{X}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{X}}^{(n)}]_{j\gamma} [g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)})]_{j\gamma} \tag{126b}$$

for some functions f and g that depend on the objective function. This is expressed in matrix form (using MATLAB notation, in which .* represents elementwise multiplication) as

$$\hat{\mathbf{W}}^{(n+1)} \leftarrow \hat{\mathbf{W}}^{(n)} .* f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)}) \tag{127a}$$

$$\hat{\mathbf{X}}^{(n+1)} \leftarrow \hat{\mathbf{X}}^{(n)} .* g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)}) . \tag{127b}$$

Because of the multiplicative form of the update rules, the estimates at any iteration will now clearly be nonnegative, provided that the initial estimates are positive. A parameter initialized to zero will stay at zero, so it is important to initialize with positive estimates. The general form of the multiplicative update algorithms is presented in Algorithm 4. The multiplicative update approach is quite common due to its ease of implementation.

Example: Maximum-likelihood estimation under a Poisson data model If there is a probabilistic model for the data, then NMF can be used to seek the maximum-likelihood (ML) estimates of $\mathbf{W}$ and $\mathbf{X}$:

$$(\hat{\mathbf{W}}, \hat{\mathbf{X}}) = \arg\max_{\mathbf{W},\mathbf{X}} p(\mathbf{Y}; \mathbf{W}, \mathbf{X}). \tag{128}$$

Algorithm 4 General form of multiplicative update NMF algorithms. M_W is the number of columns of $\mathbf{W}$, and f and g are the update equations (derived to minimize the objective function).

Input: $\mathbf{Y}, M_W$
Output: $\hat{\mathbf{W}}, \hat{\mathbf{X}}$

1: Initialize with some positive $\hat{\mathbf{W}}^{(0)}, \hat{\mathbf{X}}^{(0)}$
2: $n \leftarrow 0$
3: **while** not converged **do**
4: $\quad \hat{\mathbf{W}}^{(n+1)} \leftarrow \hat{\mathbf{W}}^{(n)} .* f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)})$
5: $\quad \hat{\mathbf{X}}^{(n+1)} \leftarrow \hat{\mathbf{X}}^{(n)} .* g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)})$
6: $\quad n \leftarrow n + 1$
7: **end while**
8: $\hat{\mathbf{W}} \leftarrow \hat{\mathbf{W}}^{(n+1)}; \hat{\mathbf{X}} \leftarrow \hat{\mathbf{X}}^{(n+1)}$

For example, consider the Poisson noise model

$$[\mathbf{Y}]_{i\gamma} \sim Pois([\mathbf{WX}]_{i\gamma}). \tag{129}$$

If the data $[\mathbf{Y}]_{i\gamma}$ are assumed to be statistically independent of each other, the PDF of the data is given by

$$p(\mathbf{Y}; \mathbf{W}, \mathbf{X}) = \prod_{i,\gamma} \frac{([\mathbf{WX}]_{i\gamma})^{[\mathbf{Y}]_{i\gamma}} \, e^{-[\mathbf{WX}]_{i\gamma}}}{[\mathbf{Y}]_{i\gamma}!}. \tag{130}$$

The ML estimate is then given by

$$(\hat{\mathbf{W}}, \hat{\mathbf{X}}) = \arg\max_{\mathbf{W},\mathbf{X}} \ln p(\mathbf{Y}; \mathbf{W}, \mathbf{X}) \qquad \text{s.t.} \quad [\mathbf{W}]_{ij} \geq 0, [\mathbf{X}]_{j\gamma} \geq 0, \ \ \forall i, j, \gamma \tag{131a}$$

$$= \arg\max_{\mathbf{W},\mathbf{X}} \sum_{i,\gamma} [\mathbf{Y}]_{i\gamma} \ln\left([\mathbf{WX}]_{i\gamma}\right) - [\mathbf{WX}]_{i\gamma} - \ln[\mathbf{Y}]_{i\gamma}! \qquad \text{s.t.} \quad [\mathbf{W}]_{ij} \geq 0, [\mathbf{X}]_{j\gamma} \geq 0, \ \ \forall i, j, \gamma \tag{131b}$$

$$= \arg\min_{\mathbf{W},\mathbf{X}} \Psi(\mathbf{W}, \mathbf{X}) \qquad \text{s.t.} \quad [\mathbf{W}]_{ij} \geq 0, [\mathbf{X}]_{j\gamma} \geq 0, \ \ \forall i, j, \gamma\,, \tag{131c}$$

which is the NMF problem with the objective function

$$\Psi(\mathbf{W}, \mathbf{X}) = \sum_{i,\gamma} [\mathbf{WX}]_{i\gamma} - [\mathbf{Y}]_{i\gamma} \ln\left([\mathbf{WX}]_{i\gamma}\right). \tag{132}$$

This is the same objective function used to minimize the I-divergence from $\mathbf{Y}$ to $\mathbf{WX}$. The resulting multiplicative update rule is given by [75, 45, 46]

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} \frac{\sum_{\gamma'} \frac{[\mathbf{Y}]_{i\gamma'}}{[\hat{\mathbf{W}}^{(n)}\hat{\mathbf{X}}^{(n)}]_{i\gamma'}} [\hat{\mathbf{X}}^{(n)}]_{j\gamma'}}{\sum_{\gamma'} [\hat{\mathbf{X}}^{(n)}]_{j\gamma'}} \tag{133a}$$

$$[\hat{\mathbf{X}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{X}}^{(n)}]_{j\gamma} \frac{\sum_{i'} \frac{[\mathbf{Y}]_{i'\gamma}}{[\hat{\mathbf{W}}^{(n)}\hat{\mathbf{X}}^{(n)}]_{i'\gamma}} [\hat{\mathbf{W}}^{(n)}]_{i'j}}{\sum_{i'} [\hat{\mathbf{W}}^{(n)}]_{i'j}}. \tag{133b}$$

This is expressed in matrix form as

$$\hat{\mathbf{W}}^{(n+1)} \leftarrow \hat{\mathbf{W}}^{(n)} .* \left[\left(\mathbf{Y} ./ (\hat{\mathbf{W}}^{(n)} \hat{\mathbf{X}}^{(n)}) \right) (\hat{\mathbf{X}}^{(n)})^T \right] ./ \left(\mathbf{1}_{N \times K} (\hat{\mathbf{X}}^{(n)})^T \right) \tag{134a}$$

$$\hat{\mathbf{X}}^{(n+1)} \leftarrow \hat{\mathbf{X}}^{(n)} .* \left[(\hat{\mathbf{W}}^{(n)})^T \left(\mathbf{Y} ./ (\hat{\mathbf{W}}^{(n)} \hat{\mathbf{X}}^{(n)}) \right) \right] ./ \left((\hat{\mathbf{W}}^{(n)})^T \mathbf{1}_{N \times K} \right) , \tag{134b}$$

where $\mathbf{1}_{N \times K}$ is an all-one matrix with the same dimensions as $\mathbf{Y}$. A small positive constant (e.g., 10^{-9}) may be added to the denominators to avoid division by zero [68, 1].

Like the squared-error objective function, the I-divergence is convex in either $\mathbf{W}$ or $\mathbf{X}$ separately, but non-convex in both $\mathbf{W}$ and $\mathbf{X}$ together [46]. This means there may be multiple local minima and/or saddle points, making it difficult to find a global minimum. The multiplicative update algorithm for the I-divergence cost function has been shown to globally converge to a stationary point, which is a necessary condition for a local minimum [20, 50]. In contrast, this property has not been shown for the multiplicative update algorithm for the squared-error cost function [1, 51, 50].

The next section will present the spectral unmixing problem, which is an example application of NMF.

4.3 Spectral Unmixing of Raman data

In many signal processing applications, the *linear mixing model* is assumed for each data vector $\mathbf{y}$:

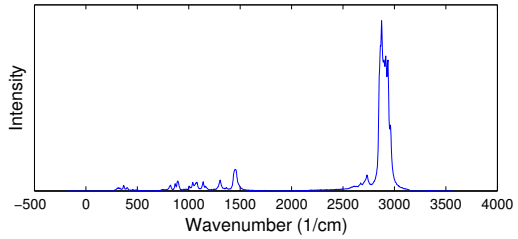
$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}, \tag{135}$$

where $\mathbf{A}$ is the component matrix, $\mathbf{x}$ is a vector of mixing coefficients, and $\mathbf{n}$ represents the measurement noise. For example, this general model is often used in spectral data analysis, in which case $\mathbf{y} \in \Re^N$ is the measured spectrum of a sample, $\mathbf{A}$ is an $N \times M$ reference library of constituent spectra $\{\mathbf{a}_1, \ldots, \mathbf{a}_M\}$, or *endmembers*, and $\mathbf{x} \in \Re^M$ is a vector of *fractional abundances* describing how much of each endmember is present in the sample mixture [38]. This thesis considers the analysis of Raman spectra, but the basic principles and techniques described here also apply more generally to other problems that use the linear mixing model. In Section 2.2, a model was presented for a dispersive Raman instrument, based on

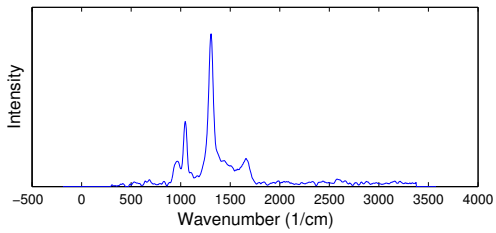
the physics of several key components of the measurement system [62, 63, 88]. A simplified form of the model[2] is given by

$$y_i \sim Pois([\mathbf{Ax}]_i), \tag{136}$$

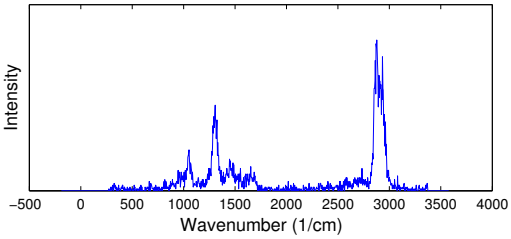
where the y_i are assumed to be statistically independent of each other. To illustrate the Poisson sensor model, Figures 17(a) and 17(b) show the clean spectra of two different substances, and Figure 17(c) shows the simulated noisy spectrum corresponding to the mixture of the two chemicals. The Raman spectra of both chemicals belong to a known reference library of 26 spectra, which was provided to us by Darren Emge of the Edgewood Chemical Biological Center.



(a) "True" spectrum of Chem. 7.



(b) "True" spectrum of Chem. 26.



(c) Simulated noisy spectrum of mixture.

Figure 17: Clean spectra of constituent materials and a simulated noisy spectrum of their mixture.

4.3.1 Supervised Approach

If the reference library $\mathbf{A}$ is assumed to be completely known, then the ML estimate

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{y}; \mathbf{x}) \tag{137}$$

[2]The other parameters in the model are easily incorporated into the work presented in this chapter, but are omitted here for simplicity.

can be found using the Richardson-Lucy (RL) algorithm[3] [73, 52], a well-known iterative method equivalent to the expectation-maximization (EM) technique [80, 88]. Each iteration of the RL algorithm is defined by

$$\hat{x}_j^{(n+1)} \leftarrow \hat{x}_j^{(n)} \left[\sum_i \frac{y_i}{[\mathbf{A}\hat{\mathbf{x}}^{(n)}]_i} [\mathbf{A}]_{ij} \right] / \sum_i [\mathbf{A}]_{ij} \,. \tag{138}$$

The likelihood function is concave in $\mathbf{x}$, and the sequence of estimates will converge toward the ML estimate. Also, as with the NMF algorithms of the previous section, the estimates at any iteration will clearly be nonnegative if the initial estimates are positive. In fact, the EM iteration (138) is equivalent to (133b); like the multiplicative update NMF algorithm, the RL algorithm can be viewed as a gradient descent method (with a particular choice of step size) to maximize the likelihood of the Poisson data. The only difference is that for the *supervised* problem, the constituent spectra are assumed to be known and only the mixing coefficients need to be estimated.

To illustrate this approach, we consider an example in which the two chemicals from Figure 17 are present in a given shot. We simulated the data according to the Poisson model (136), and each spectrum in the library was normalized to sum to 10,000.[4] Each of the 26 spectra in the reference library has 1024 frequency samples ($N = 1024$). The RL algorithm was terminated when $\|\hat{\mathbf{x}}^{(n+1)} - \hat{\mathbf{x}}^{(n)}\|_2 / \|\hat{\mathbf{x}}^{(n+1)}\|_2 <$ 1e-5 (Similarly, the NMF algorithms presented later in this chapter are terminated when $\|\hat{\mathbf{X}}^{(n+1)} - \hat{\mathbf{X}}^{(n)}\|_F / \|\hat{\mathbf{X}}^{(n+1)}\|_F <$ 1e-5).

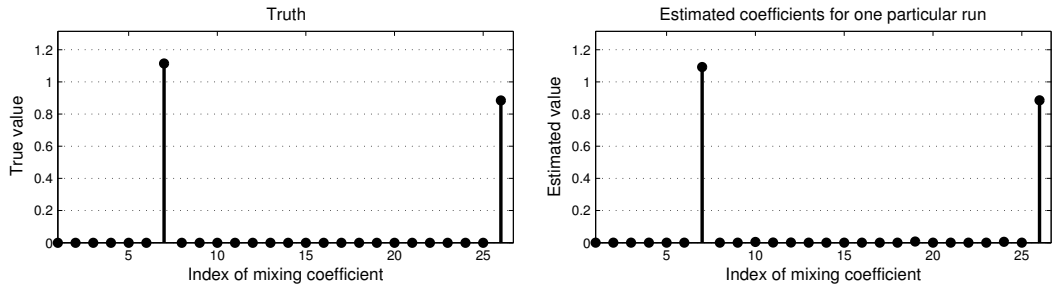
The "true" mixing vector $\mathbf{x}$ is shown in Figure 18(a), and the estimate given by the RL algorithm is shown in Figure 18(b). The estimated mixing coefficients are seen to be almost identical to the true values. To show that this result holds for other noise instantiations and not just for this particular run, we computed the RL estimate for 100 different Monte Carlo runs. The resulting sample standard deviation, sample bias, and sample root mean

[3] An alternative approach, which often requires fewer iterations, is the nonnegative iteratively reweighted least-squares (NNIRLS) algorithm discussed in Section 2.6.

[4] The choice of 10,000 is somewhat arbitrary; making another choice would simply scale the mixing coefficients. In general, mapping the coefficient values to precise physical units is difficult, involving such properties as the exposure time, laser power, atmospheric attenuation, the incident angle of the beam, and other factors, many of which may require careful calibration measurements. Following the pattern of many papers addressing signal processing algorithms for spectral unmixing, we consider these issues to be outside the scope of this thesis.

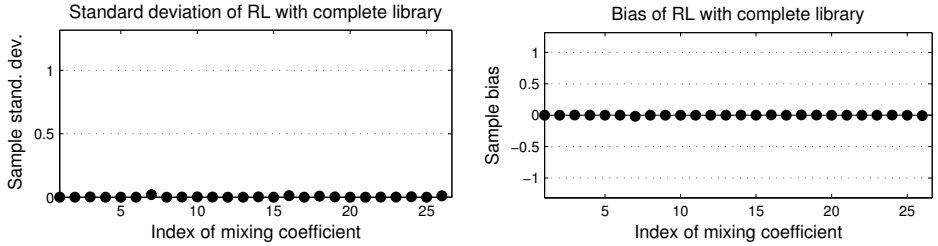
square error (RMSE) are plotted in Figure 19, where

$$\begin{aligned}\mathrm{RMSE}(\hat{x}_j) &\triangleq \sqrt{E\left[(\hat{x}_j - x_j)^2\right]} \\ &= \sqrt{\mathrm{Var}(\hat{x}_j) + [\mathrm{Bias}(\hat{x}_j)]^2}.\end{aligned}$$

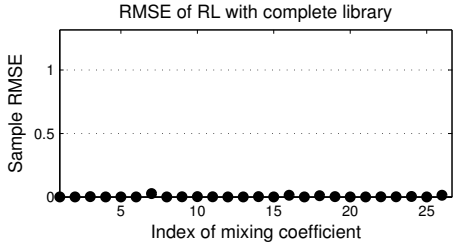


(a) "True" mixing coefficients. (b) Values estimated by the RL algorithm.

Figure 18: Estimates given by the RL algorithm when the library is comprehensive.



(a) Standard Deviation. (b) Bias.

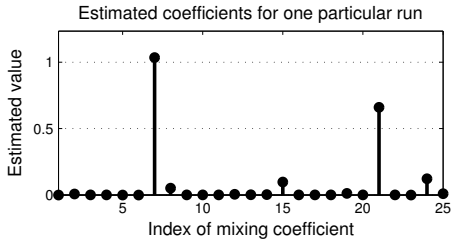


(c) RMSE.

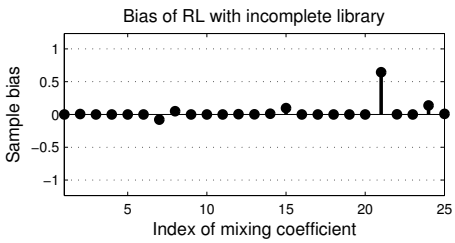
Figure 19: Sample standard deviation, bias, and RMSE of the RL algorithm when the library is comprehensive.

As shown in Figure 19(c), the RMSE is practically zero for all the mixing coefficients; this approach typically gives good performance when the reference library contains all of the constituent chemicals. The scale in Figure 19 was chosen to match that used in Figure 20 to permit easy comparison.

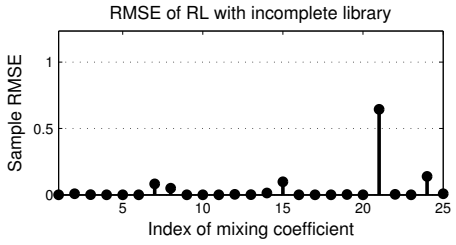
We now consider a case in which the reference library used by the RL algorithm is incomplete. For this example, the same two chemicals are present in the measurement sample as before, and the relative quantities of each are again shown in Figure 18(a). However, the library **A** used by the RL algorithm now contains only the first 25 spectra; one of the two chemicals present is not in the library. The estimate given by the RL algorithm on one particular run is shown in Figure 20(a). The algorithm confuses the "unknown" chemical with Chem. 21 because their spectra are highly correlated, as shown in Figure 21. Figures 20(b) and 20(c) contain the sample bias and sample RMSE computed using 100 Monte Carlo runs (all remaining Monte Carlo simulations in this chapter likewise use 100 runs); the variance was negligible compared to the bias and is omitted here for brevity. There is a positive bias on several mixing coefficients, particularly the 21st, because the algorithm must attribute the energy from the unknown chemical to library elements.



(a) Estimates for one particular run.



(b) Bias.

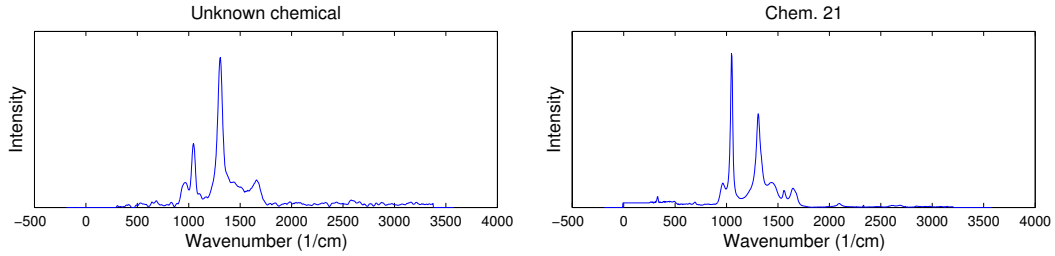


(c) RMSE.

Figure 20: Estimation performance of the RL algorithm when the library is incomplete.

4.3.2 Two-Stage NMF Approach

The major problem with the supervised approach is that it requires complete knowledge of all possible constituent spectra. Such a comprehensive library does not exist, and even if it did, the supervised algorithms would then become infeasible because of the size of the



(a) Unknown chemical. (b) Chem. 21.

Figure 21: RL algorithm confuses the "unknown" chemical with Chem. 21 because the corresponding spectra are similar.

library. As we saw in the previous section, when there are chemicals present that are not contained in the reference library, the supervised algorithms may confuse those chemicals with library members.

One way to deal with this problem is to use an unsupervised algorithm such as NMF as the first step in a two-stage scheme:

1. Given an $N \times K$ block of spectra $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_K]$ collected over K pulses, estimate both the constituent spectra ($\mathbf{W}$) and the mixing matrix ($\mathbf{X}$). For this to be possible, the chemical composition must vary from pulse to pulse, as may be the case, for example, if the sensor is moving along the ground. For the Poisson model (129), the ML estimates for $\mathbf{W}$ and $\mathbf{X}$ may be found using the multiplicative update NMF algorithm given by (133). In a language like MATLAB, the function call would have a form like $[\hat{\mathbf{W}}, \hat{\mathbf{X}}] = \text{NMF}(\mathbf{Y}, M_W)$, where M_W is the number of columns in $\mathbf{W}$.

2. If a column of $\hat{\mathbf{W}}$ is highly correlated with a column from the reference library, use the corresponding estimated mixing coefficients (from $\hat{\mathbf{X}}$) as the estimates for the mixing coefficients for that library chemical.

This basic approach is not new; the nonnegative nature of spectral mixing has inspired the application of various NMF-based approaches for spectral data analysis (e.g., [68, 75, 48]), and more sophisticated versions of this 2-stage NMF method have been presented in the literature (e.g., [68, 1]).

Even if one of the extracted spectra has a large estimated mixing coefficient, it will

still not lead to a target detection unless it correlates well with one of the spectra in the reference library. Thus, it would seem that this NMF-based approach should have fewer false alarms than the supervised unmixing methods when chemicals are present that are not in the library. However, there are several challenges in implementing this method. Deciding whether a column of $\hat{\mathbf{W}}$ is "similar enough" to a column of $\mathbf{A}$ (step 2) is typically accomplished by comparing some measure of similarity (such as the correlation or the symmetrized Kullback-Leibler divergence [68]) to a threshold. The choice of threshold is critical because it will determine the number of false alarms and missed detections. Another critical issue, which appears to be often overlooked in the literature, is how to determine an appropriate value for M_W.

Even if these parameters are chosen optimally for a given problem, we have found that this approach often performs poorly. To illustrate the shortcomings of this method, consider the example scenario represented in Figure 22. In this scenario, 40 shots of data are collected, and the two chemicals from Figure 17 are present in each shot. These pulses correspond to different physical locations; we imagine a scenario in which a truck with a Raman sensor is driving along a road. The first chemical, marked by the triangle, is the 7th element of the known library; its spectrum is shown in Figure 17(a). The second chemical, marked by the square, is not found in the library used here; its spectrum is shown in Figure 17(b). The true mixing coefficients on the 25th shot are the same as the previous example shown in Figure 18(a).

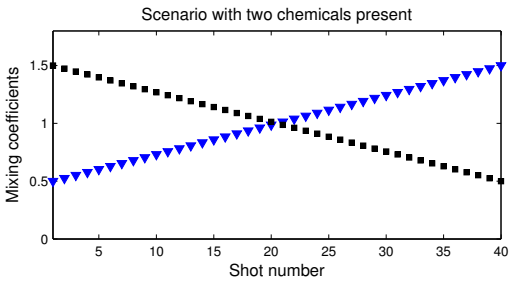
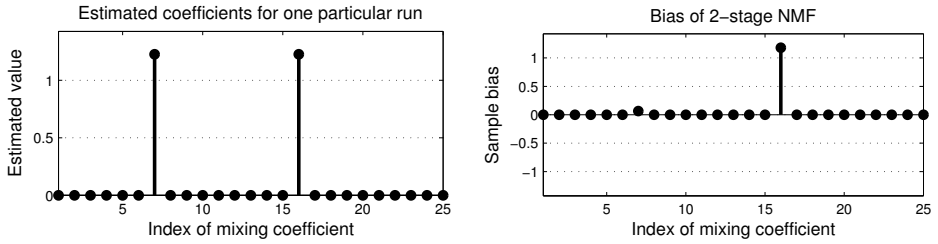


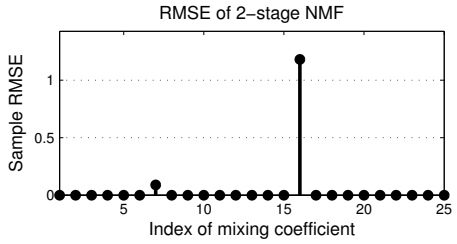
Figure 22: Scenario in which 40 shots of data are collected, and two chemicals are present in each shot. The first chemical, marked by the triangle, is the 7th element of the known library, while the second chemical, marked by the square, is the unknown chemical (the 26th element of the complete library, which is left out of the known library in this experiment).

We ran the 2-stage NMF method on this scenario, and let the algorithm "cheat" by choosing $M_W = 2$ and by choosing the correlation threshold on each shot to be the maximum threshold for which Chem. 7 passes the test of step 2. For the similarity test, we used the correlation. A more sophisticated measure such as the symmetrized Kullback-Leibler divergence may result in improved performance [68]; however, the basic limitations of the two-stage approach, described in the next few paragraphs, would still apply. After determining which components pass the similarity test, we simply used the corresponding estimated mixing coefficients (from $\hat{\mathbf{X}}$) as the estimates for the mixing coefficients for that library chemical. An alternative approach, used in [68], is to recompute the estimates for the mixing coefficients using only the components that passed the similarity test. The estimated mixing coefficients for the 25th shot (for one particular Monte Carlo run) are shown in Figure 23(a). The algorithm correctly detected Chem. 7 because one of the extracted basis vectors was highly correlated with the spectrum of Chem. 7. However, Chem. 16 was also highly correlated with the same column of $\hat{\mathbf{W}}$, which led to a false alarm. This result is consistent over the Monte Carlo runs, as shown in Figures 23(b) and 23(c).



(a) Estimates for one particular run.

(b) Bias.

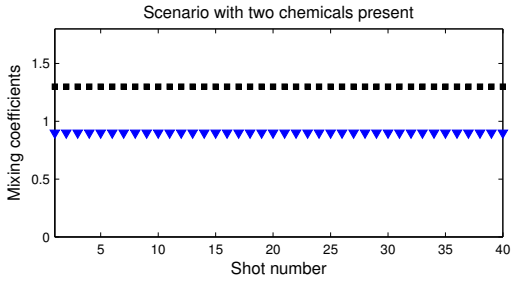


(c) RMSE.

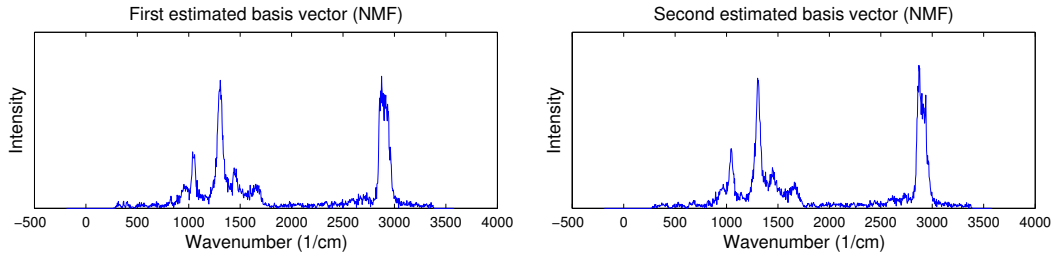
Figure 23: Performance of the 2-stage NMF approach with $M_W = 2$.

This problem might be avoided by employing a more sophisticated assignment algorithm [2] (for step 2) that incorporates one-to-one constraints between the extracted spectra

and library spectra. However, there may not be a one-to-one correspondence between the columns of $\hat{\mathbf{W}}$ and the columns of $\mathbf{A}$, especially if there is insufficient variation in the data.[5] For example, if we consider the scenario represented in Figure 24(a) in which the chemical composition is the same on every shot, the NMF algorithm in step 1 has no way of knowing that there are two chemicals present. For this example, it is impossible to predict how the NMF algorithm will choose the two basis vectors, since a single basis vector is all that is needed to obtain a good approximation of the data. In our simulation, the two extracted basis vectors are each roughly equal to the sum of the two constituent spectra, as seen in Figures 24(b) and 24(c). In general, even if we use the correct value for M_W, the basis vectors estimated by NMF will not be similar to the constituent spectra unless there is a significant amount of variation in the data. Note that the unsupervised approach depends critically on having multiple shots and processing those shots as a block; in a supervised approach, it would be sufficient to run an algorithm such as Richardson-Lucy on each shot separately.



(a) Scenario with no variation in the data.



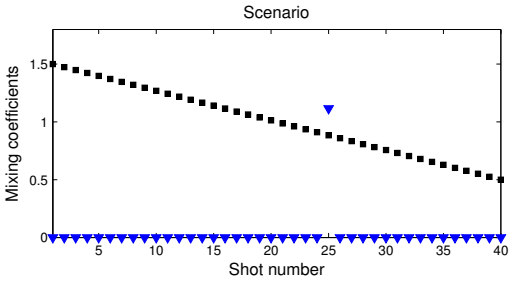
(b) First basis vector estimated by NMF.

(c) Second basis vector estimated by NMF.

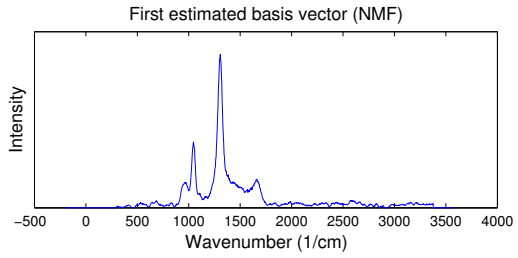
Figure 24: If there is insufficient variation in the data, NMF fails to separate the two constituent spectra.

[5]For this reason, it may be more natural to perform step 2 using canonical correlation analysis; a variation of this approach appears in [94].

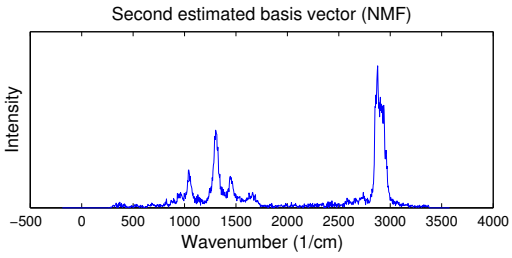
The 2-stage NMF algorithm will similarly fail if the constituent chemicals are not present in enough pulses. To demonstrate this, we now consider the scenario represented in Figure 25(a), in which the library chemical (Chem. 7) is present in only the 25th shot. As opposed to the previous example, the data matrix $\mathbf{Y}$ is no longer approximated well by a single basis vector. Thus, the two NMF basis vectors in this case will be linearly independent. However, because Chem. 7 is present on only one shot, the algorithm has no way of knowing how much of the energy from that shot should be attributed to the second basis vector vs. the first basis vector. In our simulation, almost all the energy from the 25th shot was attributed to the second basis vector, as shown in Figure 25(c) (compare with Figure 17(c)). In this case, neither of the extracted basis vectors is similar enough to the spectrum of Chem. 7 to pass the similarity test (step 2) of the 2-stage NMF algorithm. In many applications, such as the detection of hazardous chemical agents, the resulting missed detection may be unacceptable.



(a) Scenario in which the library chemical is present in only a single shot.



(b) First basis vector estimated by NMF.



(c) Second basis vector estimated by NMF.

Figure 25: If the constituent chemicals are not present in enough pulses, NMF fails to separate the two constituent spectra.

4.4 Partially-Supervised NMF (PS-NMF) Algorithm for Detection

Section 4.3.1 examined the supervised approach to spectral unmixing. In the supervised framework, the reference library is assumed to be comprehensive and only the mixing coefficients are unknown. These coefficients can be estimated using a gradient descent algorithm; in this case, the library $\mathbf{A}$ is fixed and only the mixing matrix $\mathbf{X}$ is updated on each iteration. We demonstrated a well-known shortcoming of the supervised approach: no comprehensive library exists, and when chemicals are present that are not contained in the reference library, the supervised algorithms may confuse those chemicals with library members.

Section 4.3.2 applied an unsupervised algorithm, NMF, to the Raman data and then compared the resulting estimated constituent spectra to the known library spectra. In the first step of this 2-stage scheme, both the constituent spectra $\mathbf{W}$ and mixing coefficients $\mathbf{X}$ are assumed to be unknown. The two matrices can be jointly estimated using the gradient descent technique; a certain choice of step size results in a well-known multiplicative update NMF algorithm. This 2-stage method performed poorly because knowledge of the reference library was not used in extracting the spectra. If there is a known library of elements that we are trying to detect, this knowledge should be incorporated into the NMF problem.

For these reasons, rather than assume the reference library is completely known (as in the supervised framework) or completely unknown (as in the unsupervised framework), we find it more natural to assume the library is *partially* known [67]. As in the NMF problem, we once again seek to approximate the nonnegative data matrix $\mathbf{Y}$ with a lower-rank nonnegative factorization $\mathbf{Y} \approx \mathbf{WX}$, but we now assume that the first M columns of $\mathbf{W}$ are known and only the last $m = M_W - M$ columns need to be estimated. We will call this approach "partially-supervised nonnegative matrix factorization" (PS-NMF). The number of "extra" columns, m, is specified by the user. Each iteration of the gradient

descent technique for our partially-supervised approach is given by

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} + \alpha_{ij}^{(n)} \frac{\partial}{\partial [\mathbf{W}]_{ij}} \Psi(\mathbf{W}, \mathbf{X}) \Big|_{\mathbf{W}=\hat{\mathbf{W}}^{(n)}, \mathbf{X}=\hat{\mathbf{X}}^{(n)}} \text{ for } j > M \text{ only} \tag{140a}$$

$$[\hat{\mathbf{X}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{X}}^{(n)}]_{j\gamma} + \delta_{j\gamma}^{(n)} \frac{\partial}{\partial [\mathbf{X}]_{j\gamma}} \Psi(\mathbf{W}, \mathbf{X}) \Big|_{\mathbf{W}=\hat{\mathbf{W}}^{(n)}, \mathbf{X}=\hat{\mathbf{X}}^{(n)}}. \tag{140b}$$

This differs from (125) in that (140a) is computed only for $j > M$; the first M columns of $\mathbf{W}$ are known to be equal to the reference library and do not need to be estimated. Although this idea is quite simple, we have not seen it presented elsewhere in the literature. Table 10 summarizes the differences between the supervised approach, the unsupervised approach, and the partially-supervised approach.

Table 10: Comparison of unsupervised, supervised, and partially-supervised approaches.

• *Unsupervised approach (NMF)* 1. Update $[\hat{\mathbf{W}}]_{ij}$ for **all** i, j (constituent objects are unknown) 2. Update $[\hat{\mathbf{X}}]_{j\gamma}$ for all j, γ
• *Supervised approach (RL)* 1. Update $[\hat{\mathbf{W}}]_{ij}$ for **no** i, j (constituent objects are known; $\mathbf{W} = \mathbf{A}$) 2. Update $[\hat{\mathbf{X}}]_{j\gamma}$ for all j, γ
• *Partially-supervised approach (PS-NMF)* 1. Update $[\hat{\mathbf{W}}]_{ij}$ for $j > M$ (some constituent objects known, some unknown) 2. Update $[\hat{\mathbf{X}}]_{j\gamma}$ for all j, γ

As in Section 4.2, for certain objective functions, the step sizes $\alpha_{ij}^{(n)}$ and $\delta_{j\gamma}^{(n)}$ can be chosen such that (140) becomes a multiplicative update rule:

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} [f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)})]_{ij} \quad \text{for } j > M \text{ only} \tag{141a}$$

$$[\hat{\mathbf{X}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{X}}^{(n)}]_{j\gamma} [g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)})]_{j\gamma} \tag{141b}$$

for some functions f and g. This can be implemented in matrix form as

$$\hat{\mathbf{W}}^{(n+1)} \leftarrow \hat{\mathbf{W}}^{(n)} .* f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)}) \tag{142a}$$

$$\text{Replace first } M \text{ columns of } \hat{\mathbf{W}}^{(n+1)} \text{ with the known library } \mathbf{A} \tag{142b}$$

$$\hat{\mathbf{X}}^{(n+1)} \leftarrow \hat{\mathbf{X}}^{(n)} .* g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)}) \,. \tag{142c}$$

Of course, in implementation, there is no particular reason to take the time to compute the first M columns of $\hat{\mathbf{W}}^{(n+1)}$; the formulation given in (142) is intended to illustrate how an existing NMF implementation may be quickly and easily "retrofitted" to use a PS-NMF approach.

Uses of the term "Semi-Supervised NMF" In recent years, there has been a significant amount of progress on applying nonnegative matrix factorization to the *clustering* problem. [12, 49, 39, 6, 47] Ding, et al. [12] showed that a symmetric form of NMF is equivalent to kernel K-means clustering. One advantage of the NMF framework for clustering is that "must-link" and "cannot-link" constraints (specifying pairs of objects that must belong in the same group or pairs that cannot belong in the same group[6]) are readily incorporated into the framework [49, 6]. When these kinds of pairwise constraints are added to the clustering problem, it is often known as "semi-supervised clustering," and when NMF-based algorithms are applied to this problem, they are sometimes called "semi-supervised NMF." [6, 47]

We would have preferred to call our unmixing approach "semi-supervised NMF" as well, because it seems like the most natural name, and we used this name in some of our preliminary work [65]. However, our algorithm addresses a fundamentally different problem than the "SS-NMF" algorithms discussed above and should not be confused with the NMF methods used for clustering. To try to avoid such confusion, we now choose the name "partially-supervised NMF" instead of "semi-supervised NMF."

[6] A similar problem is clustering with partially labeled data, as discussed in [47].

4.4.1 Algorithm Enhancements

Our PS-NMF algorithm generally performs well if the user guesses the correct number for m. For instance, if there is one unknown chemical present in a scenario, then PS-NMF with $m = 1$ generally gives good results. Unfortunately, the number of unknown substances will not be known in practice. If we assume too *few* unknown chemicals, then PS-NMF tends to attribute energy from the unknown chemicals to the reference spectra. In this case, it overestimates the abundances of the library chemicals, leading to false alarms. We saw an example of this in Figure 20, in which we assumed $m = 0$ (i.e., we used the RL algorithm), but there was actually one unknown substance present.

On the other hand, if we assume too *many* unknown chemicals, then PS-NMF has no reason not to attribute energy from the library chemicals to the "extra" columns of $\hat{\mathbf{W}}$. For instance, if there are actually no unknown chemicals present but we use $m = 3$, then PS-NMF may obtain a nearly perfect fit of the data using only the three extra columns. In this case, it underestimates the abundances of the library chemicals, leading to missed detections.

One way to address these issues is to choose a large number for m and to introduce a penalty term in the objective function. For example, one possible penalized form of the Poisson loglikelihood objective function (132) is

$$\Psi(\mathbf{W}, \mathbf{X}) = \sum_{i,\gamma} \left[[\mathbf{WX}]_{i\gamma} - [\mathbf{Y}]_{i\gamma} \ln \left([\mathbf{WX}]_{i\gamma} \right) \right] + \sum_{j,\gamma} [\mathbf{\Lambda}]_{j\gamma} [\mathbf{X}]_{j\gamma} \,, \tag{143}$$

where $[\mathbf{\Lambda}]_{j\gamma} = \lambda$ (for some $\lambda > 0$) if $j > M$ and $[\mathbf{\Lambda}]_{j\gamma} = 0$ otherwise. The purpose of the penalty is to encourage the algorithm to use the library spectra instead of the extra columns of $\hat{\mathbf{W}}$. It prevents data overfitting for the case where the value chosen for m is larger than the true number of unknown chemicals present. Appendix A shows that the PS-NMF update rule for the penalized cost function (143) is

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} \frac{\sum_{\gamma'} \frac{[\mathbf{Y}]_{i\gamma'}}{[\hat{\mathbf{W}}^{(n)} \hat{\mathbf{X}}^{(n)}]_{i\gamma'}} [\hat{\mathbf{X}}^{(n)}]_{j\gamma'}}{\sum_{\gamma'} [\hat{\mathbf{X}}^{(n)}]_{j\gamma'}} \quad \text{for } j > M \tag{144a}$$

$$[\hat{\mathbf{X}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{X}}^{(n)}]_{j\gamma} \frac{\sum_{i'} \frac{[\mathbf{Y}]_{i'\gamma}}{[\hat{\mathbf{W}}^{(n)} \hat{\mathbf{X}}^{(n)}]_{i'\gamma}} [\hat{\mathbf{W}}^{(n)}]_{i'j}}{\sum_{i'} [\hat{\mathbf{W}}^{(n)}]_{i'j} + [\mathbf{\Lambda}]_{j\gamma}} \,. \tag{144b}$$

Because $\lambda > 0$, nonnegativity is clearly preserved on each iteration.

Unfortunately, the update function (144) yields the same results as the unpenalized update function (133). This problem was noted by Hoyer in his similar work on nonnegative sparse coding [30] and is explained as follows. If the last m columns of $\mathbf{W}$ are scaled up and and the last m rows of $\mathbf{X}$ are scaled down by the same amount, then the main term of (143) will remain unchanged while the penalty term decreases. Thus, the optimization of (143) will result in the elements from the last m rows of $\mathbf{X}$ approaching zero while the elements in the corresponding columns of $\mathbf{W}$ blow up. The penalty term then becomes negligible and has no effect on the solution.

One possible remedy for this quandary is to add a penalty term to Ψ that constrains each column of $\mathbf{W}$ to sum to a given constant c. This would hopefully prevent the columns of $\mathbf{W}$ from growing unbounded. For example, if the extended cost function is

$$\Psi(\mathbf{W},\mathbf{X}) = \sum_{i,\gamma} \left[[\mathbf{W}\mathbf{X}]_{i\gamma} - [\mathbf{Y}]_{i\gamma} \ln \left([\mathbf{W}\mathbf{X}]_{i\gamma} \right) \right] + \sum_{j,\gamma} [\mathbf{\Lambda}]_{j\gamma} [\mathbf{X}]_{j\gamma} + \beta \frac{1}{2} \| \mathbf{1}_{1\times N} \mathbf{W} - c\mathbf{1}_{1\times M} \|^2 \tag{145}$$

for some regularization parameter β, then it is straightforward to show that the resulting PS-NMF iteration is

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} \frac{\sum_{\gamma'} \frac{[\mathbf{Y}]_{i\gamma'}}{[\hat{\mathbf{W}}^{(n)}\hat{\mathbf{X}}^{(n)}]_{i\gamma'}} [\hat{\mathbf{X}}^{(n)}]_{j\gamma'}}{\sum_{\gamma'} [\hat{\mathbf{X}}^{(n)}]_{j\gamma'} + \beta \left(\sum_{i'} [\mathbf{W}]_{i'j} - c \right)} \quad \text{for } j > M \tag{146a}$$

$$[\hat{\mathbf{X}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{X}}^{(n)}]_{j\gamma} \frac{\sum_{i'} \frac{[\mathbf{Y}]_{i'\gamma}}{[\hat{\mathbf{W}}^{(n)}\hat{\mathbf{X}}^{(n)}]_{i'\gamma}} [\hat{\mathbf{W}}^{(n)}]_{i'j}}{\sum_{i'} [\hat{\mathbf{W}}^{(n)}]_{i'j} + [\mathbf{\Lambda}]_{j\gamma}} \,. \tag{146b}$$

Unfortunately, the denominator of (146a) may now be negative, so nonnegativity is not ensured for this algorithm. Thus, instead of incorporating this normalization constraint into the objective function, we simply normalize the columns of $\mathbf{W}$ after updating them via (144a). The resulting PS-NMF algorithm is summarized in Algorithm 5.

While the convergence properties discussed in Section 4.2 may no longer hold after the inclusion of the normalization step, the PS-NMF estimates converged in all of our simulations. Other, more sophisticated, approaches to incorporate the normalization constraint on the columns of $\mathbf{W}$ may result in improved convergence properties or faster convergence.

Algorithm 5 General form of the PS-NMF algorithm. M_W is the number of columns of $\mathbf{W}$, λ is the penalty parameter, and f and g are the update equations derived to minimize the objective function. For example, for the ML approach and Poisson data, f is given by (174a) and g is given by (174c).

Input: $\mathbf{Y}, \mathbf{A}, M_W, \lambda$
Output: $\hat{\mathbf{W}}, \hat{\mathbf{X}}$

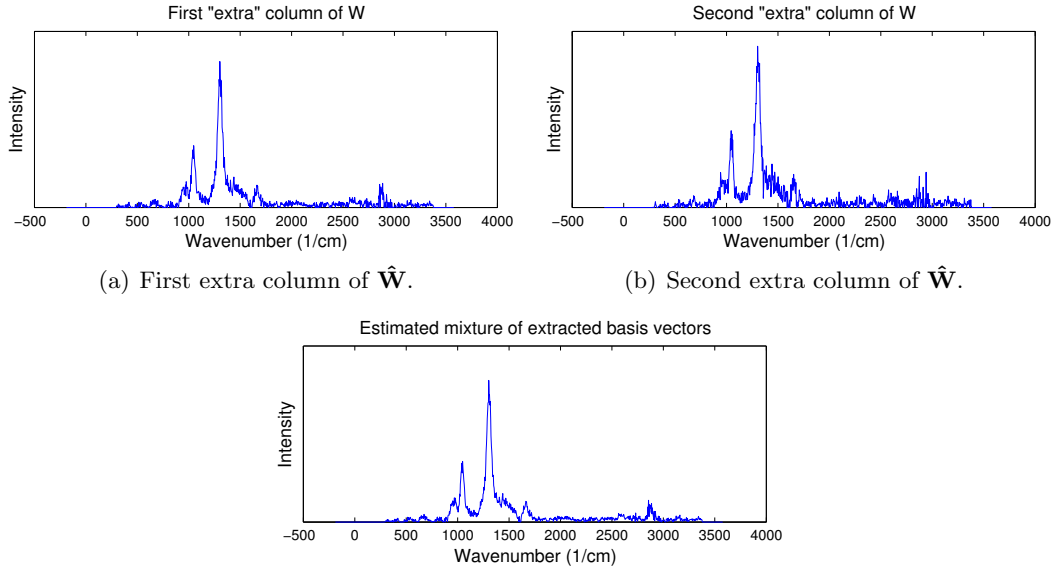
Initialize with some positive $\hat{\mathbf{W}}^{(0)}, \hat{\mathbf{X}}^{(0)}$
$n \leftarrow 0$
while not converged **do**
  $\hat{\mathbf{W}}^{(n+1)} \leftarrow \hat{\mathbf{W}}^{(n)} .* f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)}, \lambda)$
  Replace first M columns of $\hat{\mathbf{W}}^{(n+1)}$ with the known library $\mathbf{A}$
  Normalize each of the last m columns of $\mathbf{W}$
  $\hat{\mathbf{X}}^{(n+1)} \leftarrow \hat{\mathbf{X}}^{(n)} .* g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{X}}^{(n)}, \lambda)$
  $n \leftarrow n + 1$
end while
$\hat{\mathbf{W}} \leftarrow \hat{\mathbf{W}}^{(n+1)}; \hat{\mathbf{X}} \leftarrow \hat{\mathbf{X}}^{(n+1)}$

We leave this as a topic for future research.

4.4.2 PS-NMF Simulation Results

This section analyzes the estimation performance of the PS-NMF algorithm for the spectral unmixing problem discussed in Section 4.3. We begin by revisiting the scenario represented by Figure 22, in which there are two chemicals present (in varying quantities) in every shot. Because one of these chemicals is not in the reference library, both the RL algorithm and the 2-stage NMF method performed poorly for this scenario, as seen in Section 4.3.

We ran the PS-NMF algorithm with $m = 8$, or equivalently, $M_W = 33$. We used $\lambda = 50$ for all the PS-NMF simulations in this chapter. Figures 26(a) and 26(b) show the first two "extra" columns of $\hat{\mathbf{W}}$; the other six estimated basis vectors are similar and are omitted here for brevity. The estimated mixture of the eight columns, shown in Figure 26(c), is quite similar to the actual unknown spectrum shown in Figure 17(b). In general, each "extra" column of $\hat{\mathbf{W}}$ will not necessarily correspond to a different chemical; thus, these should not be referred to as "extracted spectra" or "estimated spectra." For the purpose of detecting the *library* chemicals, however, the algorithm will perform well as long as some *combination* of the extra columns of $\hat{\mathbf{W}}$ is able to properly account for the energy from the unknown spectrum.



(a) First extra column of $\hat{\mathbf{W}}$.

(b) Second extra column of $\hat{\mathbf{W}}$.

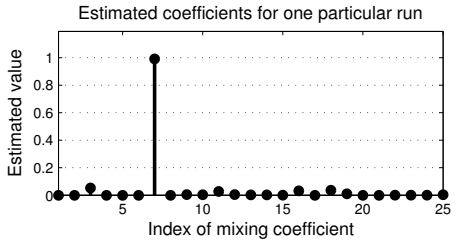
(c) Estimated mixture of extra columns of $\hat{\mathbf{W}}$.

Figure 26: First two "extra" columns of $\hat{\mathbf{W}}$ estimated by the PS-NMF algorithm, and the estimated mixture of all 8 extracted basis vectors. Compare (c) with Figure 17(b).

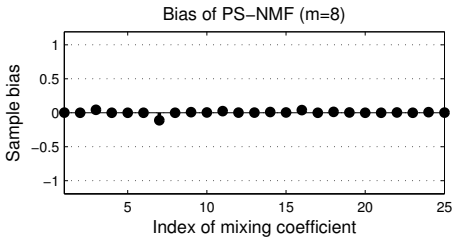
The estimated mixing coefficients for the 25th shot (for one particular Monte Carlo run) are shown in Figure 27(a); these estimates are quite close to the true values shown in Figure 18(a). The algorithm correctly detected Chem. 7 *and* correctly attributed the energy from the unknown chemical to the "extra" columns of $\hat{\mathbf{W}}$. This result is consistent over the Monte Carlo runs, as shown in Figures 27(b) and 27(c).

There is a small negative bias for Chem. 7; even with the penalty term, some of the energy from Chem. 7 is incorrectly attributed to the extra columns. This indicates that λ may be too small; increasing the penalty would lead to less error for Chem. 7. On the other hand, there is a small positive bias for the other chemicals in the library; because of the penalty term, some of the energy from the unknown chemical is incorrectly attributed to library members. This indicates that λ may be too large; decreasing λ would lead to less error for the other library chemicals. Thus, there is a tradeoff in the choice of the penalty parameter. Investigating how to select a suitable value for λ is a potential topic for future research.

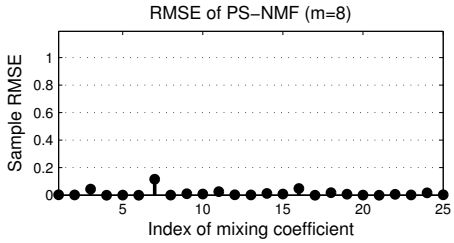
Some additional insight may be gained by revisiting the case in which the complete



(a) Estimates for one particular run.



(b) Bias.



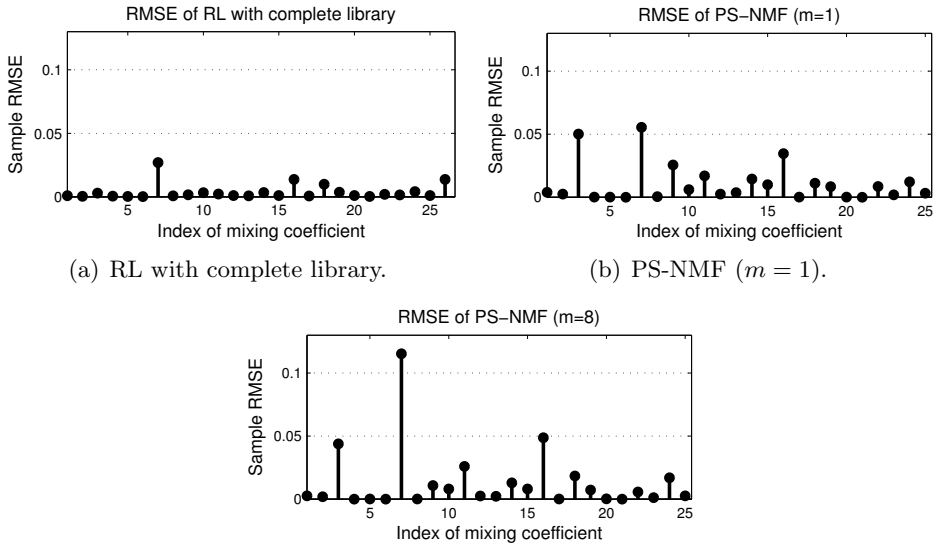
(c) RMSE.

Figure 27: Performance of the PS-NMF algorithm ($m = 8$). Compare with Figures 20 and 23.

library is known. Since all the constituent spectra are known and only the mixing coefficients need to be estimated, the RL algorithm yields relatively small errors, as shown in Figure 28(a) (a rescaled version of Figure 19(c)). If we instead have an incomplete library, but we know that the number of unknown chemicals is $m = 1$, the estimates have more error because the algorithm needs to co-estimate the values in the unknown spectrum along with the parameters of interest (the mixing coefficients corresponding to the library). The RMSE for this case is shown in Figure 28(b). The performance degrades further if we do not know the number of unknown chemicals and if we let $m = 8$; see Figure 28(c) (rescaled version of Figure 27(c)). However, even with $m = 8$, the algorithm has relatively low error and typically results in no missed detections or false alarms for this scenario.

Section 4.3.2 illustrated that one of the shortcomings of NMF, when applied to the detection of objects from a known library, is that its success is highly dependent on the amount of variation in the data. For the scenario of Figure 24(a), in which the chemical composition is the same on every shot, NMF could not successfully extract the constituent spectra.

If no penalty term is applied, the PS-NMF algorithm may similarly break down if there is insufficient variation in the data. This occurs even if the number of unknown chemicals



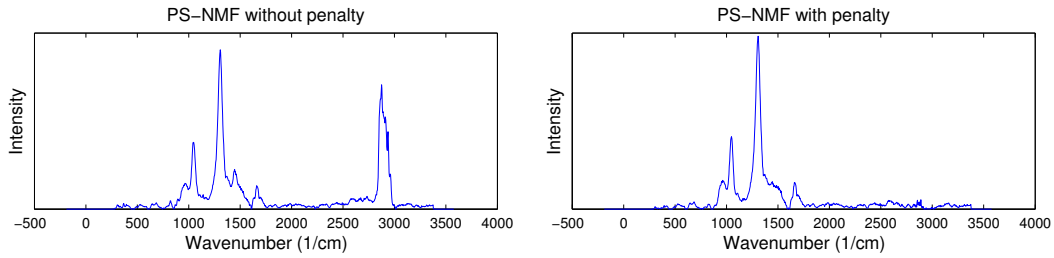
(a) RL with complete library.

(b) PS-NMF ($m = 1$).

(c) PS-NMF ($m = 8$).

Figure 28: RMSE for the case in which the library is complete, for the case in which the library is incomplete but the number of unknown objects is known, and for the case in which the library is incomplete and the number of unknown objects is unknown.

is known. For example, for the scenario of Figure 24(a), even if we correctly use $m = 1$, the PS-NMF algorithm without the penalty term has no reason not to attribute all the measured energy to the unknown column. To illustrate this, Figure 29(a) shows the extra column estimated by PS-NMF; this is roughly equal to the sum of the two constituent spectra (see Figure 17). In this case, the algorithm drastically underestimates the quantity of Chem. 7 because it attributes most of the energy from Chem. 7 to the unknown column.



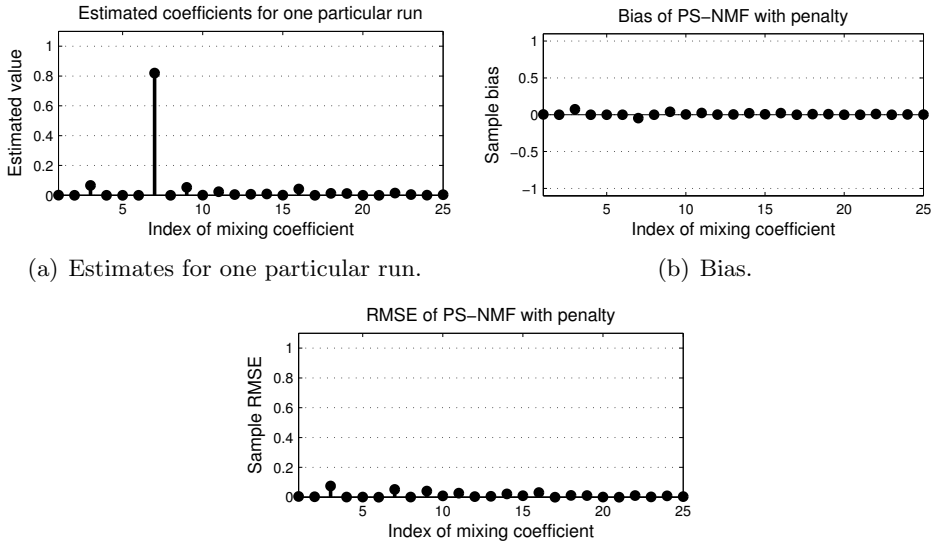
(a) PS-NMF without the penalty term.

(b) PS-NMF with the penalty term.

Figure 29: The extra column estimated by PS-NMF ($m = 1$) with and without the penalty term

If we use the penalized PS-NMF objective function, however, then this ambiguity is mostly resolved. The purpose of the penalty term in (143) is to encourage the algorithm

to prefer the library spectra over the extra columns of $\hat{\mathbf{W}}$. With the penalty, PS-NMF will attribute to the extra columns only the energy that cannot be explained almost as well by the library components. As seen in Figure 29(b), the extra column estimated by the PS-NMF algorithm with the penalty term is almost identical to the actual unknown spectrum shown in Figure 17(b). The resulting estimation performance of the algorithm is shown in Figure 31. The PS-NMF method is seen to perform quite well even when there is no variation in the data.



(a) Estimates for one particular run.

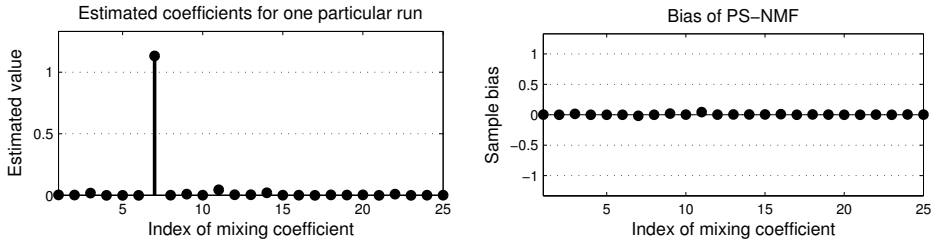
(b) Bias.

(c) RMSE.

Figure 30: Performance of the PS-NMF algorithm ($m = 1$) with the penalty term, for the scenario in which there is no variation in the data.

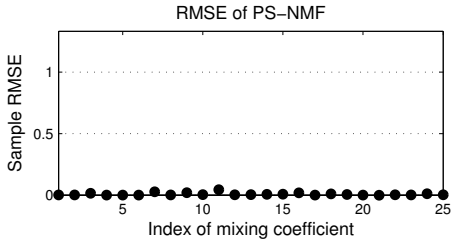
We now return to the scenario represented by Figure 25(a), in which the library chemical is present in only a single pulse. Section 4.3.2 indicated that NMF may not successfully extract the constituent spectra for cases like this; the constituent chemicals are not present in enough shots for NMF to learn their signatures accurately. The PS-NMF algorithm, on the other hand, has the constituent library spectra built into $\hat{\mathbf{W}}$ and does not need to learn them from the data. Thus, as shown in Figure 31, the PS-NMF method performs quite well for this scenario.

Finally, we push the algorithms harder by letting there be three chemicals present that



(a) Estimates for one particular run.

(b) Bias.

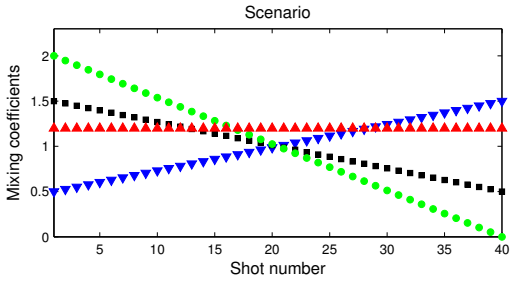


(c) RMSE.

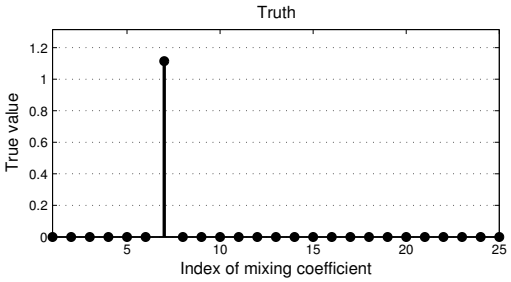
Figure 31: PS-NMF performance (for the 25th shot) for the scenario in which Chem. 7 is present in only the 25th shot.

are not in the library. Specifically, we consider the example scenario represented by Figure 32(a), where the upside-down triangle once again marks the quantity of Chem. 7 and the other three markers represent the quantities of the three unknown chemicals. The true mixing coefficients (for the 25th shot) for the reference library are shown in Figure 32(b). The estimation results for the RL algorithm are shown in Figure 33. Once again, the RL algorithm confuses the unknown chemicals for library spectra, as evidenced by the large positive bias on a number of the coefficients. The results for the PS-NMF algorithm (with $m = 8$) are plotted in Figure 34. The PS-NMF algorithm performs much worse than when only one unknown chemical was present (compare with Figure 27). It still outperforms the fully supervised method of Figure 33, however.

One disadvantage of the PS-NMF algorithm is that it is typically quite slow; each iteration of the update equation (174) requires $O(NM_WK)$ operations. The traditional multiplicative update NMF algorithm (which itself is known to be slow) also has computational complexity of $O(NM_WK)$, but the M_W for PS-NMF is typically much larger than for NMF because it includes the library; $M_W = M + m$. A more thorough analysis of comparative complexity would involve not only the cost per iteration but also the number

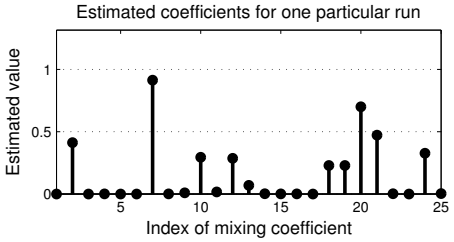


(a) Scenario with four chemicals present.

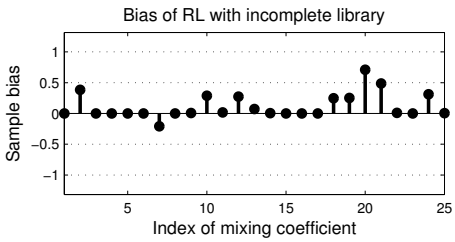


(b) True mixing coefficients on 25th shot.

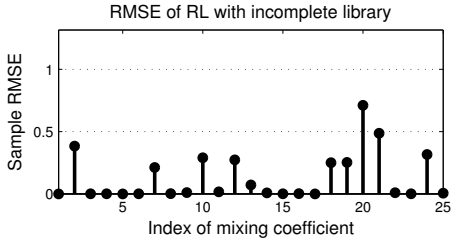
Figure 32: Scenario in which 40 shots of data are collected, and four chemicals are present in each shot. The first chemical, marked by the upside-down triangle, is the 7th element of the known library, while the other three chemicals are not in the library. The true mixing coefficients for the library chemicals on the 25th shot are shown in (b).



(a) Estimates for one particular run.

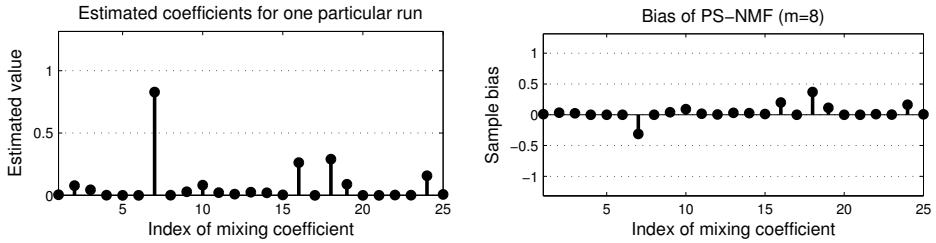


(b) Bias.



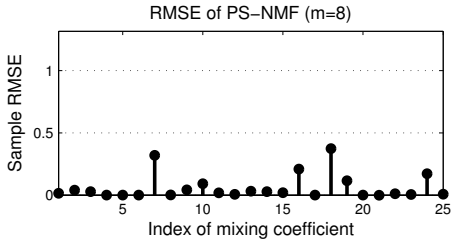
(c) RMSE.

Figure 33: Performance of the RL algorithm (for the 25th shot) when there are three unknown chemicals.



(a) Estimates for one particular run.

(b) Bias.



(c) RMSE.

Figure 34: PS-NMF ($m = 8$) performance (for the 25th shot) when there are three unknown chemicals.

of iterations typically required. This topic, as well as methods to speed up PS-NMF, are left as areas for future research.

4.5 Conclusions

This chapter presented a novel partially-supervised nonnegative matrix factorization (PS-NMF) method for the detection of objects from a known, but possibly incomplete, library. Under certain likelihood models (such as Poisson or Gaussian), this method can be viewed as a gradient descent algorithm to maximize the likelihood of the data for the case in which some, but not all, of the constituent objects are known. As an example application, we considered the spectral unmixing of Raman spectroscopy data. In our simulations, the PS-NMF algorithm gave better estimation performance than the two-stage NMF approach and the fully supervised approach when there were chemicals present that were not in the library.

CHAPTER V

ACCOUNTING FOR ERROR IN THE REFERENCE LIBRARY

5.1 Introduction

To account for the noise introduced by the measurement system, Chapter 2 presented a probabilistic model based on previous work done in the scientific imaging community. That chapter analyzed several methods for estimating the mixing coefficients under this model, and the results were compared to the Cramér-Rao lower bound. All that work assumed that the reference library spectra themselves were error-free. In reality, any measured library spectra, even if taken in a pristine laboratory environment, will be subject to noise introduced by the measurement system. This chapter expands on the work of Chapter 2 by incorporating uncertainty in the reference spectra. We begin by briefly reviewing the total least squares (TLS) problem.

5.1.1 Total Least Squares

The least-squares problem (1) of Section 2.1 can be expressed as follows:

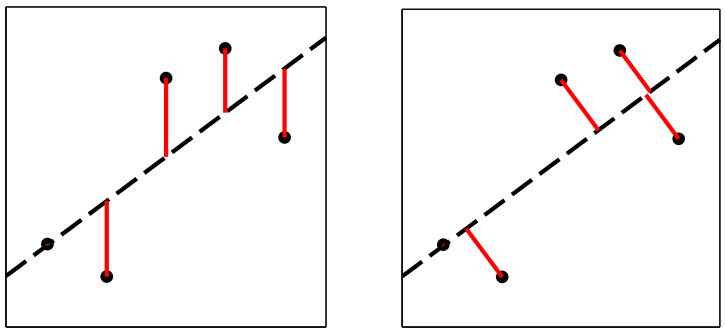
LS: Find the mixing vector $\mathbf{x}$ and error vector $\mathbf{e}$ that minimize $\|\mathbf{e}\|^2$ s.t. $\mathbf{H}\mathbf{x} = \mathbf{y} + \mathbf{e}$.

This problem accounts for the error in $\mathbf{y}$ but assumes that the mixing matrix $\mathbf{H}$ is exactly known. Perhaps the simplest and most well-known formulation that incorporates error in the model matrix is the total least squares (TLS) problem [24, 25], which in its most basic form is expressed as follows:

TLS: Find the $\mathbf{x}$, $\mathbf{e}$, and $\mathbf{E}$ that minimize $\|[\mathbf{E}\ \ \mathbf{e}]\|_F^2$ s.t. $(\mathbf{H} + \mathbf{E})\mathbf{x} = \mathbf{y} + \mathbf{e}$.

To illustrate the difference between TLS and LS, Figure 35 compares the two for the $M = 1$ case. The LS method fits the line to the data points by minimizing the error in

the second coordinate only; this error is represented by the vertical non-dashed lines in Figure 35(a). The TLS method corrects both coordinates; the approximation error for the same candidate line fit is represented by the diagonal non-dashed lines in Figure 35(b). For a more thorough overview of TLS, see [56, 91].



(a) Least squares (LS). (b) Total least squares (TLS).

Figure 35: Approximation errors minimized by LS and TLS.

5.1.2 Organization of this Chapter

This chapter is organized as follows. Section 5.2 reviews the errors-in-variables model and the well-known property that the TLS approach coincides with maximum-likelihood estimation under a Gaussian errors-in-variables (EIV) model. Section 5.2.1 presents a novel nonnegative total least squares (NNTLS) algorithm, which uses a variation of the partially-supervised nonnegative matrix factorization algorithm of Chapter 4. This same approach is then applied to a Poisson errors-in-variables model in Section 5.3. We compare the estimation results of the Poisson EIV method to the algorithms of Chapter 2 to determine how much harm is caused by ignoring the error in the reference library.

5.2 ML Estimation under an EIV Gaussian Model

There are a number of ways to model the uncertainty in the reference library. We consider the standard errors-in-variables (EIV) model [96, 22], in which $\mathbf{H}$ is considered to be a noisy observation of the deterministic unknown matrix $\mathbf{G}$. The goal is then to estimate both the mixing vector $\mathbf{x}$ and the "true" underlying library $\mathbf{G}$ from the measured spectrum $\mathbf{y}$ and

the given measured library $\mathbf{H}$. This may be accomplished using the maximum-likelihood technique:

$$(\hat{\mathbf{G}}, \hat{\mathbf{x}}) = \arg\max_{\mathbf{G},\mathbf{x}} p(\mathbf{y}, \mathbf{H}; \mathbf{G}, \mathbf{x}) \tag{147a}$$

$$= \arg\max_{\mathbf{G},\mathbf{x}} p(\mathbf{y}; \mathbf{G}, \mathbf{x}) p(\mathbf{H}; \mathbf{G}) \tag{147b}$$

$$= \arg\max_{\mathbf{G},\mathbf{x}} \left[\ln p(\mathbf{y}; \mathbf{G}, \mathbf{x}) + \ln p(\mathbf{H}; \mathbf{G})\right]. \tag{147c}$$

To illustrate this approach, we consider the classical Gaussian EIV model

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{e}, \tag{148a}$$

$$\mathbf{H} = \mathbf{G} + \mathbf{E}, \tag{148b}$$

where the elements of $\mathbf{e}$ and $\mathbf{E}$ are independent, zero-mean, and normally distributed. If we keep things simple by assuming that the elements also have equal variance σ^2, then (147c) becomes

$$(\hat{\mathbf{G}}, \hat{\mathbf{x}}) = \arg\max_{\mathbf{G},\mathbf{x}} \left[-\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{G}\mathbf{x}\|^2 - \frac{MN}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{H} - \mathbf{G}\|_F^2\right] \tag{149a}$$

$$= \arg\min_{\mathbf{G},\mathbf{x}} \left[\|\mathbf{H} - \mathbf{G}\|_F^2 + \|\mathbf{y} - \mathbf{G}\mathbf{x}\|_F^2\right] \tag{149b}$$

$$= \arg\min_{\mathbf{G},\mathbf{x}} \|[(\mathbf{H} - \mathbf{G}) \;\; (\mathbf{y} - \mathbf{G}\mathbf{x})]\|_F^2, \tag{149c}$$

which is just the TLS problem. Equation (149c) can be rearranged as

$$(\hat{\mathbf{G}}, \hat{\mathbf{x}}) = \arg\min_{\mathbf{G},\mathbf{x}} \|[\mathbf{H} \;\; \mathbf{y}] - [\mathbf{G} \;\; \mathbf{G}\mathbf{x}]\|_F^2. \tag{150}$$

Because $\mathbf{G}\mathbf{x}$ is in the range of $\mathbf{G}$, the matrix $[\mathbf{G} \;\; \mathbf{G}\mathbf{x}]$ has a rank of M, or equivalently, a nullity of 1 (assuming that $\mathbf{G}$ is full rank). Thus, the ML estimate for the EIV Gaussian model may be found by first seeking the rank-M matrix $\mathbf{F}_M$ that minimizes $\|[\mathbf{H} \;\; \mathbf{y}] - \mathbf{F}_M\|_F^2$. By the Eckart-Young Theorem [16], this low-rank approximation is found using the singular value decomposition (SVD). This property leads to the classical SVD-based approach [24, 25] for computing the TLS estimate.

5.2.1 Nonnegative Total Least Squares

For our application, a limitation of the TLS approach described above is that it does not enforce nonnegativity. If we again consider the Gaussian EIV model (148) but now apply nonnegativity constraints, the ML approach coincides with the nonnegative total least squares (NNTLS) problem:

$$(\hat{\mathbf{G}}, \hat{\mathbf{x}}) = \underset{\mathbf{G}\geq\mathbf{0},\mathbf{x}\geq\mathbf{0}}{\arg\max}\, p(\mathbf{y}, \mathbf{H}; \mathbf{G}, \mathbf{x}) \tag{151a}$$

$$= \underset{\mathbf{G}\geq\mathbf{0},\mathbf{x}\geq\mathbf{0}}{\arg\min}\, \|[(\mathbf{H}-\mathbf{G})\ \ (\mathbf{y}-\mathbf{G}\mathbf{x})]\|_F^2\,. \tag{151b}$$

The NNTLS problem has received scant attention in the literature; the only previous work of which we are aware [11, 83] has placed the nonnegativity constraint on just the mixing coefficients and not on the values of $\mathbf{G}$. Our approach in this section is to view the problem as one of low-rank approximation with nonnegativity constraints. From this perspective, the NNTLS problem is naturally addressed using a variation of the partially-supervised nonnegative matrix factorization (PS-NMF) method of Chapter 4.

As in the TLS case, (151b) may be expressed as

$$(\hat{\mathbf{G}}, \hat{\mathbf{x}}) = \underset{\mathbf{G}\geq\mathbf{0},\mathbf{x}\geq\mathbf{0}}{\arg\min}\, \|[\mathbf{H}\ \ \mathbf{y}] - [\mathbf{G}\ \ \mathbf{G}\mathbf{x}]\|_F^2 \tag{152a}$$

$$= \underset{\mathbf{G}\geq\mathbf{0},\mathbf{x}\geq\mathbf{0}}{\arg\min}\, \|[\mathbf{H}\ \ \mathbf{y}] - \mathbf{F}_M\|_F^2\,; \tag{152b}$$

the problem is once again to find the rank-M matrix $\mathbf{F}_M$ that best approximates the rank-$(M+1)$ matrix $[\mathbf{H}\ \ \mathbf{y}]$ in the Frobenius norm sense. However, for the NNTLS problem, $\mathbf{F}_M$ must satisfy two additional criteria:

1. All of the elements of $\mathbf{F}_M$ must be nonnegative.

2. The last column of $\mathbf{F}_M$ must be equal to some *nonnegative* linear combination of the first M columns.

If only the first of these two criteria needed to be satisfied, then $\mathbf{F}_M$ could be found via nonnegative matrix factorization (see Section 4.2). The problem would then be to find the (elementwise) nonnegative matrices $\hat{\mathbf{W}} \in \Re^{N\times M}$ and $\hat{\mathbf{V}} \in \Re^{M\times(M+1)}$ for which $\|[\mathbf{H}\ \ \mathbf{y}] - \mathbf{W}\mathbf{V}\|_F^2$

is minimized; we will refer to such minimizing $(\mathbf{W}, \mathbf{V})$ as $(\hat{\mathbf{W}}, \hat{\mathbf{V}})$. The product $\mathbf{F}_M = \hat{\mathbf{W}}\hat{\mathbf{V}}$ would clearly be nonnegative, satisfying the first condition above.[1] Furthermore, since $\mathbf{F}_M$ has rank M, the last column of $\mathbf{F}_M$ could be expressed as some linear combination of the first M columns. However, it would not necessarily be equal to a *nonnegative* linear combination of the first M columns. Thus, while this NMF method would ensure that $\hat{\mathbf{G}}$ is nonnegative, it would not enforce the nonnegativity of $\hat{\mathbf{x}}$.

To ensure that the second condition is satisfied, we constrain the first M columns of $\mathbf{V}$ to be equal to the identity matrix:

$$(\hat{\mathbf{W}}, \hat{\mathbf{V}}) = \underset{\mathbf{W}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}}{\arg\min} \|[\mathbf{H} \;\; \mathbf{y}] - \mathbf{W}\mathbf{V}\|_F^2 \quad \text{s.t.} \quad [\mathbf{V}]_{j\gamma} = \begin{cases} 1 & j = \gamma \\ 0 & j \neq \gamma, \gamma < M\,. \end{cases} \tag{153}$$

With this additional constraint, the last column of $\mathbf{F}_M$ is now clearly a nonnegative linear combination of the first M columns. Furthermore, because $\hat{\mathbf{W}}\hat{\mathbf{V}} = [\hat{\mathbf{G}} \;\; \hat{\mathbf{G}}\hat{\mathbf{x}}]$, it follows that $\hat{\mathbf{G}} = \hat{\mathbf{W}}$ and that $\hat{\mathbf{x}}$ is given by the last column of $\hat{\mathbf{V}}$. Our approach for NNTLS is summarized in Algorithm 6.

Algorithm 6 Overview of our nonnegative total least squares (NNTLS) method.

Input: $\mathbf{H} \in \Re^{N\times M}, \mathbf{y} \in \Re^{N\times 1}$
Output: $\hat{\mathbf{x}} \in \Re^{M\times 1}$
1: Find the $\hat{\mathbf{W}} \in \Re^{N\times M}$ and $\hat{\mathbf{V}} \in \Re^{M\times(M+1)}$ that minimize the objective function of (153).
2: $\hat{\mathbf{x}} \leftarrow$ last column of $\hat{\mathbf{V}}$.

We now discuss step 1 of Algorithm 6 in more detail. If we view the constraint in (153) as partial knowledge on $\mathbf{V}$, then (153) closely resembles the partially-supervised nonnegative matrix factorization (PS-NMF) problem of Chapter 4. We seek to approximate the nonnegative matrix $[\mathbf{H} \;\; \mathbf{y}]$ with a lower-rank nonnegative factorization $[\mathbf{H} \;\; \mathbf{y}] \approx \mathbf{W}\mathbf{V}$, but the first M columns of $\mathbf{V}$ are known and only the last column needs to be estimated. If we

[1] Note that, unlike in Chapter 4, we are here only concerned with the *product* of the two NMF factors, and not the factors themselves.

employ the gradient descent technique as in Chapter 4, then each iteration is given by

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} + \alpha_{ij}^{(n)} \frac{\partial}{\partial [\mathbf{W}]_{ij}} \Psi(\mathbf{W}, \mathbf{V}) \Big|_{\mathbf{W}=\hat{\mathbf{W}}^{(n)}, \mathbf{V}=\hat{\mathbf{V}}^{(n)}}, \tag{154a}$$

$$[\hat{\mathbf{V}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{V}}^{(n)}]_{j\gamma} + \delta_{j\gamma}^{(n)} \frac{\partial}{\partial [\mathbf{V}]_{j\gamma}} \Psi(\mathbf{W}, \mathbf{V}) \Big|_{\mathbf{W}=\hat{\mathbf{W}}^{(n)}, \mathbf{V}=\hat{\mathbf{V}}^{(n)}} \text{ for } \gamma > M \text{ only}, \tag{154b}$$

where $\Psi(\mathbf{W}, \mathbf{V})$ is the squared-error objective function. This differs from the conventional NMF approach (125) in that (154b) is computed only for $\gamma > M$; the first M columns of $\mathbf{V}$ are collectively known to be equal to the identity matrix and do not need to be estimated. The corresponding multiplicative update rule is given by

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} [f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)})]_{ij}, \tag{155a}$$

$$[\hat{\mathbf{V}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{V}}^{(n)}]_{j\gamma} [g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)})]_{j\gamma} \quad \text{for } \gamma > M \text{ only}, \tag{155b}$$

where f and g are the standard NMF update equations for the squared-error cost function [46, 1]:

$$f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)}) \triangleq [\mathbf{H} \;\; \mathbf{y}] (\hat{\mathbf{V}}^{(n)})^T \,./\, (\hat{\mathbf{W}}^{(n)} \hat{\mathbf{V}}^{(n)} (\hat{\mathbf{V}}^{(n)})^T), \tag{156a}$$

$$g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)}) \triangleq (\hat{\mathbf{W}}^{(n)})^T [\mathbf{H} \;\; \mathbf{y}] \,./\, ((\hat{\mathbf{W}}^{(n)})^T \hat{\mathbf{W}}^{(n)} \hat{\mathbf{V}}^{(n)}). \tag{156b}$$

A small positive constant (e.g., 10^{-9}) may be added to the denominators to avoid division by zero [68, 1]. Equation (155) can be implemented in matrix form as

$$\hat{\mathbf{W}}^{(n+1)} \leftarrow \hat{\mathbf{W}}^{(n)} .* f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)}), \tag{157a}$$

$$\hat{\mathbf{V}}^{(n+1)} \leftarrow \hat{\mathbf{V}}^{(n)} .* g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)}), \tag{157b}$$

$$\text{Replace first } M \text{ columns of } \hat{\mathbf{V}}^{(n+1)} \text{ with the identity matrix } \mathbf{I}_M. \tag{157c}$$

In (156) and (157), the MATLAB-style notations .* and ./ represent elementwise multiplication and division.

Of course, in implementation, there is no particular reason to take the time to compute the first M columns of $\hat{\mathbf{V}}^{(n+1)}$; the formulation given in (157) is intended to illustrate how an existing NMF implementation may be quickly and easily "retrofitted" for NNTLS.

Like TLS, this NNTLS method is easily generalized to handle multiple data vectors simultaneously; if $\mathbf{Y}$ has K columns, then $\mathbf{V}$ will have $M+K$ columns, and $\hat{\mathbf{X}}$ will be given by the last K columns of $\hat{\mathbf{V}}$. This thesis considers only the single-column case.

As discussed in Chapter 2, the Gaussian model is not generally appropriate for Raman spectroscopy data. This section considered the Gaussian EIV model merely to illustrate a connection between NNTLS and NMF that, to our knowledge, has not been noted elsewhere in the literature. The next section applies the same basic approach to the more relevant Poisson EIV model.

5.3 ML Estimation under an EIV Poisson model

This section considers the Poisson EIV model

$$y_i \sim Pois([\mathbf{G}\mathbf{x}]_i)\,, \tag{158a}$$

$$[\mathbf{H}]_{ij} \sim Pois([\mathbf{G}]_{ij})\,. \tag{158b}$$

Our goal is once again to estimate the mixing vector $\mathbf{x}$ and the "true" library $\mathbf{G}$ from the measured spectrum $\mathbf{y}$ and the given measured library $\mathbf{H}$. This may once again be accomplished using the ML technique:

$$(\hat{\mathbf{G}}, \hat{\mathbf{x}}) = \underset{\mathbf{G}\geq\mathbf{0},\mathbf{x}\geq\mathbf{0}}{\arg\max}\, p(\mathbf{y}, \mathbf{H}; \mathbf{G}, \mathbf{x}) \tag{159a}$$

$$= \underset{\mathbf{G}\geq\mathbf{0},\mathbf{x}\geq\mathbf{0}}{\arg\max}\, \left[\ln p(\mathbf{H}; \mathbf{G}) + \ln p(\mathbf{y}; \mathbf{G}, \mathbf{x})\right]. \tag{159b}$$

For the Poisson model (158), we have

$$\ln p(\mathbf{H}; \mathbf{G}) = \sum_{i,j} [\mathbf{H}]_{ij} \ln[\mathbf{G}]_{ij} - [\mathbf{G}]_{ij} - \ln([\mathbf{H}]_{ij}!)\,, \tag{160a}$$

$$\ln p(\mathbf{y}; \mathbf{G}, \mathbf{x}) = \sum_{i} y_i \ln([\mathbf{G}\mathbf{x}]_i) - [\mathbf{G}\mathbf{x}]_i - \ln(y_i!)\,, \tag{160b}$$

so (159) becomes

$$(\hat{\mathbf{G}}, \hat{\mathbf{x}}) = \underset{\mathbf{G}\geq\mathbf{0},\mathbf{x}\geq\mathbf{0}}{\arg\max} \sum_{j=1}^{M}\sum_{i=1}^{N} [\mathbf{H}]_{ij} \ln[\mathbf{G}]_{ij} - [\mathbf{G}]_{ij} + \sum_{i'=1}^{N} y_{i'} \ln([\mathbf{G}\mathbf{x}]_{i'}) - [\mathbf{G}\mathbf{x}]_{i'} \tag{161a}$$

$$= \underset{\mathbf{G}\geq\mathbf{0},\mathbf{x}\geq\mathbf{0}}{\arg\max} \sum_{j=1}^{M+1}\sum_{i=1}^{N} [\mathbf{H}\ \ \mathbf{y}]_{ij} \ln[\mathbf{G}\ \ \mathbf{G}\mathbf{x}]_{ij} - [\mathbf{G}\ \ \mathbf{G}\mathbf{x}]_{ij} \tag{161b}$$

$$= \underset{\mathbf{G}\geq\mathbf{0},\mathbf{x}\geq\mathbf{0}}{\arg\min} \sum_{i,j} [\mathbf{G}\ \ \mathbf{G}\mathbf{x}]_{ij} - [\mathbf{H}\ \ \mathbf{y}]_{ij} \ln[\mathbf{G}\ \ \mathbf{G}\mathbf{x}]_{ij}\,. \tag{161c}$$

As with the Gaussian EIV model of the previous section, (161c) seeks the nonnegative rank-M matrix $\mathbf{F}_M$ that "best" approximates $[\mathbf{H} \;\; \mathbf{y}]$, subject to the constraint that the last column of $\mathbf{F}_M$ is a nonnegative linear combination of the first M columns. However, while the Gaussian EIV model chooses the "best" approximation by minimizing the squared error, the Poisson EIV model uses the cost function of (161c), which is the same objective function used to minimize the I-divergence from $[\mathbf{H} \;\; \mathbf{y}]$ to $\mathbf{F}_M$ (see Section 2.3.2). Thus, ML estimation under a Poisson EIV model can be thought of as a "total minimum I-divergence" problem. We have not seen this problem addressed in the literature.

We proceed with the same basic approach as the previous section, and rewrite (161c) as a constrained NMF problem:

$$(\hat{\mathbf{W}}, \hat{\mathbf{V}}) = \underset{\mathbf{W}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}}{\arg\min} \; I\left([\mathbf{H} \;\; \mathbf{y}] \| \mathbf{W}\mathbf{V}\right) \qquad \text{s.t.} \;\; [\mathbf{V}]_{j\gamma} = \begin{cases} 1 & j = \gamma \\ 0 & j \neq \gamma, \gamma < M \,. \end{cases} \tag{162}$$

The resulting multiplicative update rule again has the form

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} [f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)})]_{ij} \,, \tag{163a}$$

$$[\hat{\mathbf{V}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{V}}^{(n)}]_{j\gamma} [g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)})]_{j\gamma} \quad \text{for } \gamma > M \text{ only} \,, \tag{163b}$$

but now f and g are the standard NMF update equations for the I-divergence cost function [45, 46, 75]:

$$f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)}) \triangleq \left[\left([\mathbf{H} \;\; \mathbf{y}] ./ (\hat{\mathbf{W}}^{(n)} \hat{\mathbf{V}}^{(n)}) \right) (\hat{\mathbf{V}}^{(n)})^T \right] ./ \left(\mathbf{1}_{N\times(M+1)} (\hat{\mathbf{V}}^{(n)})^T \right) \,, \tag{164a}$$

$$g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)}) \triangleq \left[(\hat{\mathbf{W}}^{(n)})^T \left([\mathbf{H} \;\; \mathbf{y}] ./ (\hat{\mathbf{W}}^{(n)} \hat{\mathbf{V}}^{(n)}) \right) \right] ./ \left((\hat{\mathbf{W}}^{(n)})^T \mathbf{1}_{N\times(M+1)} \right) \,, \tag{164b}$$

where $\mathbf{1}_{N\times(M+1)}$ is an all-one matrix with the same dimensions as $[\mathbf{H} \;\; \mathbf{y}]$. A small positive constant (e.g., 10^{-9}) may be added to the denominators to avoid division by zero [68, 1]. Equation (163) can once again be implemented in the matrix form of (157). We again note that there is no particular reason to compute the first M columns of $\hat{\mathbf{V}}^{(n+1)}$; the formulation given in (157) is intended to illustrate the connection between the EIV problem and NMF. Our approach for ML estimation under the Poisson EIV model is summarized in Algorithm 7.

Algorithm 7 Total minimum I-divergence algorithm. The update equations f and g are given by (164).

Input: $\mathbf{H} \in \Re^{N\times M}, \mathbf{y} \in \Re^{N\times 1}$
Output: $\hat{\mathbf{x}} \in \Re^{M\times 1}$

```
Initialize with some positive Ŵ^(0), V̂^(0)
n ← 0
while not converged do
    Ŵ^(n+1) ← Ŵ^(n) .* f(Ŵ^(n), V̂^(n))
    V̂^(n+1) ← V̂^(n) .* g(Ŵ^(n), V̂^(n))
    Replace first M columns of V̂^(n+1) with the identity matrix I_M.
    n ← n + 1
end while
x̂ ← last column of V̂.
```

As discussed in Section 4.2, the I-divergence cost function is convex in either $\mathbf{W}$ or $\mathbf{V}$ separately, but non-convex in both $\mathbf{W}$ and $\mathbf{V}$ together [46]. This means there may be multiple local minima and/or saddle points, making it difficult to find a global minimum. The advantage of our NMF-based approach to parameter estimation under the EIV Poisson model is that it allows us to draw from the extensive NMF literature. For example, the multiplicative update algorithm for the I-divergence cost function has been shown to globally converge to a stationary point, which is a necessary condition for a local minimum [20, 50].

It is straightforward to incorporate the "background counts" parameter (see Section 2.2.2) into the Poisson EIV model. If

$$y_i \sim Pois([\mathbf{G}\mathbf{x}]_i + [\boldsymbol{\mu}^{b-field}]_i)\,, \tag{165a}$$

$$[\mathbf{H}]_{ij} \sim Pois([\mathbf{G}]_{ij} + [\mathbf{M}^{b-lab}]_{ij})\,, \tag{165b}$$

where $\boldsymbol{\mu}^{b-field}$ and $\mathbf{M}^{b-lab}$ are the (known) parameters for the field sensor and the laboratory instrument, respectively, then the update equations are

$$f(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)}) \triangleq \left[\left([\mathbf{H}\ \ \mathbf{y}]./(\hat{\mathbf{W}}^{(n)}\hat{\mathbf{V}}^{(n)} + \mathbf{M}^b)\right)(\hat{\mathbf{V}}^{(n)})^T\right] ./ \left(\mathbf{1}_{N\times(M+1)}(\hat{\mathbf{V}}^{(n)})^T\right), \tag{166a}$$

$$g(\hat{\mathbf{W}}^{(n)}, \hat{\mathbf{V}}^{(n)}) \triangleq \left[(\hat{\mathbf{W}}^{(n)})^T\left([\mathbf{H}\ \ \mathbf{y}]./(\hat{\mathbf{W}}^{(n)}\hat{\mathbf{V}}^{(n)} + \mathbf{M}^b)\right)\right] ./ \left((\hat{\mathbf{W}}^{(n)})^T\mathbf{1}_{N\times(M+1)}\right), \tag{166b}$$

where $\mathbf{M}^b \triangleq [\mathbf{M}^{b-lab}\ \ \boldsymbol{\mu}^{b-field}]$. The derivation directly follows the derivation found in [75] and is omitted here for brevity.

5.3.1 Simulation Results

We now revisit the simulated scenario of Section 2.6.3, but this time we consider the case in which the algorithms have access to only a noisy observation of the reference library. For this simulation we let $[\mathbf{M}^{b-lab}]_{ij} = 0$, so that

$$y_i \sim Pois([\mathbf{G}\mathbf{x}]_i + [\boldsymbol{\mu}^{b-field}]_i) \tag{167a}$$

$$[\mathbf{H}]_{ij} \sim Pois([\mathbf{G}]_{ij})\,. \tag{167b}$$

We once again let $[\boldsymbol{\mu}^{b-field}]_i = 5$ for all i, and the true library $\mathbf{G}$ is the same library that was used in Section 2.6.3. We again analyze the estimation performance for the mixing coefficient of the 27^{th} chemical of the library (with the Raman spectrum shown in Figure 11) when it is the only substance present. As in Section 2.6.3, we vary the signal energy by sweeping the mixing coefficient from $x_{27} = 0.1$ to $x_{27} = 10.0$, and examine the performance of the algorithms under the different energy levels. The sample RMSE and sample bias of the algorithms are plotted in Figures 36 and 37, respectively. We used 250 Monte Carlo runs. Note that for the EIV model, each run computes a new Poisson-distributed dataset for both the sensor data *and* the library.

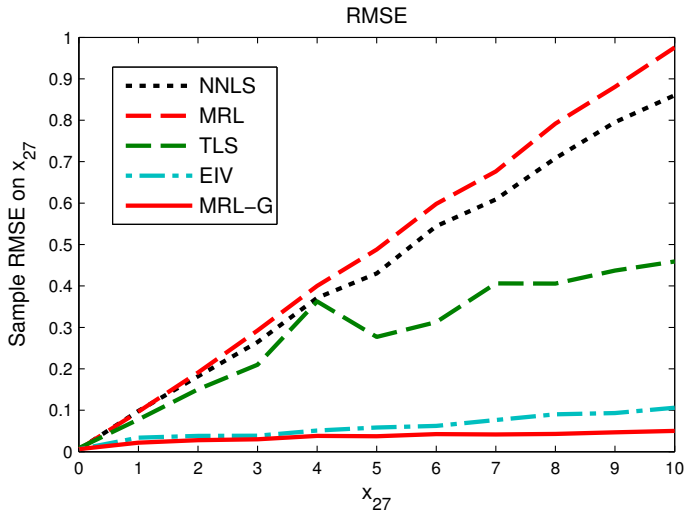


Figure 36: Sample RMSE of the algorithms. The baseline case MRL-G artificially uses the true library G. The EIV curve uses the Poisson/I-divergence formulation.

As discussed in Chapter 2, the modified Richardson-Lucy (MRL) algorithm computes

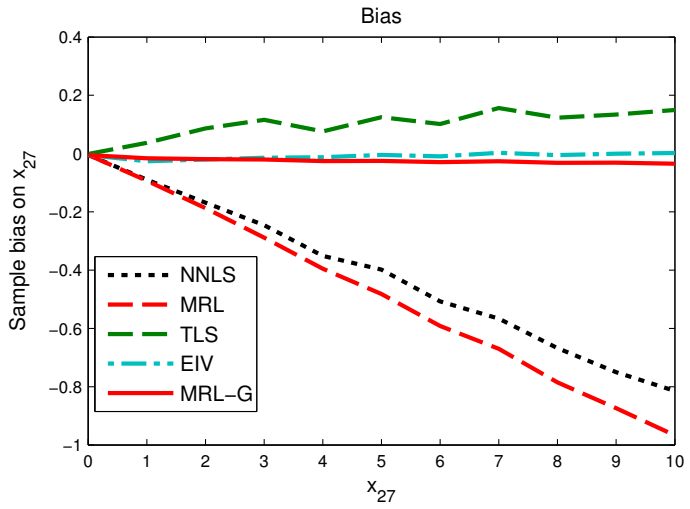


Figure 37: Sample bias of the algorithms. The baseline case MRL-G artificially uses the true library G. The EIV curve uses the Poisson/I-divergence formulation.

the ML estimate for the mixing vector under the Poisson model when it is assumed that the reference library is error-free. In this simulation, both the MRL and EIV algorithms were initialized with the NNLS estimate (any zero values in this initial estimate, however, were replaced with a small constant). The algorithms were terminated when $\|\hat{\mathbf{x}}^{(n+1)} - \hat{\mathbf{x}}^{(n)}\|_2 / \|\hat{\mathbf{x}}^{(n+1)}\|_2 <$ 1e-6. As shown in Figures 36 and 37, the MRL algorithm has much higher error (most of which is due to bias) than the EIV algorithm. This is unsurprising, as the MRL algorithm does not account for uncertainty in the reference library.

The nonnegative least squares (NNLS) and MRL approaches both have large negative bias for this particular example. This bias is partially due to the nonnegativity constraint, as explained in Section 2.6.3. However, by comparing Figure 37 with Figure 13, we see that the bias is much greater in this case; this suggests that most of the bias arises from assuming that the noisy library is noise-free.

The NNLS algorithm, which implicity assumes a Gaussian model, actually gives lower error than the MRL algorithm in this test. There is no particular reason to expect otherwise; the model used to simulate the data is different than the model for which the MRL algorithm is designed. Similarly, the TLS method, which corresponds to the EIV Gaussian model without any nonnegativity constraint, is seen to perform considerably worse than the EIV

Poisson method. However, the TLS approach still outperforms the methods that neglect the library error.

We note that Figure 36 is plotted on a different scale than Figure 12; the estimation error is much higher when $\mathbf{H}$ is used by the algorithms instead of $\mathbf{G}$. To illustrate this, we also plotted the results for the "clairvoyant" MRL algorithm, which "cheats" by using the true library $\mathbf{G}$. We call this method "MRL-G," and its performance is seen to match the corresponding curve in Figure 12, as expected. While this method obviously cannot be implemented in practice, it provides a useful baseline. The EIV method is seen to have only slightly more error than this clairvoyant baseline; even when there is a significant amount of error in the library, acceptable performance may be obtained under the EIV framework.

Many of the energy levels in the scenario analyzed above are not likely to arise in a practical setting. For example, a value of $x_{27} = 10$ means that the measured spectrum $\mathbf{y}$ has ten times as much energy as the 27^{th} spectrum in the reference library. As discussed in Section 2.2.2 (see Equation (18)), the signal-to-noise ratio (SNR) of the spectral data increases with the square root of the mean of the data. Thus, a value of $x_{27} = 10$ implies that the field data $\mathbf{y}$ is *cleaner* than the laboratory data—the ratio of the SNRs in this case would roughly be $\sqrt{10}$. This kind of situation will rarely occur in practice; the reference spectra will typically be much cleaner and stronger than the spectra measured in the field.

The simulation results above indicated that the EIV approach offers drastic improvement over MRL if the library spectra are noisier than the field data. To see if this conclusion holds for the more realistic case in which the field data is noisier than the library spectra, we next analyzed the estimation performance of the algorithms as a function of the ratio of the SNRs of the field and library spectra. These results are plotted in Figure 38. The x-axis in Figure 38 is the sum of each spectrum in the reference library (i.e., we normalize each individual spectrum such that it sums to the value shown on the x-axis; each of the library spectra sums to this same value); we consider 13 energy levels ranging from 1e4 to 1e7. In constrast, for the simulation of Figure 36, each spectrum in the library had a sum of 1e4. The value of x_{27} decreases from left to right, to maintain a constant energy level of 1e5 for $\mathbf{y}$. Thus, the left-most point in Figure 38 corresponds to the right-most point of

Figures 36 and 37. The y-axis of Figure 38 contains the *normalized* RMSE on x_{27}, which is defined as

$$\text{normalized RMSE}(\hat{x}_{27}) \triangleq \text{RMSE}(\hat{x}_{27})/x_{27}\,. \tag{168}$$

As shown in the figure, when the library spectrum has a sum of 1e5—i.e., when it has the same SNR as $\mathbf{y}$—the EIV method performs somewhat better than MRL, though the difference is not nearly as great as before. If the library spectrum has ten times as much energy as the field spectrum (i.e., if it sums to 1e6), then the benefit of EIV is practically nonexistent. It appears that the EIV approach will usually be unnecessary for this particular scenario.

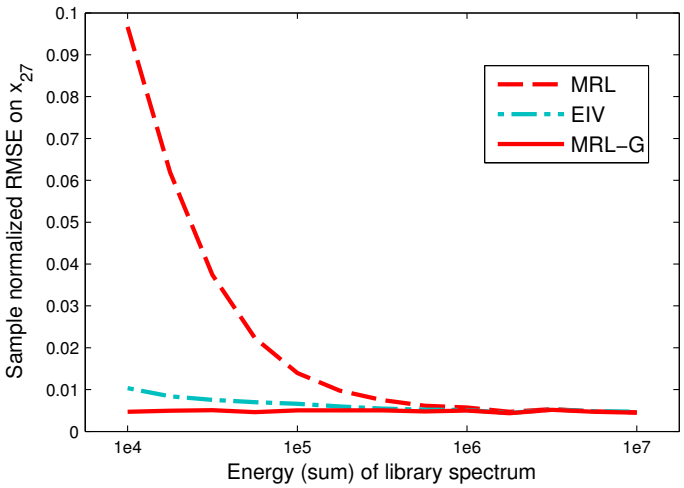


Figure 38: Normalized RMSE of the algorithms plotted vs. the energy (sum) of each spectrum in the reference library.

5.4 Conclusions

This chapter addressed the problem of spectral unmixing under an errors-in-variables (EIV) model. We showed that maximum-likelihood estimation under this model can be viewed as a problem of low-rank approximation with nonnegativity constraints. From this perspective, the problem is naturally addressed using a variation of the partially-supervised nonnegative matrix factorization (PS-NMF) method of Chapter 4. This same basic approach can be applied to the nonnegative total least squares (NNTLS) problem.

Our simulation results indicate that even when there is a significant amount of error in the reference library, acceptable performance may be obtained under the EIV framework. However, if the SNR of the library data is much greater than the SNR of the field data $\mathbf{y}$, as will typically be the case in practice, the EIV approach appears to offer little improvement over the methods that do not account for library error.

In some situations, differences between the laboratory and field instruments may be encountered that are not well understood or easily modeled. Even if the laboratory SNR is high enough that Poisson noise in the lab data is not a problem, the EIV algorithm presented in this chapter may still be reasonable if there are other discrepancies between the lab and field data that are difficult to track down. In such cases, the "total minimum I-divergence" interpretation of the EIV algorithm becomes attractive; even without a well-defined statistical model for those lab-to-field variations, the I-divergence provides a natural discrepancy measure for nonnegative quantities [10, 8].

CHAPTER VI

SUMMARY AND FUTURE RESEARCH

The objective of this dissertation was to develop models, algorithms, and performance bounds for the identification of chemicals from their Raman spectra.

6.1 Contributions

This thesis:

- Developed a probabilistic model, based on previous work in astronomical image restoration, for a basic dispersive Raman measurement system. Applied the well-known modified Richardson-Lucy algorithm (the classical expectation-maximization algorithm) to the problem of chemical mixture estimation.

- Derived performance bounds for chemical mixture estimation (in the form of Cramér-Rao lower bounds) and target detection (in the form of Neyman-Pearson bounds) under our Raman measurement model.

- Investigated the properties of several weighted least-squares and generalized least-squares methods for the case of Poisson data.

- Developed a multiple hypothesis detection (MHD) framework for the detection of chemicals, and derived an approximation to the MAP decision rule. The resulting approximation is related to the Minimum Description Length approach to inference. Showed that the common spectral unmixing approach may be viewed as an approximation to the optimal method. One of the advantages of the MHD approach is that it is naturally extended to the problem of detecting a chemical from a given set.

- Illustrated several fundamental shortcomings of the GLRT for the problem of detecting chemicals from a known reference library (discussed in more detail below). The GLRT

has been applied to this problem in the literature, but we have not previously seen these shortcomings addressed.

- Illustrated several limitations of the two-stage nonnegative matrix factorization (NMF) approach for spectral data analysis. Developed a novel partially-supervised variation of NMF for the detection of objects from an incomplete library. This algorithm is applicable to any problem in which a target is identified by comparing a block of measured data to a library of known constituent signatures.

- Developed an algorithm for ML estimation under a Poisson errors-in-variables model (which accounts for uncertainty in the reference library spectra). Showed that the same basic approach can be applied to the nonnegative total least squares (NNTLS) problem. We have not previously seen these problems addressed in the literature.

- Presented a novel probabilistic model for a dispersive Raman device featuring an *intensified* charge-coupled device (ICCD) detector (see Appendix B).

In most of our experiments, we found that choosing an appropriate general detection philosophy (e.g., spectral unmixing versus GLRT, whether or not to account for "unknown" chemicals, etc.) made a greater difference than using the "correct" noise model.

A fundamental limitation of the GLRT is that it addresses the binary hypothesis testing problem, while in many applications—such as the detection of chemicals from a known reference library—the problem may be more naturally expressed as one of multiple hypothesis detection. Our simulation results suggest that most of the error in the GLRT arises from the wrong hypotheses being tested rather than the mixing coefficients being unknown. It is thus unsurprising that the simple spectral unmixing detection approach performed better in our simulations than the GLRT. Furthermore, because of the nonnegativity of the parameters, the GLRT does not generally attain its nominal asymptotic performance; it does not generally have the CFAR property. We do not recommend the GLRT for the problem of detecting chemicals from a known reference library.

6.2 Directions for Future Work

Some possible avenues for further exploration include:

- **Development of ICCD data model:** Some Raman devices use an *intensified* charge-coupled device (ICCD) detector, which features an image intensifier coupled to a CCD array [29, 90]. Appendix B examines one way to incorporate the intensifier into our Raman measurement model. Future work could focus on improving the ICCD model presented in this appendix and on developing algorithms and performance bounds under that model.

- **Investigation of linear mixing assumption:** One of the core assumptions made in this thesis was that the constituent spectra mix linearly. While this assumption is ubiquitous in the literature, it may not always be completely appropriate for Raman spectroscopy. In general, the linear mixing model (LMM) is valid for situations in which the reflecting surface is a "checkerboard mixture" and in which a given photon interacts with only one component in the mixture [38]. This may be the case for many hyperspectral imaging scenarios in which the surface area consists of spatially segregated patterns of constituent chemicals. However, the mixtures analyzed using Raman spectroscopy are more likely to be intimately associated; the components are likely mixed on spatial scales smaller than the path length of photons in the mixture. In this scenario, an incident photon may experience multiple bounces, in which case the resulting spectrum of the mixture no longer has a linear relationship with the constituent spectra. Future work could focus on further investigating this issue.

- **Fluorescence background accommodation:** One of the main topics addressed in this dissertation was how to model and deal with the noise introduced by the sensor. Another fundamental challenge in Raman spectroscopy is the presence of background interference that is not well modeled as coming from within the sensor itself. This "interference" is the component of the measured spectrum that originates from effects other than Raman scattering. The predominant source of interference is typically fluorescence. Appendix C introduces a basic general approach for dealing with the

fluorescence background. Applying more sophisticated versions of this approach under our measurement model is left as an important topic of future research.

- **Approximations to MHD detection approach:** Because the number of hypotheses grows exponentially with the size of the reference library, the multiple hypothesis detection (MHD) approach of Chapter 3 is infeasible for the large libraries typically used in practice. As discussed in that chapter, future work could focus on developing approximate methods that do not rely on brute force enumeration of the hypotheses.

- **Combinations of different approaches for handling complicated variations in the data:** Chapter 4 presented methods that account for the presence of chemicals whose signatures are not present in the reference library. Chapter 5 developed techniques to account for errors in the reference library. Appendix C introduces a basic approach for accommodating the fluorescence background. These three thrusts, namely unknown chemicals, imperfect signatures, and fluorescence, were explored separately. Future work should consider incorporating all of these issues, particularly the tension between them. All three thrusts seek to model variability in the data. For instance, a fluorescent background signature that might be accommodated via a polynomial fit might also be accommodated by the "slack" in the library signatures provided by the EIV approaches of Chapter 5, or misinterpreted as an "unknown chemical" by the PS-NMF methods of Chapter 4, or some combination of both. All of these methods seek to provide additional "wiggle room" in fitting the data; the addition of variables such as polynomial coefficients and the elements of "unknown chemical" vectors runs the risk of overfitting the data. This observation leads us to conjecture that Minimum Description Length methods, as mentioned in Chapter 3, might have a role in unifying many of the methods explored in this thesis. If fluorescence models, PS-NMF approaches, and EIV techniques are all required to pay a price for offering tighter fits to the data, that may provide a natural way to balance these approaches against one another.

APPENDIX A

PS-NMF UPDATE EQUATION FOR PENALIZED COST FUNCTION

This appendix derives the multiplicative update algorithm for the objective function

$$\Psi(\mathbf{W},\mathbf{X}) = \sum_{i,\gamma} \left[[\mathbf{WX}]_{i\gamma} - [\mathbf{Y}]_{i\gamma} \ln \left([\mathbf{WX}]_{i\gamma} \right) \right] + \sum_{j,\gamma} [\mathbf{\Lambda}]_{j\gamma} [\mathbf{X}]_{j\gamma} \,. \tag{169}$$

The derivative of Ψ with respect to $[\mathbf{W}]_{ij}$ is given by

$$\begin{aligned} \frac{\partial}{\partial [\mathbf{W}]_{ij}} \Psi &= \sum_{\gamma'} \sum_{i'} \frac{\partial}{\partial [\mathbf{W}]_{ij}} \left([\mathbf{WX}]_{i'\gamma'} \right) - \frac{[\mathbf{Y}]_{i'\gamma'}}{[\mathbf{WX}]_{i'\gamma'}} \frac{\partial}{\partial [\mathbf{W}]_{ij}} \left([\mathbf{WX}]_{i'\gamma'} \right) \\ &= \sum_{\gamma'} [\mathbf{X}]_{j\gamma'} - \frac{[\mathbf{Y}]_{i\gamma'}}{[\mathbf{WX}]_{i\gamma'}} [\mathbf{X}]_{j\gamma'} . \end{aligned} \tag{170}$$

The derivative with respect to $[\mathbf{X}]_{j\gamma}$ is similarly given by

$$\frac{\partial}{\partial [\mathbf{X}]_{j\gamma}} \Psi = \sum_{i'} \left([\mathbf{W}]_{i'j} - \frac{[\mathbf{Y}]_{i'\gamma}}{[\mathbf{WX}]_{i'\gamma}} [\mathbf{W}]_{i'j} \right) + [\mathbf{\Lambda}]_{j\gamma} \,. \tag{171}$$

The general gradient descent technique is therefore given by

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} - \alpha_{ij}^{(n)} \left(\sum_{\gamma'} [\hat{\mathbf{X}}^{(n)}]_{j\gamma'} - \sum_{\gamma'} \frac{[\mathbf{Y}]_{i\gamma'}}{[\hat{\mathbf{W}}^{(n)} \hat{\mathbf{X}}^{(n)}]_{i\gamma'}} [\hat{\mathbf{X}}^{(n)}]_{j\gamma'} \right) \quad \text{for } j > M \tag{172a}$$

$$[\hat{\mathbf{X}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{X}}^{(n)}]_{j\gamma} - \delta_{j\gamma}^{(n)} \left(\sum_{i'} [\hat{\mathbf{W}}^{(n)}]_{i'j} + [\mathbf{\Lambda}]_{j\gamma} - \sum_{i'} \frac{[\mathbf{Y}]_{i'\gamma}}{[\hat{\mathbf{W}}^{(n)} \hat{\mathbf{X}}^{(n)}]_{i'\gamma}} [\hat{\mathbf{W}}^{(n)}]_{i'j} \right) . \tag{172b}$$

Here, the first M columns of $\mathbf{W}$ are known and do not need to be estimated in (172a). If the step sizes are chosen by $\alpha_{ij}^{(n)} = \frac{[\hat{\mathbf{W}}^{(n)}]_{ij}}{\sum_{\gamma'} [\hat{\mathbf{X}}^{(n)}]_{j\gamma'}}$ and $\delta_{j\gamma}^{(n)} = \frac{[\hat{\mathbf{X}}^{(n)}]_{j\gamma}}{\sum_{i'} [\hat{\mathbf{W}}^{(n)}]_{i'j} + [\mathbf{\Lambda}]_{j\gamma}}$, then (172) becomes the multiplicative update rule

$$[\hat{\mathbf{W}}^{(n+1)}]_{ij} \leftarrow [\hat{\mathbf{W}}^{(n)}]_{ij} \frac{\sum_{\gamma'} \frac{[\mathbf{Y}]_{i\gamma'}}{[\hat{\mathbf{W}}^{(n)} \hat{\mathbf{X}}^{(n)}]_{i\gamma'}} [\hat{\mathbf{X}}^{(n)}]_{j\gamma'}}{\sum_{\gamma'} [\hat{\mathbf{X}}^{(n)}]_{j\gamma'}} \quad \text{for } j > M \tag{173a}$$

$$[\hat{\mathbf{X}}^{(n+1)}]_{j\gamma} \leftarrow [\hat{\mathbf{X}}^{(n)}]_{j\gamma} \frac{\sum_{i'} \frac{[\mathbf{Y}]_{i'\gamma}}{[\hat{\mathbf{W}}^{(n)} \hat{\mathbf{X}}^{(n)}]_{i'\gamma}} [\hat{\mathbf{W}}^{(n)}]_{i'j}}{\sum_{i'} [\hat{\mathbf{W}}^{(n)}]_{i'j} + [\mathbf{\Lambda}]_{j\gamma}} \,. \tag{173b}$$

This is expressed in matrix form as

$$\hat{\mathbf{W}}^{(n+1)} \leftarrow \hat{\mathbf{W}}^{(n)} .* \left[\left(\mathbf{Y} ./ (\hat{\mathbf{W}}^{(n)} \hat{\mathbf{X}}^{(n)}) \right) (\hat{\mathbf{X}}^{(n)})^T \right] ./ \left(\mathbf{1}_{N \times K} (\hat{\mathbf{X}}^{(n)})^T \right) \tag{174a}$$

$$\text{Replace first } M \text{ columns of } \hat{\mathbf{W}}^{(n+1)} \text{ with the known library } \mathbf{A} \tag{174b}$$

$$\hat{\mathbf{X}}^{(n+1)} \leftarrow \hat{\mathbf{X}}^{(n)} .* \left[(\hat{\mathbf{W}}^{(n)})^T \left(\mathbf{Y} ./ (\hat{\mathbf{W}}^{(n)} \hat{\mathbf{X}}^{(n)}) \right) \right] ./ \left((\hat{\mathbf{W}}^{(n)})^T \mathbf{1}_{N \times K} + \mathbf{\Lambda} \right) . \tag{174c}$$

In (174), the MATLAB-style notations .* and ./ represent elementwise multiplication and division.

APPENDIX B

INCORPORATING THE INTENSIFIER INTO THE RAMAN DATA MODEL

Chapter 2 presented a data model for a dispersive Raman instrument based on the physics of two key components: the spectrometer and the charge-coupled device (CCD) detector (see Figure 5). Some Raman devices (e.g., [28]) use an *intensified* charge-coupled device (ICCD) detector, which features an image intensifier coupled to a CCD array [29, 90]. The resulting system is diagrammed in Figure 39. The intensifier amplifies the number of photons internally and thus allows detection of weaker signals than is possible with a non-intensified CCD. Perhaps more importantly, the gain of the intensifier is controlled by adjusting a voltage applied to the device, and temporal gating/shuttering may therefore be accomplished by simply turning off the control voltage. On the other hand, the ICCD typically has worse spatial resolution than its non-intensified counterpart, and at moderate-to-high light levels the ICCD typically yields lower signal-to-noise ratio. This appendix examines how to incorporate the intensifier into the model of Chapter 2. Further development and analysis of the ICCD model and corresponding algorithms are left as an important topic for future research.

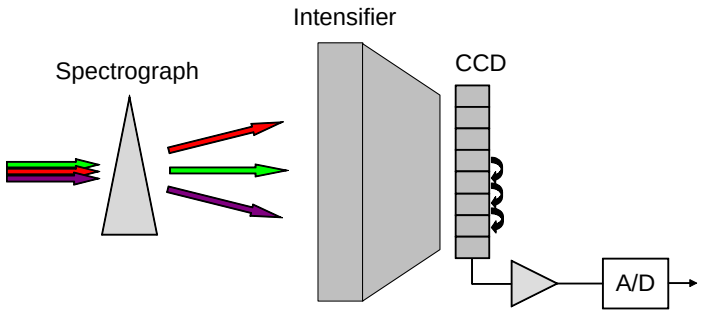


Figure 39: Key components of a dispersive Raman system featuring an intensified charge-coupled device (ICCD) detector.

The three main components of a typical intensifier are illustrated in Figure 40. The

photocathode converts the incident photons into electrons, and this image of electrons is then focused onto the microchannel plate (MCP). The MCP, which is an array of thousands of miniature parallel glass hollow tubes, multiplies the number of incoming electrons by a gain determined by the "bias voltage" applied to it. The resulting electrons are then converted back to photons by the phosphor screen. The output signal is typically linked to the CCD via a fiberoptic bundle, although lens-coupled systems have also been used.

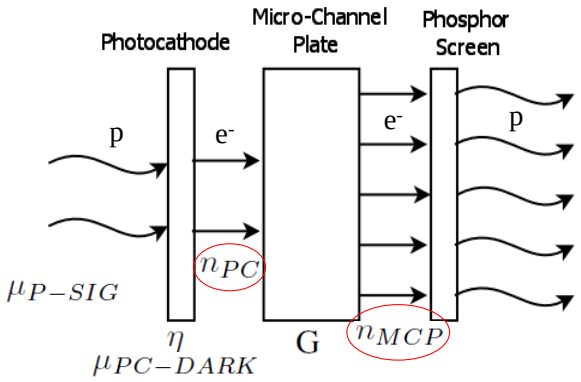


Figure 40: Three main components of the intensifier.

Various authors have analyzed the signal-to-noise ratio (SNR) of the ICCD and compared it with the SNR of the non-intensified CCD under different operating regimes [29, 23, 15]. However, we have seen little work on developing statistical models for ICCD data. The model presented here is somewhat similar to that presented by Sandel and Broadfoot in [76]; both rely on the key experimental observation [27, 100] that the MCP has roughly an exponentially-distributed gain when a single electron is input. The main differences are that [76] models the A/D quantization error, which we neglect, while our model includes several terms that were not included in [76], such as the efficiency and dark current of the photocathode, the point-spread function of the spectrograph, and the efficiency and dark current of the CCD. Another difference is that we use the geometric distribution, and not the exponential, to model the MCP.

For notational simplicity, we first consider the signal at just a single frequency. The "ideal noiseless signal" (in photons) incident on the photocathode is denoted by μ_{P-SIG}. The incident signal is subject to the usual Poisson statistics, and the photocathode has

associated with it a quantum efficiency η and Poisson-distributed dark noise (with mean $\mu_{PC-DARK}$) analogous to the CCD parameters $\boldsymbol{\beta}$ and $\boldsymbol{\mu^b}$ discussed in Chapter 2. The "binomial selection theorem" described in Section 2.2.2 once again enables the quantum efficiency parameter to be incorporated into the Poisson distribution, so that the number of electrons generated on the photocathode is given by

$$n_{PC} \sim Pois(\eta\mu_{P-SIG} + \mu_{PC-DARK}). \tag{175}$$

These parameters appear as labels in Figure 40.

The output of the microchannel plate, n_{MCP}, is not simply the input n_{PC} multiplied by the nominal gain G; the statistics of the MCP must be characterized. Both Wiza [100] and Guest [27] experimentally observed that the MCP gain appears to be approximately exponentially distributed when a single electron is input. However, the exponential distribution, being continuous, does not seem entirely appropriate here, so we instead use its discrete counterpart, the geometric distribution. A geometrically-distributed random variable n with parameter α is characterized by the probability mass function

$$p(n) = (1-\alpha)^n\alpha. \tag{176}$$

If a single electron is input (i.e., if $n_{PC} = 1$), then the MCP output is modeled by

$$n_{MCP} \sim Geom(1/[G+1]). \tag{177}$$

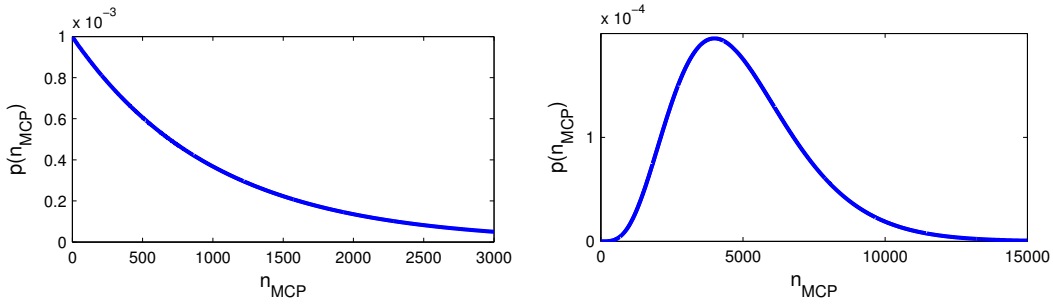
The parameter of the geometric distribution is set to $1/(G+1)$ so that that the mean of the output will match the nominal gain:

$$E(n_{MCP}) = G. \tag{178}$$

The distribution of n_{MCP} for the case in which $n_{PC} = 1$ and $G = 1000$ is plotted in Figure 41(a).

If n_{PC} electrons are input, then the output is the sum of n_{PC} geometrically-distributed random variables, each with a mean of G (and parameter of 1/[G+1]). If these are assumed to be independent, then the sum is characterized by the negative binomial distribution

$$n_{MCP} \sim NB(n_{PC}, 1/[G+1]). \tag{179}$$



(a) $G = 1000, n_{PC} = 1$. (b) $G = 1000, n_{PC} = 5$.

Figure 41: Negative binomial model for microchannel plate.

The distribution of n_{MCP} for the case in which $n_{PC} = 5$ and $G = 1000$ is plotted in Figure 41(b). The conditional mean and variance of n_{MCP} are given by

$$E(n_{MCP}|n_{PC}) = \frac{n_{PC}(1 - 1/[G+1])}{1/[G+1]} \tag{180a}$$

$$= n_{PC}G \tag{180b}$$

and

$$Var(n_{MCP}|n_{PC}) = \frac{n_{PC}(1 - 1/[G+1])}{1/[G+1]^2} \tag{181a}$$

$$= n_{PC}[(G+1)^2 - (G+1)] \tag{181b}$$

$$= n_{PC}(G^2 + G). \tag{181c}$$

By the law of total expectation, we have

$$E(n_{MCP}) = E[E(n_{MCP}|n_{PC})] \tag{182a}$$

$$= E(n_{PC}G) \tag{182b}$$

$$= G(\eta\mu_{P-SIG} + \mu_{PC-DARK}). \tag{182c}$$

Similarly, by the law of total variance,

$$\begin{aligned}
\text{Var}(n_{MCP}) &= E[\text{Var}(n_{MCP}|n_{PC})] + \text{Var}[E(n_{MCP}|n_{PC})] && \text{(183a)}\\
&= E[n_{PC}(G^2+G)] + \text{Var}[n_{PC}G] && \text{(183b)}\\
&= (G^2+G)(\eta\mu_{P-SIG}+\mu_{PC-DARK}) + G^2(\eta\mu_{P-SIG}+\mu_{PC-DARK}) && \text{(183c)}\\
&= (2G^2+G)(\eta\mu_{P-SIG}+\mu_{PC-DARK}) && \text{(183d)}\\
&\approx 2G^2(\eta\mu_{P-SIG}+\mu_{PC-DARK}). && \text{(183e)}
\end{aligned}$$

We note that the accuracy of the approximation made between (183d) and (183e) improves as G increases, because G^2 becomes more significant relative to G. Under this approximation, the variance is twice what it would be if a deterministic MCP gain of G were assumed. Interestingly, some SNR analyses in the literature, without actually specifying a statistical model for the MCP, include an "SNR degradation" term or "noise factor" which is usually around 2. For example, both [29] and [23] assume an MCP noise factor of 1.8, while [15] assumes a noise factor of 1.6.

Finally, as in Chapter 2, we account for the quantum efficiency β and dark current (with mean $\mu_{CCD-DARK}$) on the CCD:

$$y = n_{CCD} + n_{CCD-DARK}, \tag{184}$$

where

$$n_{CCD} \sim \text{Binom}(n_{MCP}, \beta) \tag{185}$$

and

$$n_{CCD-DARK} \sim Pois(\mu_{CCD-DARK}). \tag{186}$$

Again applying the laws of total expectation and total variance, it is straightforward to show that

$$E(y) = \beta G(\eta\mu_{P-SIG}+\mu_{PC-DARK}) + \mu_{CCD-DARK} \tag{187}$$

and that

$$\begin{aligned}
\text{Var}(y) &= 2G^2\beta^2(\eta\mu_{P-SIG}+\mu_{PC-DARK})[1+1/(2\beta G)] + \mu_{CCD-DARK} && \text{(188a)}\\
&\approx 2G^2\beta^2(\eta\mu_{P-SIG}+\mu_{PC-DARK}) + \mu_{CCD-DARK} && \text{(188b)}
\end{aligned}$$

if $G \gg 1/\beta$. Our proposed ICCD model for the ith frequency bin is then summarized as follows:

$$y_i = n_i^{CCD} + n_i^{CCD-DARK} \tag{189a}$$

$$n_i^{CCD-DARK} \sim Pois(\mu_i^{CCD-DARK}) \tag{189b}$$

$$n_i^{CCD} \sim \text{Binom}(n_i^{MCP}, \beta_i) \tag{189c}$$

$$n_i^{MCP} \sim NB(n_i^{PC}, 1/[G+1]) \tag{189d}$$

$$n_i^{PC} \sim Pois(\eta_i \mu_i^{P-SIG} + \mu_i^{PC-DARK}) \tag{189e}$$

$$\mu_i^{P-SIG} = [\mathbf{PAx}]_i, \tag{189f}$$

where, as in Chapter 2, $\mathbf{P}$ is the point-spread function (PSF) of the diffraction grating, $\mathbf{A}$ is the reference library, and $\mathbf{x}$ is the vector of mixing coefficients. Notice that if n_i^{MCP} were Poisson distributed, then n_i^{CCD} would also be Poisson distributed with mean $\beta_i \mu_i^{MCP}$ by the binomial selection theorem; unfortunately, the negative binomial nature of n_i^{MCP} prevents us from making such a convenient simplification.

Of course, the work presented in this appendix is far from complete; future work could focus on improving this ICCD model. For example, the "SNR degradation" term, or "noise factor," is known to typically vary with the bias voltage applied to the MCP [29, 15]; this effect is not explained by our model, which predicts a constant noise factor of 2. Perhaps more importantly, different authors include different system parameters (e.g., the efficiency of the phosphor screen, the loss of the fiberoptic bundle, etc.) in their SNR analyses [29, 23, 15]; future work could determine which terms are most dominant and include only those in our model. Future work could also focus on developing algorithms under this model. For example, for the purpose of estimation, the generalized least-squares methods (and in particular, the nonnegative iteratively reweighted least-squares [NNIRLS] algorithm) of Chapter 2 may be applicable.

Such approximate algorithms may be particularly attractive since merely *writing* the complete equation for the probability density defined by (189a) is a cumbersome task, let alone developing algorithms to maximize the resulting complicated likelihood functions. Equation (189a) is a sum of a Poisson distribution with a binomial distribution, where one

of the parameters of that binomial distribution has a negative binomial distribution, and one of the parameters of that negative binomial distribution is Poisson distributed. The resulting probability density of y_i is likely to contain several infinite sums that do not lead to "clean" expressions. We conjecture that techniques such as saddle-point approximation, which have proven useful in approximating the densities of random variables that are the sum of a Gaussian and a Poisson random variable [86], might be helpful in tackling (189a).

APPENDIX C

A GENERAL APPROACH FOR FLUORESCENCE BACKGROUND ACCOMMODATION

One of the main topics addressed in this dissertation is how to model and deal with the noise introduced by the sensor. Another fundamental challenge in Raman spectroscopy is the presence of background interference that is not well modeled as coming from within the sensor itself. This "interference" is the component of the measured spectrum that originates from effects other than Raman scattering. The predominant source of interference is typically fluorescence. This appendix describes a basic general approach for dealing with the fluorescence background. Applying more sophisticated versions of this approach under our measurement model is left as an important topic of future research.

The typical method to deal with background interference is to first estimate the background and then to subtract it from the measured spectrum as a *preprocessing* step. For example, one common crude technique is to model the slowly-varying background with a polynomial, typically of 5th or 6th order [40, 101, 4, 102]. Under this polynomial background model, the preprocessing approach is illustrated in Figures 42-44. Figure 42 contains the four spectra in the reference library ($M = 4$) of this example. Figure 43(a) shows the (simulated) fluorescence background $\mathbf{f}$, which is a 6th order polynomial. To keep this example simple, we use an AWGN measurement model:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{f} + \mathbf{w} \tag{190a}$$

$$= [\mathbf{A}\ \mathbf{V}] \begin{bmatrix} \mathbf{x} \\ \\ \mathbf{p} \end{bmatrix} + \mathbf{w}\,, \tag{190b}$$

where $\mathbf{V}$ is an $N \times 7$ Vandermonde matrix [89], $\mathbf{p}$ is a 7×1 vector of polynomial coefficients, and $\mathbf{w}$ is white Gaussian noise. The measurement vector $\mathbf{y}$ is plotted in Figure 43(b) for the case in which $x_j = 1/4$ for all j.

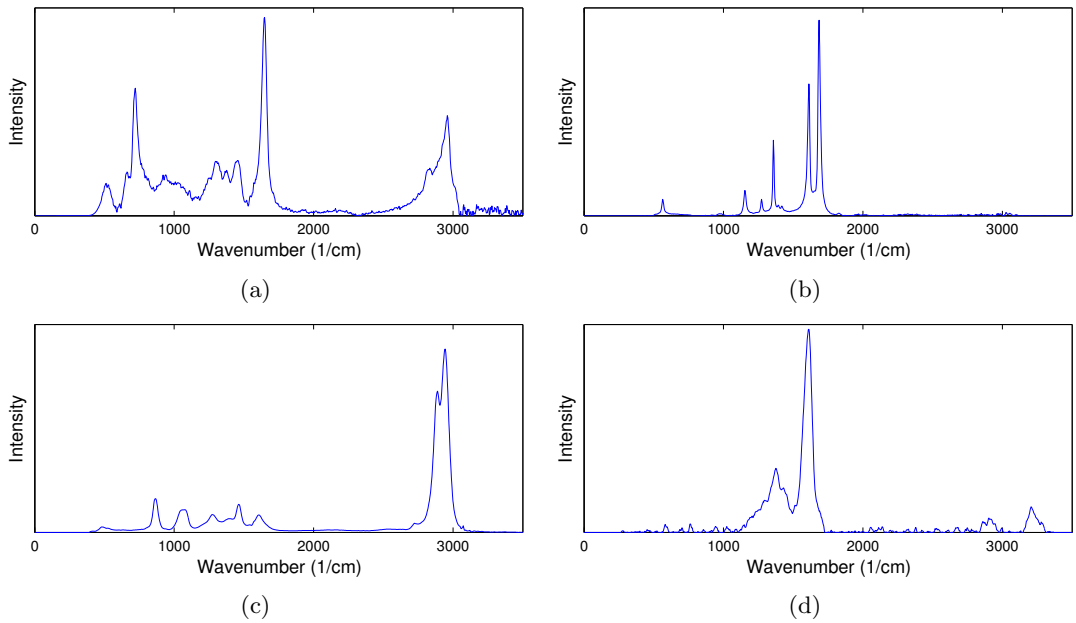
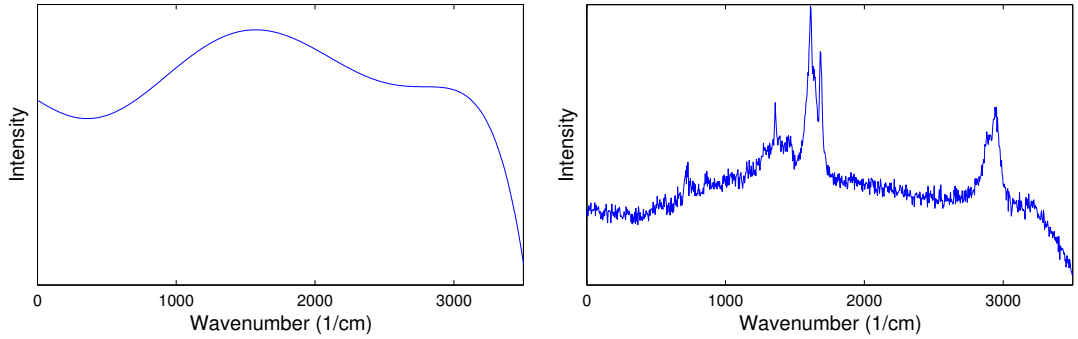


Figure 42: Reference library ($M = 4$).

The preprocessing approach estimates the fluorescence background by fitting a polynomial to the data. The resulting estimated background is plotted for different polynomial orders in Figure 44. When the wrong order is assumed, the performance is degraded. Even when the correct polynomial order is asssumed, the estimated background still has error, since this technique does not account for the contribution from the library spectra.

Our suggested approach is to parametrically model the background and to jointly estimate the nuisance parameters corresponding to the background along with the the parameters of interest (the mixing coefficients). For example, for the simple polynomial background model of (190), we estimate the entire vector $[\mathbf{x}^T \ \mathbf{p}^T]^T$ as suggested by (190b). The resulting estimated background is plotted for different polynomial orders in Figure 45. Since this method co-estimates the mixing coefficients along with the background parameters, the estimated background is much more accurate than that given by the preprocessing approach.

The presence of nuisance parameters decreases the estimation performance for the parameters of interest—the background interference will hurt any processing algorithm, not just algorithms that co-estimate the background. We may quantify this harmful effect by



(a) Fluorescence background, simulated using 6th order polynomial.

(b) Measured spectrum $\mathbf{y}$.

Figure 43: Background interference and measured spectrum.

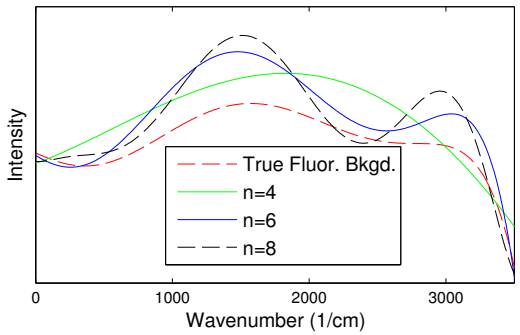


Figure 44: Background estimated by preprocessing approach using polynomials of 4th, 6th, and 8th order.

computing the Cramér-Rao Bound and comparing it to the case where the background parameters are not included in the model. For example, the standard deviation CRLB for the above AWGN model is shown for both cases in Table 11. For the first row, the CRLB uses (4) instead of (190b), leaving the fluorescence background out of the model. The bounds in the first row are smaller than in the second, indicating that if the fluorescence is ignored, we will be overconfident in how well we can estimate the mixing coefficients. For this particular example, the effect is small.

Table 11: Effect of nuisance parameters on the CRLB.

	x_1	x_2	x_3	x_4
CRLB without fluorescence:	0.0057	0.0049	0.0049	0.0054
CRLB with fluorescence:	0.0072	0.0050	0.0054	0.0058

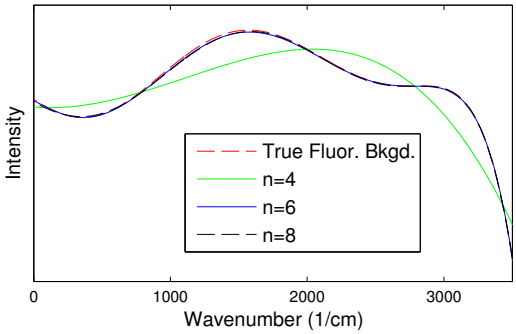


Figure 45: Background estimated by suggested method using polynomials of 4th, 6th, and 8th order.

We conjecture that examples such as this can be improved on by incorporating the nuisance parameters into our measurement model of Section 2.2.2. One challenge is that while the mixing coefficients are constrained to be nonnegative, the polynomial coefficients for the background should be allowed to be negative as well as positive; the resulting polynomial background, however, should be nonnegative [31]. This places some complicated constraints on the coefficients. One approach would be to use polynomials involving sums of squares [43, 3]; another might be use a real-valued polynomial model for the *logarithm* of the fluorescence background. Powell-Sabin splines [98] might also be useful. We leave this as a topic for future research.

Another, related, task for future research is to explore more sophisticated methods to parametrically model the background.

REFERENCES

[1] Berry., M. W., Browne, M., Langville, A. N., Pauca, V. P., and Plemmons, R. J., "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics & Data Analysis*, vol. 52, pp. 155–173, September 2007.

[2] Blackman, S. and Popoli, R., *Design and Analysis of Modern Tracking Systems*. Artech House, 1999.

[3] Brickman, L., "On nonnegative polynomials," *The American Mathematical Monthly*, vol. 69, no. 3, pp. 218–221, 1962.

[4] Cao, A., Pandya, A. K., Serhatkulu, G. K., Weber, R. E., Dai, H., Thakur, J. S., Naik, V. M., Naik, R., Auner, G. W., Rabah, R., and Freeman, D. C., "A robust method for automated background subtraction of tissue fluorescence," *Journal of Raman Spectroscopy*, vol. 38, no. 9, pp. 1199–1205, 2007.

[5] Carroll, R. J. and Ruppert, D., *Transformation and Weighting in Regression*. Chapman and Hall, 1988.

[6] Chen, Y., Rege, M., Dong, M., and Hua, J., "Non-negative matrix factorization for semi-supervised data clustering," *Journal of Knowledge and Information Systems*, vol. 17, no. 3, pp. 355–379, 2008.

[7] Chernoff, H., "On the distribution of the likelihood ratio," *The Annals of Mathematical Statistics*, vol. 25, no. 3, pp. 573–578, 1954.

[8] Choi, K., *Minimum I-divergence Methods for Inverse Problems*. PhD thesis, Georgia Institute of Technology, Dec. 2005.

[9] Cichocki, A., Zdunek, R., and Amari, S., "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Signal Separation*, vol. 3889, (Charleston, SC, USA), pp. 32–39, March 2006.

[10] Csiszár, I., "Why least squares and maximum entropy? - an axiomatic approach to inverse problems," *The Annals of Statistics*, vol. 19, pp. 2033–2066, December 1991.

[11] De Moor, B. L. R., "Total linear least squares with inequality constraints," ESAT-SISTA report 1990-02, Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium, March 1990.

[12] Ding, C., He, X., and Simon, H. D., "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proceedings of the SIAM Data Mining Conference*, pp. 606–610, 2005.

[13] Dixon, D. D., Bhattacharya, D., O'Neill, T. J., Tümer, O. T., White, R. S., Zych, A. D., and Wheaton, W. A., "Constrained linear algebraic deconvolution of poisson data," *Astrophysical Journal*, vol. 457, pp. 789–797, Feb. 1996.

[14] Downs, R. T., "The RRUFF project: an integrated study of the chemistry, crystallography, Raman and infrared spectroscopy of minerals," in *Program and Abstracts of the 19th General Meeting of the International Mineralogical Association*, (Kobe, Japan), 2006. O03-13.

[15] Dussault, D. and Hoess, P., "Noise performance comparison of ICCD with CCD and EMCCD cameras," in *Infrared systems and photoelectronic technology* (Dereniak, E. L., Sampson, R. E., and Johnson, C. B., eds.), vol. SPIE Proc. 5563, pp. 195–204, 2004.

[16] Eckart, C. and Young, G., "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[17] Farrell, M. D., *Analysis of Modeling, Training, and Dimension Reduction Approaches for Target Detection in Hyperspectral Imagery.* PhD thesis, Georgia Institute of Technology, Dec. 2005.

[18] Ferraro, J. R., Nakamoto, K., and Brown, C. W., *Introductory Raman Spectroscopy.* Academic Press, 2nd ed., 2003.

[19] Fessler, J. A., "Iterative methods for image reconstruction." IEEE International Symposium on Biomedical Engineering, May 2008. Available at http://www.eecs.umich.edu/ fessler/papers/files/talk/08/isbi-notes.pdf.

[20] Finesso, L. and Spreij, P., "Nonnegative matrix factorization and I-divergence alternating minimization," *Linear Algebra and its Applications*, vol. 416, pp. 270 – 287, 2006.

[21] Fuhrmann, D., Preza, C., O'Sullivan, J., Snyder, D., and Smith, W., "Spectrum estimation from quantum-limited interferograms," *IEEE Transactions on Signal Processing*, vol. 52, pp. 950–961, April 2004.

[22] Fuller, W. A., *Measurement Error Models.* New York: John Wiley and Sons, 1987.

[23] Glasgow, B. B., Glaser, M. S., and Whitley, R. H., "Remote imaging in the ultraviolet using intensified and non-intensified CCDs," in *Image Acquisition and Scientific Imaging Systems* (Titus, H. C. and Waks, A., eds.), vol. SPIE Proc. 2173, pp. 85–96, 1994.

[24] Golub, G. H. and Van Loan, C. F., "An analysis of the total least squares problem," *SIAM J. Numer. Anal.*, vol. 17, pp. 883–893, Dec. 1980.

[25] Golub, G. H. and Van Loan, C. F., *Matrix Computations.* Baltimore: The Johns Hopkins Univ. Press, 1983.

[26] Grünwald, P. D., Myung, I. J., and Pitt, M. A., *Advances in Minimum Description Length: Theory and Applications (Neural Information Processing).* The MIT Press, 2005.

[27] Guest, A. J., "A computer model of channel multiplier plate performance," *Acta Electronica*, vol. 14, no. 1, pp. 79–97, 1971.

[28] Higdon, N. S., Chyba, T. H., Richter, D. A., Ponsardin, P. L., Armstrong, W. T., Lobb, C. T., Kelly, B. T., Babnick, R. D., and Sedlacek III, A. J., "Laser interrogation of surface agents (LISA) for chemical agent reconnaissance," vol. 4722, pp. 50–59, SPIE, 2002.

[29] Holst, G. C., *CCD Arrays, Cameras, and Displays*. SPIE Press, 2nd ed., 1998.

[30] Hoyer, P., "Non-negative sparse coding," in *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 557 – 565, 2002.

[31] Hudson, D. J., "Least-squares fitting of a polynomial constrained to be either non-negative, non-decreasing or convex," *Journal of the Royal Statistical Society B*, vol. 31, no. 1, pp. 113–118, 1969.

[32] Hyvärinen, A. and Oja, E., "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.

[33] Janesick, J. R., *Scientific Charge-Coupled Devices*. SPIE Press, 2001.

[34] Kay, S., Xu, C., and Emge, D., "Chemical detection and classification in Raman spectra," in *Signal and Data Processing of Small Targets* (Drummond, O. E., ed.), vol. SPIE Proc. 6969, (Orlando, FL), p. 696904, 2008.

[35] Kay, S. M., *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.

[36] Kay, S. M., *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall, 1998.

[37] Kay, S. M., "The multifamily likelihood ratio test for multiple signal model detection," *IEEE Signal Processing Letters*, vol. 12, pp. 369–371, May 2005.

[38] Keshava, N. and Mustard, J. F., "Spectral unmixing," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 44 –57, 2002.

[39] Kim, J. and Park, H., "Sparse nonnegative matrix factorization for clustering," technical report, Georgia Institute of Technology, 2008. Available at http://hdl.handle.net/1853/20058.

[40] Koo, T. W., *Measurement of blood analytes in turbid biological tissue using near-infrared Raman spectroscopy*. PhD thesis, Massachusetts Institute of Technology, Aug. 2001.

[41] Kupinski, M. A. and Barrett, H. H., *Small-animal SPECT imaging*. New York: Springer Science+Business Media, Inc., 2005.

[42] Lanterman, A. D., "Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model order estimation," *International Statistical Review*, vol. 69, pp. 185–212, 2001.

[43] Lasserre, J. B., "A sum of squares approximation of nonnegative polynomials," *SIAM Review*, vol. 49, 2007.

[44] Lawson, C. L. and Hanson, R. J., *Solving Least Squares Problems*. Prentice Hall, 1974.

[45] Lee, D. D. and Seung, H. S., "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[46] Lee, D. D. and Seung, H. S., "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2000.

[47] Lee, H., Yoo, J., and Choi, S., "Semi-supervised nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 17, pp. 4–7, Jan. 2010.

[48] Li, H., Adali, T., Wang, W., and Emge, D., "Non-negative matrix factorization with orthogonality constraints and its application to Raman spectroscopy," *Journal of VLSI Signal Processing*, vol. 48, pp. 83–97, 2007.

[49] Li, T., Ding, C., and Jordan, M. I., "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization," in *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, (Washington, DC, USA), pp. 577–582, IEEE Computer Society, 2007.

[50] Lin, C. J., "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Transactions on Neural Networks*, vol. 18, pp. 1589–1596, Nov. 2007.

[51] Lin, C. J., "Projected gradient methods for nonnegative matrix factorization," *Neural Computation*, vol. 19, pp. 2756 – 2779, 2007.

[52] Lucy, L. B., "An iterative technique for the rectification of observed distributions," *Astronomical Journal*, vol. 79, pp. 745–754, June 1974.

[53] Mackay, D. J. C., *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[54] Mallick, M., Drake, B., Park, H., Register, A., Blair, D., West, P., Palkki, R., Lanterman, A., and Emge, D., "Comparison of Raman spectra estimation algorithms," in *Proceedings of the Twelfth International Conference on Information Fusion*, (Seattle, WA), July 2009.

[55] Manolakis, D., Marden, D., and Shaw, G. A., "Hyperspectral image processing for automatic target detection applications," *The Lincoln Laboratory Journal*, vol. 14, no. 1, pp. 79–116, 2003.

[56] Markovsky, I. and Van Huffel, S., "Overview of total least squares methods," *Signal Processing*, vol. 87, pp. 2283–2302, October 2007.

[57] Mayo, D. W., Miller, F. A., and Hannah, R. W., *Course Notes on the Interpretation of Infrared and Raman Spectra*. John Wiley & Sons, Inc., 2004.

[58] McCain, S. T., Willett, R. M., and Brady, D. J., "Multi-excitation Raman spectroscopy technique for fluorescence rejection," *Opt. Express*, vol. 16, no. 15, pp. 10975–10991, 2008.

[59] McCullagh, P. and Nelder, J. A., *Generalized linear models*. London: Chapman & Hall, 2nd ed., 1989.

[60] Nelder, J. A. and Wedderburn, R. W. M., "Generalized linear models," *Journal of the Royal Statistical Society, Series A, General*, vol. 135, pp. 370–384, 1972.

[61] Paatero, P. and Tapper, U., "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, pp. 111–126, 1994.

[62] Palkki, R. D. and Lanterman, A. D., "Algorithms and performance bounds for chemical identification under a poisson model for Raman spectroscopy," in *Proceedings of the Twelfth International Conference on Information Fusion*, (Seattle, WA), July 2009.

[63] Palkki, R. D. and Lanterman, A. D., "Chemical mixture estimation under a poisson Raman spectroscopy model," *Optical Engineering*, vol. 49, p. 113601, Nov. 2010.

[64] Palkki, R. D. and Lanterman, A. D., "Identifying chemicals from their Raman spectra using minimum description length," in *Signal and Data Processing of Small Targets* (Drummond, O., ed.), vol. SPIE Proc. 7698, (Orlando, FL), p. 769807, 2010.

[65] Palkki, R. D. and Lanterman, A. D., "A non-negative matrix factorization algorithm for the detection of chemicals from an incomplete Raman library," in *Chemical, Biological, Radiological, Nuclear, and Explosives (CBRNE) Sensing XI*, vol. SPIE Proc. 7665, (Orlando, FL), p. 766519, 2010.

[66] Palkki, R. D. and Lanterman, A. D., "A minimum description length approach to detecting chemicals via their Raman spectra," *Optical Engineering*, vol. 50, p. 083601, Aug. 2011.

[67] Palkki, R. D. and Lanterman, A. D., "Detecting constituent objects using partially-supervised nonnegative matrix factorization," *Digital Signal Processing*, (submitted March 2011).

[68] Pauca, V. P., Piper, J., and Plemmons, R. J., "Nonnegative matrix factorization for spectral data analysis," *Linear Algebra and its Applications*, vol. 416, pp. 29 – 47, 2006.

[69] Pitt, G. D., Batchelder, D. N., Bennett, R., Bormett, R. W., Hayward, I. P., Smith, B. J. E., Williams, K. P. J., Yang, Y. Y., Baldwin, K. J., and Webster, S., "Engineering aspects and applications of the new Raman instrumentation," *IEE Proceedings - Science, Measurement and Technology*, vol. 152, pp. 241–318, Nov. 2005.

[70] Ponsardin, P. L., Higdon, N. S., Chyba, T. H., Armstrong, W. T., III, A. J. S., Christesen, S. D., and Wong, A., "Expanding applications for surface-contaminant sensing using the laser interrogation of surface agents (LISA) technique," vol. 5268, pp. 321–327, SPIE, 2004.

[71] Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V. L., and Siemiginowska, A., "Statistics, Handle with Care: Detecting Multiple Model Components with the Likelihood Ratio Test," *Astrophysical Journal*, vol. 571, pp. 545–559, May 2002.

[72] Raman, C. V. and Krishnan, K. S., "A new type of secondary radiation," *Nature*, vol. 121, pp. 501–502, 1928.

[73] Richardson, W. H., "Bayesian-based iterative method of image restoration," *Journal of the Optical Society of America*, vol. 62, pp. 55–59, January 1972.

[74] Rissanen, J., "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[75] Sajda, P., Du, S., and Parra, L., "Recovery of constituent spectra using non-negative matrix factorization," in *Wavelets: Applications in Signal and Image Processing X*, vol. SPIE Proc. 5207, pp. 321–331, 2003.

[76] Sandel, B. R. and Broadfoot, A. L., "Statistical performance of the intensified charge coupled device," *Applied Optics*, vol. 25, pp. 4135–4140, Nov. 1986.

[77] Scharf, L. L. and Friedlander, B., "Matched subspace detectors," *IEEE Trans. on Signal Processing*, vol. 42, pp. 2146–2157, Aug. 1994.

[78] Scharlemann, E. T., "Modeling chemical detection sensitivities of active and passive remote sensing systems," in *Lidar Remote Sensing for Environmental Monitoring IV* (Singh, U. N., ed.), vol. SPIE Proc. 5154, pp. 126–137, 2003.

[79] Schwarz, G. E., "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[80] Shepp, L. A. and Vardi, Y., "Maximum-likelihood reconstruction for emission tomography," *IEEE Trans. on Medical Imaging*, vol. 1, pp. 113–121, 1982.

[81] Shlens, J., "A tutorial on principal component analysis," tech. rep., Systems Neurobiology Laboratory, Salk Insitute for Biological Studies, Dec. 2005.

[82] Sigurdsson, S., Larsen, J., Philipsen, P. A., Gniadecka, M., Wulf, H. C., and Hansen, L. K., "Estimating and suppressing background in Raman spectra with an artificial neural network," technical report, Technical University of Denmark, Lyngby, Denmark, Nov. 2003.

[83] Sirkeci, B., "Total least squares approaches for spectral unmixing of hyperspectral imagery," Master's thesis, Northeastern University, Boston, MA, 2001.

[84] Slamani, M., Chyba, T. H., LaValley, H., and Emge, D., "Spectral unmixing of agents on surfaces for the joint contaminated surface detector (JCSD)," in *Signal and Data Processing of Small Targets* (Drummond, O. E., ed.), vol. SPIE Proc. 6699, p. 66991B, 2007.

[85] Slamani, M., Fisk, B., Chyba, T., Emge, D., and Waugh, S., "An algorithm benchmark data suite for chemical and biological (chem/bio) defense applications," in *Signal and Data Processing of Small Targets*, vol. SPIE Proc. 6969, (Orlando, FL), p. 696903, 2008.

[86] Snyder, D. L., Helstrom, C. W., Lanterman, A. D., Faisal, M., and White, R. L., "Compensation for readout noise in CCD images," *J. Optical Society of America A*, vol. 12, pp. 272–283, February 1995.

[87] Snyder, D. L. and Miller, M. I., *Random Point Processes in Time and Space*. Springer-Verlag, 2nd ed., 1991.

[88] Snyder, D., Hammoud, A., and White, R., "Image recovery from data acquired with a charge-coupled-device camera," *Journal of the Optical Society of America A*, vol. 10, pp. 1014–1023, May 1993.

[89] Strang, G., *Linear Algebra and its Applications*. Harcourt Brace Jovanovich, Inc., 1988.

[90] Sweedler, J. V., Ratzlaff, K. L., and Denton, M. B., *Charge-Transfer Devices in Spectroscopy*. VCH Publishers, Inc., 1994.

[91] Van Huffel, S. and Vandewalle, J., *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM Publications, 1991.

[92] Van Trees, H. L., *Detection, Estimation and Modulation Theory, Part I*. New York: John Wiley and Sons, 1968.

[93] Wang, W., Adali, T., Li, H., and Emge, D., "Detection using correlation bound and its application to Raman spectroscopy," in *2005 IEEE Workshop on Machine Learning for Signal Processing*, (Minnesota), pp. 259–264, IEEE, Sep. 2005.

[94] Wang, W., Adali, T., and Emge, D., "Unsupervised target detection using canonical correlation analysis and its application to raman spectroscopy," in *2007 IEEE Workshop on Machine Learning for Signal Processing*, pp. 247 –252, 2007.

[95] Wheaton, W. A., Dunklee, A. L., Jacobson, A. S., Ling, J. C., Mahoney, W. A., and Radocinski, R. G., "Multiparameter linear least-squares fitting to poisson data one count at a time," *Astrophysical Journal*, vol. 438, pp. 322–340, Jan. 1995.

[96] Wiesel, A., Eldar, Y. C., and Yeredor, A., "Linear regression with gaussian model uncertainty: Algorithms and bounds," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2194–2205, June 2008.

[97] Wilks, S. S., "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[98] Willemans, K. and Dierckx, P., "Nonnegative surface fitting with Powell-Sabin splines," *Numerical Algorithms*, vol. 9, no. 2, pp. 263–276, 1995.

[99] Willett, R., "Multiscale reconstruction for photon-limited shifted excitation Raman spectroscopy," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. III–833 –III–836, 2007.

[100] Wiza, J. L., "Microchannel plate detectors," *Nuclear Instruments and Methods*, vol. 162, pp. 587–601, 1979.

[101] Zhao, J., Lui, H., Mclean, D. I., and Zeng, H., "Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy," *Applied Spectroscopy*, vol. 61, pp. 1225–1232, 2007.

[102] Zhu, Q., Quivey, R. G., and Berger, A. J., "Raman spectroscopic measurement of relative concentrations in mixtures of oral bacteria," *Applied Spectroscopy*, vol. 61, pp. 1233–1237, 2007.

VITA

Ryan Palkki was born and raised in Milwaukie, Oregon. He received his B.S. in electrical engineering from Brigham Young University in 2004. He then joined the Center for Signal and Image Processing at the Georgia Institute of Technology, where his research in target tracking led to an M.S. in electrical engineering in August 2006. Since March 2011 Ryan has been working full-time at MIT Lincoln Laboratory on signal processing for airborne radar.