Purdue University
Purdue e-Pubs

Open Access Dissertations

Theses and Dissertations

8-2016

Multivariate Statistical Methods that Enable Fast Raman Spectroscopy

Owen G. Rehrauer
Purdue University

Follow this and additional works at: https://docs.lib.purdue.edu/open_access_dissertations

Part of the Chemistry Commons

Recommended Citation

Rehrauer, Owen G., "Multivariate Statistical Methods that Enable Fast Raman Spectroscopy" (2016). *Open Access Dissertations*. 836.
https://docs.lib.purdue.edu/open_access_dissertations/836

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Graduate School Form 30
(Updated 11/20/2014)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Owen G. Rehrauer

Entitled MULTIVARIATE STATISTICAL METHODS
THAT ENABLE FAST RAMAN SPECTROSCOPY

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Dor Ben-Amotz

Mary J. Wirth

Garth J. Simpson

Shelley Claridge

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Dor Ben-Amotz

Approved by: Timothy Zwier 04/28/2016

Head of the Department Graduate Program Date

MULTIVARIATE STATISTICAL METHODS THAT ENABLE FAST RAMAN SPECTROSCOPY

A Dissertation
Submitted to the Faculty
of
Purdue University
by
Owen G. Rehrauer

In Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

August 2016
Purdue University
West Lafayette, Indiana

I dedicate this work to the memory of Winthrop H.K. Dahl.

Carpe diem, quam minimum credula postero.

ACKNOWLEDGMENTS

I would like to acknowledge the support provided by my family, particularly my sisters, Meg and Emily, mother, father, and my grandparents: Harold and Ruth and Charles and Virginia. I would also like to acknowledge my close friends Gavin Jones and Terry Villarreal, who have provided me with emotional support and, more critically, have bought me many a conciliatory drink after long weeks in lab. Lastly, I would like to give thanks for the support of Gavin Jones' cat, Meiko, who has dutifully woken me up at 7:30 AM for a year. I am touched she is so committed to my success in lab.

I am forever indebted to Robert Herbert and You Seok Hwang, without whom I would have never made the decision to attend graduate school. I am extremely grateful that Carol Russell, Robert Beck, Larry Simonson, Richard Andrews, Peyton Beales, Ian Parsons, and Charles Phoebe were willing to discuss strategies for selecting a graduate school as well as assist me in applying for graduate school.

I am grateful for the mentorship of David Wilcox, who introduced me to chemometrics and the joys of Matlab and LaTex. I would like acknowledge the many wonderful conversations I have had with Robert Birdsall, Ben Rogers, Blake Rankin, Sam Zukowski, Bharat Mankani, Joshua Long, and Sarah Matt regarding research. Many thanks to Brad Lucier and Greg Buzzard for their guidance with regards to mathematics and statistics as well as their seemingly infinite patience.

Last, but certainly not least; words are inadequate to express my gratitude to Dor Ben-Amotz for teaching more than I ever thought I could know about research and for making me a better scientist.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Rehrauer, Owen G. Ph.D., Purdue University, August 2016. Multivariate Statistical Methods that Enable Fast Raman Spectroscopy. Major Professor: Dor Ben-Amotz.

Raman spectroscopy is a useful tool in investigating inter- and intra-molecular interactions as well as classifying and quantifying chemical species in a sample. Many materials of societal interest, such as proteins and pharmaceuticals, have distinctive Raman spectra with sharp features. However, the adoption of Raman spectra has been hindered by the low rate of Raman scattering, interference from fluorescence, and high spectrometer costs. This work demonstrates multivariate stastical methods that enable fast Raman measurements, despite the low rate of Raman scattering. These methods include a novel type of spectrometer, that uses computer-controlled optical filters to efficiently capture Raman photons and multiplex them onto either one or two photon counting detector(s). This method, referred to as optimal-binary compressive detection (OB-CD), allows for the collection of chemical information in 10's of microseconds, rather than milliseconds as might be common for Raman spectroscopy performed using a multichannel detector. A method for orthogonalizing moderate amounts of fluorescence from Raman signal in OB-CD is presented. Fast imaging, with speeds as high as 2.5 frames-per-second, is demonstrated and algorithms for image denoising are discussed. Lastly, methods that enables Raman classification using minimal computation time and a technique for accurately processing Raman thermometry data are presented.

1. INTRODUCTION

This work describes multivariate statistical methods that either enable microsecond Raman spectroscopic measurements or efficiently convert Raman spectra into low-dimensional chemical information. While this will, necessarily, require a discussion of the mathematics needed to perform such tasks, an emphasis will also be given to concerns more familiar to an analytical chemist. These include: applications to chemical problems, effects of instrumentation, and practical concerns when handling samples. It is my hope that this document can be a "starting-off point" for aspiring chemometricians in the Ben-Amotz lab. While this text will avoid adding rigorous descriptions of common spectral unmixing techniques; for instance, multivariate curve resolution, it is hoped that seeing the application of such methods will inspire graduate students to learn more about this exciting field. I highly recommend that such students consult the book "Introduction to Multivariate Statistical Analysis in Chemometrics" by Varmuza and Filzmoser as a reference on chemometrics. Readers already familiar with such methods will hopefully be able to use this as a quick reference for my work in Raman spectral unmixing, classification, and optimal-binary compressive detection filter generation.

2. OPTIMAL-BINARY COMPRESSIVE DETECTION - OVERVIEW OF ALGORITHM AND DESCRIPTION OF INSTRUMENTATION

2.1 Abstract

This chapter provides an summary of the optimal-binary compressive detection (OB-CD) filter generation algorithm used to generate the binary filters used in OB-CD Raman spectroscopy. Additionally, the spectrometers used to make OB-CD Raman measurements in future chapters are described. The objective of this chapter is to provide readers with a straight-forward description of the generation of OB-CD filters, while referencing papers that provide a more in-depth discussion of this material.

2.2 Background

Raman scattering is a low-rate process when compared to other optical processes such as fluorescence emission or infrared adsorption. As a consequence of this, it can takes on the order of milliseconds (if not seconds) to collect Raman spectra of signal-to-noise ratios (SNR) needed for many applications. This limitation hinders the use of Raman spectroscopy in hyperspectral imaging and hinders the use of Raman spectroscopy in applications where temporal resolution greater 1 ms in required (e.g., reaction kinetics monitoring). Fundamentally, this is due to the low SNR of fast Raman measurements, which can be attributed to the low-rate of Raman scattering (signal) and to the read noise of multichannel detectors (noise), typically a charged couple device (CCD) camera or a complimentary metal-oxide semiconductor (CMOS) camera. Other strategies to improve the SNR of fast Raman measurements include increasing the rate of Raman scattering, such as by using surface-enhanced Raman scattering (SERS) substrates. Unfortunately, many analysis types are not amenable to use of SERS particles (such as the real-time monitoring of pharmaceuticals), finding an appropriate SERS particle for an analyte can prove difficult, and, empirically, the SERS effect can prove difficult to repeatably implement (as per anecdotes in the Ben-Amotz lab). Other strategies for increasing signal include using a high power laser, which can prove troublesome for samples that are sensitive to heating or burning, and the use of an excitation laser with a visible wavelength (since Raman scattering scales with the reciprocal of λ^4), which unfortunately causes many samples to exhibit fluorescent emission.

This then leads one to consider ways to reduce detector read noise. One method to reduce detector read noise is cooling the CCD or CMOS chip. This, unfortunately, is already factored into the 1 ms minimum integration time considered above. Moreover, CCD cameras have a read noise that increases with increased read times. Electronic delays in commercial CCD cameras also limit the speed at which signal can be collected from a CCD chip (in the experience of the Ben-Amotz lab, to hundreds of microseconds). While CMOS cameras can be read quite quickly (and do not suffer a read-speed dependent noise), they have low quantum efficiencies in the red and NIR wavelengths. As a result, use of CMOS cameras are limited to the to applications where an excitation laser with a visible wavelength can be used. As noted, the use of such an excitation laser can lead to more frequently observing fluorescence in a sample. If one considers fluorescence to be a source of noise (or a nuisance signal), then avoiding visible excitation is necessary for many applications (as will be further discussed in Chapter 3).

Given the above, an alternative, fast, low-noise Raman photon detection strategy was developed. A summary of this strategy, referred to as OB-CD, is described below.

2.3 Overview of OB-CD Raman Spectroscopy

This section summarizes previous work [1], [2], [3] done by a variety of graduate students in the Ben-Amotz lab, notably David Wilcox, in collaboration with Brad

Lucier and Gregory Buzzard. This section is intended to familiarize readers with the OB-CD filter generation algorithm; please see the cited references for a more thorough description and validation of OB-CD.

2.3.1 Model for Filter Generation

Assume that a Raman spectral analysis is performed where the identity (and spectra) of all potential n chemicals is known. Let P be an i by j matrix, with each column containing the area-normalized spectra of some $n = 1...j$ number of known chemicals, such that P_{ij} represents the probability of a photon from chemical j falling in energy bin i. Let $\bar{\Lambda}$ be a j by 1 column-vector containing the scattering rate (in photons per seconds) for a given measurement. Thus, any spectra in the sample under analysis can be represented as $P\bar{\Lambda}$.

The objective of the OB-CD algorithm is to build an i by j matrix of filters, F, where F_{ij} is the probability of a photon from chemical j landing on energy bin i being transmitted to a single channel detector (here, we'll assume that we will build a number of filters equal to n, the number of chemical species in the system under analysis). We, explicitly, do not assume that each filter in F will be applied for an equal portion of the total measurement time, τ. Instead, the percentage of measurement time each filter is applied for is represented as a matrix, T, where the diagonal, T_{jj}, represents the portion of the measurement time each filter is applied (all other components in T are 0). Making the assumption that the OB-CD spectrometer will utilize a noise-free photon-counting detector, the Raman scattering rates for each chemical in the system, $\hat{\Lambda}$ are estimated from the number of photon counts through each filter, $\hat{x}$, in measurement time τ, using the following relationship:

$$\hat{\Lambda} = (F^T P)^{-1} \tau^{-1} T^{-1} \hat{x}$$

For simplicity, let $A = (F^T P)^{-1}$ and $BA = I$, which makes it possible to rewrite the above as:

$$\hat{\Lambda} = B\tau^{-1}T^{-1}\hat{x}$$

To build "optimal" filters, the goal of the OB-CD filter generation algorithm is build filters that minimize the error in the estimate of $\bar{\Lambda}$. Explicitly, this corresponds to a minimization of the following quantity:

$$E \parallel \hat{\Lambda} - B\tau^{-1}T^{-1}\hat{x} \parallel ^2$$

Minimizing the square of this quantity over all F is not computationally efficient. Instead, this formula is manipulated into a more computationally tractable form.

Let e_j denote a column vector whose j^{th} entry is 1 and whose remaining entries are 0. The j^th component of $\hat{x}$, $x_j = e_j^T\hat{x}$, can be shown to have a mean (and variance, as it is a Poisson variable) of:

$$T_{jj}e_j^T P\bar{\Lambda} = T_{jj}(A\bar{\Lambda})_j$$

We then define the variable ϵ for the j^{th} chemical under analysis to be:

$$\epsilon_j = \frac{x_j}{T_{jj}} - (A\bar{\Lambda})_j$$

Thus, ϵ can be defined as:

$$\epsilon = T^{-1}\hat{x} - A\bar{\Lambda}$$

Here, each j^{th} entry of ϵ has a mean of 0 and a variance of $T_{jj}^{-1}(A\bar{\Lambda})_j$. After some rearrangement and with the relationship $BA = I$ in mind, it can be shown that:

$$B\epsilon = BT^{-1}\hat{x} - \bar{\Lambda}$$

Where the i^th entry in $B\epsilon$, $\sum_j b_{ij}\epsilon_j$, has a mean of 0 and and a variance of:

$$\sum_i b_{ij}^2 Var(\epsilon_j) = \sum_j b_{ij}^2 T_j^{-1} j(A\bar{\Lambda})_j$$

This equation describes the variance of a Poisson count variable, $\hat{x}$, and is equal to the square error in the estimate of $\bar{\Lambda}$ that we're trying to minimize. While this is very useful in its own right, its also worth noting that this minimization can be performed over some limited set of n. This is a fact that will be exploited when minimizing the error in the estimates of a subset of chemicals in the training set, P, as further described in Chapter 3.

Taking the equation above, setting it equal to $E \parallel \hat{\Lambda} - B\tau^{-1}T^{-1}\hat{x} \parallel {}^2$, and summing over all i nets the following relationship:

$$E \parallel \hat{\Lambda} - B\tau^{-1}T^{-1}\hat{x} \parallel {}^2 = \sum_i \sum_j b_{ij}^2 T_j^{-1} j(A\bar{\Lambda})_j$$

$$\sum_i \sum_j b_{ij}^2 T_j^{-1} j(A\bar{\Lambda})_j = \sum_j T_j^{-1} j(A\bar{\Lambda})_j \sum_i b_{ij}^2 = \sum_j \frac{1}{T_{jj}}(A\bar{\Lambda})_j \|Be_j\|^2$$

Note that this equation depends on both B and T, both of which are matrices that need to be solved for. For a fixed B and $\bar{\Lambda}$, T can be solved for using the Cauchy-Schwarz inequality. This is done by solving for a scaling factor, a, that solves for the equality of spaces. This method is much further expanded on in previous work by Buzzard and Lucier [3]. Using this equality, it can be shown:

$$T_{ii} = \frac{\|Be_i\|\sqrt{(A\bar{\Lambda})_i}}{\sum_j \|Be_j\|\sqrt{(A\bar{\Lambda})_j}}$$

Given this definition of T, it possible to make an initial “guess” of T for a fixed B. To do so, F is assumed to initially be equal to P. This allows T, and therefor B, to be solved for. However, once a B (and therefor, F) is solved for, T will again be calculated.

To solve for B, we use the definition of T_{ii} above as well as the equality $E \parallel \hat{\Lambda} - B\tau^{-1}T^{-1}\hat{x} \parallel {}^2 = \sum_j \frac{1}{T_{jj}}(A\bar{\Lambda})_j\|Be_j\|^2$ to find that:

$$E \parallel \hat{\Lambda} - B\tau^{-1}T^{-1}\hat{x} \parallel {}^2 = (\sum_i \|Be_i\|\sqrt{(A\bar{\Lambda})_i})^2$$

Note that objective of this algorithm is to solve for the minimum of both sides of the equality. Since the right side of the equality is going to be solved for and we assume non-negativity in F, P, and $\bar{\Lambda}$, it is possible to solve for the value itself rather than its square.

As demonstrated in [3], minimizing this quantity results in optimal filters that can be expressed as binary, or nearly-binary, with at most $n - 1$ non-binary entries. In practice, the non-binary entries are rounded to be binary. Hence, the filters generated are described as optimal (or nearly-optimal) binary compressive filters.

Minimizing this quantity is done by scaling rows of A, rather than empirically solving over all F. While this problem should be computationally difficult to solve for a number of reasons (e.g., it is non-linear and non-convex), it has been empirically found that this quantity can be minimized in seconds to minutes (depending on the value of n) algorithmically using a linear programming function for $n \leq 20$.

Once A (and therefor, F) have been solved for, B is calculated using generalized least squares (GLS). If D is defined as the diagonal entries of $A\Lambda$, then an optimal B can be solved for as:

$$B = (A^T D^{-1} T A)^{-1} A^T D^{-1}$$

Once this B is solved for, T is recalculated, followed by B_optmal (in other words, the GLS for B is rerun, not the entire algorithm). All of this calculations have been written as Matlab script, as described in [1].

2.4 OB-CD Raman Spectrometers

The OB-CD spectrometers used in Chapter 3 are described here. Each system is described further detail in [2] (for the system with the 785 nm excitation laser) and [4] (for the system with the 514 nm excitation laser).

2.4.1 OB-CD Spectrometer with 785 nm Excitation Laser

An OB-CD spectrometer using a 785 nm laser for excitation whose schematic is shown in Fig. 1 is shown below. Note that this schematic is not drawn to optical scale. This microscope is configured to collect the backscattered Raman signal with the same objective lens that is used to focus the laser (Innovative Photonic Solutions 100 mW) onto the sample. The laser passes through a laser-line bandpass filer (Semrock LL01-785-12.5) before it is focused onto the sample using a microscope objective (Olympus, LMPlan, IR, 20x, 0.4 NA), and unless indicated otherwise the laser power at the sample was about 85 mW.. The backscattered light is collected and then separated from the laser Rayleigh scattering using a dichroic mirror (Semrock LPD01-785RS-25) and a 785 nm notch filter (Semrock, NF03-785E-25). Then, the Raman scattered light is sent to the spectrometer (right portion of Fig. 1), where it is filtered first using a 900 nm shortpass filter (Thorlabs, FES0900), followed by passing through a volume holographic grating (Edmund Optics, ∼1200 lines mm^{-1}, center wavelength 830 nm). This light is then dispersed onto the DMD (Texas Instruments, DLP D400, 1080 × 1920 mirror array with 13.4 μm mirror pitch). The spectral window in this system is ∼200–1700 cm^{-1}. For all data collected in this paper, we binned fifteen columns of adjacent DMD mirrors together, yielding a total of 128 “bins” with each energy bin corresponding to ∼30 cm^{-1}. Light from the DMD is then focused onto a photon-counting avalanche photodiode (APD) module (PerkinElmer, SPCMD2969PE) with a dark count rate of ∼200 photons s^{-1}. TTL pulses from the APD are counted using a USB data acquisition (DAQ) card (National Instruments, USB-6212BNC). The system is controlled with interface software written in Labview 2013. Binary filter generation is performed as described above in algorithm written in Matlab (Matlab 7.13 R2011b).

2.4.2 OB-CD Spectrometer with 514 nm Excitation Laser

An OB-CD spectrometer using a 514 nm laser for excitation whose schematic is shown in Fig. 2 is shown below. Much like the previously described system, our microscope is configured to collect the backscattered Raman signal with the same objective lens that is used to focus the argon ion laser (Modu Laser Stellar Pro L 100 mW) onto the sample. The laser passes through a laser-line bandpass filer (Semrock RazorEdge) before it is focused onto the sample using a microscope objective (Nikon MPlan, 20x, 0.4 NA), and unless indicated otherwise the laser power at the sample was about 12 mW. The backscattered light is collected and then separated from the laser Rayleigh scattering using a dichroic mirror (Semrock RazorEdge). Then, the Raman scattered light is sent to the spectrometer (right portion of Fig. 2), where it is filtered first using a long pass (edge) filter (Semrock RazorEdge), followed by passing through a volume holographic grating (Wasatch Photonics, ∼1000 lines mm^{-1}).

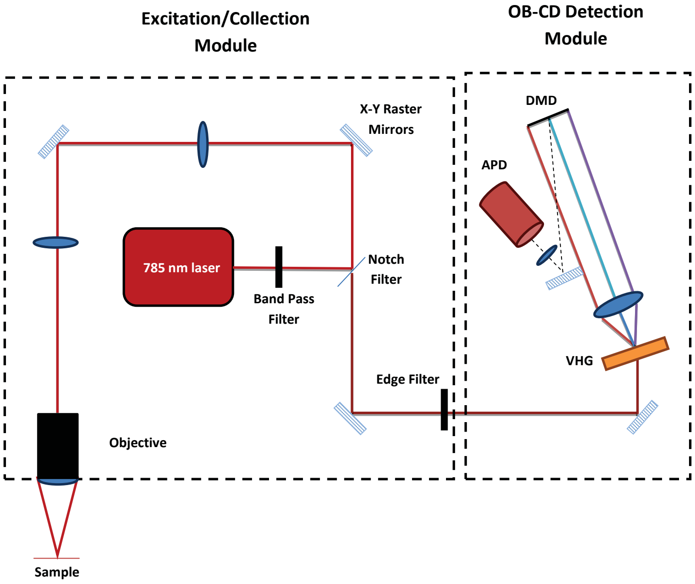


Figure 2.1. Schematic of the OB-CD Raman system based upon a 785 nm laser excitation source.

This light is then dispersed onto the DMD (Texas Instruments, DLP3000, 608 × 684 mirror array with 10.8 μm mirror pitch). The spectral window in this system is ∼200–4100 cm^{-1}. For all data collected in this paper, we binned two columns of adjacent DMD mirrors together, yielding a total of 342 "bins" with each energy bin corresponding to ∼12 cm^{-1}. Light from the DMD is then focused onto a photon-counting photomultiplier tube (PMT) (Hamamatsu model #H10682-01) with a dark count rate of ∼500 photons s^{-1}. TTL pulses from the PMT are counted using a USB data acquisition (DAQ) card (National Instruments, USB-6212BNC). The system is controlled with interface software written in Labview 2013. As above, binary filter generation is performed as described above in algorithm written in Matlab (Matlab 7.13 R2011b).

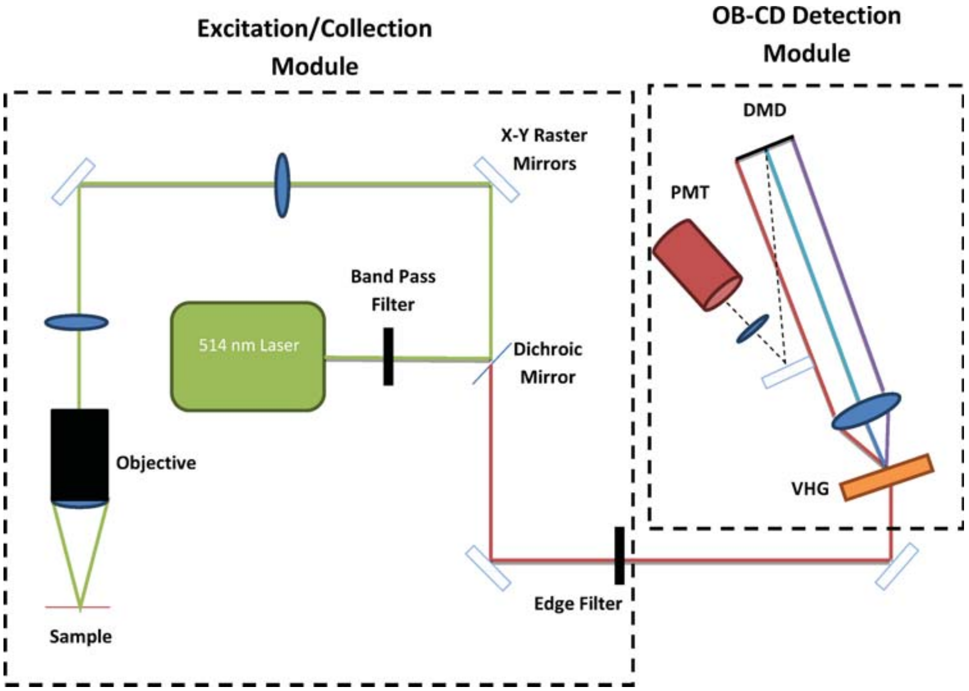


Figure 2.2. Schematic of the OB-CD Raman system based upon a 514 nm laser excitation source.

2.5 Conclusion

This chapter provides an overview of the OB-CD filter generation algorithm and a summary of the OB-CD spectrometers used to generate data in future chapters. While not exhaustive, this chapter also provides readers references for more thorough descriptions of the OB-CD algorithm and spectrometers.

3. FLUORESCENCE MITIGATION IN OB-CD

3.1 Abstract

The recently-developed optimized binary compressive detection (OB-CD) strategy has been shown to be capable of using Raman spectral signatures to rapidly classify and quantify liquid samples and to image solid samples. Here I demonstrate that OB-CD can also be used to quantitatively separate Raman and fluorescence features, and thus facilitate Raman-based chemical analyses in the presence of fluorescence background. I describe a general strategy for fitting and suppressing fluorescence background using OB-CD filters trained on third-degree Bernstein polynomials. I present results that demonstrate the utility of this strategy by comparing classification and quantitation results obtained from liquids and powdered mixtures, both with and without fluorescence. These results demonstrate high-speed Raman-based quantitation in the presence of moderate fluorescence. Moreover, I show that this OB-CD based method is effective in suppressing fluorescence of variable shape, as well as fluorescence that changes during the measurement process, as a result of photobleaching.

3.2 Background

Current spectroscopic chemical analysis instruments are capable of generating very large data sets. However, to derive information from this data, it is often necessary to transform the data to a lower-dimensional space, using methods such as principal component analysis (PCA) [5] or partial-least squares (PLS) [6] to facilitate subsequent chemical classification and quantitation. Moreover, because the collection of high-dimensional data is often the slowest step in the process, a number of compressive detection strategies [7–11] have been introduced with the goal of increasing data collection speed by making measurements only in the low-dimensional space containing the information of interest. One such method is the previously described optimized binary compressive detection (OB-CD) strategy, in which OB filters are applied to a digital mirror microarray (DMD) to redirect or collect photons of (multiple) selected colors, for detection using a single channel detector, such as a photon counting photomultiplier tube (PMT) or an avalanche photodiode (APD) [1, 2].

Previously [1–3], it was demonstrated that the OB-CD strategy enabled high-speed chemical classification, quantitation, and imaging. Here I demonstrate an extension of the OB-CD method that facilitates Raman classification and quantitation in the presence of a fluorescence background. The key advantages of this method, relative to conventional fluorescence subtraction strategies [12–17], include its compatibility with automated high-speed chemical analysis in the presence of variable fluorescence backgrounds and compatibility with OB-CD spectroscopic instrumentation. Here I show that fluorescence backgrounds may be quantified and subtracted on-the-fly by including Bernstein polynomial spectral functions in the OB-CD training set, along with Raman spectra of the components of interest. In other words, I augment the Raman spectral training set with Bernstein polynomial spectral shapes to model fluorescence and thus obtain OB-CD filters that are used to quantify both the chemical components of interest and the fluorescence background spectra.

3.2.1 Nuisance Parameters

It often happens that the photon emission rates of some chemical species are of more interest than others. One might have a contaminant with a broad (known) spectrum; while this may be one of the chemical species S_j, the accuracy of estimation of that particular Λ_j is unimportant. In the field of Optimal Design of Experiments, such variables Λ_j are known as *nuisance parameters*: They are a necessary part of the model, but the accuracy of their estimates is unimportant except insofar as it affects the estimates of the other variables [18].

It is assumed that there are only $n' < n$ chemical species whose photon emission rates are of interest. For such systems, the variances of the estimates of Λ_j for $n' < j \leq n$ are not of interest and so instead of performing the minimization shown in Chapter 2, the sum of the variances of the estimates of only the first n' emission rates are minimized:

$$\sum_{j=1}^{n'} E\big(|(BT^{-1}\hat{x})_j - \bar{\Lambda}_j|^2\big) = \sum_{j=1}^{n'} \sum_{k=1}^{m} b_{jk}^2 T_{kk}^{-1} (F^T P \bar{\Lambda})_k$$

In effect, Λ_j is estimated for $n' < j \leq n$ only well enough to minimize the error of the sum of variances of the first n' emission rates.

3.2.2 Estimating Fluorescence Using Bernstein Polynomials

A fluorescent spectrum is generally smooth and broad, in contrast to the narrow peaks found in Raman spectra; it may vary from one sample to another, or from time to time for the same sample.

One can reduce the amount of fluorescence in a sample by photobleaching the sample prior to analysis [19]. Another commonly-used procedures for subtracting such fluorescence backgrounds from Raman spectra is to fit the fluorescent spectrum to a polynomial [12,13]; in effect, this amounts to using a fixed polynomial of specified degree to model a fluorescent spectrum. This may work well if in a set of samples the fluorescence doesn't vary over space or time; this is often not the case, however, in Raman (and particularly Raman imaging) applications, for which one must estimate the shape of the fluorescence dynamically.

Instead of somehow applying conventional fluorescence fitting and subtraction strategies to OB-CD measurements, I use OB-CD filters that are derived using either actual fluorescence spectra of known shape or a family of polynomials that models general fluorescence spectra.

More specifically, if the sample of interest contains a fixed fluorescent signal of known shape, the fluorescence spectrum is simply added as an extra column of P with associated rate variable Λ_j, and then treat Λ_j as a nuisance parameter, not adding its variance to the sum of variances being minimized.

If the fluorescence has a spectrum that varies over space or time, however, the previous procedure cannot be applied and must model the fluorescence dynamically, as follows.

Since fluorescence backgrounds can often be fit reasonably well by a cubic polynomial, there should be a polynomial basis for cubic polynomials such that (1) all basis elements are nonnegative on $[0, 1]$ and (2) every nonnegative polynomial can be written as a linear combination of the basis elements with nonnegative coefficients. Note that negative coefficients, which are supposed to model rates, are nonphysical and increase the variance of our estimates. Unfortunately, for $r > 1$, no such basis exists.

Of interest, however, is the Bernstein basis [18, 20] of polynomials of degree r, given by:

$$B_{\nu,r}(x) = \binom{r}{\nu} x^{\nu}(1-x)^{r-\nu}, \quad \nu = 0, 1, \ldots, r,$$

These polynomials are nonnegative on $[0, 1]$ (so any linear combination of them with nonnegative coefficients is nonnegative on $[0, 1]$). They have other nice properties: They form a basis for the space of polynomials of degree r; they have optimal stability is some sense [18]; and "many" if not "most" nonnegative polynomials that come up in practice as models for fluorescent spectra have nonnegative coefficients in the Bernstein basis (or at least, any negative coefficients are not "overly" large). The Bernstein polynomials even resemble single-peak spectra.

In the specific case $r = 3$ the Bernstein polynomials appear as $B_{0,3}(x) = (1-x)^3$, $B_{1,3}(x) = 3x(1-x)^2$, $B_{2,3}(x) = 3x^2(1-x)$, $B_{3,3}(x) = x^3$. These polynomials are shown in Fig. 1.

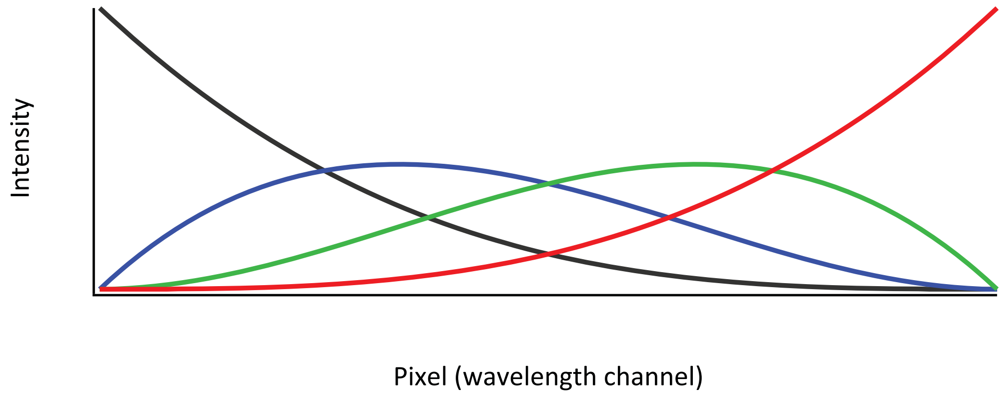


Figure 3.1. Plot of the four degree-three Bernstein polynomials as a function of wavelength channel (scaled to be over the interval $[0, 1]$.) The colors denote the various polynomials: black, $B_{3,0}(x)$; blue, $B_{3,1}(x)$; green, $B_{3,2}(x)$; and red, $B_{3,3}(x)$.

In this work, fluorescence in Raman spectra is modeled with linear combinations of $B_{\nu,3}(x)$, $\nu = 0, 1, 2, 3$, and the Bernstein coefficients are often treated as nuisance parameters. Because fluorescent spectra are not precisely polynomials, this introduces model error, which I discuss in the Results and Discussion section. We have also found that for some purposes using quartic polynomials ($r = 4$) gives more accurate results.

3.3 Methods and Materials

The following sections describes the methods and materials used in these experiments.

3.3.1 OB-CD Spectrometers

The instruments used in this work are described in Chapter 2. Note that the data from these instruments was further processed using Igor Pro version 6.04.

3.3.2 Chemicals Used in Classification/Quantitation

Acetone and benzene were purchased from Macron (batch #0000070736) and OmniSolv (lot #42282), respectively. Hexanes were acquired from Baxter (lot #901141). Methylcyclohexane was acquired from Mallinckroft (lot #1906 KCBN). Aniline and toluene were acquired from Aldrich (batch #05925CB) and Mallinckroft (lot #8608 X14752), respectively. Aniline was purified via distillation by heating aniline to 190° C in a round bottom flask, which was connected to a chilled condenser. Ethanol was acquired from Koptec (200 Proof, lot #225411) and water was ultra-purified in our lab (Milli-Q UF Plus, 18.2 mΩ cm, Millipore). The overhead transparency was 3M brand (model #PP2500).

3.4 Results and Discussion

Here I demonstrate how the OB-CD detection strategy is influenced by whether the photon emission rates of some of the spectra used for training OB-CD filters are treated as nuisance parameters. Such spectra are designated as *nuisance spectra.* To show this, I generated an OB-CD training set containing the spectra of hexane, methylcyclohexane, a spectral feature arising from the NIR objective in the 785 nm OB-CD system, and the four Bernstein polynomials shown in Fig. 1. Using this training set, I calculated two sets of OB-CD filters (explicitly, both the matrix of binary filters, F, the measurement time matrix, T, and the the left-inverse of dot product of the normalized spectra, P, and the transpose of F, which is designated B). One set of filters was generated in which no spectral features were considered to be nuisance spectra and a second set of filters was generated in which the spectral feature arising from the NIR objective and the Bernstein polynomials were considered to be nuisance spectra. The binary filters turned out to be identical in both cases (although that need not in general be the case), while the measurement time matrices, T, were quite different as shown in Fig. 2. Notice that the binary filters with significant overlap with the NIR objective and the Bernstein polynomials peaks trained assuming no spectra are nuisance spectra are turned on for longer percentages of the total measurement time relative to OB-CD filters for the same spectral components when no portion of the training set is considered a nuisance spectra.

One might expect that additional time spent measuring the non-nuisance spectra would result in lower variance in the recovered Raman rates for these components. To test this, the two sets of OB-CD filters (and the associated measurement time matrices) shown in Fig. 2 were used to classify hexane and methylcyclohexane (explicitly, there was no added fluorescence in these samples despite including Bernstein polynomials in both OB-CD training sets). Each chemical was measured using each of the sets of OB-CD filters for 1,000 measurement with 10 ms total integration time. The results of these measurements, shown in Fig. 3, clearly reveal the reduced variance (smaller 95% confidence bands) obtained when treating only the two components of interest as non-nuisance spectra. Note that the mean recovered Raman rates for hexane and methylcyclohexane (the centers of the ellipses in Fig 3) generated from both sets of OB-CD filters differed very little.

3.4.1 Validation of Bernstein Polynomials

To test and validate this OB-CD fluorescence suppression strategy, I used both a white light source to simulate fluorescence and samples containing fluorescent components. The results described in the section below were obtained using a white light illuminator as a convenient fluorescence mimic, as its intensity can readily be varied, and its shape resembles typical fluorescence backgrounds. It also has a different shape in the 514 nm and 785 nm spectral region, which is useful in confirming our fluorescence mitigation strategy is appropriate for multiple fluorescent peak shapes.

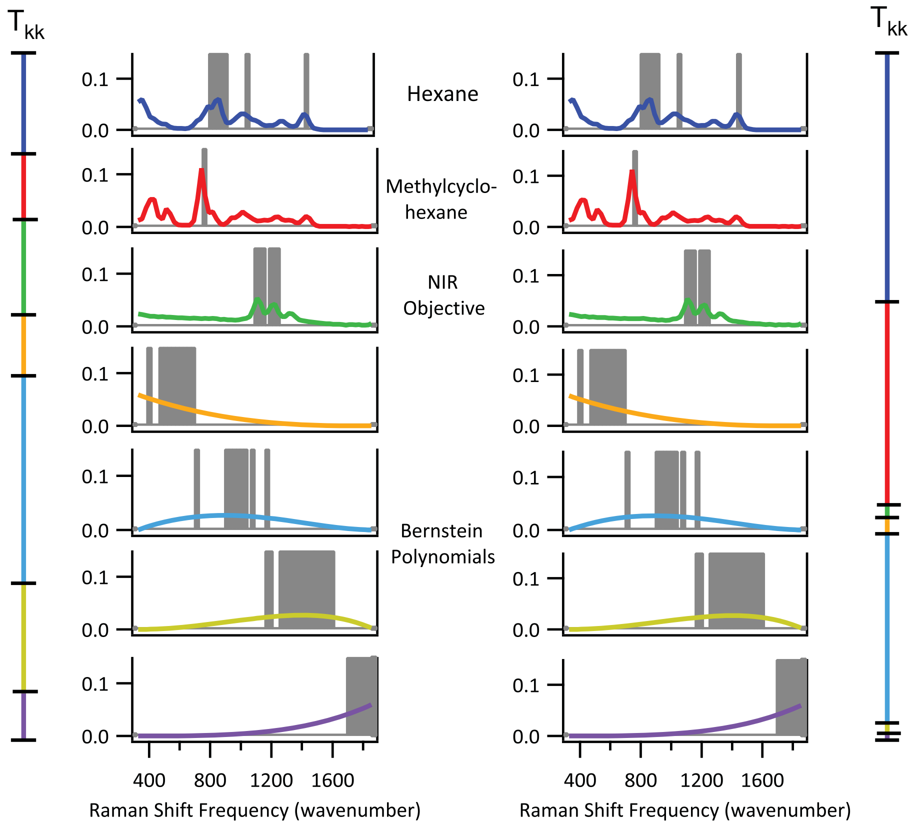


Figure 3.2. The colored curves are training spectra (each normalized to unit area) and the gray bands indicate regions in which the OB-CD filters are on (i.e., direct light towards the detector). The Raman spectra were obtained with a spectral resolution of 30 cm^{-1}. The lines labeled T_{kk} correspond to the fraction of the total measurement time that data is collected using the filter associated with each spectral component (denoted by color). The OB-CD filter and T_{kk} results on the left were obtained without considering any components to be nuisance spectra, while those on the right were obtained when considering the NIR objective and Bernstein polynomials to be nuisance spectra.

In subsequent subsections, I describe results obtained using samples with fluorescent impurities, rather than white light, to validate our OB-CD strategy. Here I produced OB-CD filters by training using Bernstein polynomials as well as the Raman spectra of n-hexane and methycyclohexane (and a spectral component arising from the NIR

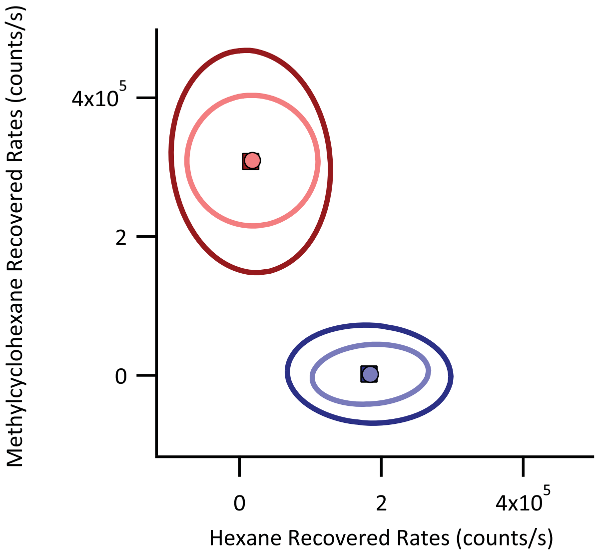


Figure 3.3. Recovered Raman rates for hexane (blue) and methylyclohexane (red) generated using OB-CD filters that considered components of the training set to be nuisance spectra (dark blue and dark red) and OB-CD filters that considered the spectral component arising from the NIR objective and the four Bernstein polynomials to be nuisance spectra (light blue and light red). In all cases, 1,000 OB-CD measurements were taken with a total integration time of 10 ms. The ellipses represent the 95% confidence interval of the recovered Rates for each sample. The large markers in the center of each ellipse represent the mean recovered Raman rates.

objective for the 785 nm system). I treat the coefficients of the Bernstein polynomials and the spectral component arising from the NIR objective for the 785 nm system as nuisance parameters. The resulting filters and training spectra from the 785 nm system are virtually identical to those shown in Fig. 2 (and are provided in the Appendix 1, along with the filters generated for the 514 nm OB-CD system).

The following results were obtained by holding the white light at constant intensity (of about 4 million counts per second) such that the total Raman/fluorescence signal intensity never exceeds 5 million counts per second (the limit of linearity for the APD in the 785 nm excitation system). Then, I varied the Raman excitation laser intensity (using neutral density filters) in order to vary the relative amount of Raman and fluorescence in the measured spectra and OB-CD signals, and thus determine how the relative fluorescence background to Raman intensity ratio influenced our the recovery of apparent Raman rates using OB-CD.

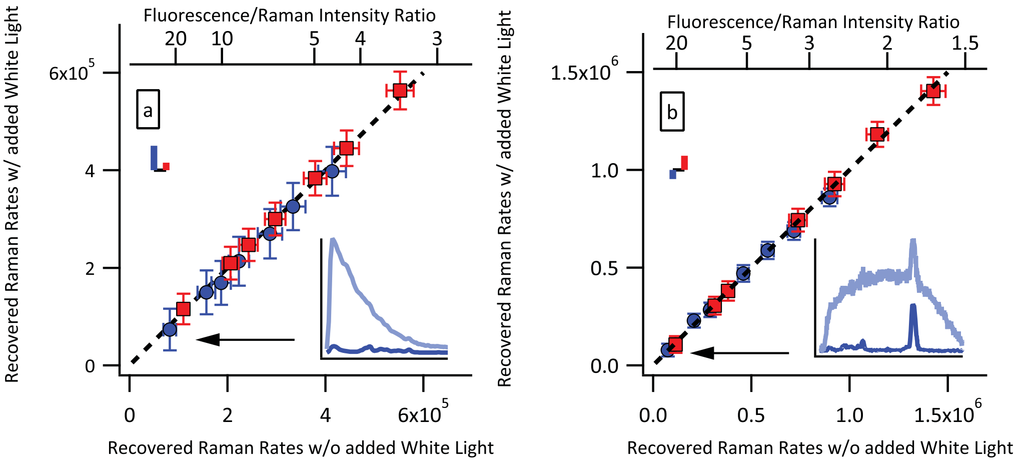


Figure 3.4. Plot of hexane (blue) and methylcyclohexane (red) recovered Raman rates measured without added white light versus recovered Raman rates measured with added white light on the (a) 785 nm laser excitation system and (b) 514 nm laser excitation system. Rates have each been corrected by removing a small constant vertical offset (modeling error) whose magnitude was determined by measuring the apparent recovered Raman rates obtained in measurements performed on white light without Raman. The magnitude of this correction is represented by the colored bars in the upper left of each plot. Each point represents the means of 1,000 measurements (each obtained using a 30 ms total integration time) with error bars representing 1 standard deviation. Top axis denotes the ratio of the total (integrated) number of the white light/Raman photons. The inset spectra were obtained from hexane with and without added white light with ~1 OD neutral density filter and correspond to measurements made at the points denoted by the arrows.

At each Raman signal intensity, estimated Raman scattering rates were recovered using from OB-CD measurements as described in Chapter 2, both with and without the white light background. This allowed for the determination of the error in recovered Raman rates as a result of adding fluorescent (white light) background. Fig. 4

compares the Raman rates recovered with (y-axis) and without (x-axis) added white light when using 30 ms total integration time for each measurement. The number at the top indicates the ratio of the integrated area of the fluorescence and Raman signals and the error bars represent the standard deviations of the Raman rates for each component.

If the fluorescence is not perfectly modeled, then the recovered Raman rates can contain a systematic modeling error whose magnitude increases with fluorescent intensity. When using a constant intensity white light to model fluorescence such modeling error introduces an approximately constant offset to the recovered Raman rates. The magnitude of this offset can be determined by measuring the apparent recovered rates of the Raman components obtained when measuring only white light (containing no Raman photons). For the measurements shown in Fig. 4, these modeling errors were relatively small and have magnitudes indicated by the bars in the upper left of each plot. These modeling errors have been subtracted from each of the points in Fig. 4. In other words, before correcting for this modeling error, all of the recovered Raman rates were slightly offset from the dashed diagonal line (of slope 1).

If the spectrum of white light was modeled perfectly by a degree-three polynomial (or was corrected for modeling error, as described above), the mean recovered Raman rates for samples with added white light would be expected to not significantly differ from the mean recovered Raman scattering rates without added white light. In other words, it would be expected that the points would lie on a line with slope one, which are indicated by the dashed lines in Fig. 4. The agreement between the points and dashed line in Fig. 4 clearly demonstrate that Raman components can be quantified accurately in the presence of fluorescence backgrounds whose integrated intensity is up to 20 times that of the Raman component of interest. Note that the factor of 20 is obtained from the results shown in Fig. 4, as this is when the Raman signal-to-noise approaches 1:1.

Samples with larger fluorescence/Raman intensity ratios can in principle be accurately analyzed using OB-CD with longer measurement times. However, when the integration time approaches one second it may be appropriate to use conventional full spectral measurements and fluorescence subtraction procedures as the OB-CD strategy is primarily advantageous for performing high speed (or low light level) measurements that are not compatible with CCD detection. Additionally, performing OB-CD measurements on samples in which fluorescence is more than 20 times as intense as the Raman signal of interest would require careful modeling error correction (as the modeling error would become large relative to the Raman intensities). These results demonstrate the accuracy with which high speed OB-CD Raman classification and quantitative measurements may be performed without correcting for model error so long as the fluorescence background intensity does not exceed 20 times the Raman signal intensity.

3.4.2 Raman Quantitation and Classification of Fluorescent Samples

3.4.2.1 Toluene and Fluorescent Aniline

The following results were obtained using liquid mixtures of toluene and an aniline sample that was partially oxidized and, as a result, developed a fluorescent impurity that could be removed by distillation. We selected these two liquids because of their significant spectral overlap (the dot product of the two normalized spectral vectors is 0.91) and thus successful classification of aniline/toluene mixtures may be used to demonstrate that our fluorescent mitigation strategy is compatible with the Raman-based quantification of mixtures of spectrally similar chemicals.

I trained OB-CD filters using 785 nm spectra obtained from distilled aniline, toluene, the NIR objective spectrum describe above, as well as the four Bernstein polynomials shown in Fig. 1 (all of the resulting spectra and OB-CD filter functions are given in the Appendix 1). For this experiment, the spectral component arising from the NIR objective and the four Bernstein polynomials are treated as nuisance spectra. We used OB-CD to recover Raman rates for toluene, distilled aniline, fluorescent aniline, and mixtures of both types of aniline and toluene. Using these recovered Raman rates, I calculated apparent volume fractions of aniline and toluene as follows:

$$\chi_i = \frac{w_i \hat{\Lambda}_i}{\sum_i w_i \hat{\Lambda}_i},$$

where w_i is equal to $M_i/\hat{\Lambda}_i^{\mathrm{Max}}$, M_i is equal to molarity of the i^{th} pure liquid, and $\hat{\Lambda}_i^{\mathrm{Max}}$ is equal to the mean recovered Raman scattering rate for the i^{th} pure liquid as previously reported [2]. I then estimated the apparent volume fraction, (Φ), for aniline and toluene in each sample. We did this by dividing the apparent mole fraction (for either aniline or toluene) by the molarity of each pure liquid as follows:

$$\Phi_i = \frac{M_i \chi_i}{\sum_i M_i \chi_i}.$$

Fig. 5 plots the resulting apparent volume fractions of toluene and aniline as well as mixtures of the two. The results in Fig. 5 demonstrate that the mean recovered Raman rates are insensitive to the fluorescence arising from the impure aniline sample. The variance of the measurements without fluorescence, however, is less than the variance of samples with fluorescence (as a result of additional shot noise in the latter measurements).

3.4.2.2 Aqueous Ethanol and Fluorescent Tequila

Previous work [21, 22] has demonstrated that Raman spectroscopy can be used to quantify the volume percentage of ethanol in tequila samples and qualitatively distinguish distilled (“silver”) and highly fluorescent, aged (or “golden”) tequilas, even in the presence of fluorescence, which is more common in aged, so-called “golden” tequila. Here I show that our OB-CD fluorescence mitigation strategy can be used

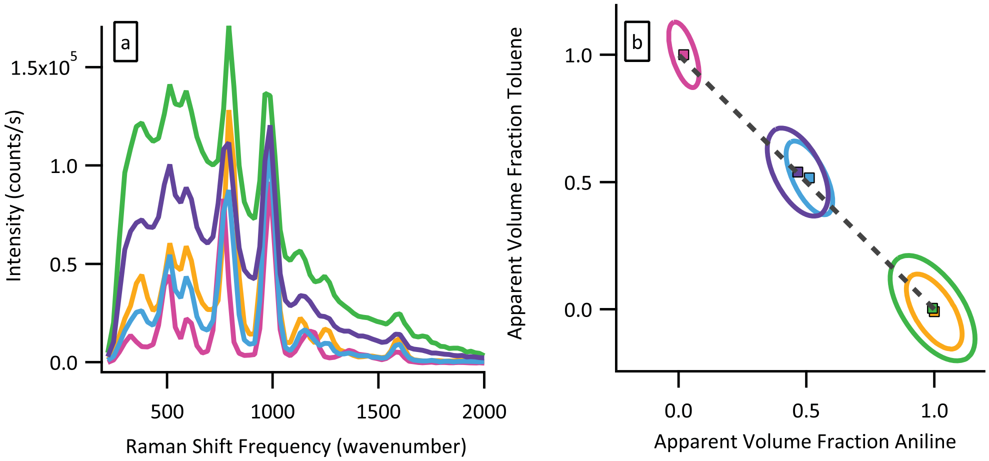


Figure 3.5. (a) Spectra of distilled aniline (orange), toluene (magenta), fluorescent aniline (green), a 47:53 volume-by-volume mixture of distilled aniline and toluene (dark green), and a 52:48 mixture of fluorescent aniline and toluene (cyan) measured on the 785 nm OB-CD system. (b) Apparent volume fractions of distilled aniline (orange), toluene (magenta), fluorescent aniline (green), a 47:53 volume-by-volume mixture of distilled aniline and toluene (dark green), and a 52:48 mixture of fluorescent aniline and toluene (cyan). Each chemical was sampled 1,000 times at 20 ms per experiment. Ellipses correspond to the 95% confidence interval of the recovered rates for each sample. The large squares with black borders in the center of each ellipse represent the mean of each sample.

to quantify the volume percentage of ethanol in tequila, even for “golden” tequila samples at speeds much greater than those previously reported for this application.

I used the 514 nm laser-based OB-CD system for these studies as the C-H and O-H stretch vibrational bands are not readily detectable using the 785 nm system. Spectra were collected and OB-CD filters were trained using ethanol, water, and Bernstein polynomials (as shown in the Appendix 1). Note that the four Bernstein polynomials were treated as nuisance spectra for OB-CD filter generation. After this, OB-CD was used to recover Raman scattering rates for ethanol and water in a silver tequila (“Arandas” brand) and a golden tequila (“Casamigos” brand). The corresponding apparent volume fractions were obtained from the recovered rates as shown in Fig. 6.

In order to keep the fluorescence photon rates within the linear regime of the PMT detector (below 5 million counts per second), the laser intensity at the sample was reduced to ∼2 mW using a neutral density filter placed in front of excitation laser and the integration time per sample was increased to 100 ms.

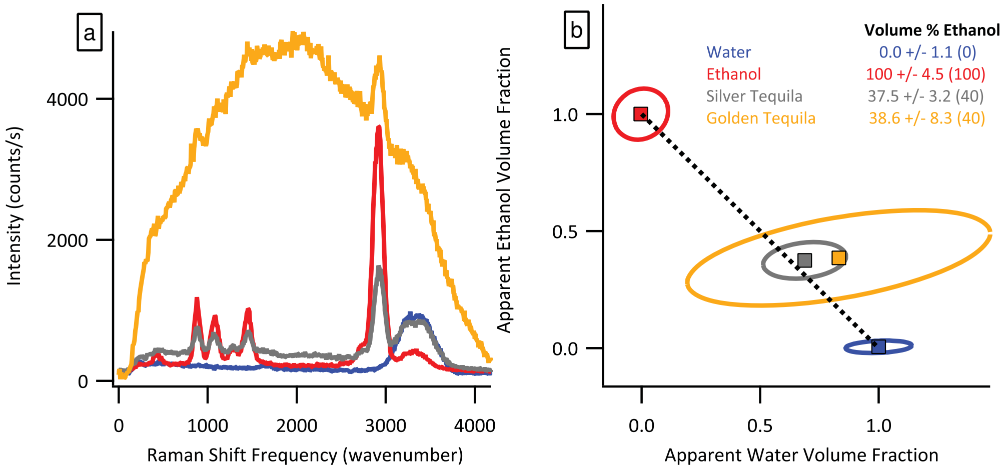


Figure 3.6. (a) Spectra of water (blue), ethanol (red), Arandas brand silver tequila (gray), and Casamigos brand golden tequila (dark yellow) measured on the 514 nm OB-CD system (b) Apparent volume fractions of water (blue), ethanol (red), silver tequila (gray), and golden tequila (dark yellow) are compared with the nominal volume fractions (as obtained from the label on the tequila bottles). Each chemical was sampled 1,000 times at 100 ms per OB-CD measurement. Ellipses correspond to the 95% confidence interval of the recovered rates for each sample. Large squares with black borders represent the mean of each sample. The dashed line corresponds to line with slope −1. Inset table reports the mean apparent volume fraction of ethanol (plus/minus 1 standard deviation) for each sample and then, parenthetically, the label ethanol volume percentage for each sample.

The inset table in Fig. 6(b) shows the mean apparent ethanol volume fractions obtained in a measurement times of 100 ms and, parenthetically, the label volume percent ethanol for each sample. As can be seen, our predicted volume percentages of ethanol very nearly match the label percentage. However, the variance of the fluorescent “golden” tequila measurements is much greater than that of the “silver” tequila measurements due to the increased shot noise resulting from the fluorescent background. In spite of this, it should be noted that by using OB-CD, it is possible to accurately predict the volume percentage of ethanol in tequila samples even when the integrated intensity of the fluorescence of the sample is 20 times larger than the integrated Raman signal. Since the signal-to-noise of such measurements is typically limited by photon (shot) noise rather than read noise, the total time required to obtain a given precision depends on the available laser power and thus is expected to be comparable to that obtained using full spectral measurements (under otherwise identical conditions).

3.4.2.3 Fluorescent Plastic Film Photobleaching

Here I show an imaging application of our OB-CD fluorescence modeling technique to demonstrate that the recovery of Raman scattering rates is unaffected by fluorescence photobleaching. For this purpose, I used a clear, overhead transparency (a type of plastic film) of ∼1.7 mm thickness. This sample was chosen as it was found to contain both Raman and fluorescence signals when excited at 785 nm and the fluorescence could be very quickly photobleached by exposure to the excitation laser at full power. Additionally, the film exhibited fluorescence with an integrated intensity 10 times that of the integrated intensity Raman features.

While photobleaching decreased the fluorescence background intensity of the sample by ∼50%, the remaining fluorescence could not readily be further photobleached. Thus, for OB-CD training purposes, I generated a Raman spectrum of the plastic by manually performing a polynomial background subtraction from a spectrum of photobleached plastic in the 328 nm^{-1} to 2057 nm^{-1} region. More specifically, the polynomial subtraction was performed using the “backcor” MATLAB algorithm (Vincent Mazet, 2010), using an Asymmetric Huber cost function, a threshold of 0.01, and four third-degree Bernstein polynomials as a basis. The resulting background subtracted Raman spectrum, as well as the spectra of the overhead transparency before and after photobleaching are shown in Fig. 7.

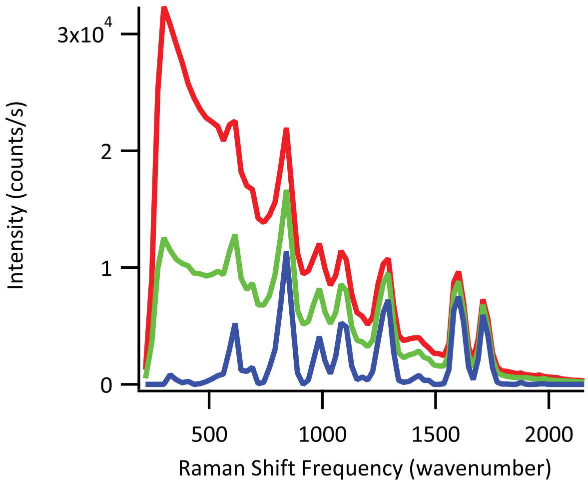


Figure 3.7. The measured spectra of a cellulose acetate overhead transparency are plotted before photobleaching (red) and after 20 minutes of photobleaching (green). The output of the polynomial baseline subtraction is also plotted (blue).

Next, OB-CD filters were calculated by training on the fluorescence-subtracted Raman features of the plastic film, the spectral component arising from the NIR objective, and the four third-degree Bernstein polynomials (and the resulting training spectra are provided in Appendix 1). Note that unlike the filters constructed for previous examples, no components were considered nuisance spectra when calculating OB-CD filters, as I wanted to accurately estimate the intensity of the fluorescence before and after photobleaching. Using these filters, a 200 pixel × 200 pixel region of the plastic film (approximately 1 mm × 1 mm) was imaged with an integration time of 10 ms per pixel. Once this image was collected, two lines were photobleached in the transparency to form a photobleached "+" pattern near the center of the field of view (as shown in Fig. 10). Each line, consisting of 50 pixels, was photobleached for 10 minutes by scanning the laser repeatedly over the "+" pattern at a rate of 1 second per pixel. After photobleaching, the same field of view was reimaged and OB-CD was used to recover the Raman and fluorescence rates. Images were generated using the recovered rates using a method that was previously described [2]. The fluorescence intensity in this image, was determined from the sum of the recovered rates for all four Bernstein polynomials. Note that several spots on the plastic film were highly fluorescent (likely due to a fluorescent impurity, or dust particle, in the film), with counts well outside the linear region of the PMT and a fluorescence background intensity much greater than 20 times the average Raman signal. These pixels also had unusually high recovered apparent Raman rates, which I attributed to model error. These points were removed from the image, as indicated by black dots in the images shown in Fig. 8.

The upper two panels in Fig. 8 show the apparent recovered Raman rates before (left) and after (right) photobleaching, while the lower two panels show the corresponding apparent recovered fluorescence rates. Note that there is no evidence of a "+" pattern in the upper right panel; this indicates that photobleaching did not alter the apparent recovered Raman rates obtained from the film. There was a small (∼10%) decrease in average fluorescence intensity after photobleaching. I attributed this to the photobleaching that occurred while scanning the laser over the entire region during the OB-CD measurement. The similarity of the two upper images in Fig. 9 clearly demonstrates that it possible to obtain high-speed Raman intensity measurements using OB-CD filters in the presence of a fluorescent background with variable intensity.

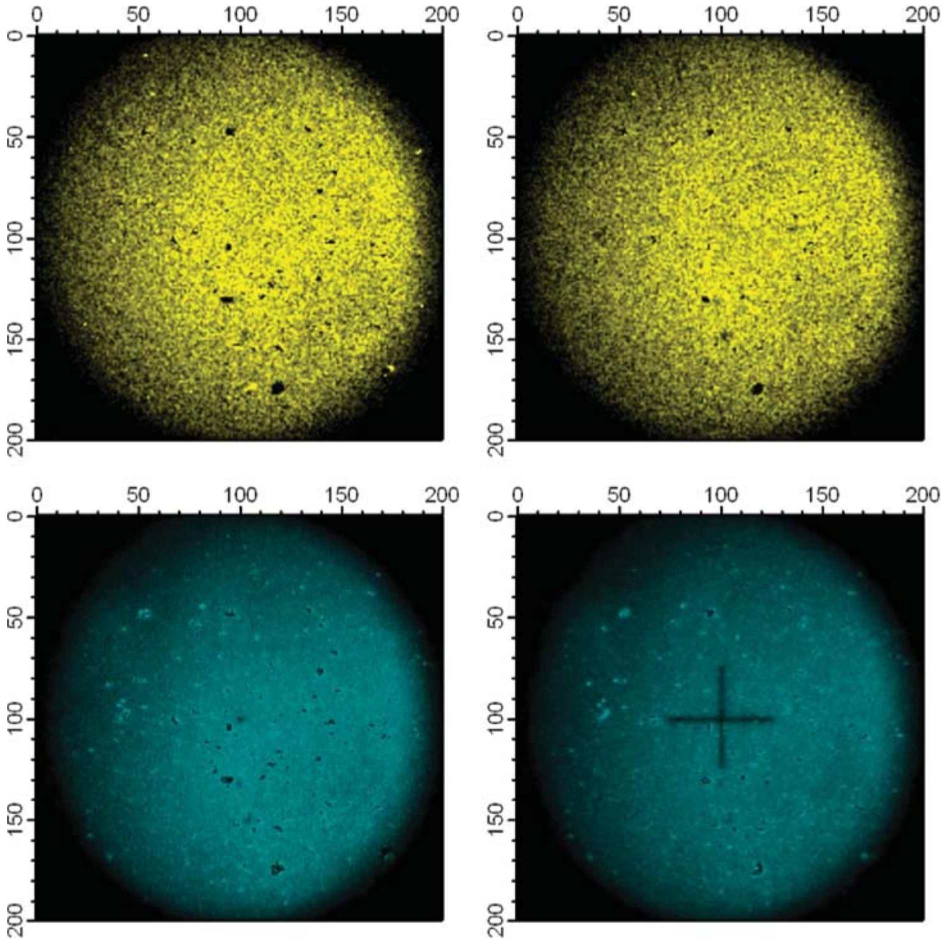


Figure 3.8. Images showing the recovered Raman (yellow) and fluorescence (cyan) rates of a cellulose acetate overhead transparency before (images on the left) and after (images on the right) photobleaching a "+" pattern into the center of the imaged area. All images were collected with an integration time of 10 ms per pixel. The circular nature of these images arises from the field of view of the objective, as the images were obtained by raster-scanning the angle of the laser as it enters the back of the objective (while remaining centered in the objective).

3.5 Conclusion

I have demonstrated an OB-CD fluorescence mitigation strategy that can be used to accurately recover Raman rates from samples with moderate fluorescence intensity (that is, up to 20 times more intense than that of the integrated Raman signal).

These results were achieved by quantifying fluorescence using OB-CD filters trained on cubic Bernstein polynomials. I have validated this strategy using both white light as a fluorescence mimic, as well as using fluorescent liquid and solid samples. The present results demonstrate the feasibility of fast (sub-second) OB-CD based Raman classification and quantitation of moderately fluorescent samples. This approach can be extended to systems with a fluorescence/Raman intensity ratios greater than 20:1, but would likely require turning down the laser intensity (to avoid detector saturation) and using much longer integration times. Thus, the presented OB-CD strategy is expected to be most useful in applications requiring fast analysis of liquid and solid samples whose fluorescence does not overwhelm the underlying Raman chemical fingerprints. This is consistent with previous results [3], which indicated that the trade-off between higher read-noise and higher spectral information content of full-spectral CCD measurements relative to the OB-CD detection strategy would indicate that OB-CD is most advantageous (relative to CCD measurements) in fast (low-signal) applications that unaccessible to CCD-based measurements.

4. BINARY-COMPLIMENTARY COMPRESSIVE FILTERS FOR RAMAN SPECTROSCOPY

4.1 Abstract

The previously described optimal-binary compressive detection strategy has been shown to enable fast Raman spectroscopic measurements for systems containing a small number of chemical components. However, during data collection, only a fraction of the scattered photons are transmitted through each OB-CD filter. Thus, a large percentage of photons are not detected during the OB-CD measurement process. Here I show an alternative binary compressive filter strategy and a novel compressive spectrometer, wherein binary filters are generated in pairs, such that when one OB-CD filter is applied to a digital micromirror device (DMD), the exact complement of that of that filter is, itself, also a filter. As a result, all scattered photons are either reflected positive 12 degrees or negative 12 degrees relative to the surface of the DMD, where these photons are then detected by photon counting detectors. Using this new spectrometer and filtering strategy, all of the scattered Raman photons are measured for the entirety of the total measurement time. This new filtering strategy, henceforth referred to as “OB-CD2”, is shown to result in approximately $\sqrt{2}$ lower variance in the estimate of Raman scattering rates of chemicals when compared to the original optimal-binary compressive detection strategy. Using OB-CD2, chemical classification at speeds up to 3 μs per measurement and Raman imaging as fast 2.5 frames-per-second are demonstrated.

4.2 Background

Raman spectroscopy takes advantage of the unique energies at which different chemicals inelastically scatter light. This method can be useful in chemical classification and quantitation. However, Raman spectroscopy is a relatively low scattering rate process and, as a result, collecting a Raman spectra at an adequate signal-to-noise takes seconds to minutes. As a result, it is difficult-to-impossible to extend Raman spectroscopy to applications that require measurement times on the microsecond time scale, such as chemical imaging or monitoring reaction kinetics. One strategy that enables fast Raman measurements is multivariate optical filtering. These methods utilize a computer-controlled optical filter, such as a digital micromirror device (DMD) [1] or liquid-crystal tunable filter (LCT) [23], to multiplex photons onto a single channel detector. This is advantageous owing to the fast read times, low read noise of single channel detectors when compared to multichannel detectors, such as a charge-coupled device (CCD) camera. The previously described optimal-binary compressive detection (OB-CD) photon collection strategy [1–4] is one example of a multivariate optical filtering method.

OB-CD has been shown to enable fast Raman measurements for applications such as fast liquid classification and fast imaging of mixtures of powders. It should be noted that OB-CD is similar to a number of other previously described methods [7–11] in that it utilizes a DMD to multiplex photons from different bins onto a single photon counting detector. Where OB-CD differs from these other strategies is how the OB-CD algorithm selects energy bins to multiplex, or in other words, how OB-CD optical filters are generated. It has been previously shown that OB-CD filters outperform filters built using total least squares (TLS) [1] and that OB-CD filters can be used to classify ternary mixtures of xylenes at speeds up to 100 times faster than previously described methods using a similar spectrometer design [2].

However, the OB-CD detection strategy may have difficulty operating at integration times on the order of microseconds. This is due to an optical inefficiency of the OB-CD spectrometer: while Raman photons transmitted by an OB-CD filter are being measured, all other Raman photons are being sent to a photodump and, as a result, are not detected. This is in contrast with a multichannel detector, such as a charge-coupled device (CCD) camera, where all Raman photons would be measured for the entirety of the total measurement time. To overcome this inefficiency, here I present an alternative OB-CD filter generation algorithm. This new algorithm generates pairs of complimentary binary filters, such that when one filter is applied to the DMD, its binary compliment is also a filter (and thus also applied to the DMD). Photons transmitted by one filter can be collected by a photon-counting photodiode as in OB-CD, while the photons that would be normally sent to a photodump are instead collected by a second photon-counting photodiode as they also correspond to photons transmitted through another filter. I refer to this filter generation strategy as “OB-CD2”, which reflects that measurements are now being made on two photon-counting photodiodes. Accordingly, a second photon-counting detector was installed in an OB-CD spectrometer, which is described in Section 4.3.1.

4.2.1 Model for Binary-Complimentary Filter Generation

OB-CD2 is a derivative on OB-CD and, as a result, many of the assumptions made in OB-CD filter generation hold true for OB-CD2. For a review of OB-CD filter generation, please refer to Chapter 2. Here we will note assumptions made that are shared in OB-CD and OB-CD2, though in limited detail, and highlight differences between OB-CD and OB-CD2.

Let P be a matrix with each column containing the area-normalized spectra of some n known chemicals and $\bar{\Lambda}$ be a (column) vector of mean photon emission rates of all chemical species under consideration. We assume that m independent measurements are made and that in the k^{th} measurement, the probability of detecting a photon with energy falling in the i^{th} bin is given by F_{ik}, the ik-entry of the matrix of filters F.

As with OBCD, given the vector $\hat{x}$ containing the empirical photon counts transmitted through each filter in F, we estimate $\bar{\Lambda}$ by the unbiased estimator $\hat{\Lambda} = B\tau^{-1}T^{-1}\hat{x}$, where T is a matrix whose diagonal is the percentage of the total measurement each filter in F is applied, τ is the total measurement time, and $B = (F^T P)^{-1}$.

The objective of the OB-CD2 is to construct a matrix of filters that allows for the estimation of $\bar{\Lambda}$ by minimizing the determinant of the variance-covariance matrix of $\hat{\Lambda}$:

$$Q(B, T, F, \bar{\Lambda}) = \det(B\tau^{-1}T^{-1}\text{diag}(TF^T P\bar{\Lambda})T^{-1}\tau^{-1}B^T)$$

Moreover, we also want to constrain F to consist of pairs of binary-complementary filters, that is, the absolute difference between each entry in a pair of filters is 1 and each entry in F_{ij} is either 0 or 1. This formulation (for OB-CD2) corresponds to the D-optimality criteria of experimental design, in contrast to the E-optimality criteria considered in the setting of OB-CD, which minimizes the summation of eigenvalues of the variance-covariance matrix. For more discussions about the comparison of D-optimality and E-optimality criteria, see Pukelsheim [24].

The formulation of complementary binary filters also relaxes the constraint that the number of filters should be equal to the number of chemicals under analysis (n). Instead, we build $m = (n-1)$ pairs of filters (that is, $2m$ filters in total), as this ensures that there are even number of filters constructed and that enough filters are generated to successfully classify/quantify the chemicals of interest.

The construction of the OB-CD2 optimal filter design is supported by the following observations. First, we note that, as with OB-CD, the value of Q obtained by taking $\bar{\Lambda} = \Lambda_0$, the vector of the form $(1, \ldots, 1)$ of length n, provides a reasonable bound for any value of $\bar{\Lambda}$. Second, while optimizing Q over all variables is difficult, this problem can be solved efficiently when certain dependencies between the variables are imposed. Specifically, for fixed values of T and $A = F^T P$, the unique optimal design matrix B is given by $B = (A^T T D^{-1} T A)^{-1} A^T T D^{-1} T$ where $D = diag(T_i A \Lambda_0)$.

Similarly, for a fixed value of F, if B depends on A and T as described above, then the unique normalized time matrix T^* satisfies $T^*_{kk} = (1/n)[AB]_{kk}$.

This enables us to propose the following procedure to construct the optimal filter design and implement it in Matlab. For a fixed matrix of filters F, characterized by m base filters and m corresponding complementary filters (calculated by 1 minus the value of the base filters), we compute the optimal normalized time matrix T_F and design matrix B_F by iterating their values, starting at the initial guess $T_0 = \text{diag}(1/n)$, using the formula

$$[T_{i+1}]_{kk} = \frac{1}{2}\left([AB_i]_{kk} + [AB_i]_{k'k'}\right)$$

where $A = F^T P$, $k' = 2m - k$, Λ_0 is the vector of the form $(1, \ldots, 1)$ of length n, $D = diag(T_i A \Lambda_0)$ and

$$B_i = (A^T T_i D^{-1} T_i A)^{-1} A^T T_i D^{-1} T_i$$

The values of T_{i+1} are then normalized to ensure that the summation of its components is equal to 1. This routine is iterated for $N = 100$ steps to obtain the optimal values of T_F and B_F. We note that at every step of this procedure, the value of the time matrix for the k^{th}-filter mirrors that of the $(2m-k)^{th}$-filter to ensure equal integration times on pairs complementary filters.

Finally, the function $Q(B_F, T_F, F)$ is then optimized over all possible values of the matrix of filters F. This is done by using a linear programming algorithm, similar to the algorithm used in OB-CD, which allows the optimal matrix of filters F^* to admit non-binary values. As with OB-CD, the optimal non-binary matrix of filters are nearly binary, in the sense that few entries of F^* are strictly between 0 and 1. A correction step is then performed to produce the optimal binary matrix (as with

OB-CD, this consists of rounding non-binary values to 0 or 1) of filters F^* and the corresponding optimal normalized time matrix T^* and design matrix B^*.

4.3 Materials and Methods

This section describes an OB-CD2 spectrometer built in conjunction with Laserlabs, LLC as well as the chemicals used to perform the work performed in section 4.4. Note the description of the OB-CD2 spectrometer will be described more fully (particularly with regards to optical design considerations) in a manuscript by Bharat Mankani that is currently in preparation.

4.3.1 OB-CD Spectrometer with 532 nm Excitation Laser

An OB-CD spectrometer using a 532 nm laser for excitation whose schematic is shown in Fig. 1 is shown below. Much like the previously described system, our microscope is configured to collect the backscattered Raman signal with the same objective lens that is used to focus the argon ion laser (Sapphire SD 532–150 CDRH Laser, 150 mW) onto the sample. The laser passes through a laser-line bandpass filer (Semrock RazorEdge LWP Filter, U–grade-25 mm) before it is focused onto the sample using a microscope objective (Olympus MPlan, 20x, 0.4 NA), and unless indicated otherwise the laser power at the sample was about 15 mW. The backscattered light is collected and then separated from the laser Rayleigh scattering using a dichroic mirror. Then, the Raman scattered light is sent to the spectrometer (right portion of Fig. 2), where it is filtered first using a long pass (edge) filter (Semrock RazorEdge, S–grade-12.5 mm), followed by passing through a volume holographic grating (Wasatch Photonics, $\sim$600 lines mm^{-1}). This light is then dispersed onto the DMD (Texas Instruments, DLP3000, 608 × 684 mirror array with 10.8 μm mirror pitch). The spectral window in this system is $\sim$200–4100 cm^{-1}. For all data collected in this paper, we binned two columns of adjacent DMD mirrors together, yielding a total of 342 “bins” with each energy bin corresponding to $\sim$12 cm^{-1}. Light from the

DMD is then focused onto a photon-counting photomultiplier tube (PMT) (Hamamatsu model #H10682-01) with a dark count rate of ∼500 photons s^{-1}. TTL pulses from the PMT are counted using a USB data acquisition (DAQ) card (National Instruments, USB-6212BNC). The system is controlled with interface software written in Labview 2013. As above, binary filter generation is performed as described above in algorithm written in Matlab (Matlab 7.13 R2011b).

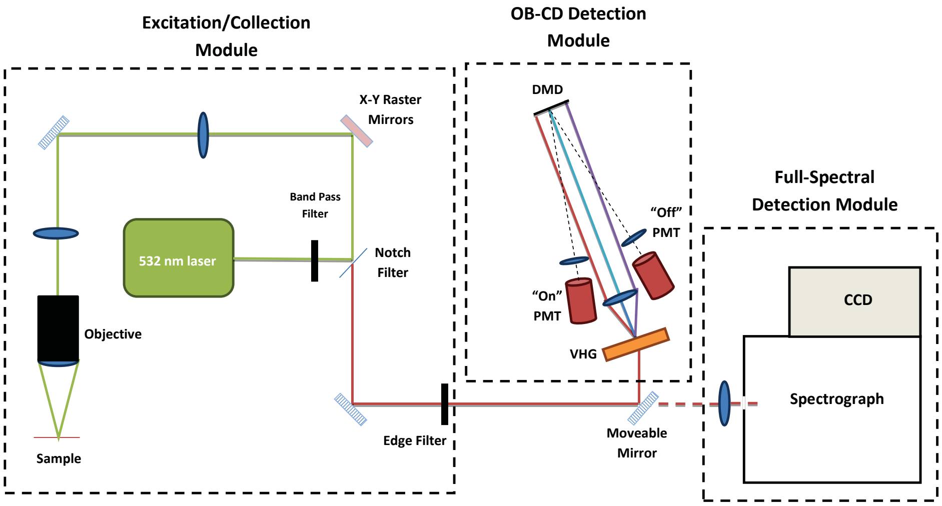


Figure 4.1. Schematic of the OB-CD2 Raman system based upon a 532 nm laser excitation source. Note that the system includes a CCD camera, though this was not used in this work.

4.3.2 Chemicals Used in Classification/Quantitation

Acetone and benzene were purchased from Macron (batch #0000070736) and OmniSolv (lot #42282), respectively. Hexanes were acquired from Baxter (lot #901141). Methylcyclohexane was acquired from Mallinckroft (lot #1906 KCBN). Benzoic Acid and acetaminophen were obtained from Sigma (lot #26115MA and lot #SLBC6391V, respectively). The benzoic acid was recrystallized in methanol overnight to produce large crystals and to remove a fluorescent impurity.

4.4 Results and Discussions

The following section summarizes the differences between estimated Raman scattering rates obtaining from OB-CD and OB-CD2. Additionally, examples of fast Raman imaging of solids enabled by OB-CD2 are provided.

4.4.1 Comparison to OB-CD for Binary Classifications

In previous work [1], acetone and benzene were classified in 30 μs using OB-CD Raman spectroscopy. Here I show the classification of acetone and benzene using the OB-CD2 spectrometer described above, using filters trained using either OB-CD or OB-CD2. Fig 2 shows the training spectra with the resulting OB-CD (left) or the OB-CD2 (right) filters overlaid over them. These spectra, and the resulting recovered rates, were collected using 65 mW of laser power at the sample. In the bottom portion of Fig 2, I show the resulting Raman scattering rate estimates for pure acetone and pure benzene at a 3 μs total integration time using either OB-CD or OB-CD2. Note that the resulting 95 % confidence interval ellipses for acetone and benzene are significantly overlapped for the OB-CD recovered rates, but not overlapped in the case of the OB-CD2 recovered rates. This demonstrates that at a given total integration time, Raman scattering rates recovered through OB-CD2 filters have lower variance, as all of the Raman photons are measured for the entirety of the filter measurement process. If the T matrix for both samples were identical, we would expect the OB-CD2 filters to give recovered rates with approximately a factor of $\sqrt{2}$ lower variance than the recovered rates (for the same chemicals measured at the same total integration time) determined using OB-CD. Here, the estimated Raman scattering rates for acetone and benzene recovered using OB-CD2 filters have a 35 % (or, in other words, on the order of a factor of a $\sqrt{2}$ reduction) lower variance than the OB-CD recovered rates.

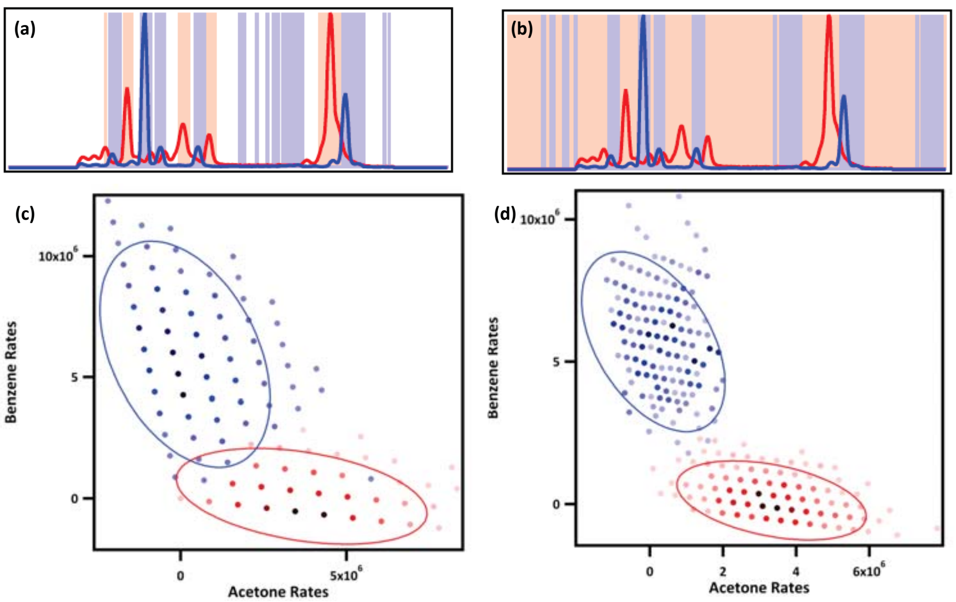


Figure 4.2. Comparison of a the recovered Raman scattering rates of acetone (red) and benzene (blue) estimated using (a) OB-CD and (b) OB-CD2. The two OB-CD or OB-CD2 filters are shown in plots above each rate scatterplot overlaid over the spectra of acetone and benzene. The 95 % confidence intervals of the recovered Raman scattering rates for each sample were calculated and are plotted as the ellipses in both (a) and (b). Note that the percent of the total measurement time for the OB-CD filters was 52 % for the filter shown in red and 48 % for filter shown in blue. Each OB-CD2 filter was applied for 50 % of the total measurement time.

4.4.2 Comparison to OB-CD for Ternary and Quaternary Classifications

Here I compare the recovered Raman scattering rates estimated using OB-CD and OB-CD2 filters trained using chemicals ($n = 3$): benzene, hexane, and methylcyclohexane. Note that for training sets with $n = 2$ chemicals, OB-CD2 produces the same number of filters as there are chemicals. However, for system with $n \geq 3$ chemicals, OB-CD2 produces more filters than chemicals in the training set (as noted above, $2(n-1)$). Fig 3a shows the training spectra and resulting OB-CD filters as a heat map, where the inner product between each filter vector and each normalized spectra is represented using a color (as denoted by the colorbars). Fig 3b shows the OB-CD2 filters for the same spectral training set. Note that for this work, the laser power at the sample was 25 mW.

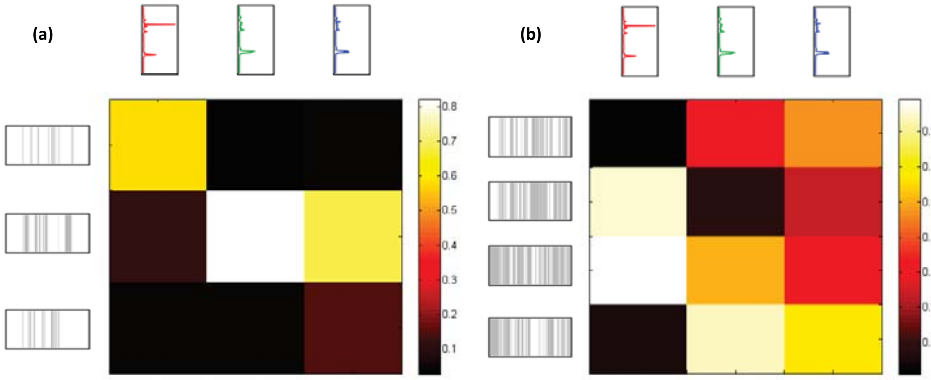


Figure 4.3. (a) Heat map of the overlap between OB-CD filters and the spectra used to train said OB-CD filters. The red trace is the spectra of benzene, the green trace is hexane, and the blue trace is the spectra of methylcyclohexane. A color bar is presented to the right of the heat map. Note that the diagonal contains the largest overlap values for each filter, though not the largest absolute overlap values (which are found for all spectra with the second filter). The measurement time for each filter was 16.4 %, 22.6 %, and 60.9 %, respectively. (b) Heat map of the overlap between OB-CD2 filters and the spectra used to train said OB-CD2 filters. A color bar is presented to the right of the heat map. The measurement time for each pair of filters was 50 % of the total time.

Once filters were calculated, the Raman scattering rates of pure benzene, hexane, and methylcyclohexane were estimated using OB-CD and OB-CD2 filters, with a total integration time of 1 ms. Fig 4 summarizes these results using a scatter plot matrix.

As seen in Fig 4, the size of the recovered Raman scattering rate ellipses for OB-CD2 are smaller than those in OB-CD. Note that the shape of the recovered Raman rate ellipses is often similar, as seen in the shape of the hexane and methylcyclohexane recovered Raman scattering rate distributions, but can differ significantly, as seen in the difference between the shape of the 95 % confidence intervals for benzene. This has been attributed to the difference in the minimization performed in OB-CD versus OB-CD2. In OB-CD, the filter generation algorithm minimizes the sum of the squares of all directions of the recovered Raman scattering rate ellipses, whereas in OB-CD2, the algorithm minimizes the area of the recovered Raman scattering rate ellipses.

Table 1 summarizes the variance of the estimated recovered Raman scattering rates for each sample. Note that the total variance for OB-CD is approximately 35 % larger than the of OB-CD2. This result was attributed to more light being measured during the same experiment time in OB-CD2 when compared to OB-CD.

Table 4.1.
Variance of Recovered Raman Scattering Rates for Ternary Classifications

Sample	OB-CD (10^9 counts/s)	OB-CD2 (10^9 counts/s)
Benzene	9.34	6.92
Hexane	4.64	3.87
Methylcyclohexane	13.9	6.84
Total	27.7	17.6

In addition this, OB-CD and OB-CD2 filters were trained on a set of four chemicals (acetone, benzene, hexane, and methylcyclohexane). The training spectra and filters are summarized in Fig 5. Note that, as with the classifications performed above, the laser power at the sample for this work was 25 mW. However, the total integration time was increased to 10 ms for all measurements.

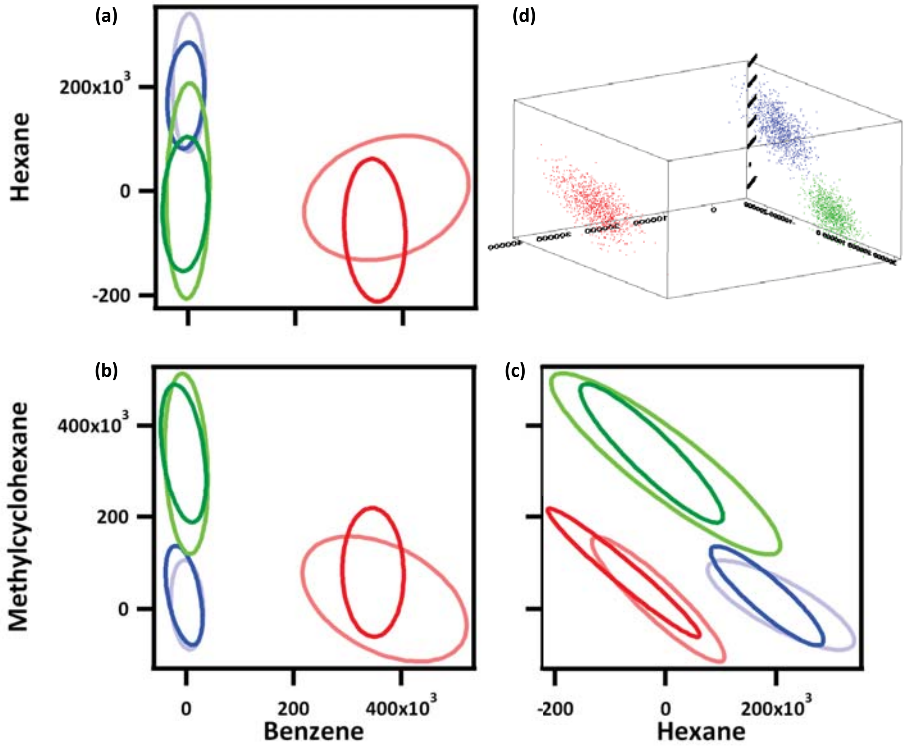


Figure 4.4. (a–c) Scatter-plot histograms of the OB-CD (lighter colors) and OB-CD2 (darker colors) recovered Raman scattering rates for samples of benzene (red), hexane (blue), and methylcyclohexane (green) estimated using a total measurement time of 1 ms in both cases. The ellipses shown correspond to the 95 % confidence intervals for the recovered Raman scattering rates of each sample. Note that each scatter-plot shows two-dimensions in the three-dimensional classification space. For example, in (b), the scatter-plot shows the recovered Raman scattering rates for benzene and methylcyclohexane. Note that the benzene and methylcyclohexane samples have mean non-zero rates in one dimension and near mean zero recovered rates in another dimension. The samples of hexane, however, have near mean-zero recovered rates for both methylcyclohexane and benzene, as would be expected for a sample of pure hexane (which should only have position, non-zero recovered rates in the hexane dimension). (d) A three dimensional scatter-plot of the same OB-CD2 recovered Raman scattering rates shown in a–c.

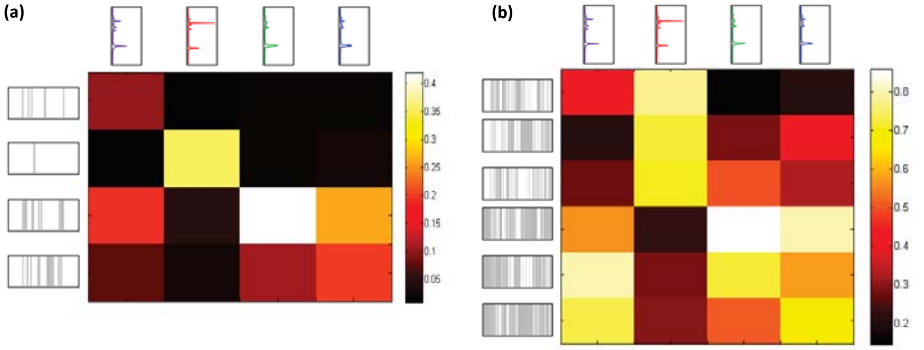


Figure 4.5. (a) Heat map of the overlap between OB-CD filters and the spectra used to train said OB-CD filters. The purple trace is the spectra of acetone, the red trace is the spectra of benzene, the green trace is hexane, and the blue trace is the spectra of methylcyclohexane. A color bar is presented to the right of each heat map. Note that the diagonal contains the largest overlap values for each filter, though not the largest absolute overlap values (which are found for all spectra with the second filter). The measurement time for each filter was 28.0 %, 11.4 %, 24.9 %, and 35.7 %, respectively. (b) Heat map of the overlap between OB-CD2 filters and the spectra used to train said OB-CD2 filters. A color bar is presented to the right of the heat map. The measurement time for each pair of filters was 33.333 % of the total time.

The resulting Raman scattering rates estimated using OB-CD and OB-CD2 are summarized as scatter plot matrix in Fig 6. Note that, as with the ternary classification, the recovered Raman scattering rate ellipses were generally smaller for OB-CD2 than OB-CD. The variances of the recovered Raman scattering rates are summarized in Table 2. The total variance, as well as the variance of acetone, hexane, and methylcyclohexane, is reduced in the recovered OB-CD2 rates when compared to the recovered OB-CD rates. However, the benzene recovered Raman scattering rates in OB-CD and OB-CD2 are almost identical. While it is difficult to definitely assign a cause to this, it is my suspicion that this is a result of the comparatively high scattering rate of benzene relative to all other compounds in the training set (nearly double the other chemicals, per Fig 6). As a result of this high scattering rate, which is also mostly in two energy bins (the two large peaks of benzene, as seen in the red spectra in Fig 5), there are a large number of photons associated with benzene measured in each pair of OB-CD2 filters. This is a necessity of making complimentary filters. As a result of this, there is a large shot noise associated with benzene (again, owing to the higher rates) even through filters not "mostly" measuring benzene. Due to this, it is possible that, for the benzene sample, the recovered rates not associated with benzene have great variance owing to the higher shot noise. Future work will investigate this further.

Table 4.2.
Variance of Recovered Raman Scattering Rates for Quaternary Classifications

Sample	OB-CD (10^9 counts/s)	OB-CD2 (10^9 counts/s)
Acetone	17.9	0.847
Benzene	1.52	1.53
Hexane	1.34	0.875
Methylcyclohexane	2.31	1.42
Total	6.96	4.68

4.4.3 Fast Raman Imaging Using OB-CD2

The objective of the OB-CD2 algorithm is to enable very fast Raman spectroscopic measurements. One application that would benefit from faster Raman measurement times is Raman hyperspectral imaging. Here I demonstrate fast Raman imaging of two pharmaceutical-like compounds.

Large benzoic acid and acetminophen crystals were gently ground into smaller crystals using a spatula. Then, small piles of these smaller crystals were placed on a gold microscope slide. These piles were flattened to produce an even surface for

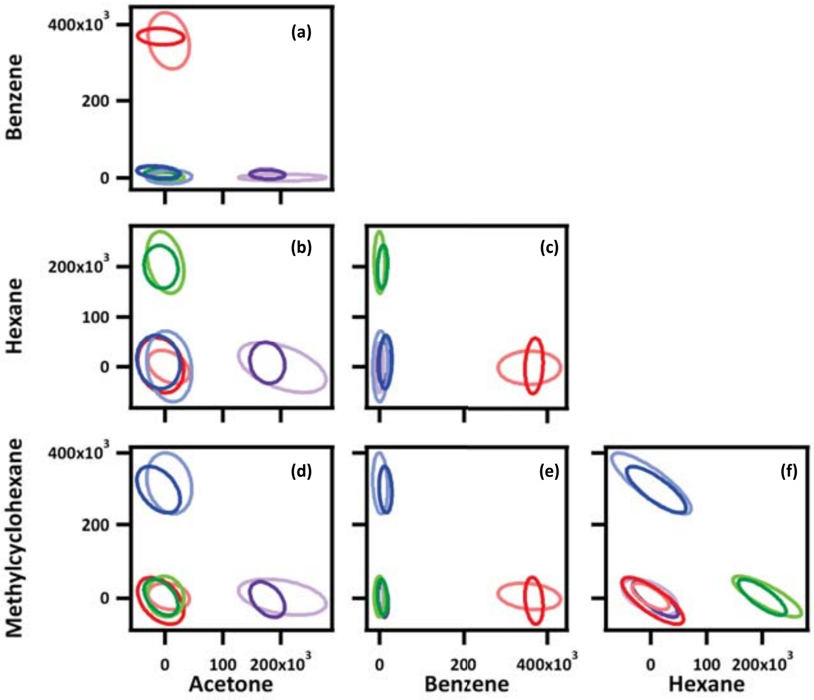


Figure 4.6. (a–f) Scatter-plot histograms of the OB-CD (lighter colors) and OB-CD2 (darker colors) recovered Raman scattering rates for samples of acetone (purple), benzene (red), hexane (blue), and methylcyclohexane (green) estimated using a total measurement time of 10 ms in both cases. The ellipses shown correspond to the 95 % confidence intervals for the recovered Raman scattering rates of each sample. Note that each scatter-plot shows two-dimensions in the four-dimensional classification space.

imaging. Then, the piles of acetaminophen and benzoic acid were pushed against each other.

Training spectra were generated by measuring the spectra at six different points over the surface each pure powder. The mean of these spectra was used for training. This was done as small differences were observed in the relative intensity of Raman peaks taken at different points on the surface of pure acetaminophen and benzoic acid. This was attributed to differences in crystal orientation, resulting in slightly different Raman spectra. It has been empirically found that taking the mean of spectra of such anisotropic crystals yields robust spectra for training OB-CD and OB-CD2 filters. OB-CD2 filters for the mean spectra of acetaminophen and benzoic acid are shown in Fig 7.

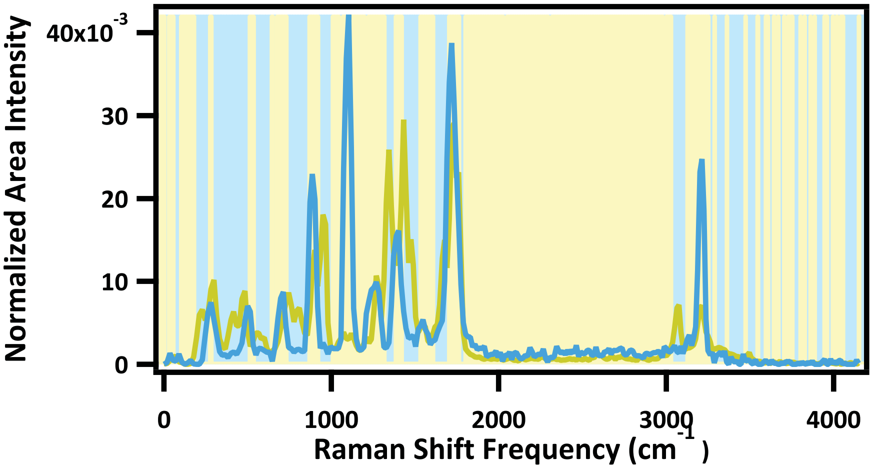


Figure 4.7. OB-CD2 filters generated for acetaminophen (yellow) and benzoic acid (cyan) are shown overlaid over the area normalized spectra of acetaminophen and benzoic acid. Each OB-CD2 filter was applied for 50 % of the total measurement time. For this work, the laser power at the sample was set to 25 mW.

Once OB-CD2 filters were trained, a 500 mm by 500 mm area, consisting of 240 by 240 pixels, at the interface between acetaminophen and benzoic acid was imaged at varying integration times. Two of these images, taken at integration times of 100 μs/pixel and 10 μs/pixel can been seen in Fig 8a. Note that 40 pixels at the top and left of these images has been truncated in order to minimize a hysteresis in the raster scanning mirrors when moving at high speeds. These integration times correspond to a total image time of 5.76 s and 0.576 s, respectively.

The images in Fig 8a associated with a 100 μs/pixel and 10 μs/pixel integration times are quite noisy. To minimize the visual effect of this noise, these images were denoised using the following procedure:

The photon counts measured through each filter were assumed to have Poisson (shot) noise but no other sources of noise. While Poisson denoising is possible, it can be computationally expensive. As an alternative, we have transformed the Poisson data into Gaussian-like data using the Anscombe transform [25], as Gaussian noise is constant and there are many, freely available, fast algorithms for Gaussian denoising. Here, we have selected the "BM3D" algorithm [26] and processed the Anscombe-transformed counts through each filter using a 5-by-5 pixel window. Then, the denoised counts are retransformed back to a Poisson distribution using an "optimal" inverse Anscombe transform (as described by [27]). These denoised counts are then transformed into recovered Raman scattering rates as described in Chapter 2. The resulting denoised images are shown in Fig 8b. As can be seen, these images appear less noisy to the eye.

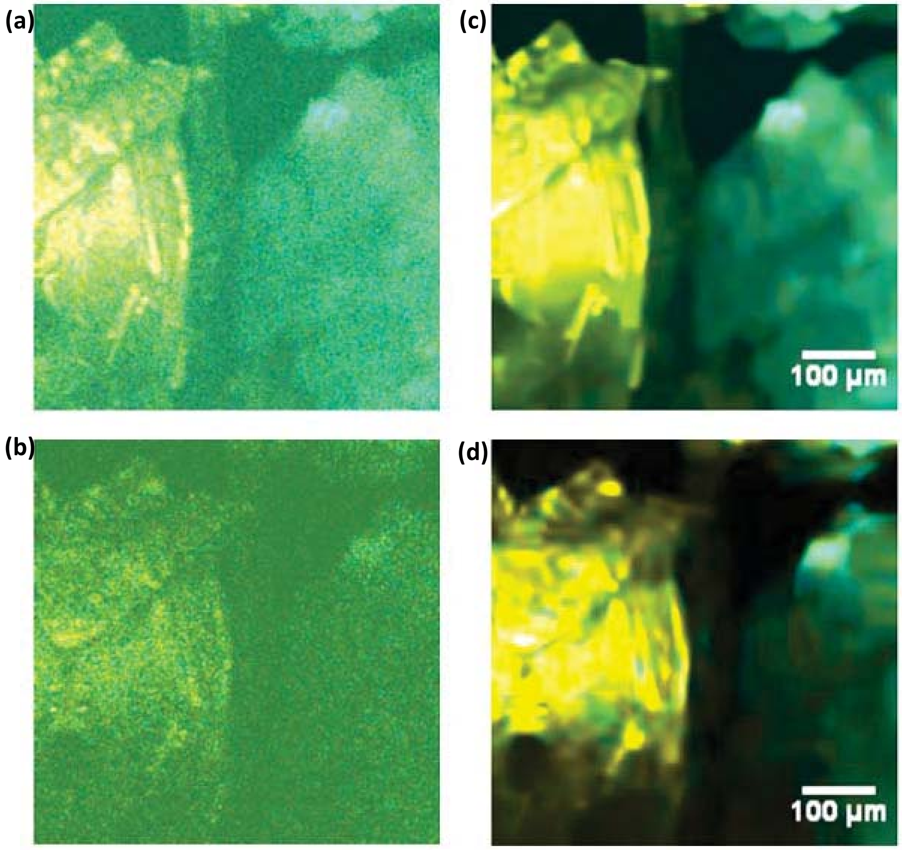


Figure 4.8. (a–b) Images of the interface of piles of acetaminophen (yellow) and benzoic acid (cyan) powders taken at 100 μs (a) and 10 μs total integration time. Note that despite being the same imaged area, the objects move to the right in the frame. This was attributed to a hysteresis of the scanning mirrors. (c–d) Images (a) and (b) after denoising, as described above.

4.5 Conclusion

I have demonstrated a new variant of the previously described OB-CD filtering strategy that takes advantage of a second photon-counting detector and a new filter generation algorithm. This new strategy, referred to as OB-CD2, has been shown to result in lower variance in the recovered Raman scattering rate estimates when compared to OB-CD performed at the same total integration time. As a result of this, OB-CD2 enables faster classifications than OB-CD. Additionally, OB-CD2 may enable new applications for OB-CD for a more mundane reason: the filter image applied to the DMD need not be change to apply two different OB-CD2 filters when $n = 2$. For the same system, it would not be possible to measure photons through two filters in OB-CD simultaneously, as they must be applied sequentially on the DMD. Since the application of the filter to the DMD takes on the order of hundreds of milliseconds, not needing to change filters in OB-CD2 enables applications that require constant measurements over a period time, such as determination of kinetic rate constants. Future work will focus on exploring such applications.

5. RAMAN THERMOMETRY USING NON-NEGATIVE MATRIX FACTORIZATION DERIVED FACTORS

5.1 Abstract

Raman thermometry is the practice of using Raman spectral information to predict temperature of a sample. This technique is useful when monitoring the temperature of systems that would be perturbed by the addition of a thermometer or thermocouple. Several implementations of Raman thermometry have been previously described. Of these, techniques that use changes in relative Raman peak areas to predict temperature are the most commonly used technique for aqueous solutions applications. This is because the OH- stretch relative peak areas of water are quite sensitive to temperature change. Recent implementations of Raman thermometry reported measurement precision of ±1 ° Celsius in a measurement time of ∼1 second. While quite fast, this is still slow compared to other optical thermometry techniques and limits the applications of Raman thermometry. Here I describe a new spectral processing method of for obtaining temperature estimates of aqueous solutions. Rather than determining Raman peak area using isosbestic point of water's OH- stretch, as many existing methods do, I use non-negative matrix factorization (NNMF) to derive factors that account for the change in OH- peak shape of water over a temperature range. I then use a least squares fit to estimate areas for the two spectral factors, which I then correlate to temperature. Using a calibration curve based upon a three-term fit, I am able to predict the temperature of aqueous solutions with a ±1 ° C precision in 10 ms rather than ∼ 1 second, as previous methods have reported. I demonstrate this approach is appropriate for aqueous solutions with different types of spectral interference, such as fluorescence and spectral overlap from

CH- stretch peaks, and I demonstrate it is possible to obtain temperature precision similar to that of water in the same integration time.

5.2 Background

Optical thermometry is a family of techniques wherein the temperature a sample is estimated by correlating temperature to changes in sample's optical properties. Raman thermometry is the process of using temperature-dependent changes in Raman scattering to predict temperature [28–33]. Raman thermometry has the advantage of being appropriate for use on a variety of materials and being applicable to micron-scale systems [34]. However, Raman thermometry has the disadvantage of being slow compared to other optical thermometry techniques (assuming the same measurement precision is desired) [35], as it typically needs seconds of acquisition time to generate temperature predictions with a precision of ± 1 ° C.

Several different methods of performing Raman thermometric have been reported; for instance: the use of the ratio of Stokes vs. anti-Stokes peaks [36], peak shifts [31], and relative changes in peak area [29, 33]. Of these, the technique that utilizes changes in relative peaks areas has most commonly been used for aqueous solutions, as the OH- bands of water have noticeable temperature-dependent changes. This technique, originally described by d'Ararigo et. al [28], has a precision of about ± 1 °C with measurement times of approximately 1 second [33]. In this method [28], it was reported that the OH- bands of water has an isosbestic point at approximately 3400 cm^{-1}. Rather than make assumptions about the shape of the temperature bands, integration is performed to the left and the to the right of isosbestic point and a ratio of these peak areas is correlated to temperature as measured by another, calibrated device (e.g., a thermocouple or thermometer).

The approach described here is an alternative to integration based upon an isosbestic point. Instead, I use a spectral unmixing method called non-negative matrix factorization (NNMF) to extract two factors that, when taken as a linear combination, describe the shape of all the spectra measured over a temperature range. Once these two NNMF factors are calculated, we use least-squares fitting to calculate weights (areas) for the two factors for each measured spectrum. Then we take the ratio of the areas of the NNMF factors and correlate the area-ratio to temperature as measured by a thermocouple in our Raman system. Fig 1a shows the spectra of water at 20–80 °C and Fig 1b shows the resulting NNMF-derived factors.

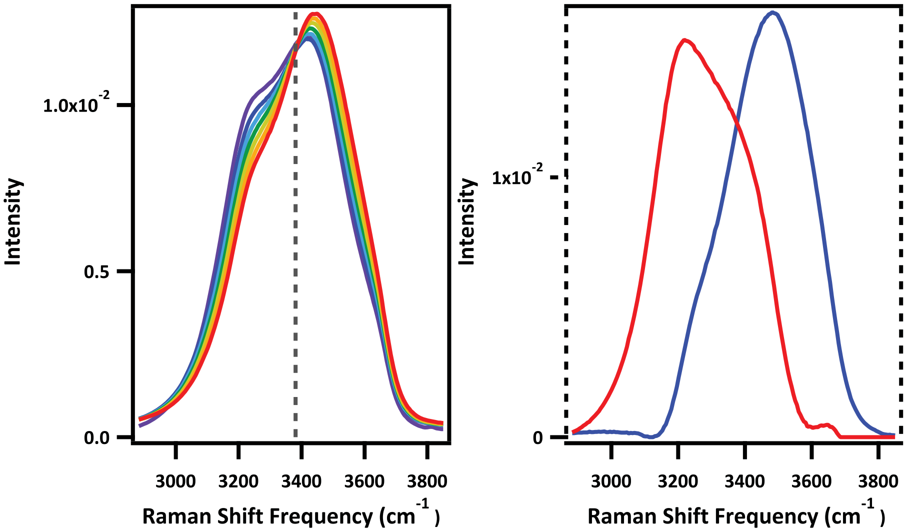


Figure 5.1. (a) Raman spectra of water from 2883–4073 cm^{-1} taken at (purple) 20 °C, (blue) 30 °C, (cyan) 40 °C, (green) 50 °C, (yellow) 60 °C, (orange) 70 °C, and (red) 80 °C. The isosbestic point of water is located at the hashed line. (b) The two NNMF factors derived from the spectra shown in part (a).

5.3 Materials and Methods

This section describes the Raman microscope used to perform these experiments as well as the chemicals used in these experiments.

5.3.1 Raman Microscope

Raman spectra were collected, using a 514 nm excitation laser system that was previously described([37,38]). This system has a sample holder that can be set from -15 °C to 100 °C using a temperature (to the nearest 0.01 °C) controlled spectroscopic cell holder (LC600, Quantum Northwest), and a home-built Raman instrument (utilizing a 514.5 nm argon ion excitation laser with a power of $\sim$15 mW at the sample).

5.3.2 Temperature Calibration

The following section describes a method for constructing a temperature calibration curve and its validation. The Raman spectra of pure Water (Milli-Q UF Plus, 18.2 Mcm, Millipore) was measured using a 100 ms integration time, with 10 accumulations taken, at 10 °C increments from 20 °C to 80 °C, with the temperature controlled by the Raman system as described above. The temperature of the sample was allowed to equilibrate for 15 minutes prior to collecting its Raman spectrum. Ten replicates were taken at each temperature.

Once collected, these spectra were loaded into Matlab (version R2013b, Mathworks). Once in Matlab, the region of the collected spectra corresponding to the OH-stretch of water (2883–4073 cm^{-1}) were added, as column vectors, to a matrix D. This matrix was used as the input for non-negative matrix factorization (NNMF), using the “nnmf” command in Matlab’s Statistics and Machine Learning Suite with the number of significant components set to “2”. It was assumed that the spectra of water can be described by a linear combination of the NNMF derived factors, with single-value decomposition indicating that two factors were adequate to describe over 95% of the variance of the spectra. To handle any variability in the baseline intensity of non-fluorescent samples, I also included a line (normalized to unit area) in our least-squares fitting matrix, such that any spectra can be considered a linear combination of the two NNMF factors and a constant term. This matrix was used to determine concentrations (areas) for each NNMF factor using a least squares fit.

Note that the two NNMF derived factors look quite similar to the "hydrogen bonding" and "hydrogen bonding" shapes for water's OH- as described by d'Arrigo [28], as seen in Fig 1b. Keeping with notation, we will refer to the NNMF derived factors as "hydrogen bonding" (red trace in Fig 1b) and "non-hydrogen bonding" (blue trace in Fig 1b).

To create a calibration curve, the natural log of the mean ratio of the peak areas, calculated as the ratio of the area$_{non-hydrogenbonding}$ factor to the area$_{hydrogenbonding}$ factor (henceforth, "K"), at each temperature were plotted against the reciprocal of temperature (in Kelvin). These values (ln(K) and 1/T) are then plotted as a so-called van't Hoff plot. In a van't Hoff plot, it is typically assumed that the two populations of OH- stretches represent a two-state system. Thus, the ratio of the hydrogen and non-hydrogen bonded peaks areas can be thought of as a equilibrium constant, hence the use of the symbol "K". Using the fit of the data points in the van't Hoff plot, temperature can be correlated to the log of the ratio of the two water peak areas.

5.3.3 Aqueous Solution Preparation

Aqueous solutions were prepared using ultrapurified water (Milli-Q UF Plus, 18.2 Mcm, Millipore). A 50 % volume-by-volume solution of tert-butanol (TBA) was prepared using tert-butanol from Sigma-Aldrich (lot # SHBD8139V). A 1 M sodium chloride (NaCl) solution was prepared using sodium chloride from Mallinckrodt (lot # H43615). A 0.25 nM solution of rhodamine 6G was prepared using rhodamine from Eastman (lot # 11927).

5.3.4 Collection of Spectra of Aqueous Solutions

All solutions discussed in section 5.4 were collected under the following conditions: 100 ms integration time with 10 accumulations and 10 replicates collected and 100 ms integration time with 1 accumulation and 100 replicates.

5.4 Results and Discussions

This section presents the results of the experiments described above and puts them in context of the existing thermometry literature.

5.4.1 Comparison of Different Fit Models

Other authors have demonstrated Raman thermometry of aqueous solutions by linearly correlating the ratio of non-hydrogen bonded/hydrogen bonded water peaks with temperature, typically with goodness-of-fit values around $R^2 = 0.999$ [39, 40]. The temperature accuracy as a result of these fits is ±1 °C with integration times of approximately 1 s. Based upon the data generated by our Raman system, I have found that ln(K) and 1/T do not have an extremely strong linear relationship. Future work will focus on better understanding the nature of this discrepancy. However, there was a visibly non-linear relationship between the ln(K) and reciprocal temperature, non-linear fits were considered. To this end, a quadratic fit was tested. Additionally, based upon previous work done in the Ben-Amotz lab [41], a linear enthalpy dependence with temperature would give the following relationship for ln(K) and 1/T:

$$ln(K) = a + b/T + c * ln(1/T)$$

Where a, b, and c are fit constants and b and c describe the dependence of enthalpy on temperature. To generate a goodness-of-fit value for this fit type, the “polyfit” Matlab command was used (with a custom fit type). We found the ln(K) values for water correlated with 1/T in a manner described by the fit above with a $R^2 = 1.0000$. The result of these fits can be seen in Fig 2, where the difference between the known-true temperature (per the thermocouple) and the predicted value for a linear and two three-term fit models (quadratic and the model described above) are compared.

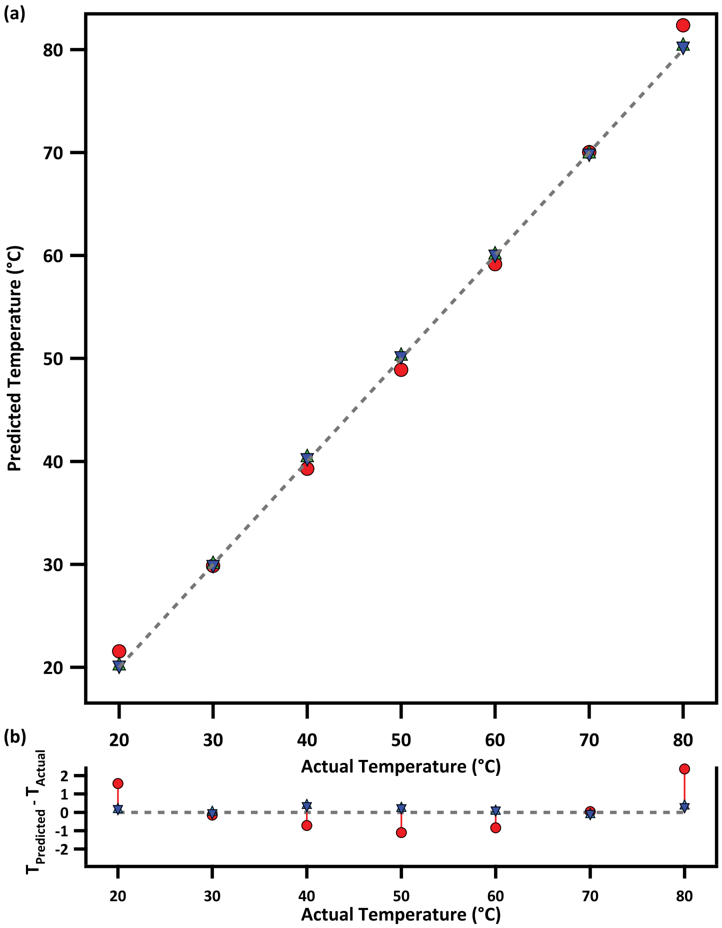


Figure 5.2. (a) Plot of the natural log of the peak areas of the NNMF derived factors for water (as seen in Fig 1b) versus the reciprocal of temperature in Kelvin for water whose Raman spectrum was collected at 20–80 °C (as seen in Fig 1a). The measured data points are pointed as solid black circles, with three different fits overlaid over this data: fitting the points to a linear fit (red), a quadratic fit (blue), and the three-term fit based upon a constant enthalpy described in this section (green) (b) The difference between the measured temperature values and the estimated temperature from each fit, for each temperature. The mean difference between the estimated temperatures calculated using a linear fit and the thermocouple measured temperature was 0.78 °C and the mean difference between the estimated temperatures calculated using both three term fits and the thermocouple measured temperature was 0.15 °C.

Note that the peak area values used here were calculated using a least-squares fit to the NNMF derived factors. The sum of the magnitudes of the difference between the thermocouple and the predicted value for both fit types is presented in Fig 2b. As can be seen, the three-term fit models resulted in a significant (factor of 5) improvement in temperature prediction accuracy.

5.4.2 Comparison of NNMF and Isosbestic Integration

In order to compare these results to those previously reported, I separated factors that could be attributed to the high signal-to-noise of the Ben-Amotz lab's Raman system and factors arising from how I processed Raman spectra (i.e., our fit model). To do this, I compared our the results of using a least-squares fit using the NNMF derived factors (henceforth, "NNMF integration") to the results of using peak area integration relative to the isosbestic point of the OH- stretch of water to calculate a peak ratio (henceforth, "isosbestic integration"). To do this, I selected the same spectral region (2883–4073 cm^{-1}) for both analyses. I identified an isosbestic point in our data at approximately 3410 cm^{-1}, as seen in Fig 1a. NNMF integration was performed as described in the subsection 5.3.2. Isosbestic integration was performed by summing the intensity of the energy bins to the left (in the case of hydrogen bonding) and to the right (in the case of non-hydrogen bonding) of the isosbestic point. Note that in both cases, the minimum value of the wavenumber range was subtracted from each spectra to correct for baseline (to minimize baseline noise in the isosbestic integration). For these analyses, the three-term fit model described in subsection 5.4.1 was used for calibration.

These spectra were then processed into a van't Hoff plot, as described above. The resulting van't Hoff plot is in shown in Fig 3a. Fig 3b highlights the difference between the measured temperature values and the estimated temperature from each integration method. Note that the mean model error for the isosbestic integration is qualitatively similar, if marginally worse, than the NNMF integration method. As I will demonstrate in the subsection 5.4.3, the NNMF integration method, combined with the three-term fit described above, can be easily used for a variety of samples, even samples that might be difficult to use with the isosbestic integration method.

5.4.3 Precision of Temperature Estimate Versus Signal-to-Noise

Here I derive a correlation between signal-to-noise of the OH- stretch in a Raman spectra of water and the variance of the resulting temperature estimate. Signal-to-noise was selected as the independent variable for this analysis, rather than integration time or laser power, as signal-to-noise is applicable to any Raman system. Here signal-to-noise is calculated as the area of the Water OH- stretch (the sum of the signal from 3025–3800 cm^{-1}) minus the noise of the baseline (calculated at ∼4100-4200 cm^{-1}) divided by the noise of the baseline. It would be expected that the variance in the input spectra would linearly correlate with the variance in the temperature estimate. Since signal-to-noise scaled inversely with the variance of OH- stretch area, it would be expected that a plot of the inverse of the signal-to-noise of the OH- stretch area versus the standard deviation of the estimated temperature would be linear. As can be seen in Fig 4, such a plot is linear.

The plot in Fig 4 was generated by adding Poisson noise to the spectra of water measured without any laser attenuation and an integration time of 1 s. The addition of Poisson noise by dividing the initial water spectra by a constant, then running the "poissrnd" function in Matlab, which adds Poisson noise (of the appropriate mean) to the input spectra. The spectra of water measured at 1s was divided by increasingly raised powers of 2, starting at the power of 1 and ending at the power of 13 (i.e., ranging from a factor of 2 to a factor 8192).

Using the fit shown in Fig 4, a signal-to-noise ratio of ∼200 corresponds to a temperature estimate with a precision of ±1 °C and a signal-to-noise ratio of ∼30 is need for a temperature estimate with a precision of ±10 °C. These signal-to-noise values correlate with an integration time 10 ms and 1 ms, respectively, using the

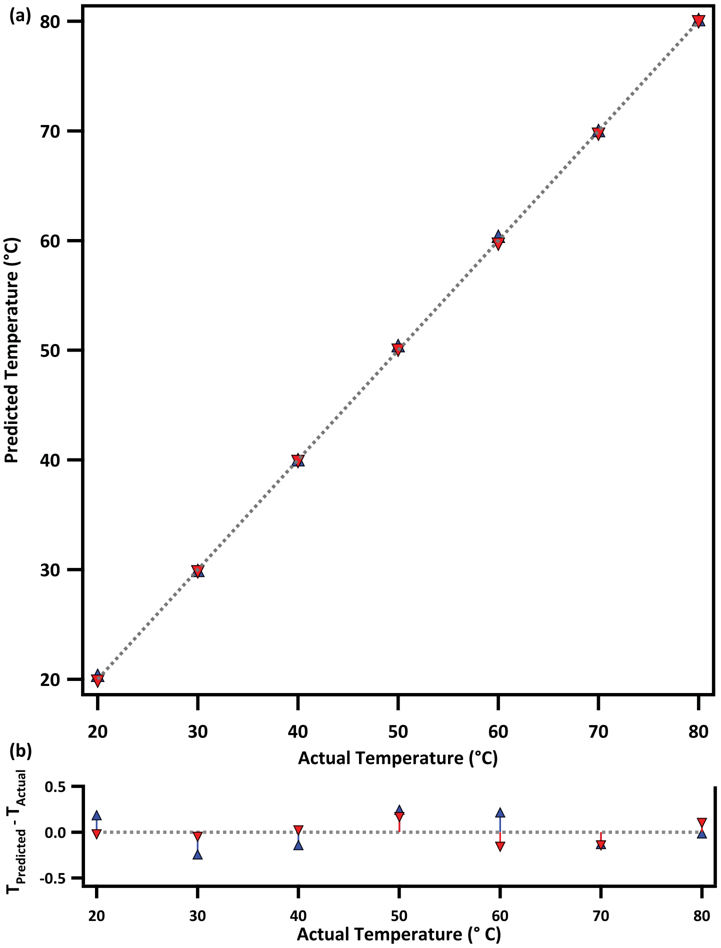


Figure 5.3. (a) Plot of the thermocouple measured temperature for water versus the estimated temperature of water over the range of 20–80 °C. Estimated temperatures calculated using NNMF integration are shown using inverted red triangles, estimated temperatures calculated using isosbestic integration are shown using blue triangles, and a line of slope 1 is plotted in hashed grey. (b) The difference between the measured temperature values and the estimated temperature values from each integration method at each temperature. The mean difference between the estimated temperatures calculated using NNMF integration and the thermocouple measured temperature was 0.15 °C and the mean difference between the estimated temperatures calculated using isosbestic integration and the thermocouple measured temperature was 0.20 °C.

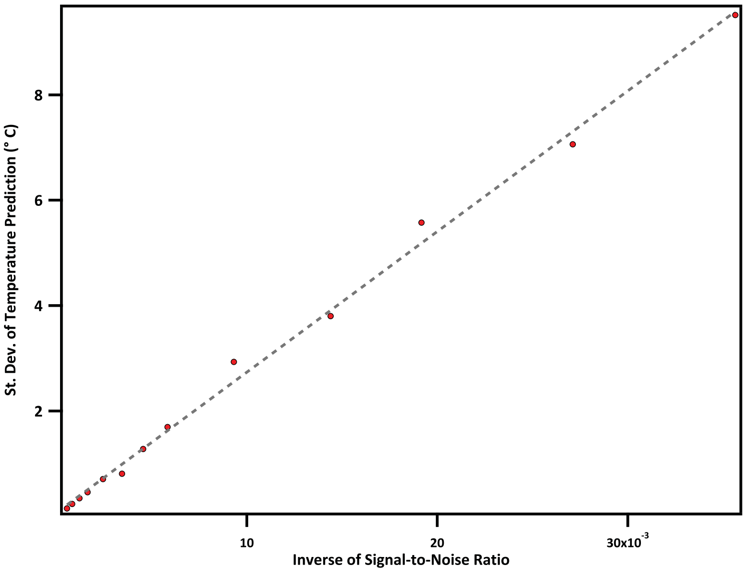


Figure 5.4. Plot of the inverse of the signal-to-noise of the water OH-stretch versus the standard deviation of the temperature estimate of the same spectra. This data was generated by corrupting the spectra of Water, measured at 20 degrees C, by dividing all intensities in the spectra by a constant and adding Poisson noise (using the "poissrnd" function in Matlab). The hashed-line is a linear fit ($R^2 = 0.99$) of the data and the red circles correspond to spectra at varying levels of added noise.

Raman microscope described above with ∼15 mW of laser power at the sample. Future work will attempt to compare these values with experimentally derived values. Careful consideration will be given to issues of baseline correction and detector noise, which can cause a non-negligible non-linearity in plots like Fig 4.

5.4.4 Results of Aqueous Solution Temperature Prediction

Here I report the accuracy and precision of estimated temperature for a variety of aqueous solutions. These estimated temperatures were calculated using a calibration curve calculated using a peak area derived from NNMF integration and the three-term fit described in subsection 5.4.2. The first system I report is a 50 % volume-by-volume mixture of water and TBA. This system is notable in so far as it contains a significant amount of a chemical containing CH- bonds. Since the CH- stretch feature overlaps with the OH- stretch in Raman spectra, this sample demonstrates that NNMF integration can still be used over a limited range of the OH- stretch spectra (explicitly, the region that does not contain the CH- stretch spectral feature). The inset in Fig 5b shows the difference between the NNMF derived factors for the water:TBA mixture and pure water, which look noticeably different. This is not surprising, as it has been previously demonstrated that the structure of water around alcohols differs from that of pure water. Even so, using NNMF derived factors, it is possible to accurately predict the temperature of water:TBA mixtures, as shown in Fig 5b.

Next, I demonstrate the applicability of the use of NNMF integration to samples that have perturbed water OH- structure (and therefor, OH- stretch Raman features), but lack an interfering spectral feature such as a CH- stretch. One example of this type of aqueous mixture is aqueous sodium chloride, which has perturbed OH- stretch feature when compared to pure water as seen in the inset of Fig 6a. NNMF derived factors were calculated using the spectra shown in Fig 6a and used to estimate the temperature of 1 M NaCl solutions. Despite looking quite similar to the water NNMF factors, using the NNMF derived from water for NaCl resulted in more error in temperature estimate than using the NNMF factors derived from NaCl. The resulting temperature estimated are shown in Fig 6b, which are qualitatively similar, in terms of accuracy and precision, to the temperature estimates for pure water under similar conditions. Thus, it has been demonstrated that NNMF is able to derive factors from aqueous solutions with perturbed OH- stretch features, such as salt solutions.

Fluorescence is a common nuisance component in Raman spectroscopy and I would be remiss if I did not demonstrate the accuracy and precision of this Raman thermometry model for fluorescent solutions. Fluorescence arising from solutes/impurities in aqueous solutions is common, but broad spectral features arising from glass in the optically probed volume can be treated similarly (owing to the broadness of glass Raman spectral features). Raman thermometry is a subject of great interest to the microfluidics applications [34], which cannot easily attach a thermocouple to their systems without perturbing them. It is not uncommon for the Raman spectra of materials within microfluidic devices to also contain significant glass signal. Thus, I have demonstrated the applicability of NNMF integration/three-term fitting to systems that include glass signal. While I have not measured the Raman spectra of aqueous solutions within a microfluid device, I have collected the spectra of water when the sample holder on our Raman system was moved to include a significant contribution of glass signal. To remove the broad glass features, I performed a background subtraction on each full spectra, as described in Chapter 3, prior to NNMF integration. Of note, the inset in Fig 7a shows the difference between NNMF derived factors calculated using the spectra of pure water and NNMF derived factors calculated using the spectra shown in Fig 7a after baseline subtraction. These NNMF derived factors looked quite similar and, as a result, the NNMF derived factors for pure water were used to perform NNMF integration on the baseline-subtracted water-with-glass spectra. The estimated temperatures resulting from this are shown in Fig 7b. The accuracy and precision of these results are worse than those of pure water, though this is not unexpected given the interference in the OH- stretch signal from the glass. Thus, it has been shown that, by using baseline subtraction, temperatures of aqueous solutions containing significant contributions from glass can still be accurately estimated using NNMF integration.

Lastly, the temperature of a fluorescent sample was estimated. The spectra of an aqueous 0.25 nM R6G solution over a 60 °C temperature range are shown in Fig 8. As with the spectra containing glass signal, baseline subtraction was performed prior to NNMF integration. Also, as with the spectra containing glass signal, the NNMF derived factors for baseline subtracted aqueous R6G looked very similar to those from pure water, as shown in the inset in Fig 8a. Unlike the spectra shown in Fig 7a, the aqueous R6G spectra contained a much greater relative contribution from fluorescence, as can be seen in Fig 8a. Even so, the accuracy and precision of resulting temperature estimated are not greatly impacted relative to water, as shown in Fig 8b. Given the frequency at which fluorescent interference is encountered in Raman spectrometry, as discussed in Chapter 3, it is advantageous that Raman thermometry using NNMF integration does not suffer a loss of accuracy or precision in the presence of fluorescence.

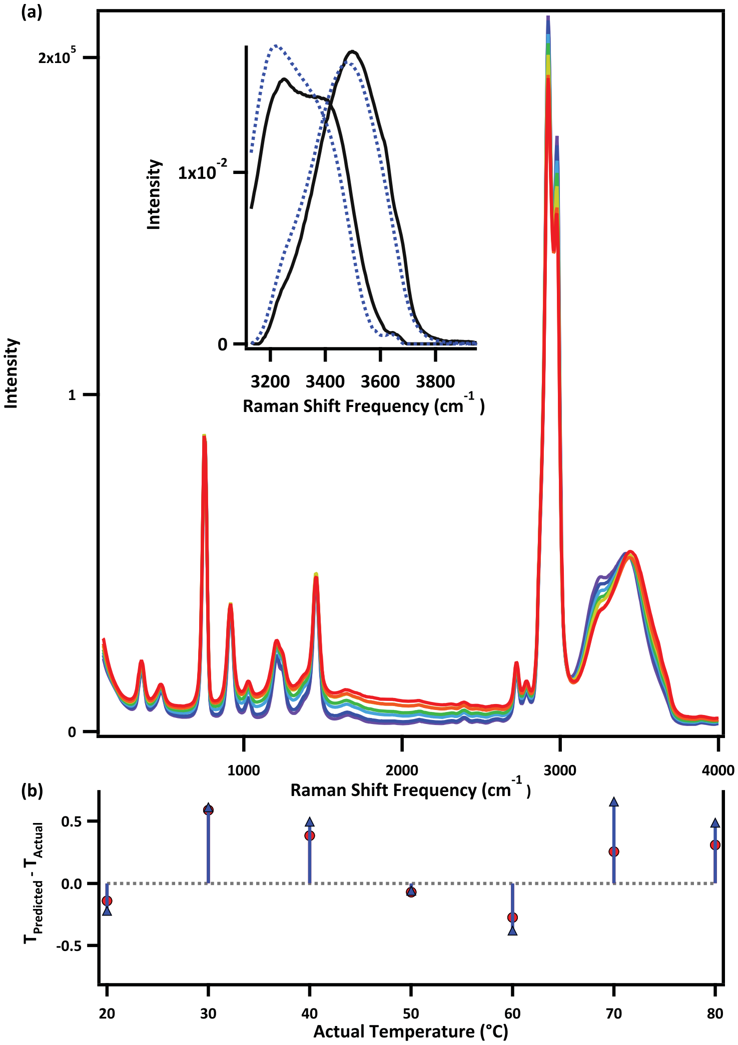


Figure 5.5. (a) Plot of the spectra of a 50 % volume-by-volume mixture of water and TBA taken at (purple) 20 °C, (blue) 30 °C, (cyan) 40 °C, (green) 50 °C, (yellow) 60 °C, (orange) 70 °C, and (red) 80 °C. The inset shows the difference between the NNMF derived factors for pure water (blue) and the water:TBA mixture (red) calculated over 3025–4073 cm^{-1}. (b) The difference between the measured temperature values and the estimated temperature values from each integration method at each temperature for 1 s integration (red circles) and 100 ms integration time (blue triangles). The mean difference between the estimated temperatures calculated at 1 s integration was 0.15 °C and at 100 ms integration time was 0.22 °C. The mean standard deviation of the estimated temperatures calculated at 1 s integration was 0.10 °C and at 100 ms integration time was 0.31 °C.

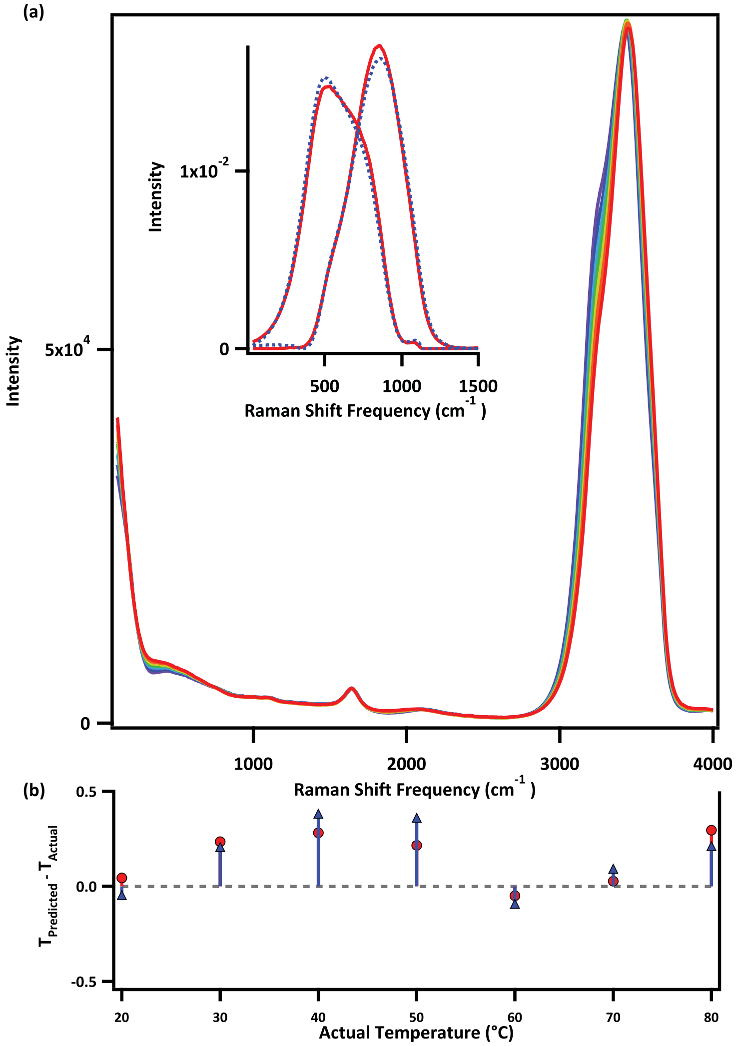


Figure 5.6. (a) Plot of the spectra of a 1 M solution of sodium chloride taken at (purple) 20 °C, (blue) 30 °C, (cyan) 40 °C, (green) 50 °C, (yellow) 60 °C, (orange) 70 °C, and (red) 80 °C. The inset shows the difference between the spectra of pure water (blue) and the NaCl solution (red) calculated over the same wavenumber region. (b) The difference between the measured temperature values and the estimated temperature values from each integration method at each temperature for 1 s integration (red circles) and 100 ms integration time (blue triangles). The mean difference between the estimated temperatures calculated at 1 s integration was 0.15 °C and at 100 ms integration time was 0.15 °C. The mean standard deviation of the estimated temperatures calculated at 1 s integration was 0.09 °C and at 100 ms integration time was 0.24 °C.

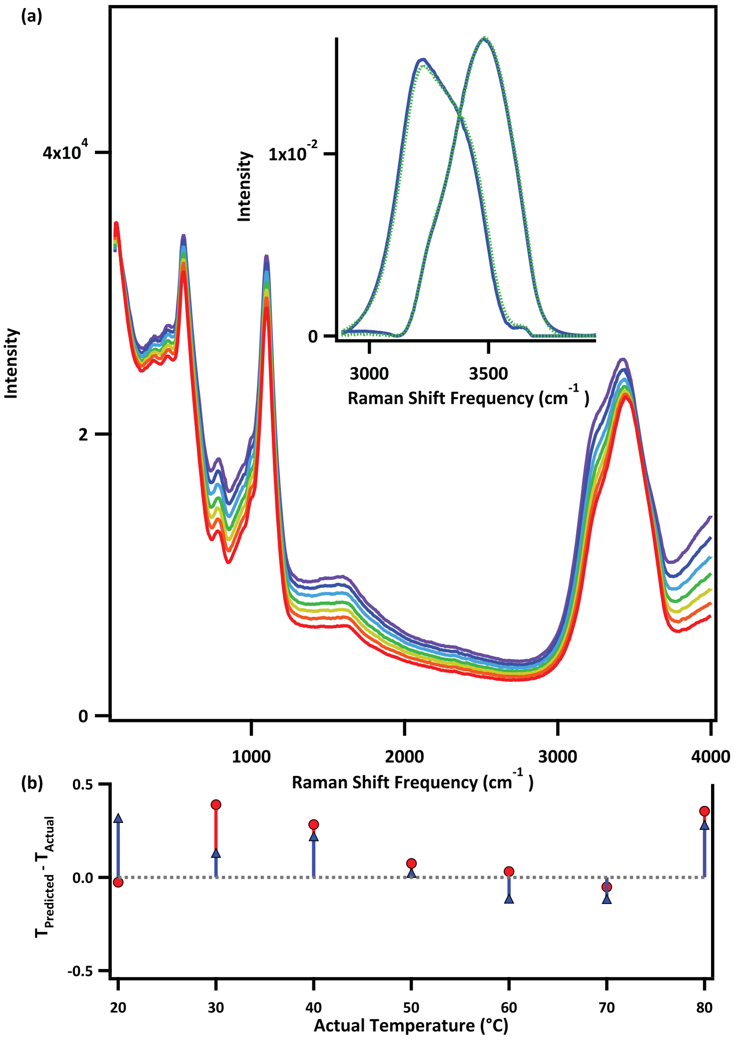


Figure 5.7. (a) Plot of the spectra of water with cuvette glass in the optically probed volume taken at (purple) 20 °C, (blue) 30 °C, (cyan) 40 °C, (green) 50 °C, (yellow) 60 °C, (orange) 70 °C, and (red) 80 °C. The inset shows the difference between the NNMF derived factors for pure water (blue) and the water with cuvette glass in the optically probed volume with a polynomial subtraction to remove the glass signal (hashed grey) calculated over the same wavenumber region. (b) The difference between the measured temperature values and the estimated temperature values from each integration method at each temperature for 1 s integration (red circles) and 100 ms integration time (blue triangles). The mean difference between the estimated temperatures calculated at 1 s integration was 0.15 °C and at 100 ms integration time was 0.10 °C. The mean standard deviation of the estimated temperatures calculated at 1 s integration was 0.22 °C and at 100 ms integration time was 0.69 °C.

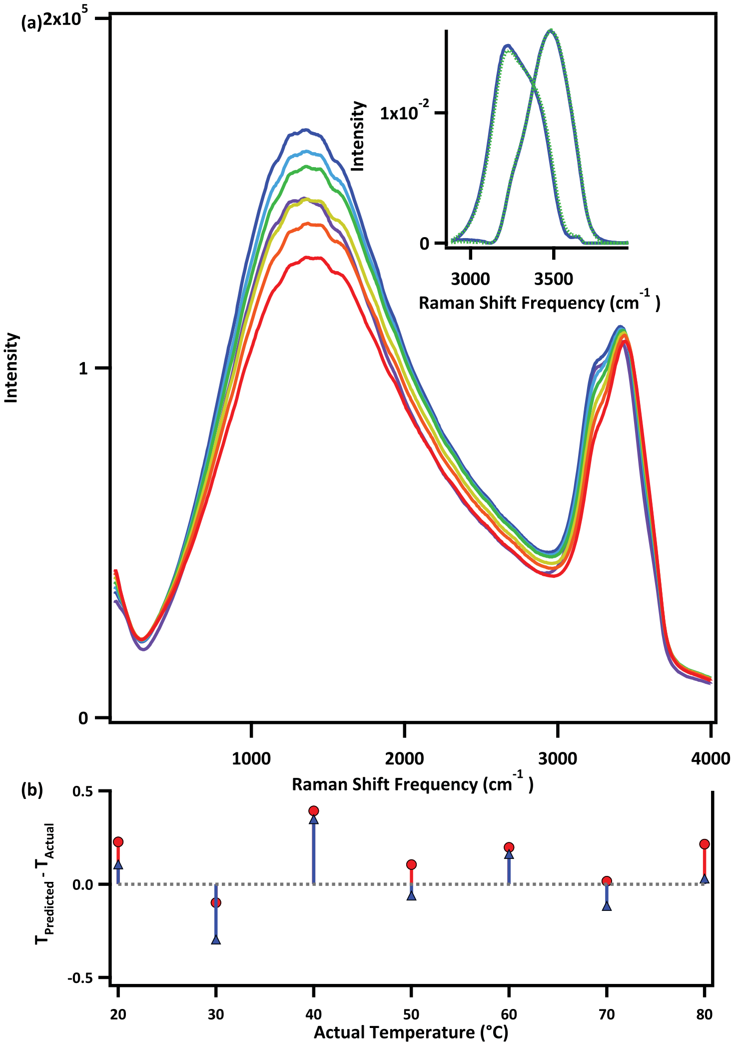


Figure 5.8. (a) Plot of the spectra of a 0.25 nM R6G solution taken at (purple) 20 °C, (blue) 30 °C, (cyan) 40 °C, (green) 50 °C, (yellow) 60 °C, (orange) 70 °C, and (red) 80 °C. The inset shows the difference between the NNMF derived factors for pure water (blue) and the 0.25 nM R6G solution with a polynomial subtraction to remove fluorescence (hashed green) calculated over the same wavenumber region. (b) The difference between the measured temperature values and the estimated temperature values from each integration method at each temperature for 1 s integration (red circles) and 100 ms integration time (blue triangles). The mean difference between the estimated temperatures calculated at 1 s integration was 0.15 °C and at 100 ms integration time was 0.02 °C. The mean standard deviation of the estimated temperatures calculated at 1 s integration was 0.17 °C and at 100 ms integration time was 0.37 °C.

5.5 Conclusion

A novel model for building a calibration curve for Raman thermometry applied to aqueous samples has been presented. Using NNMF integration and a three-term fit for the construction of a van't Hoff plot, rather than isosbestic integration and a linear fit, results in an reduction in measurement on the order of a factor of 10 to get the same level of precision. It should be noted that this improvement strictly comes from data processing; this method can be immediately applied to existing Raman thermometry hardware. This method has been shown to be appropriate for a variety of aqueous solutions, including aqueous alcohol mixtures, saline solutions, fluorescent samples, and water samples with glass features in their spectra.

It should be noted that there is interest in monitoring the temperature of water droplets [29, 40, 42], particularly under conditions that result in sub 273 K temperatures using Raman thermometry. These experiments measure the temperatures of water droplets falling in a vacuum while they are rapidly cooling. These experiment may benefit from the data processing methods described here, as these methods typically need to signal averaging over many droplets to estimate temperature. Future work may explore the application of this Raman thermometric data processing method on supercooled water droplets.

6. BINARY SPECTRAL HISTOGRAMS OF SMOOTHED, SECOND DERIVATIVE RAMAN SPECTRA ENABLE FAST SPECTRAL LIBRARY LOOK-UP

6.1 Abstract

Identifying a material by matching its Raman spectrum to a spectrum in a library of known spectra is one of the many applications of Raman spectroscopy. This application of spectroscopy, which will be referred to as "library look-up" henceforth, has recently been extended to previously inaccessible chemical problems as a result of the advent of low-cost, portable Raman spectrometers. These spectrometers are typically either hand-held or briefcase-sized and are controlled by low-cost microcomputers (with minimal computational power) or by cellular phones. As a result, calculating the spectral distance (typically expressed as a correlation coefficient) between a measured spectra and each library spectrum can take an undesirably long time (e.g., many seconds). Here I show a new method for performing Raman spectral library look-up that is optimized for computers with low computational power. This reduction in required computational resources is achieved by comparing spectra, appropriately preprocessed, that are then converted to binary spectral histograms. The Hamming distance between each level in the binary histogram of a measured spectrum and a library of binary spectral histograms is then computed. This process is order of magnitudes faster than calculating a correlation coefficient and thus is appropriate for performing library look-up on a portable Raman system. I demonstrate this by comparing the speed of calculating an R^2 to a Hamming distance for a library of Raman spectra collected using an InnoOptics portable Raman spectrometer.

6.2 Background

Identifying an unknown sample based upon its Raman spectrum, using a library of known Raman spectra, is a routine application of Raman spectroscopy in the pharmaceutical industry [23]. This work flow is also used by law enforcement to identify substances of interest, such as narcotics [43] and counterfeit products [44]. Both of these applications have recently adopted hand-held Raman instruments to enable analysis outside of an analytical laboratory. This has been enabled by the development of hand-held, or bench-top portable, Raman spectrometers. These portable systems are either controlled by an on-board microcomputer or by a laptop/cellular phone. In either case, the processing power of the computers controlling a portable Raman system is greatly diminished relative to a desktop computer. As a result, running moderately computationally expensive algorithms on these Raman systems can take longer than it would on a desktop computer. One such computation is the calculation of a correlation coefficient between a measured Raman spectrum, with several hundred measured wavelengths, and a library of a hundred, or more, previously measured Raman spectra. While on a desktop, this calculation may take seconds, or less, it may take tens of seconds on a cellular phone, which is an unacceptably long period of time for applications like forensics. Thus, computationally inexpensive alternatives to a correlation coefficient distance metric are desirable.

Several authors have demonstrated the use of dimension-reductions techniques like principal component analysis (PCA), partial-least squares (PLS), and other methods [43–45] for classification using portable Raman spectrometers. This is done to reduce spectra consisting of hundreds-to-thousands of wavelengths to a handful of dimensions. The reduced dimensional scores are then stored as a library and a measured spectrum's PCA/PLS scores are calculated using (precalculated) loadings. This is a very efficient way to handle the issue of computational costs, but has limited scalability. Specifically, it is impossible for users to add spectra to a library without rerunning PCA/PLS, something many end-users may be unfamiliar with or unwilling to spend the time required to do. Here I show an alternative strategy for reducing the computational cost of calculating the spectral distance between a measured spectra and spectra in a library. This technique is quite similar to a method described by Fang et. al [46], however differs in its application to Raman spectroscopy and in spectral preprocessing. It also shares similarities with the hyperspctral image processing methods using in the infrared geophysical image classification [47, 48], though the spectral preprocessing and histogram generation methods demonstrated here differ significantly from the cited methods.

6.3 Methods and Materials

The following section describes the Raman spectra and spectral processing methods used to perform fast Raman spectral library look-up.

6.3.1 Raman Spectra

A library of 157 Raman spectra were provided to the Ben-Amotz lab by Jason Jiang of InnoOptics. These spectra were collected at an InnoOptics site in China using a hand-held InnoOptics Raman spectrometer, with a 785 nm excitation wavelength and a CCD camera with 1024 wavelength channels. Unfortunately, other critical data acquisition parameters (e.g., integration time, laser power) as well as compound identity were not provided consistently by InnoOptics, though file names would indicate that the majority of these spectra were collected using 80 mW laser power and 200-5000 ms of integration time (varying from sample to sample). At the request of InnoOptics, I cannot make this library of spectra publicly available; however, four spectra from the library are shown in Fig 1a. Note that 2 pairs of spectra in the library were found to be identical and 5 spectra in the library had spectra that were entirely dominated by fluorescence. These spectra were omitted from the work

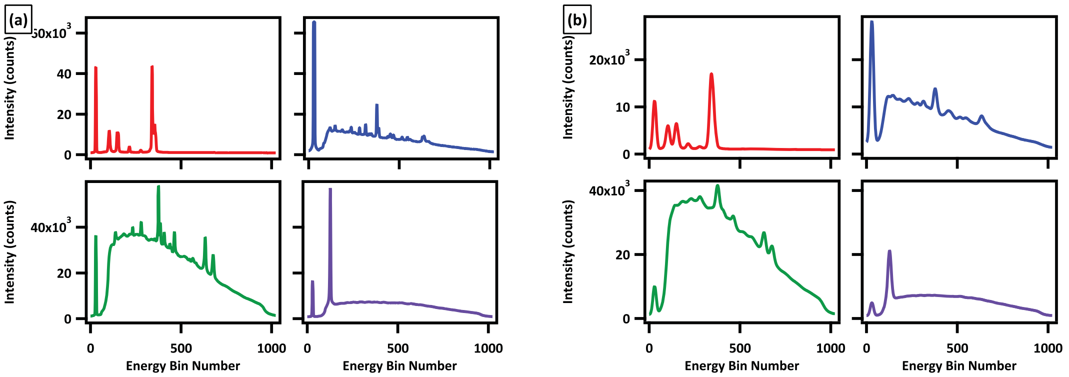


Figure 6.1. (a) Four spectra from the spectral library provided by InnoOptics. (b) The same spectra shown in (a) after binomial smoothing.

shown here as it was not possible to discriminate between them. Thus, there were 150 spectra in the library used in this work.

Future work in this area will consist of acquiring a library of approximately ∼100 Raman spectra, collected under known acquisition conditions.

6.3.2 Spectral Preprocessing

The spectral library described above was imported into Matlab (R2014a) for further manipulation. The spectra were smoothed using a 50 pixel width binomial smoothing filter with a width of 50. This was done to prevent high-frequency, baseline noise from affecting spectral distance measurements. The effects of smoothing are shown in Fig 1b. Next, the second-derivative of each smoothed spectrum was calculated in order to minimize the effect of broad fluorescence peaks in the Raman spectrum. Using smooth, second-derivative spectra for classification has previously been shown to be an effective when the Raman spectra contain fluorescence [16]. Examples of the spectra shown in Fig 1 processed in this manner are shown in Fig 2a.

After smoothing and taking the second-derivative of each spectra in the library, the spectra are normalized to area 1 over the range of energy bin 100 - energy bin 740

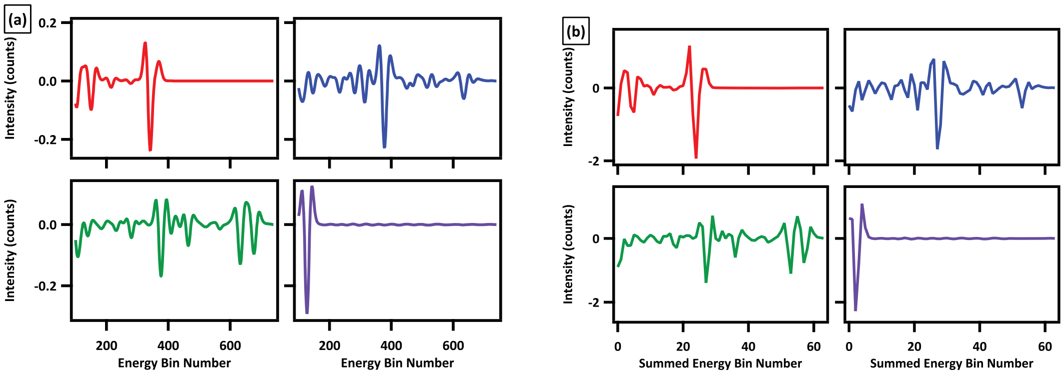


Figure 6.2. (a) Second-derivative of the spectra from Fig 1b truncated to energy bins 100–740 and normalized to unit area. (b) Spectra from (a) binned to have a total of 64 energy bins.

(unfortunately, a method to convert energy bins to wavenumbers were not provided by InnoOptics). The range was chosen as all spectra had the same peak below at ∼energy bin 30, with no characteristic peaks below energy bin 100. Likewise, above energy bin 740, there was almost no spectral information. After this, the normalized, smoothed, second-derivative spectra were binned into 64 bin spectra. Since the spectra after truncation to energy bins 100–740 consist of 640 energy bins, this can be thought of binning each 10 energy bins in the spectra. The effect of this can be seen in Fig 2b.

6.3.3 Binary Histograms

These 64-pixel spectra are then split into 2, 32-pixel spectra (so that 32 bit binary strings can be generated). These 32-pixel spectra are thresholded at 5 different levels (five was empirically determined to be an optimal number of thresholds). These thresholds were selected to be -0.2, -0.4, -0.6, -0.8, -1.0. At each threshold, a 32-bit binary word was generated, where each bit in the word was set to 0 if the pixel in the 32-bit word did not have signal below the threshold (since these are second-derivative spectra, more intensity corresponds to a more negative peaks) and set to 1 if the pixel had intensity below the threshold. This was repeated for both 32-pixel spectra, at each threshold, for each spectra in the library. The resulting 10, 32-bit binary strings generated for each spectra were used as a library of "known spectra". To validate this model, Hamming distance between each spectrum's 10, 32-bit binary strings and the library of binary strings were calculated. Hamming distance [49] between two binary strings, A and B, is defined as follows:

$$\mathrm{HammingDistance} = \frac{\mathrm{A} \wedge \mathrm{B}}{\mathrm{A} \vee \mathrm{B}}$$

where $\wedge$ is the "AND" operation and $\vee$ is the "OR" operation. This operation can be thought of as akin to a correlation coefficient for binary strings. The spectrum in the library that has the largest sum of Hamming distances, over all levels in the binary histogram, to a measured spectrum is most the spectra in the library most likely to correspond to the measured spectrum. In subsection 6.4.1, the accuracy of this approach will be demonstrated.

Despite the relatively low reduction in overall dimensionality ($32*10 = 320$ binary dimensions versus 1024 in the raw spectra), the Hamming distance is significantly easier to compute. This is due to the comparatively low computational cost of Boolean operations on digital numbers compared to algebraic operations using real numbers. This will be demonstrated in subsection 6.4.2 as a comparison in the computation time required to perform a correlation coefficient measurement to the binary look-up routine described above.

6.4 Results and Discussions

The following section describes the accuracy of the new library look-up algorithm under various condition as well as the comparison of computational costs for this method versus using a full-spectral correlation coefficient for library look-up

6.4.1 Accuracy of Binary Histogram Library Look-up

The accuracy of using the binary spectra above, with either a single cut-off or five cut-offs, for library look-up is discussed below. In addition to using the spectra as they were provided, Poisson (shot) noise was added to each spectra to simulate shorter integration times or having limited amount of sample to measure (for instance, find a single crystal of a compound in a forensic environment). This was done by dividing each spectra by some value, then adding Poisson noise at every pixel to the spectra divided by the constant. The effects of adding Poisson noise to the input spectra is shown in Fig 3.

Table 1 summarizes the results of this testing. The classification accuracy values reported for the test spectra with added Poisson noise are the means of 10 repeated classifications. As can be seen, evening reducing the signal by a factor of 100, the library look-up algorithm remains fairly accurate (90.9% classification accuracy). This compares to the use of a correlation coefficient on the smoothed, second-derivative, 64 energy bin spectra, which has a correlation coefficient of 66.7 % (though this includes positive and negative data unlike the binary histograms).

Table 6.1.
Classification Accuracy of Binary Histogram Library Look-up Accuracy

Single Cut-Off	Five Cut-Offs	10x Reduction	100x Reduction
100	100%	97.7%	90.9%

To further test robustness, the position of the Raman spectral features were shifted by ± 1 or ± 2 energy bins to test the robustness of this model under slight system misalignment. Being robust to small misalignments would be ideal for portable instrumentation used by non-expert users, as it might not be treated as gently as laboratory instrumentation commonly is. Additionally, an algorithm that is robust to slight misalignments minimized the need to include some sort of spectral standard (e.g., a Neon light source) for calibration in samples. Table 2 shows the effect of

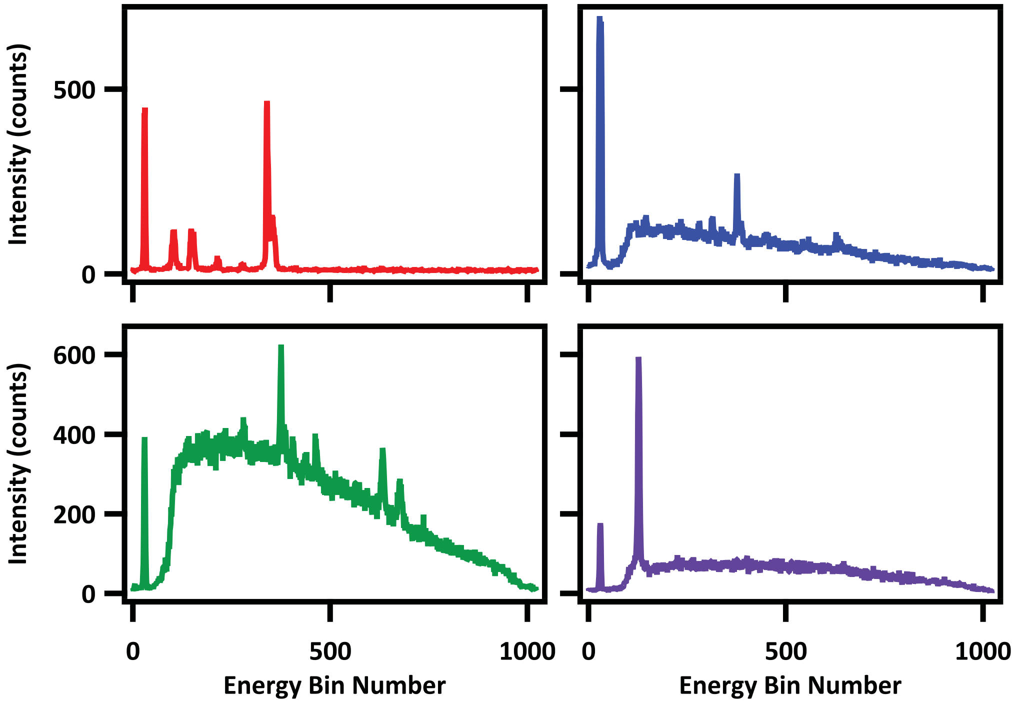


Figure 6.3. Spectra from Fig 1a, divided by 100, with added Poisson noise. Note that when compared to Fig 1a, the blue and green spectra suffer from the most loss of information,as they had the lowest signal-to-noise of the input spectra. The resulting error propagates to misclassification, as summarized in Table 1.

shifting spectral features by different amounts. Unfortunately, these results indicate that the algorithm is more robust to noise (or low-signal) than it is to energy bin drift. It would be desirable for instruments utilizing this algorithm to minimize the drift of the location of peaks on its CCD detector.

Table 6.2.
Classification Accuracy of Binary Histogram Algorithm with Peak-Shifted Input Spectra

+1 bin shift	-1 bin shift	+2 bin shift	-2 bin shift
76.7	82.0%	56.7%	66.0%

Lastly, I compared the effect of adding Poisson noise to trial spectra divided by a factor of 100 (as seen in Fig 4) and shifting the Raman spectral features by 1-2 energy bins (as seen in Table 2). The resulting classification accuracies, for 10 repeated classifications, are reported in Table 3. The classification accuracy under these conditions is poor, but it should be noted that the input spectra were significantly corrupted.

Table 6.3.
Classification Accuracy of Binary Histogram Algorithm with Peak-Shifted, Noisy (100x) Input Spectra

+1 bin shift	-1 bin shift	+2 bin shift	-2 bin shift
73.4	77.1%	55.1%	60.3%

6.4.2 Computational Efficiency

The binary histogram algorithm's computational cost was compared to that of the use a full spectral distance metric (correlation coefficient). The computational cost of performing full spectral library look-up, as well as library look-up using the binned, smoothed, secondary derivative spectra (BSSD) (as seen in Fig 3), are compared to the cost of library look-up using the binary histogram algorithm described above. All of this work was performed on a home-built computer running 64-bit Ubuntu 14.04, a 3.8 GHz AMD FX–4300 Quad-Processor, 8 GB RAM). All computations were performed in Matlab. Note that the time to run a single threshold level of the histogram (explicitly, for two of the 32-bit binary strings) is reported below, rather than the full, 5 level histogram. I have reported the computational costs in this manner as the number of "cut-offs" in the binary histogram algorithm is user-controllable (and, explicitly, scales linearly with the number of cut-offs). As the number of cut-offs increases, so does the robustness of the algorithm. However, users may wish to weigh the accuracy of their model to its computational costs.

A comparison of the average time taken to run each algorithm 100 times is reported in Table 4. As can be seen here, using the binary histogram library look-up method improves library look-up speed by an order of magnitude compared to the full spectral method and a factor of five compared to BSSD. It should be noted that the binary histograms with 5 cut-off levels shown in subsection 6.4.1 take approximately same time run as the BSSD spectra, despite the binary histogram having five times the dimensionality of the BSSD spectra.

Table 6.4.
Average Time to Compute a Spectral Distance for a Single Level in a Binary Histogram, Full Spectra, and Binned, Smoothed, Second-Derivative Spectra (BSSD)

Binary Spectra (ms)	Full Spectra (ms)	BSSD (ms)
0.25	1.9	1.2

It should be noted that because these algorithms were run in Matlab, the absolute computational costs may have added costs owing to the computational cost of running an interpretative language. Even so, it is clear from the results in Table 4 that, relative to using a full-spectral distance metric, the binary histogram algorithm is significantly more computationally efficient. This is true even for spectra processed using the same spectral preprocessing methods. In fact, the time it takes to determine the distance to a histogram with five cut-offs is equivalent to the time it takes to calculate the correlation coefficient of the BSSD spectra.

6.5 Conclusions

I have demonstrated a novel, computationally efficient method for classifying Raman spectra using a library of known-spectra. This method takes advantage of the comparatively low computational costs of bit counting binary strings relative to the computational costs of performing the algebraic operations necessary to compute a correlation coefficient. Spectral preprocessing is performed to minimize the contributions of (high frequency) noise and fluorescence. Even accounting for the dimensional reduction performed in the spectral preprocessing of the binary histogram algorithm, the algorithm demonstrated above is more than 5 times less computationally costly than a full-spectral library look-up method (as seen in Table 1).

This method has been adopted by InnoOptics in their “Narcotics Inspection APK” product, which is a commercial product containing the Raman spectra used for library look-up in the proceeding sections. According to Jason Jiang at InnoOptics, the binary histogram library look-up method runs in 233 ms on a Samsung T310 Tablet (Android 4.2.2 Jelly Bean operating system, 1.5GHz Dual Core processing).

REFERENCES

REFERENCES

[1] David S. Wilcox, Gregory T. Buzzard, Brad J. Lucier, Ping Wang, and Dor Ben-Amotz. Photon level chemical classification using digital compressive detection. *Analytica Chimica Acta*, 755:17–27, 2012.

[2] David S. Wilcox, Gregory T. Buzzard, Brad J. Lucier, Owen G. Rehrauer, Ping Wang, and Dor Ben-Amotz. Digital compressive chemical quantitation and hyperspectral imaging. *The Analyst*, 18657:865707–865717, 2013.

[3] Gregory T. Buzzard and Brad J. Lucier. Optimal filters for high-speed compressive detection in spectroscopy. *SPIE Proceedings*, 138:4982–4990, 2013.

[4] Owen G. Rehrauer, Bharat Mankani, Gregory T. Buzzard, Brad J. Lucier, and Dor Ben-Amotz. Fluorescence modeling for optimized-binary compressive detection raman spectroscopy. *Optics Express*, 23:423935–23951, 2015.

[5] James W. Chan, Douglas S. Taylor, Theodore Zwerdling, Stephen M. Lane, Ko Ihara, and Thomas Huser. Micro-raman spectroscopy detects individual neoplastic and normal hematopoietic cells. *Biophysical Journal*, 90:648–656, 2006.

[6] Annika M. K. Enejder, Tae-Woong Koo, Jeankun Oh, Martin Hunter, Slobodan Sasic, Michael S. Feld, and Gary L. Horowitz. Blood analysis by raman spectroscopy. *Optics Letters*, 27:2004–2006, 2002.

[7] Olusola O. Soyemi, Frederick G. Haiback, Paul J. Gemperline, and Michael L. Myrick. Nonlinear optimization algorithm for multivariate optical element design. *Applied Spectroscopy*, 56:477–487, 2002.

[8] Marco F. Duarte, Mark A.!Davenport, Dharmpal Tarkhar, Jason N. Laska, Ting Sun, Kevin F. Kelly, and Richard G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25:83–101, 1998.

[9] William C. Sweatt, Clinton A. A. Boye, Stephen M. Gentry, Michael R. Descour, Brian R. Stallard, and Carter L. Grotbeck. Isis: an information-efficient spectral imaging system. *SPIE Proceedings*, 3438:98–106, 1998.

[10] Natalilia Uzunbajakava, Peter de Peinder, Gert W. t Hooft, and Antonius T. M. van Gogh. Low-cost spectroscopy with a variable multivariate optical element. *Analytical Chemistry*, 78:7302–7308, 2006.

[11] Jospeh E. Vornehm, Ava Jingwen Dong, Robert W. Boyd, and Zhimin Shi. Multiple-output multivariate optical computing for spectrum recognition. *Optics Express*, 21:25005–25014, 2014.

[12] Chad A. Lieber and Anita Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological raman spectra. *Applied Spectroscopy*, 57:1363–1367, 2003.

[13] Jianhua Zhao, Harvey Lui, David I. McLean, , and Haishan Zeng. Automated autofluorescence background subtraction algorithm for biomedical raman spectroscopy. *Applied Spectroscopy*, 61:1225–1232, 2007.

[14] N. Everall, R. W. Jackson, J. Howard, and K. Hutchinson. Fluorescence rejection in Raman spectroscopy using a gated intensified diode array detector. *Journal of Raman Spectroscopy*, 17:415–423, 1986.

[15] N. J. Everall, J. P. Partanen, J. R.M. Barr, and M. J. Shaw. Threshold measurements of stimulated Raman scattering in gases using picosecond KrF laser pulses. *Optics Communications*, 64:393–397, 1987.

[16] Dongmao Zhang and Dor Ben-Amotz. Enhanced chemical classification of Raman images in the presence of strong fluorescence interference. *Applied Spectroscopy*, 54:1379–1383, 2000.

[17] D. V Martyshkin, R. C. Ahuja, A. Kudriavtsev, and S. B. Mirov. Effective suppression of fluorescence light in Raman measurements using ultrafasttime gated charge coupled device camera. *Review of Scientific Instrumentation*, 75:630–635, 2004.

[18] Rida T. Farouki. The bernstein polynomial basis: a centennial retrospective. *Computer Aided Geometric Design*, 29:379–419, 2012.

[19] Kurtulus Golcuk, Gurjit S. Mandair, Andrew F. Callender, Nadder Sahar, David H. Kohn, and Michael D. Morris. Is photobleaching necessary for raman imaging of bone tissue using a green laser? *Biochimica et Biophysica Acta - Biomembranes*, 1758:868–873, 2006.

[20] Sergei Bernstein. Demonstration du theoreme de weierstrass fondee sur le calcul des probabilities. *Communications de la Societ Mathmatique de Kharkoff*, 13:1–2, 1912.

[21] Claudio Frausto-Reyes, Cirilo Medina-Gutiérrez, R. Ysacc Sato-Berrú, and L. R. Sahagún. Qualitative study of ethanol content in tequilas by Raman spectroscopy and principal component analysis. *Spectrochimica Acta A*, 61:2657–2662, 2005.

[22] R. Ysacc Sato-Berrú, Jorge Medina-Valtierra, Cirilo Medina-Gutiérrez, and Claudio Frausto-Reyes. Quantitative NIR Raman analysis in liquid mixtures. *Spectrochimica Acta A*, 60:2225–2229, 2004.

[23] Derya Cebeci Maltas, Kaho Kwok, Ping Wang, Lynne S. Taylor, and Dor Ben-Amotz. Rapid classification of pharmaceutical ingredients with raman spectroscopy using compressive detection strategy with pls-da multivariate filters. *Journal of pharmaceutical and biomedical analysis*, 80:63–68, 2013.

[24] Friedrich Pukelsheim. *Optimal design of experiments,* , Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., 1993.

[25] F. J. Anscombe. The transform of poisson, binomial and negative-binomial data. *Biometrika*, 35:246–254, 1948.

[26] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *Proceedings of SPIE Electronic Imaging 2006*, 6064, 2006.

[27] Markku Mäkitalo and Alessandro Foi. Optimal inversion of the anscombe transform in low-count poisson image denoising. *IEEE Transactions on Image Processing*, 20, 2011.

[28] G. D. D'Arrigo, G. Maisano, F. Mallamace, P. Migliardo, and F. Wanderlingh. Raman scattering and structure of normal and supercooled water. *Journal of Chemical Physics*, 75:4264–4270, 1981.

[29] R. Vehring and G. Schweiger. Optical determination of the temperature of transparent microparticles. *Applied Spectroscopy*, 46:25–27, 1992.

[30] D. E. Hare and C. M. Sorensen. Raman spectroscopic study of bulk water supercooled to -33 degrees c. *Journal of Chemical Physics*, 93:25–33, 1990.

[31] V. Zhelyaskov, G Georgiev, and Z. Nickolov. Temperature study of intra- and inter-molecular coupling and fermi resonance constants in the raman spectra of liquid water using fourier deconvolution. *Journal of Raman Spectroscopy*, 19:405–412, 1988.

[32] G. E. Walrafen, M. S. Hokmabadi, and W. H. Yang. Raman isosbestic points from liquid water. *Journal of Chemical Physics*, 85:6964–6969, 1986.

[33] Kevin L. Davis, Kei-Lee K. Liu, Maureen Lanan, and Michael D. Morris. Spatially resolved temperature measurements in electrophoresis capillaries by raman thermometry. *Analytical Chemistry*, 65:293–298, 1993.

[34] Soo Ho Kim, Jermim Noh, Min Ku Jeon, Ki Woong Kim, Luke P. Lee, and Seong Ihl Woo. Micro-raman thermometry for measuring the temperature distribution inside the microchannel of a polymerase chain reaction chip. *Journal of Micromechanics and Microengineering*, 16:526–530, 2006.

[35] P. R.N. Childs, J. R. Greenwood, and C. A. Long. Review of temperature measurement. *Review of Scientific Instruments*, 71:2959–2978, 2000.

[36] Fred LaPlant, George Laurence, and Dor Ben-Amotz. Theoretical and experimental uncertainty in temperature measurement of materials by raman spectroscopy. *Applied Spectroscopy*, 50:1034–1038, 1996.

[37] Joel G. Davis, Kamil P. Gierszal, Ping Wang, and Dor Ben-Amotz. Water structural transformation at molecular hydrophobic interfaces. *Nature*, 491:582–585, 2012.

[38] Kamil P. Gierszal, Joel G. Davis, Mike D. Hands, David S. Wilcox, Lyudmila V. Slipchenko, and Dor Ben-Amotz. Hydrogen bonding in liquid water. *Journal of Physical Chemistry Letters*, 2:2930–2933, 2011.

[39] Jared D. Smith, Christopher D. Cappa, Kevin R. Wilson, Ronald C. Cohen, Phillip L. Geissler, Ronald C. Cohen, and Richard J. Saykally. Unified description of temperature-dependent hydrogen-bond rearrangements in liquid water. *Proceedings of the National Academy of Sciences*, 102:14171–14174, 2005.

[40] Jared D. Smith, Christopher D. Cappa, Walter S. Drisdell, Ronald C. Cohen, and Richard J. Saykally. Raman thermometry measurements of free evaporation from liquid water droplets. *Journal of American Chemical Society*, 128:12892–12898, 2006.

[41] Dor Ben-Amotz. Global thermodynamics of hydrophobic cavitation, dewetting, and hydration. *Journal of Chemical Physics*, 123:184504–1–184504–8, 2005.

[42] Kevin R. Wilson, Bruce S. Rude, Jared Smith, Chris Cappa, D. T. Co, R. D. Schaller, M. Larsson, T. Catalano, and R. J. Saykally. Investigation of volatile liquid surfaces by synchrotron x-ray spectroscopy of liquid microjets. *Review of Scientific Instruments*, 75:725–736, 2004.

[43] Kathryn Y. Noonan, Lindsey A. Tonge, Owen S. Fenton, David B. Damiano, and Kimberley A. Frederick. Rapid classification of simulated street drug mixtures using raman spectroscopy and principal component analysis. *Applied Spectroscopy*, 63:742–747, 2009.

[44] Mariana R. de Almeida, Deleon N. Correa, Werickson F. C. Rocha, Francisco J. O. Scafi, and Ronei J. Poppi. Discrimination between authentic and counterfeit banknotes using raman spectroscopy and pls-da with uncertainty estimation. *Microchemical Journal*, 109:170–177, 2013.

[45] Connie M. Gryniewicz-Ruzicka, Jason D. Rodriguez, Sergey Arzhantsev, Lucinda F. Buhse, and John F. Kauffman. Libraries, classifiers, and quantifiers: A comparison of chemometric methods for the analysis of raman spectra of contaminated pharmaceutical materials. *Journal of Pharmaceutical and Biomedical Analysis*, 61:191–198, 2012.

[46] Yu Fang, Yong Ma, Hao Li, Kun Liang, Shengqing Wang, and Hongyuan Wang. Rapid ultraspectral matching method based on crosscut feature extraction. *Optical Review*, 20:259–265, 2013.

[47] Xiuping Jia and J. A. Richards. Binary coding of imaging spectrometer data for fast spectral matching and classification. *Remote Sensing of Environment*, 43:47–53, 1993.

[48] Huan Xie and Xiaohua Tong. A probability-based improved binary encoding algorithm for classification of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7:2108–2118, 2014.

[49] Richard W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 29:147–160, 1950.

APPENDIX

A. APPENDIX

A.1 Appendix 1: Supplemental Figures from Chapter 3

The following section provides additional figures that not included, but referenced in, Chapter 3.

A.1.1 Training Spectra and Filters

Area-normalized spectra plotted against Raman shift frequency (wavenumber) shown here pertain to results obtained using a 785 nm and 514 nm excitation laser as noted above each subfigure. Note that the 785 nm spectra were measured with a resolution of 30 cm^{-1} and the 514 nm spectra were measured with a resolution of 12 cm^{-1}. Unless otherwise noted, the spectral component arising from the NIR objective and the Bernstein polynomials were always considered nuisance spectra.

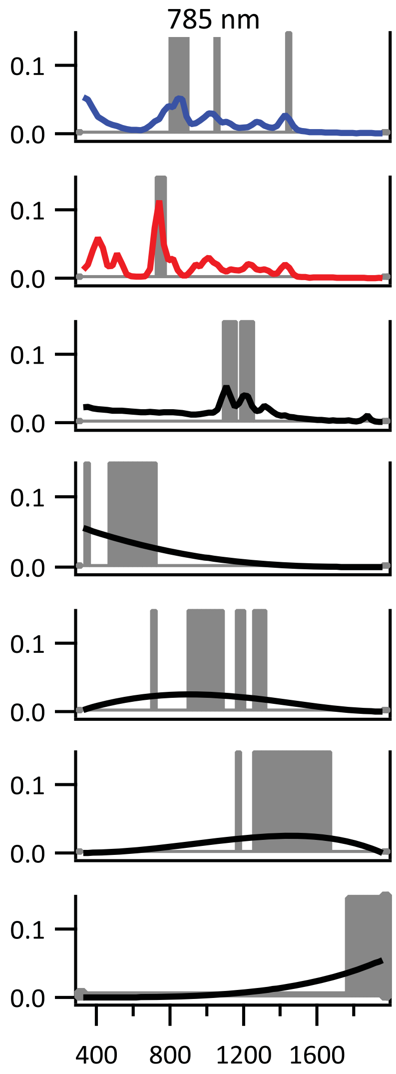


Figure A.1. Spectra and the resulting OB-CD filters for (in order from top down): n-hexane, methylcyclohexane, the spectral component arising from the NIR objective and the four degree-three Bernstein polynomials. The fraction of the total measurement time that each filter was collecting was 0.385, 0.219, 0.055, 0.034, 0.268, 0.031, and 0.009, respectively.

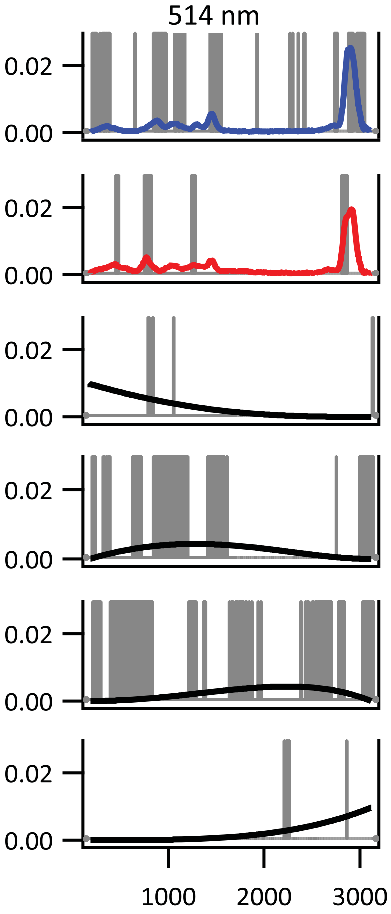


Figure A.2. Spectra and the resulting OB-CD filters for (in order from top down): n-hexane, methylcyclohexane, and the four degree-three Bernstein polynomials. The fraction of the total measurement time that each filter was collecting was 0.248, 0.388, 0.010, 0.032, 0.277 and 0.045, respectively.

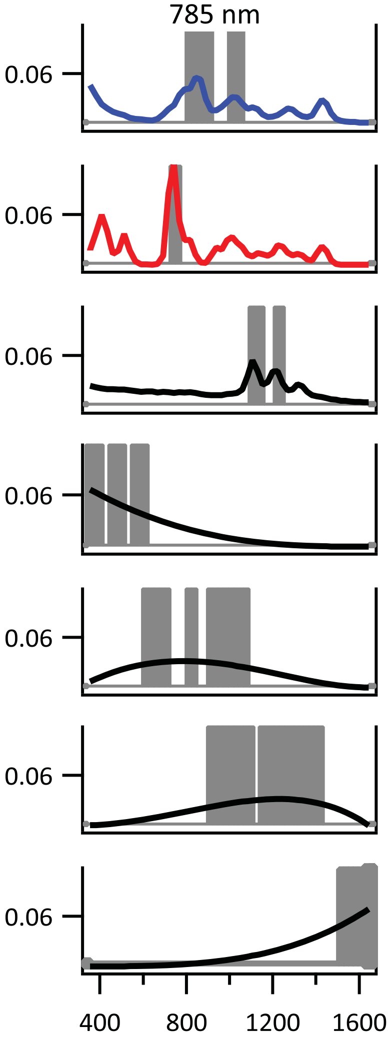


Figure A.3. Spectra and the resulting OB-CD filters for (in order from top down): aniline, toluene, spectral component arising from the NIR objective, and the four degree-three Bernstein polynomials. The fraction of the total measurement time that each filter was collecting was 0.484, 0.265, 0.025, 0.003, 0.154, 0.059 and 0.010, respectively.

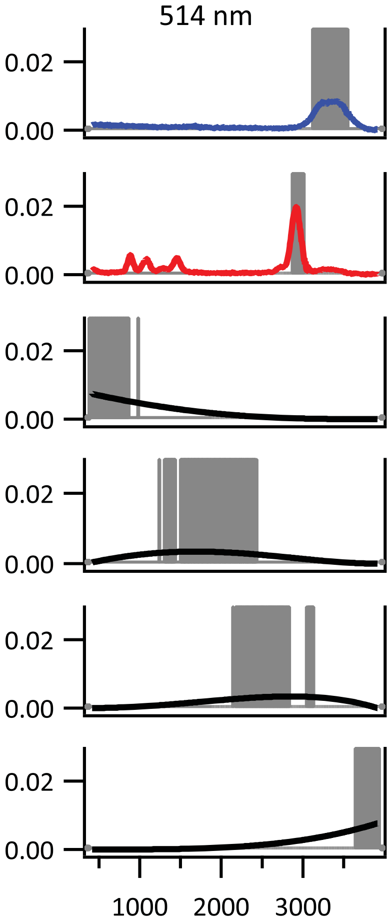


Figure A.4. Spectra and the resulting OB-CD filters for (in order from top down): ethanol, water, and the four degree-three Bernstein polynomials. The fraction of the total measurement time that each filter was collecting was 0.276, 0.261, 0.027, 0.116, 0.232, and 0.088 , respectively.

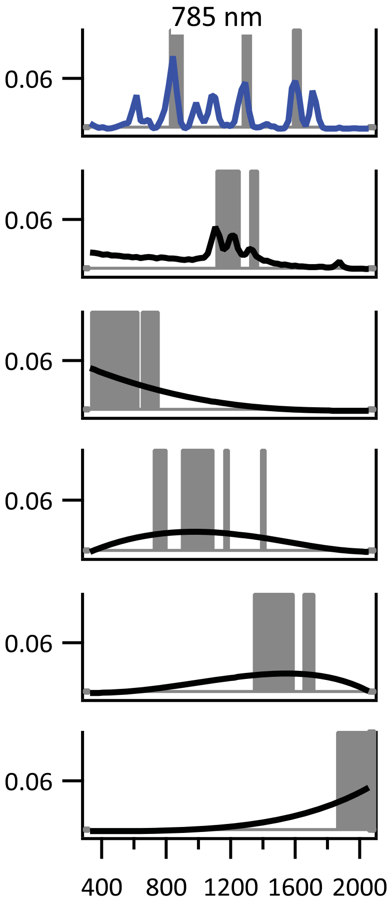


Figure A.5. Spectra and the resulting OB-CD filters for (in order from top down): Raman features of the plastic film, the spectral component arising from the NIR objective, and the four degree-three Bernstein polynomials. The fraction of the total measurement time that each filter was collecting was 0.099, 0.210, 0.099, 0.271, 0.212, and 0.109, respectively. No components were considered nuisance spectra, as I wanted to accurately estimate the intensity of the fluorescence before and after photobleaching.

A.2 Appendix 2: Binary-Complementary Compressive Filter Generation Code

The following section provides the Matlab code used to generate the binary-complementary filters described in Chapter 4. Comments here (normally denoted by the % symbol) are instead shown within parentheses in normal facing text, whereas code is shown in math (italics) text.

A.2.1 Binary-Complementary Optimal Filter Code

```
function [F, T, B, A, P] = Make_OBCD2_Filters(spectra1, spectra2)
(Step 1. User input: (1) spectra and (2) number of pairs of filters)
[p, pairs] = loadinput(spectra1, spectra2);

(Step 2: Optimal filters)
N = size(p, 1); (length of spectra)
n = size(p, 2); (number of spectra)

f = rand(pairs, N);
x = f * p;
y0 = reshape(x, 1, pairs * n);
f = @(y) - Comp(p, pairs, n, y);
X = [ ];
b = [ ];
Aeq = [ ];
beq = [ ];
lb = zeros(1, pairs * n);
ub = ones(1, pairs * n);
vec = fmincon(f, y0, X, b, Aeq, beq, lb, ub);
val = -f(vec);
[Ades, T, fdes] = FindAT(p, pairs, n, vec);
```

$Tdes = diag(T);$
$lambda = (ones(1,n)')/n;$
$D = diag(Tdes * Ades * lambda);$
$Bdes = inv(Ades' * Tdes * inv(D) * Tdes * Ades) * Ades' * Tdes * inv(D) * Tdes;$

$fdes_1 = fdes;$
$fdes_1(fdes_1 < 1) = 0;$
$fdes_1(fdes_1 > 1) = 1;$

(Writes the files into the folder)
$dlmwrite('P.txt', p,'\backslash t');$
$dlmwrite('B.txt', Bdes,'\ \backslash t');$
$dlmwrite('F.txt', fdes_1,'\backslash t');$
$dlmwrite('T.txt', Tdes,'\backslash t');$

(Convert intermediate variable names to final variable names)
$F = fdes_1;$
$P = p;$
$B = Bdes;$
$T = Tdes;$
$A = Ades;$

$lengthfdes = length(fdes_1);$
$nmirror = 2;$
$Q = fdes_1';$
$[nrows, ncol] = size(Q);$
$Q2 = zeros(nmirror * nrows, ncol);$
$j = 1;$
$for\ k = 1 : ncol;$

```
        for i = 1 : nrows;
            Q2(j : j + (nmirror - 1), k) = Q(i, k);
            j = j + nmirror;
        end
        j = 1;
    end
    Q2 = Q2';
    Q2 = fliplr(Q2);

     [n, m] = size(Q2);
    for ii = 1 : n;
        X = zeros(1, m);
        Y = zeros(1, m);
        X = Q2(ii, :);
        Y = X;
        for jj = 1 : 607;
            Y = cat(1, X, Y);
        end
        out = Y;
    imwrite(logical(out), strcat('F', num2str(ii), '.bmp'));
    end
    clear X Y n m i j k ii jj Q Q2 nrows ncol nmirror lengthF;

%%%%%%%%%%%%%%%%%%%%%%%
```

(note: The following code is written for n=2 spectra. Notes are included to modify for n≥2 spectra.)

function $[q,\ pairs]$ = loadinput$(spectra1, spectra2)$

$[m, \sim] = size(spectra1);$

$q = zeros(m, 2);$ (Change 2 to the number of spectra in the training set)
$q(:, 1) = norm_area(spectra1);$
$q(:, 2) = norm_area(spectra2);$
(Add additional spectra here)

$q(q \leq 0) = 0;$

$pairs = 1;$ (Modify to be n-1 pairs)

%%%%%%%%%%%%%%%%%%%%%%%%%

function val =Comp$(p, pairs, n, y)$

$N = size(p, 1);$
$x = reshape(y, pairs, n);$
$x1 = 1 - x;$
$x = [x; x1];$
$m = 2 * pairs;$
$r = zeros(1, m);$

$u = zeros(m, n);$
$A = zeros(m, n);$

(Compute A)
$for\ i = 1 : m$

```
    u(i,:) = x(i,:)/norm(x(i,:));
end
f = zeros(m, N);
for i = 1 : m
    [r(i)f(i,:)] = OptScale(p, u(i,:));
    A(i,:) = f(i,:) * p;
end

(Compute T)
T = ones(1, m)/m;
T = MaxT(A, T, p, m);

(Compute L)
D = zeros(m, m);
for i = 1 : m
    D(i, i) = T(i)/sum(A(i,:));
end

L = A' * D * A;
if(abs(det(L)) ≤ 1e − 20)
    val = log(abs(det(L)));
else
    val = −20;
end

%%%%%%%%%%%%%%%%%%%%%%%

function [Aout, Tout, fout] =FindAT(p, pairs, n, y)
```

```
N = size(p, 1);
x = reshape(y, pairs, n);
r = zeros(1, pairs);
x1 = 1 − x;
x = [x; x1];
m = 2 ∗ pairs;
u = zeros(pairs, n);
A = zeros(pairs, n);

fori = 1 : pairs
    u(i, :) = x(i, :)/norm(x(i, :));
end

f = zeros(pairs, N);
fdes = zeros(pairs, N);

for i = 1 : pairs
    [r(i)f(i, :)] = OptScale(p, u(i, :));
    fdes(i, :) = round(f(i, :));
    A(i, :) = fdes(i, :) ∗ p;
end
fadd = 1 − fdes;
Aadd = 1 − A;
fout = [fdes; fadd];
Aout = [A; Aadd];

T = ones(1, m)/m;
Tout = MaxT(Aout, T, p, m);
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%

function Tnew =Iter(T, A, p, m)

S = diag(T);
n = size(p, 2);
lambda = ones(1, n)′;
D = diag(S ∗ A ∗ lambda);
pairs = m/2;
Tnew = zeros(1, m);
if (abs(det(D)) ≤ 1e − 20)
    tem = A′ ∗ S ∗ inv(D) ∗ S ∗ A;
        if (abs(det(tem)) ≤ 1e − 5)
            B = inv(A′ ∗ S ∗ inv(D) ∗ S ∗ A) ∗ A′ ∗ S ∗ inv(D) ∗ S;
            Snew = A ∗ B;
            for i = 1 : pairs
                Tnew(i) = Snew(i, i) + Snew(pairs + i, pairs + i);
            end
            for i = (pairs + 1) : m
                Tnew(i) = Snew(i, i) + Snew(−pairs + i, −pairs + i);
            end
            Tnew = Tnew/sum(Tnew);
        else
            Tnew = T;
        end
else
    Tnew = T;
end
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%

function T =MaxT(A, T, p, m)

    for i = 1 : 100
    T = Iter(T, A, p, m);
end

%%%%%%%%%%%%%%%%%%%%%%%%%

function [r, f] =OptScale(p, u)

N = size(p, 1);
n = size(p, 2);

A = [];
b = [];
lb = zeros(1, N);
ub = ones(1, N);
C = null(u);

Aeq = C' * p';
beq = zeros(1, n - 1);

G = -u * p';
options = optimset('Display', 'none');

f = linprog(G, A, b, Aeq, beq, lb, ub, [], options);
r = -G * f;
```

A.3 Appendix 3: Binary Library Look-up Algorithm

The following section provides the Matlab code used to perform library look-up on binned smoothed, second-derivative Raman spectra that are then converted to binary histograms, as described in Chapter 6. Comments here (normally denoted by the % symbol) are instead shown within parentheses in normal facing text, whereas code is shown in math (*italics*) text.

A.3.1 Binary-Complementary Optimal Filter Code

```
function [CCarray] =BSSG_Lookup(TestSpectrum, cI, cP, library64, cutoff)

(cI is a measure of noise)
(cP is a shift amount, simulating would happen if peaks shifted in pixel location on the CCD.)
(TestSpectrum is, unsurprising, a spectrum to test against the library.)

As = 100; (First pixel in spec library to look-up. Use to remove low-information regions of spectra.)
Bs = 740; (Last pixel in spec library to look-up. Use to remove low-information regions of spectra.)
[m, n] = size(TestSpectrum);

if cI == 1
    Stest = TestSpectrum;
else
    Stest = (TestSpectrum/cI) + poissrnd(TestSpectrum/cI);
    Stest(Stest ≤ 0) = 0;
end
```

```
ifcP == 0
    Stest = Stest; elseif cP == -1
    Stest2(1,1) = Stest(1,1);
    for a = 2 : m
        Stest2(a,1) = Stest(a - 1,1);
    end
    Stest = Stest2;
elseif cP == 1
    Stest2(m,1) = Stest(m,1);
    for a = 1 : m - 1
        Stest2(a,1) = Stest(a + 1,1);
    end
    Stest = Stest2;
elseif cP == -2
    Stest2(1 : 2,1) = Stest(1 : 2,1);
    for a = 3 : m
        Stest2(a,1) = Stest(a - 2,1);
    end
    Stest = Stest2;
elseif cP == 2
    Stest2(m - 1 : m,1) = Stest(m - 1 : m,1);
    for a = 1 : m - 2
        Stest2(a,1) = Stest(a + 2,1);
    end
    Stest = Stest2;
end

Bin64 = MakeBinaryTestVector2(Stest,As,Bs,cutoff);
```

```
[CCarray] = Bin64CC2(Bin64, library64);

clear a

%%%%%%%%%%%%%%%%%%%%%%%%

function [BinaryWord] =MakeBinaryTestVector2(TestSpectrum, As, Bs, cutoff)

(Take a spectrum, mask from As to Bs, take a SGSD, split into two-halves (32 bits each), then make a binary word.)

(Step 1: Let's make TestSpectrum a SGSD spectrum)

[m, n] = size(TestSpectrum);
x = [1 : m]';

sgspec = Binomial_Smoothing(TestSpectrum, 50);
sgspec = Binomial_Smoothing(sgspec, 50);
sgspec = Binomial_Smoothing(sgspec, 50);

dy = diff(sgspec);
dx = diff(x);
dydx = dy./dx;
dx2 = (dx(1 : end - 1) + dx(2 : end))/2;
s2d = diff(dydx)./dx2;

s2d(1 : As, 1) = 0; (zeros masked values)
s2d(Bs : m, 1) = 0;
```

```
$s2d\_norm = DorsNorm(s2d, As, Bs);$
```

(Step 2: Convert the SGSD spectrum into a 64 pixel SGSD spectrum called s64d)

```
$count10 = 0;$
$count64 = 0;$
$s64d = zeros(64, 1);$

$while\ count64 \leq 64$
    $whilecount10 \leq 10$
        $s64d(count64+1,1) = s64d(count64+1,1) + s2d\_norm((As+10*count64+count10),1);$
        $count10 = count10 + 1;$
    $end$
    $count10 = 0;$
    $count64 = count64 + 1;$
$end$

$BinaryWord = BinaryVector(s64d, cutoff);$

%%%%%%%%%%%%%%%%%%%%%%

function $[bin\_filt]$ =Binomial_Smoothing($input\_spec, smooth\_fact$)
```

(Np is the smoothing parameter - assumed to be 20 to match Dor)

```
$[m, \sim] = size(input\_spec);$
$bin\_filt = zeros(m, 1);$
```

```
Smooth = diag(rot90(pascal((2 * smooth_fact) + 1)))';
Smooth_corrected = (Smooth/(4^smooth_fact));

for i = 1 : m
    if le(i, smooth_fact)
        bin_filt(i, 1) = input_spec(i, 1);
    elseif ge(i + smooth_fact, m)
        bin_filt(i, 1) = input_spec(i, 1);
    else
        bin_filt(i, 1) = (Smooth_corrected * input_spec(i - smooth_fact : i + smooth_fact, 1));
    end
end

%%%%%%%%%%%%%%%%%%%%%%%%%

function [normed_fxn] =DorsNorm(fxn, As, Bs)
```

(As and Bs are the range of rows to normalize over.)

```
[m, ~] = size(fxn);

product_fxn = zeros(m, 1);

for i = 1 : Bs
    product_fxn(i + As - 1, 1) = fxn(i + As - 1, 1) * fxn(i + As - 1, 1);
end
```

$Sum\_product\_fxn = sum(product\_fxn(1:m,1));$
$Sum\_product\_fxn = sqrt(Sum\_product\_fxn);$

$normed\_fxn = zeros(m,1);$

$for\ j = 1:Bs$
    $normed\_fxn(j + As - 1, 1) = fxn(j + As - 1, 1)/Sum\_product\_fxn;$
end

%%%%%%%%%%%%%%%%%%%%%%%%%

function $[binout]$ =BinaryVector($Test64spectrum2d, cutoff$)

(Converts a 64-bit spectrum into two 32 bit binary words)

(Test64spectrum2d is a 64 bin spectra, usually created in ”MakeBinaryTestVector”)

(cutoff is the cutoff for a 64 bin Savitzky-Golay smoothed second-derivative spectra (binned in a way described in ”MakeBinaryTestVector”)).

(word is for 32 bit systems. Who even uses those anymore? This is code that I will probably cut as I'm running this on 64-bit Linux. Dor originally coded this to run in Igor, which is limited to 32 bits.)

$wn = 1;$
$count64 = 0;$
$binout = 0;$

```
$whilecount64 < 64$
    $if\ Test64spectrum2d(wn + count64, 1) \leq cutoff$
        $binout = binout + 2\hat{}(count64);$
    $end$
    $count64 = count64 + 1;$
$end$
```

%%%%%%%%%%%%%%%%%%%%%%%%%

function $[CCarray]$ =Bin64CC2$(b64, library64)$

(Calculates distance measurements between binary words)

$[\sim, n] = size(library64);$ (n=number of spectra in library.)

$CCarray = zeros(1, n);$

```
$for\ i = 1 : n$
    $CCarray(1, i) = (bitand(b64, library64(1, i)))/(bitor(b64, library64(1, i)));$
$end$
```

VITA

VITA

Owen Rehrauer was born in Troy, MO on September 16th, 1982. He graduated from Nashoba Regional High School in Bolton, Massachusetts and proceeded to attend Framingham State College, where he graduated *cum laude* in May, 2005 with a B.S. in Food Science with minors in Biology and Physics. From 2005–2012, he worked as an analytical chemist at several corporations in the greater Boston area, including Waters Corporation, Ameridose LLC, and Cubist Pharmaceuticals. In 2012, he enrolled as a graduate student at Purdue University. He joined the Ben-Amotz lab and graduated with his PhD in Physical Chemistry in August, 2016.

PUBLICATION

Fluorescence modeling for optimized-binary compressive detection Raman spectroscopy

Owen G. Rehrauer,[1] **Bharat R. Mankani,**[1] **Gregery T. Buzzard,**[2]
Bradley J. Lucier,[3] **and Dor Ben-Amotz**[1,*]

[1]*Purdue University, Department of Chemistry, West Lafayette, IN, USA*
[2] *Purdue University, Department of Mathematics, West Lafayette, IN, USA*
[3]*Purdue University, Departments of Mathematics and Computer Science, West Lafayette, IN, USA*
[*]*bendor@purdue.edu*

Abstract: The recently-developed optimized binary compressive detection (OB-CD) strategy has been shown to be capable of using Raman spectral signatures to rapidly classify and quantify liquid samples and to image solid samples. Here we demonstrate that OB-CD can also be used to quantitatively separate Raman and fluorescence features, and thus facilitate Raman-based chemical analyses in the presence of fluorescence background. More specifically, we describe a general strategy for fitting and suppressing fluorescence background using OB-CD filters trained on third-degree Bernstein polynomials. We present results that demonstrate the utility of this strategy by comparing classification and quantitation results obtained from liquids and powdered mixtures, both with and without fluorescence. Our results demonstrate high-speed Raman-based quantitation in the presence of moderate fluorescence. Moreover, we show that this OB-CD based method is effective in suppressing fluorescence of variable shape, as well as fluorescence that changes during the measurement process, as a result of photobleaching.

© 2015 Optical Society of America

OCIS codes: (300.6450) Spectroscopy, Raman; (120.6200) Spectrometers and spectroscopic instrumentation; (070.6120) Spatial light modulators.

References and links

1. J. W. Chan, D. S. Taylor, T. Zwerdling, S. M. Lane, K. Ihara, and T. Huser, “Micro-Raman spectroscopy detects individual neoplastic and normal hematopoietic cells,” Biophys. J. **90**, 648–656 (2006).
2. A. M. K. Enejder, T.-W. Koo, J. Oh, M. Hunter, S. Sasic, M. S. Feld, and G. L. Horowitz, “Blood analysis by Raman spectroscopy,” Opt. Lett. **27**, 2004–2006 (2002).
3. O. O. Soyemi, F. G. Haiback, P. J. Gemperline, and M. L. Myrick, “Nonlinear optimization algorithm for multivariate optical element design,” Appl. Spectros. **56**, 477–487 (2002).
4. M. F. Duarte, M. A. Davenport, D. Tarkhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, “Single-pixel imaging via compressive sampling,” IEEE Signal Process. Mag. **25**, 83–101 (2008).
5. W. C. Sweatt, C. A. A. Boye, S. M. Gentry, M. R. Descour, B. R. Stallard, and C. L. Grotbeck, “ISIS: an information-efficient spectral imaging system,” Proc. SPIE **3438**, 98–106 (1998).
6. N. Uzunbajakava, P. de Peinder, G. W. t Hooft, and A. T. M. van Gogh, “Low-cost spectroscopy with a variable multivariate optical element,” Anal. Chem. **78**, 7302–7308 (2006).
7. J. E. Vornehm, A. J. Dong, R. W. Boyd, and Z. Shi, “Multiple-output multivariate optical computing for spectrum recognition,” Opt. Express **21**, 25005–25014 (2014).

8. D. S. Wilcox, G. T. Buzzard, B. J. Lucier, P. Wang, and D. Ben-Amotz, “Photon level chemical classification using digital compressive detection,” Anal. Chim. Acta **755**, 17–27 (2012).
9. D. S. Wilcox, G. T. Buzzard, B. J. Lucier, O. G. Rehrauer, P. Wang, and D. Ben-Amotz, “Digital compressive chemical quantitation and hyperspectral imaging,” Analyst **138**, 4982–4990 (2013).
10. G. T. Buzzard and B. J. Lucier, “Optimal filters for high-speed compressive detection in spectroscopy,” Proc. SPIE **8657**, 865707 (2013).
11. J. Zhao, H. Lui, D. I. McLean, and H. Zeng, “Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy,” Appl. Spectrosc. **61**, 1225–1232 (2007).
12. C. A. Lieber and A. Mahadevan-Jansen, “Automated method for subtraction of fluorescence from biological Raman spectra,” Appl. Spectrosc. **57**, 1363–1367 (2003).
13. D. V. Martyshkin, R. C. Ahuja, A. Kudriavtsev, and S. B. Mirov, “Effective suppression of fluorescence light in Raman measurements using ultrafasttime gated charge coupled device camera,” Rev. Sci. Instrum. **75** 630–635 (2004).
14. N. Everall, R. W. Jackson, J. Howard, and K. Hutchinson, “Fluorescence rejection in Raman spectroscopy using a gated intensified diode array detector,” J. Raman Spectrosc. **17** 415–423 (1986).
15. N. J. Everall, J. P. Partanen, J. R.M. Barr and M. J. Shaw, “Threshold measurements of stimulated Raman scattering in gases using picosecond KrF laser pulses,” Opt. Commun. **64** 393–397 (1987).
16. D. Zhang and D. Ben-Amotz “Enhanced chemical classification of Raman images in the presence of strong fluorescence interference,” Appl. Spectrosc. **54** 1379–1383 (2000).
17. F. Pukelsheim, *Optimal design of experiments*, (John Wiley & Sons Inc. 1993).
18. A. C. Atkinson, A. N. Donev, and R. D. Tobias, *Optimum experimental designs with SAS*, (Oxford University Press, 2007).
19. K. Golcuk, G. S. Mandair, A. F. Callender, N. Sahar and D. H. Kohn and M. D. Morris, “Is photobleaching necessary for Raman imaging of bone tissue using a green laser?” BBA - Biomembranes **1758**, 868–873 (2006).
20. S. Bernstein, “Demonstration du theoreme de weierstrass fondee sur le calcul des probabilities,” Comm. Soc. Math. Kharkov **13**, 1–2 (1912).
21. R. T. Farouki, “The Bernstein polynomial basis: a centennial retrospective,” Comput. Aided Geom. Design **29**, 379–419 (2012).
22. C. Frausto-Reyes, C. Medina-Gutiérrez, R. Sato-Berrú, and L. R. Sahagún, “Qualitative study of ethanol content in tequilas by Raman spectroscopy and principal component analysis,” Spectrochim. Acta A **61**, 2657–2662 (2005).
23. R. Sato-Berrú, J. Medina-Valtierra, C. Medina-Gutiérrez, and C. Frausto-Reyes, “Quantitative NIR Raman analysis in liquid mixtures,” Spectrochim. Acta A **60**, 2225–2229 (2004).

---

1. Introduction

Current spectroscopic chemical analysis instruments are capable of generating data sets that are so large that they require transforming the data to a lower-dimensional space, using methods such as principal component analysis (PCA) [1] or partial-least squares (PLS) [2] to facilitate subsequent automated chemical classification and quantitation. Moreover, because the collection of high-dimensional data is often the slowest step in the process, a number of compressive detection strategies [3–7] have been introduced with the goal of increasing data collection speed by making measurements only in the low-dimensional space containing the information of interest. One such method is our previously described optimized binary compressive detection (OB-CD) strategy, in which OB filters are applied to a digital mirror microarray (DMD) to redirect or collect photons of (multiple) selected colors, for detection using a single channel detector, such as a photon counting photomultiplier tube (PMT) or an avalanche photodiode (APD) [8, 9].

Previously [8–10], we demonstrated that the OB-CD strategy enabled high-speed chemical classification, quantitation, and imaging. Here we demonstrate an extension of the OB-CD method that facilitates Raman classification and quantitation in the presence of fluorescence background. Key advantages of this method, relative to conventional fluorescence subtraction strategies [11–16], include its compatibility with automated high-speed chemical analysis in the presence of variable fluorescence backgrounds. Here we show that fluorescence backgrounds may be quantified and subtracted on-the-fly by including Bernstein polynomial spectral functions in the OB-CD training set, along with Raman spectra of the components of interest. In other words, we augment the Raman spectral training set with Bernstein polynomial spectral shapes to model fluorescence and thus obtain OB-CD filters that are used to quantify both the chemical components of interest and the fluorescence background spectra.

2. Theory

We begin by giving a short summary of the OB-CD mathematical framework [8]; some of the wording is taken from the cited paper and we refer the reader there for more details and discussion, including our assumption of linear additive spectra, which justifies in part the form of our model. This framework is based on the subfield of statistics known as “Optimal Design of Experiments” [17]. We then discuss the notion of *nuisance parameters* and how this notion can be exploited to reduce the variance of the estimated photon emission rates pertaining to the (non-nuisance) components of interest. Finally, we introduce a new strategy for modeling, quantifying, and suppressing fluorescence background signals in Raman spectra.

2.1. Review of mathematical model

We assume that our chemical sample consists of a mixture of n known chemical species S_j, $j = 1, \dots, n$. In a particular sample, the species S_j emits photons at a rate Λ_j, so the number of photons emitted in time t is a Poisson random variable with mean $t\Lambda_j$; our goal is to estimate these rates as accurately as possible, in order to quantify the composition of a sample containing such components.

The wavelength, or energy, of each photon observed in the experiments can be labeled with an integer $i \in \{1, \dots, N\}$. Assume that we know the shape of the spectrum associated with species S_j; denote the probability that a photon from species S_j has label i by P_{ij}, so $\sum_{i=1}^{N} P_{ij} = 1$. In other words, the P_{ij}, $i = 1, \dots, N$, form the spectrum of the jth compound, normalized so that the sum is 1. Thus the stream of labeled photons emanating from a sample is modeled by a vector Poisson process with rates $P\Lambda$, where $\Lambda = (\Lambda_1, \dots, \Lambda_n)^T$, and $P = (P_{ij})_{N\times n}$. (Here and later, superscript T denotes “transpose.”) If we measure the number of photons that arrive in each energy bin for time t then the number of photons with label i entering our instrument from all chemical species has a Poisson distribution with mean $t(P\Lambda)_i = t\sum_{j=1}^{n} P_{ij}\Lambda_j$. We assume that the number of wavelength channels, N, is greater than the number of chemical species n, and that the columns of P are linearly independent, i.e., P has full rank. (In other words, we assume that no spectrum can be written as a linear combination of the other spectra.)

We consider taking m independent measurements with $m \geq n$. In the kth measurement, we set in our *optical filter* the transmittance of all photons with energy level i to be a number F_{ik} with $0 \leq F_{ik} \leq 1$; i.e., the probability that in the kth measurement a photon with energy label i is counted is F_{ik}. Our observation in the kth measurement is the total photon count, summed over all energy levels i, from observing the photon stream for time T_{kk}, which will be a Poisson random variable with mean

$$T_{kk}\sum_{i=1}^{N} F_{ik}\left(\sum_{j=1}^{n} P_{ij}\Lambda_j\right) = T_{kk}\sum_{i=1}^{N}\sum_{j=1}^{n} F_{ik}P_{ij}\Lambda_j.$$

(We use a double subscript on T_{kk} because we will make these numbers the diagonal of a matrix T.) We refer to the columns of the matrix $F = (F_{ik})_{N\times m}$ as *filters*, and the entries of F can be chosen as we wish, since they are parameters of our measuring device. Based on [17] and as in [8], we choose filters F and corresponding measurement times T to minimize the variance in estimated rates, subject to a total time constraint expressed as $\sum_{k=1}^{m} T_{kk} = 1$.

For a DMD, we can choose only $F_{ik} = 0$ or $F_{ik} = 1$, while for an analog spatial light modulator (SLM) we can in principle choose any $0 \leq F_{ik} \leq 1$. (Moreover, one could use a DMD as an analogue filter by varying time for which each individual energy bin in a given filter is turned on, although we have not done so in the present studies.)

We denote by $\mathbf{x}$ our complete observation, a vector of m independent Poisson random variables with means and variances given by the vector $TF^TP\Lambda$, where T is the $m \times m$ diagonal matrix with diagonal entries $T_{11}, \ldots, T_{mm}$ and F^TP is an $m \times n$ matrix. We assume that F is chosen so that F^TP has rank n (which is possible since P has rank n and $m \geq n$).

If we denote by $\hat{\mathbf{x}}$ a sample from this random variable, then our estimate $\hat{\Lambda}$ of the true rates Λ is given by $\hat{\Lambda} = BT^{-1}\hat{\mathbf{x}}$, where the $n \times m$ matrix $B = (b_{ik})_{n\times m}$ is a *left inverse* of F^TP, i.e., $B(F^TP) = I$, the $n \times n$ identity matrix. (If $n = m$ then F^TP is a square matrix and B is simply $(F^TP)^{-1}$.) We note that the expected value of $\hat{\Lambda}$ satisfies $E(\hat{\Lambda}) = BT^{-1}E(\hat{\mathbf{x}}) = BT^{-1}(TF^TP\Lambda) = \Lambda$, so $\hat{\Lambda}$ is an unbiased estimator of Λ.

It was shown earlier [8] that the variance of the estimate $\hat{\Lambda}_j$ of the jth rate Λ_j is given by

$$E\left(|(BT^{-1}\mathbf{x})_j - \Lambda_j|^2\right) = \sum_{k=1}^{m} b_{jk}^2 T_{kk}^{-1}(F^TP\Lambda)_k.$$

In the cited paper the variances were summed over *all* Λ_j to derive

$$E\left(\|BT^{-1}\mathbf{x} - \Lambda\|^2\right) = \sum_{j=1}^{n}\sum_{k=1}^{m} b_{jk}^2 T_{kk}^{-1}(F^TP\Lambda)_k = \sum_{k=1}^{m} \frac{((F^TP)\Lambda)_k}{T_{kk}}\|B\mathbf{e}_k\|^2. \tag{1}$$

Here $\mathbf{e}_k$ is a vector whose components are zero except for a 1 in the kth component and $\|y\|^2 = \sum_{j=1}^{n} y_j^2$.

We now deviate a bit from the earlier exposition [8], where it is perhaps not stated clearly that because the sum of variances (1) depends on the unknown rates Λ, we cannot choose m, F, T, and B to minimize (1) for *all* Λ simultaneously. Therefore we pick a single $\overline{\Lambda} = (1, \ldots, 1)^T$ and choose m, F, T, and B to minimize

$$E\left(\|BT^{-1}\mathbf{x} - \overline{\Lambda}\|^2\right) = \sum_{k=1}^{m} \frac{((F^TP)\overline{\Lambda})_k}{T_{kk}}\|B\mathbf{e}_k\|^2 \tag{2}$$

under the constraints $m \geq n$; $0 \leq F_{ik} \leq 1$, $1 \leq i \leq n$, $1 \leq k \leq m$; $0 < T_{kk}$, $1 \leq k \leq m$, and $\sum_{k=1}^{m} T_{kk} = 1$; and $B(F^TP) = I$. This choice of $\overline{\Lambda}$ has proved useful absent additional information [8]. Our experience indicates that the resulting filters and measurement times are relatively insensitive to changing individual coefficients in $\overline{\Lambda}$ by up to a factor of 100.

2.2. *Nuisance parameters*

It often happens that the photon emission rates of some chemical species are of more interest than others. One might have a contaminant with a broad (known) spectrum; while this may be one of the chemical species S_j, we don't really care about the accuracy of estimation of that particular Λ_j. In the field of Optimal Design of Experiments, such variables Λ_j are known as *nuisance parameters*: They are a necessary part of the model, but we don't care about the accuracy of their estimates except insofar as it affects the estimates of the other variables [18].

We assume that we are truly interested in the photon emission rates of only the first $n' < n$ chemical species. For such systems, we don't care about the variances of the estimates of Λ_j for $n' < j \leq n$ and so instead of minimizing (2) we minimize the sum of the variances of of the estimates of only the first n' emission rates:

$$\sum_{j=1}^{n'} E\left(|(BT^{-1}\mathbf{x})_j - \overline{\Lambda}_j|^2\right) = \sum_{j=1}^{n'}\sum_{k=1}^{m} b_{jk}^2 T_{kk}^{-1}(F^TP\overline{\Lambda})_k.$$

In effect, we estimate Λ_j for $n' < j \leq n$ only well enough to minimize the error of the sum of variances of the first n' emission rates.

2.3. Estimating fluorescence using Bernstein polynomials

A fluorescent spectrum is generally smooth and broad, in contrast to the narrow peaks found in Raman spectra; it may vary from one sample to another, or from time to time for the same sample.

One can reduce the amount of fluorescence in a sample by photobleaching the sample prior to analysis [19]. Another commonly-used procedures for subtracting such fluorescence backgrounds from Raman spectra is to fit the fluorescent spectrum to a polynomial [11,12]; in effect, this amounts to using a fixed polynomial of specified degree to model a fluorescent spectrum. This may work well if in a set of samples the fluorescence doesn't vary over space or time; this is often not the case, however, in Raman (and particularly Raman imaging) applications, for which one must estimate the shape of the fluorescence dynamically.

Instead of somehow applying conventional fluorescence fitting and subtraction strategies to OB-CD measurements, we use OB-CD filters that are derived using either actual fluorescence spectra of known shape or a family of polynomials that models general fluorescence spectra.

More specifically, if the sample of interest contains a fixed fluorescent signal of known shape, we do the following. We simply add the fluorescence spectrum as an extra column of P with associated rate variable Λ_j, and then treat Λ_j as a nuisance parameter, not adding its variance to the sum of variances being minimized.

If the fluorescence has a spectrum that varies over space or time, however, we cannot apply the previous procedure and must model the fluorescence dynamically, as follows.

Since fluorescence backgrounds can often be fit reasonably well by a cubic polynomial, we would like to identify a polynomial basis for cubic polynomials such that (1) all basis elements are nonnegative on $[0,1]$ and (2) every nonnegative polynomial can be written as a linear combination of the basis elements with nonnegative coefficients. (Negative coefficients, which are supposed to model rates, are nonphysical and increase the variance of our estimates.) Unfortunately, for $r > 1$, no such basis exists.

Of interest, however, is the *Bernstein basis* [20,21] of polynomials of degree r, given by

$$B_{\nu,r}(x) = \binom{r}{\nu} x^{\nu}(1-x)^{r-\nu}, \quad \nu = 0,1,\ldots,r,$$

are nonnegative on $[0,1]$ (so any linear combination of them with nonnegative coefficients is nonnegative on $[0,1]$). They have other nice properties: They form a basis for the space of polynomials of degree r; they have optimal stability is some sense [21]; and "many" if not "most" nonnegative polynomials that come up in practice as models for fluorescent spectra have nonnegative coefficients in the Bernstein basis (or at least, any negative coefficients are not "overly" large). The Bernstein polynomials even resemble single-peak spectra.

In the specific case $r = 3$ we have $B_{0,3}(x) = (1-x)^3$, $B_{1,3}(x) = 3x(1-x)^2$, $B_{2,3}(x) = 3x^2(1-x)$, $B_{3,3}(x) = x^3$. These polynomials are shown in Fig. 1.

In this work we model fluorescent spectra with linear combinations of $B_{\nu,3}(x)$, $\nu = 0,1,2,3$, and often treat the Bernstein coefficients as nuisance parameters. Because fluorescent spectra are not precisely polynomials, this introduces *model error*, which we discuss in the Results and Discussion section. We have also found that for some purposes using quartic polynomials ($r = 4$) gives more accurate results.

3. Methods and materials

3.1. OB-CD spectrometers

We have previously described an OB-CD spectrometer using a 785 nm laser for excitation with 75 mW at the sample [8]. In the present studies we have utilized that spectrometer as well as

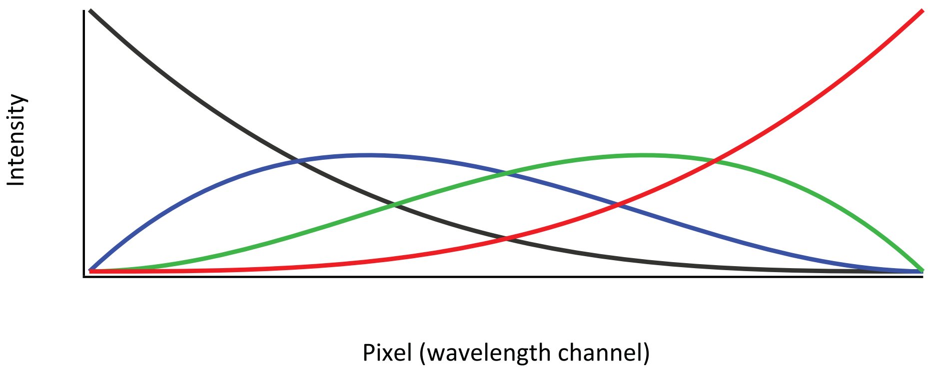


Fig. 1. Plot of the four degree-three Bernstein polynomials as a function of wavelength channel (scaled to be over the interval $[0, 1]$.) The colors denote the various polynomials: black, $B_{3,0}(x)$; blue, $B_{3,1}(x)$; green, $B_{3,2}(x)$; and red, $B_{3,3}(x)$.

a new OB-CD spectrometer using a 514 nm laser for excitation whose schematic is shown in Fig. 2 (the 785 nm excitation laser system is similar in design, as previously described in [8]). Much like the previously described system, our microscope is configured to collect the back-scattered Raman signal with the same objective lens that is used to focus the argon ion laser (Modu Laser Stellar Pro L 100 mW) onto the sample. The laser passes through a laser-line bandpass filer (Semrock RazorEdge) before it is focused onto the sample using a microscope objective (Nikon MPlan, 20x, 0.4 NA), and unless indicated otherwise the laser power at the sample was about 12 mW for all experiments described in this paper. The backscattered light is collected and then separated from the laser Rayleigh scattering using a dichroic mirror (Semrock RazorEdge). Then, the Raman scattered light is sent to the spectrometer (right portion of Fig. 2), where it is filtered first using a long pass (edge) filter (Semrock RazorEdge), followed by passing through a volume holographic grating (Wasatch Photonics, $\sim$1000 lines mm^{-1}). This light is then dispersed onto the DMD (Texas Instruments, DLP3000, 608 $\times$ 684 mirror array with 10.8 μm mirror pitch). The spectral window in this system is $\sim$200–4100 cm^{-1}. For all data collected in this paper, we binned two columns of adjacent DMD mirrors together, yielding a total of 342 “bins” with each energy bin corresponding to $\sim$12 cm^{-1}. Light from the DMD is then focused onto a photon-counting photomultiplier tube (PMT) (Hamamatsu model #H10682-01) with a dark count rate of $\sim$500 photons s^{-1}. TTL pulses from the PMT are counted using a USB data acquisition (DAQ) card (National Instruments, USB-6212BNC). The system is controlled with interface software written in Labview 2013. Binary filter generation is performed as previously described using Matlab (Matlab 7.13 R2011b) [8]. Data was in some cases further processed and manipulated using Igor Pro 6.04.

3.2. Chemicals used in classification/quantitation

Acetone and benzene were purchased from Macron (batch #0000070736) and OmniSolv (lot #42282), respectively. Hexanes were acquired from Baxter (lot #901141). Methylcyclohexane was acquired from Mallinckroft (lot #1906 KCBN). Aniline and toluene were acquired from Aldrich (batch #05925CB) and Mallinckroft (lot #8608 X14752), respectively. Aniline was purified via distillation by heating aniline to 190° C in a round bottom flask, which was connected to a chilled condenser. Ethanol was acquired from Koptec (200 Proof, lot #225411) and water was ultrapurified in our lab (Milli-Q UF Plus, 18.2 mΩ cm, Millipore). The overhead

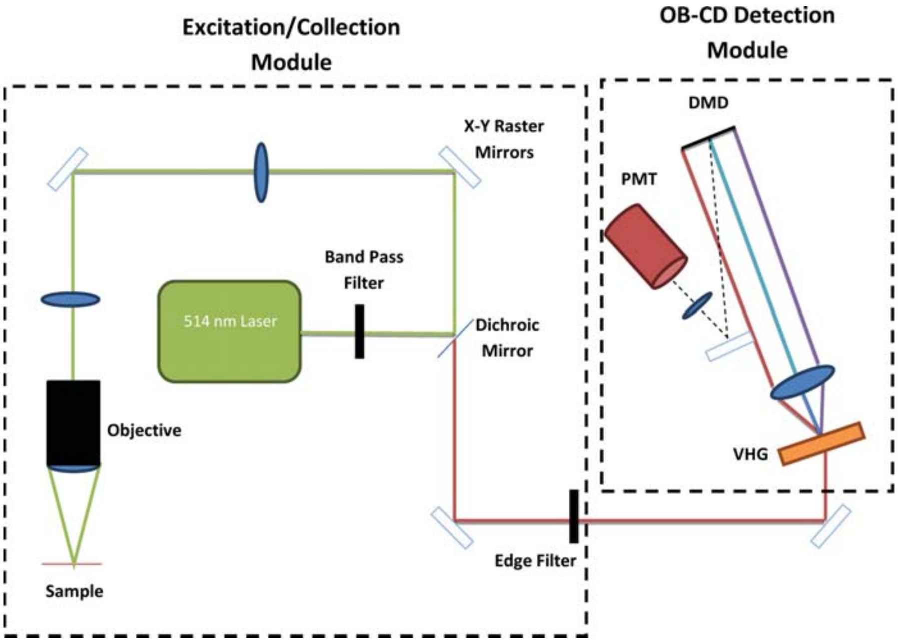


Fig. 2. Schematic of the OB-CD Raman system based upon a 514 nm laser excitation source.

transparency was 3M brand (model #PP2500).

4. Results and discussion

4.1. *Treatment of nuisance parameters*

Here we demonstrate how the OB-CD detection strategy is influenced by whether the photon emission rates of some of the spectra used for training OB-CD filters are treated as nuisance parameters; we designate such spectra as *nuisance spectra*. More specifically, we generated an OB-CD training set containing the spectra of hexane, methylcyclohexane, a spectral feature arising from the NIR objective in the 785 nm OB-CD system, and the four Bernstein polynomials shown in Fig. 1. Using this training set, we calculated two sets of OB-CD filters (explicitly, both the matrix of binary filters, F, and the measurement time matrix, T): one set for which no spectral features were considered to be nuisance spectra and a second set for which the spectral feature arising from the NIR objective and the Bernstein polynomials were considered to be nuisance spectra. The optimal filters turned out to be identical in both cases (although that need not in general be the case), while the measurement time matrices, T, were quite different as shown in Fig. 3. Notice that the OB-CD filters associated with non-nuisance spectra are turned on for longer percentages of the total measurement time relative to OB-CD filters for the same spectral components when no portion of the training set is considered a nuisance spectra.

One might expect that additional time spent measuring the non-nuisance spectra would result in lower variance in the recovered Raman rates for these components. To test this, the two sets of OB-CD filters (and the associated measurement time matrices) shown in Fig. 3 were used to classify hexane and methylcyclohexane (explicitly, there was no added fluorescence in these samples despite including Bernstein polynomials in both OB-CD training sets). Each chemical was measured using each of the sets of OB-CD filters for 1,000 measurement with 10 ms total

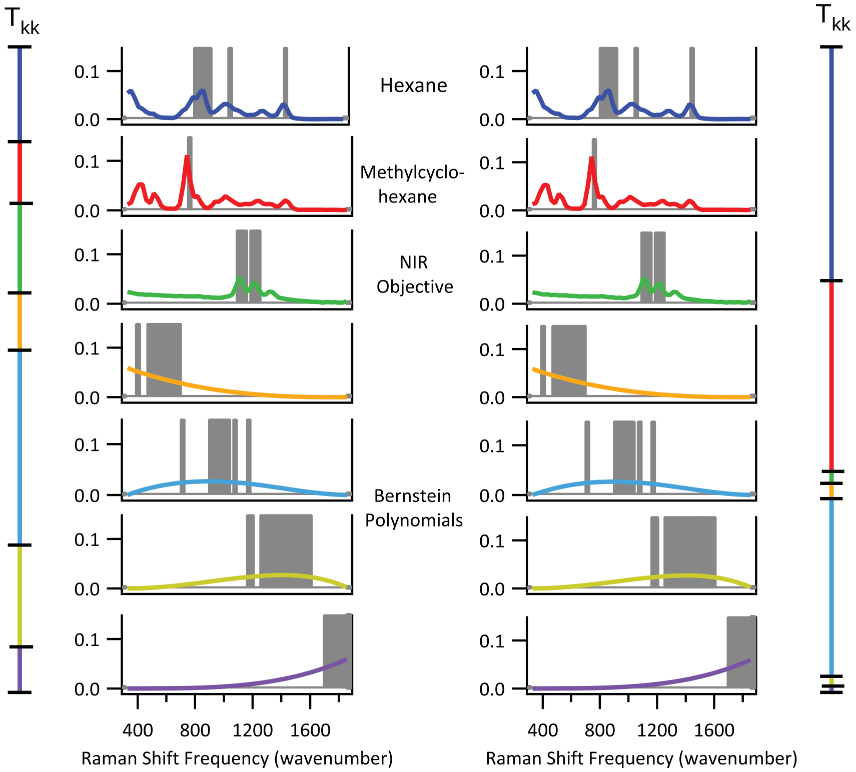


Fig. 3. The colored curves are training spectra (each normalized to unit area) and the gray bands indicate regions in which the OB-CD filters are on (i.e., direct light towards the detector). The Raman spectra were obtained with a spectral resolution of 30 cm^{-1}. The lines labeled T_{kk} correspond to the fraction of the total measurement time that data is collected using the filter associated with each spectral component (denoted by color). The OB-CD filter and T_{kk} results on the left were obtained without considering any components to be nuisance spectra, while those on the right were obtained when considering the NIR objective and Bernstein polynomials to be nuisance spectra.

integration time. The results of these measurements, shown in Fig. 4, clearly reveal the reduced variance (smaller 95% confidence bands) obtained when treating only the two components of interest as non-nuisance spectra. Note in particular that the mean recovered Raman rates for hexane and methylcyclohexane generated from both sets of OB-CD filters differed very little.

4.2. Validation of Bernstein polynomials

To test and validate our OB-CD fluorescence suppression strategy we either used a white light source to simulate fluorescence or samples containing fluorescent components. The results described in this section were obtained using a white light illuminator as a convenient fluorescence mimic, as its intensity can readily be varied, and its shape resembles typical fluorescence backgrounds (and has a different shape in the 514 nm and 785 nm spectral region). In subsequent subsections, we describe results obtained using samples with fluorescent impurities, rather than white light, to validate our OB-CD strategy. Here we produced OB-CD filters by training using Bernstein polynomials as well as the Raman spectra of n-hexane and methycyclohexane (and a

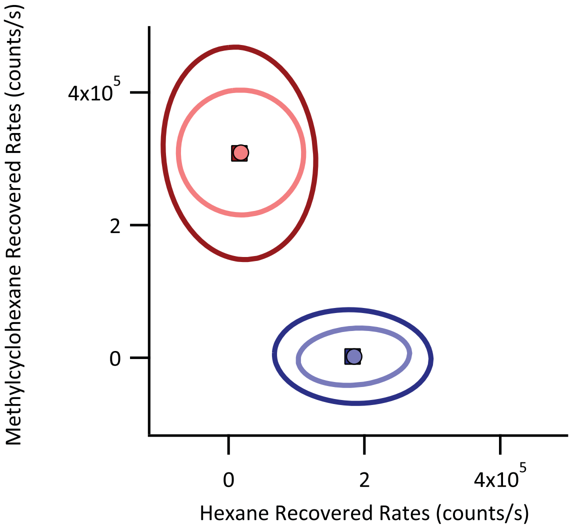


Fig. 4. Recovered Raman rates for hexane (blue) and methylyclohexane (red) generated using OB-CD filters that considered components of the training set to be nuisance spectra (dark blue and dark red) and OB-CD filters that considered the spectral component arising from the NIR objective and the four Bernstein polynomials to be nuisance spectra (light blue and light red). In all cases, 1,000 OB-CD measurements were taken with a total integration time of 10 ms. The ellipses represent the 95% confidence interval of the recovered Rates for each sample. The large markers in the center of each ellipse represent the mean recovered Raman rates.

spectral component arising from the NIR objective for the 785 nm system). We treat the coefficients of the Bernstein polynomials and the spectral component arising from the NIR objective for the 785 nm system as nuisance parameters. The resulting filters and training spectra from the 785 nm system are virtually identical to those shown in Fig. 3 (and are provided in the Appendix, along with the filters generated for the 514 nm OB-CD system).

The following results were obtained by holding the white light at constant intensity (of about 4 million counts per second) such that the total Raman/fluorescence signal intensity never exceeds 5 million counts per second. We then varied the Raman excitation laser intensity (using neutral density filters) in order to vary the relative amount of Raman and fluorescence in the measured spectra and OB-CD signals, and thus determine how fluorescence background influenced our the recovery of apparent Raman rates using OB-CD.

At each Raman signal intensity, we recovered Raman rates using the OB-CD filters as described in Section 2.1, both with and without the white light background. This allowed us to determine the error in recovered Raman rates as a result of adding fluorescent (white light) background. Figure 5 compares the Raman rates recovered with (*y*-axis) and without (*x*-axis) added white light when using 30 ms total integration time for each measurement. The number at the top indicates the ratio of the integrated area of the fluorescence and Raman signals and the error bars represent the standard deviations of the Raman rates for each component (see the figure caption for further details).

If fluorescence is not perfectly modeled, then the recovered Raman rates can contain a systematic modeling error whose magnitude increases with fluorescent intensity. When using a constant intensity white light to model fluorescence such modeling error introduces an approximately constant offset to the recovered Raman rates. The magnitude of this offset can be de-

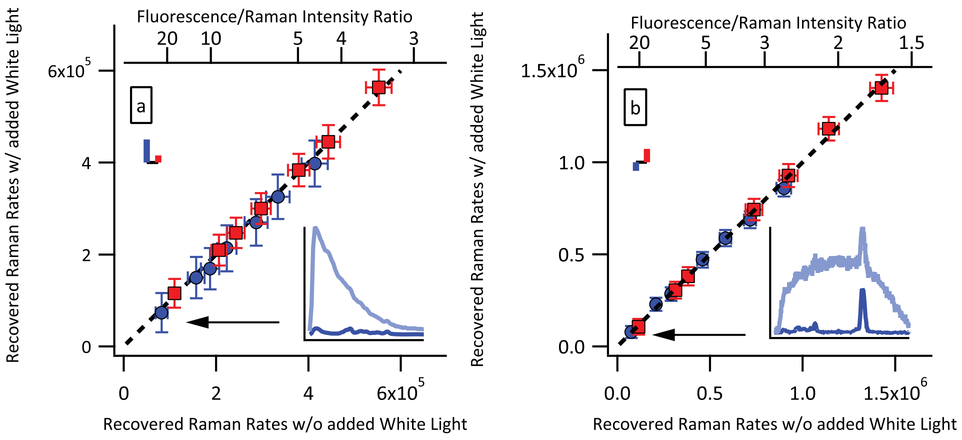


Fig. 5. Plot of hexane (blue) and methylcyclohexane (red) recovered Raman rates measured without added white light versus recovered Raman rates measured with added white light on the (a) 785 nm laser excitation system and (b) 514 nm laser excitation system. Rates have each been corrected by removing a small constant vertical offset (modeling error) whose magnitude was determined by measuring the apparent recovered Raman rates obtained in measurements performed on white light without Raman. The magnitude of this correction is represented by the colored bars in the upper left of each plot. Each point represents the means of 1,000 measurements (each obtained using a 30 ms total integration time) with error bars representing 1 standard deviation. Top axis denotes the ratio of the total (integrated) number of the white light/Raman photons. The inset spectra were obtained from hexane with and without added white light with ∼1 OD neutral density filter and correspond to measurements made at the points denoted by the arrows.

termined by measuring the apparent recovered rates of the Raman components obtained when measuring only white light (containing no Raman photons). For the measurements shown in Fig. 5, these modeling errors were relatively small and have magnitudes indicated by the bars in the upper left of each plot. These modeling errors have been subtracted from each of the points in Fig. 5. In other words, before correcting for this modeling error, all of the recovered Raman rates were slightly offset from the dashed diagonal line (of slope 1).

If the spectrum of white light was modeled perfectly by a degree-three polynomial (or was corrected for modeling error, as described above), we would expect that the mean recovered Raman rates for samples with added white light would not significantly differ from the mean recovered Raman scattering rates without added white light. In other words, we would expect the points to lie on a line with slope one as indicated by the dashed lines in Fig. 5. Thus, the agreement between the points and dashed line in Fig. 5 clearly demonstrate that Raman components can be quantified accurately in the presence of fluorescence backgrounds whose integrated intensity is up to 20 times that of the Raman component of interest. Note that the factor of 20 is obtained from the results shown in Fig. 5, as this is when the Raman signal-to-noise approaches 1:1.

Samples with larger fluorescence/Raman intensity ratios can in principle be accurately analyzed using OB-CD with longer measurement times. However, when the integration time approaches one second it may be appropriate to use conventional full spectral measurements and fluorescence subtraction procedures as the OB-CD strategy is primarily advantageous for performing high speed (or low light level) measurements that are not compatible with CCD detection. Additionally, performing OB-CD measurements on samples in which fluorescence is more than 20 times as intense as the Raman signal of interest would require careful modeling error correction (as the modeling error would become large relative to the Raman intensities). The results presented in Section 4.3 demonstrate the accuracy with which high speed OB-CD Raman classification and quantitative measurements may be performed without correcting for model error so long as the fluorescence background intensity does not exceed 20 times the Raman signal intensity.

4.3. Raman quantitation and classification of fluorescent samples

4.3.1. Toluene and fluorescent aniline

The following results were obtained using liquid mixtures of toluene and an aniline sample that was partially oxidized and, as a result, developed a fluorescent impurity that could be removed by distillation. We selected these two liquids because of their significant spectral overlap (the dot product of the two normalized spectral vectors is 0.91) and thus successful classification of aniline/toluene mixtures may be used to demonstrate that our fluorescent mitigation strategy is compatible with the Raman-based quantification of such spectrally overlapped mixtures.

We trained OB-CD filters using 785 nm spectra obtained from distilled aniline, toluene, the NIR objective spectrum describe above, as well as the four Bernstein polynomials shown in Fig. 1 (all of the resulting spectra and OB-CD filter functions are given in the Appendix). For this experiment, we treated the spectral component arising from the NIR objective and the four Bernstein polynomials as nuisance spectra. We used OB-CD to recover Raman rates for toluene, distilled aniline, fluorescent aniline, and mixtures of both types of aniline and toluene. Using these recovered Raman rates, we calculated apparent volume fractions of aniline and toluene as follows:

$$\chi_i = \frac{w_i \hat{\Lambda}_i}{\sum_i w_i \hat{\Lambda}_i},$$

where w_i is equal to $M_i/\hat{\Lambda}_i^{\mathrm{Max}}$, M_i is equal to molarity of the i^{th} pure liquid, and $\hat{\Lambda}_i^{\mathrm{Max}}$ is equal to the mean recovered Raman scattering rate for the i^{th} pure liquid as previously reported [9]. We then estimated the apparent volume fraction (Φ) for aniline and toluene in each sample. We did this by dividing the apparent mole fraction (for either aniline or toluene) by the molarity of each pure liquid as follows:

$$\Phi_i = \frac{M_i \chi_i}{\sum_i M_i \chi_i}.$$

Figure 6 plots the resulting apparent volume fractions of toluene and aniline as well as mixtures of the two. The results in Fig. 6 demonstrate that the mean recovered Raman rates are insensitive to the fluorescence arising from the impure aniline sample. The variance of the measurements without fluorescence, however, is less than the variance of samples with fluorescence (as a result of additional shot noise in the latter measurements).

4.3.2. Aqueous ethanol and fluorescent tequila

Previous work [22, 23] has demonstrated that Raman spectroscopy can be used to quantify the volume percentage of ethanol in tequila samples and qualitatively distinguish distilled (“silver”) and highly fluorescent, aged (or “golden”) tequilas, even in the presence of fluorescence (more common in aged, so-called “golden” tequila). Here we show that our OB-CD fluorescence mitigation strategy can be used to quantify the volume percentage of ethanol in tequila, even for fluorescent “golden” tequila samples at speeds much greater than those previously reported for this application.

We used the 514 nm laser-based OB-CD system for these studies, both because there was more fluorescence produced at this wavelength than when using the 785 nm excitation and because the C-H and O-H stretch vibrational bands are not readily detectable using the 785 nm

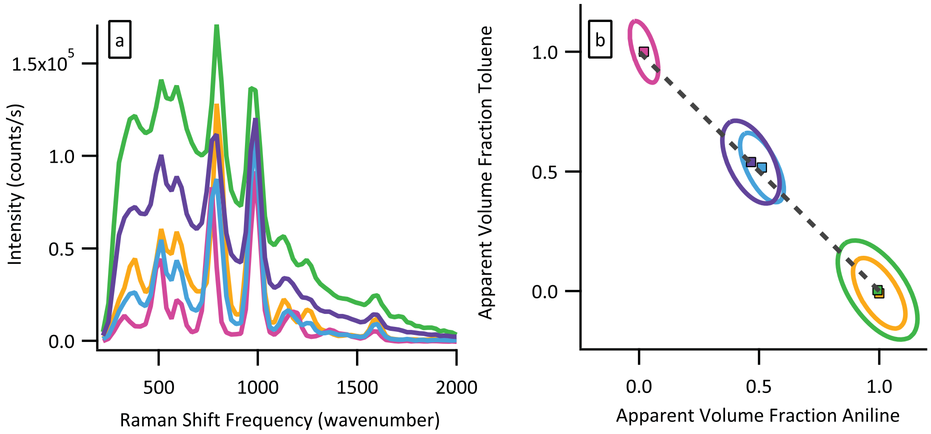


Fig. 6. (a) Spectra of distilled aniline (orange), toluene (magenta), fluorescent aniline (green), a 47:53 volume-by-volume mixture of distilled aniline and toluene (dark green), and a 52:48 mixture of fluorescent aniline and toluene (cyan) measured on the 785 nm OB-CD system. (b) Apparent volume fractions of distilled aniline (orange), toluene (magenta), fluorescent aniline (green), a 47:53 volume-by-volume mixture of distilled aniline and toluene (dark green), and a 52:48 mixture of fluorescent aniline and toluene (cyan). Each chemical was sampled 1,000 times at 20 ms per experiment. Ellipses correspond to the 95% confidence interval of the recovered rates for each sample. The large squares with black borders in the center of each ellipse represent the mean of each sample.

system. Spectra were collected and OB-CD filters were trained using ethanol, water, and Bernstein polynomials (shown in the Appendix). Note that we treated the four Bernstein polynomials as nuisance spectra for OB-CD filter generation. After this, OB-CD was used to recover Raman rates for ethanol and water in a silver tequila (“Arandas” brand) and a golden tequila (“Casamigos” brand). The corresponding apparent volume fractions were obtained from the recovered rates as shown in Fig. 7. In order to keep the fluorescence photon rates within the linear regime of the PMT detector, the laser intensity at the sample was reduced to ∼2 mW using a neutral density filter placed in front of excitation laser and the integration time per sample was increased to 100 ms.

The inset table in Fig. 7(b) shows the mean apparent ethanol volume fractions obtained in a measurement times of 100 ms and, parenthetically, the label volume percent ethanol for each sample. As can be seen, our predicted volume percentages of ethanol very nearly match the label percentage. However, the variance of the fluorescent “golden” tequila measurements is much greater than that of the “silver” tequila measurements due to the increased shot noise resulting from the fluorescence background. In spite of this, we note that by using OB-CD, we can accurately predict the volume percentage of ethanol in tequila samples even when the integrated intensity of the fluorescence of the sample is 20 times larger than the integrated Raman signal. Since the signal-to-noise of such measurements is typically limited by photon (shot) noise rather than read noise, the total time required to obtain a given precision depends on the available laser power and thus is expected to be comparable to that obtained using full spectral measurements (under otherwise identical conditions).

4.3.3. Fluorescent plastic film photobleaching

Here we show an imaging application of our OB-CD fluorescence modeling technique to demonstrate that the recovery of Raman scattering rates is unaffected by fluorescence pho-

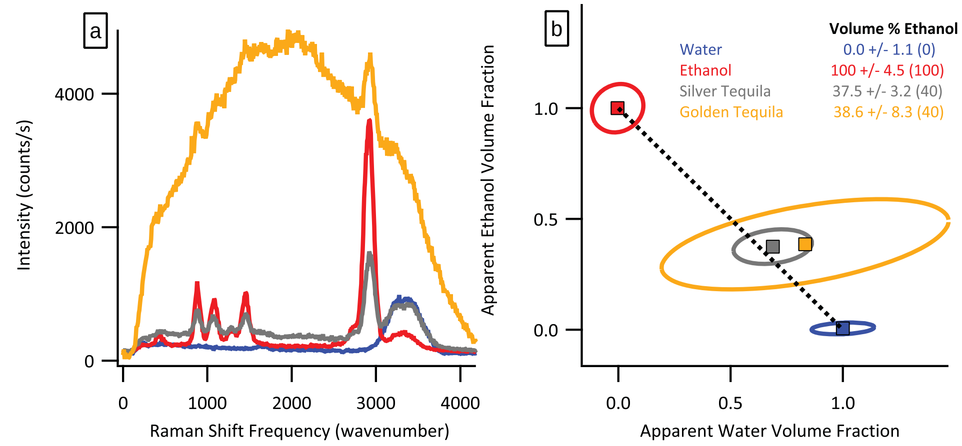


Fig. 7. (a) Spectra of water (blue), ethanol (red), Arandas brand silver tequila (gray), and Casamigos brand golden tequila (dark yellow) measured on the 514 nm OB-CD system (b) Apparent volume fractions of water (blue), ethanol (red), silver tequila (gray), and golden tequila (dark yellow) are compared with the nominal volume fractions (as obtained from the label on the tequila bottles). Each chemical was sampled 1,000 times at 100 ms per OB-CD measurement. Ellipses correspond to the 95% confidence interval of the recovered rates for each sample. Large squares with black borders represent the mean of each sample. The dashed line corresponds to line with slope -1. Inset table reports the mean apparent volume fraction of ethanol (plus/minus 1 standard deviation) for each sample and then, parenthetically, the label ethanol volume percentage for each sample.

tobleaching. For this purpose, we used a plastic film sample consisting of a clear overhead transparency of ∼1.7 mm thickness. This sample was chosen as it was found to contain both Raman and fluorescence signals when excited at 785 nm and the fluorescence could be photobleached by exposure to the excitation laser. Additionally, the film exhibited fluorescence with an integrated intensity 10 times that of the integrated intensity Raman features.

While photobleaching decreased the fluorescence background intensity of the sample by ∼50%, the remaining fluorescence could not readily be further photobleached. Thus, for OB-CD training purposes, we generated a Raman spectrum of the plastic by manually performing a polynomial background subtraction from a spectrum of photobleached plastic in the 328 nm^{-1} to 2057 nm^{-1} region. More specifically, the polynomial subtraction was performed using the “backcor” MATLAB algorithm (Vincent Mazet, 2010), using an Asymmetric Huber cost function, a threshold of 0.01, and four third-degree Bernstein polynomials as a basis. The resulting background subtracted Raman spectrum, as well as the spectra of the overhead transparency before and after photobleaching are shown in Fig. 8.

Next, OB-CD filters were calculated by training on the fluorescence-subtracted Raman features of the plastic film, the spectral component arising from the NIR objective, and the four third-degree Bernstein polynomials (and the resulting training spectra are provided in supplementary material). Note that unlike the filters constructed for previous examples, no components were considered nuisance spectra when calculating OB-CD filters, as we wanted to accurately estimate the intensity of the fluorescence before and after photobleaching. Using these filters, a 200 pixel × 200 pixel region of the plastic film (approximately 1 mm × 1 mm) was imaged with an integration time of 10 ms per pixel. Once this image was collected, two lines were photobleached in the transparency to form a photobleached “+” pattern near the center of the field of view (as shown in Fig. 10). Each line, consisting of 50 pixels, was photobleached for

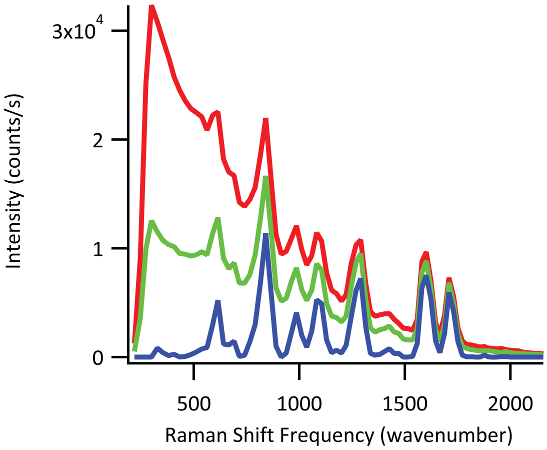


Fig. 8. The measured spectra of a cellulose acetate overhead transparency are plotted before photobleaching (red) and after 20 minutes of photobleaching (green). The output of the polynomial baseline subtraction is also plotted (blue).

10 minutes by scanning the laser repeatedly over the "+" pattern at a rate of 1 second per pixel. After photobleaching, the same field of view was reimaged and OB-CD was used to recover the Raman and fluorescence rates. Images were generated using the recovered rates using a method we have previously described [9]. The fluorescence intensity in this image, was determined from the sum of the recovered rates for all four Bernstein polynomials. Note that several spots on the plastic film were highly fluorescent (likely due to a fluorescent impurity, or dust particle, in the film), with counts well outside the linear region of the PMT and a fluorescence background intensity much greater than 20 times the average Raman signal. These pixels also had unusually high recovered apparent Raman rates, which we attributed to model error. These points were removed from the image, as indicated by black dots in the images shown in Fig. 9.

The upper two panels in Fig. 9 show the apparent recovered Raman rates before (left) and after (right) photobleaching, while the lower two panels show the corresponding apparent recovered fluorescence rates. Note that there is no evidence of a "+" pattern in the upper right panel; this indicates that photobleaching did not alter the apparent recovered Raman rates obtained from the film. There was a small (∼10%) decrease in average fluorescence intensity after photobleaching. We attributed this to the photobleaching that occurred while scanning the laser over the entire region during the OB-CD measurement. The similarity of the two upper images in Fig. 9 clearly demonstrates that we are able to obtain high-speed Raman intensity measurements in the presence of a fluorescent background with variable intensity.

5. Conclusions

We have demonstrated an OB-CD fluorescence mitigation strategy that can be used to accurately recover Raman rates from samples with moderate fluorescence intensity (that is, up to 20 times more intense than that of the integrated Raman signal). These results were achieved by quantifying fluorescence using OB-CD filters trained on cubic Bernstein polynomials. We have validated this strategy using both white light as a fluorescence mimic, as well as using fluorescent liquid and solid samples. Thus, the present results demonstrate the feasibility of fast (sub-second) OB-CD based Raman classification and quantitation of moderately fluorescent

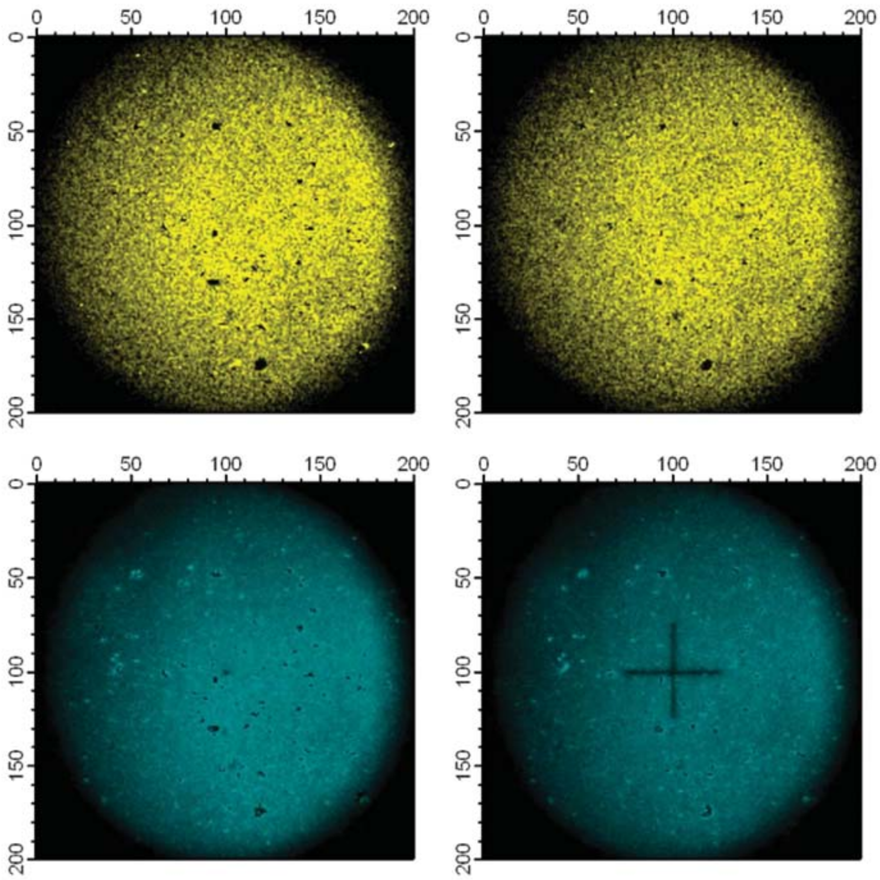
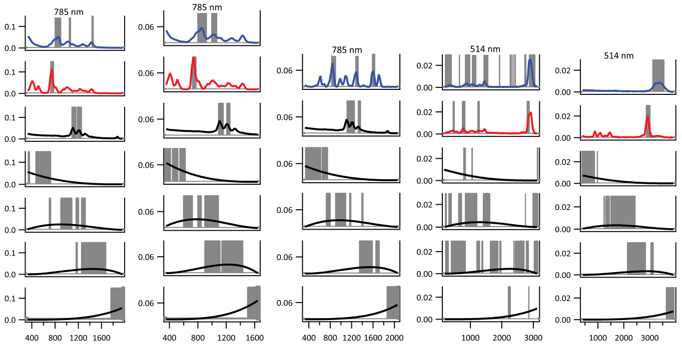


Fig. 9. Images showing the recovered Raman (yellow) and fluorescence (cyan) rates of a cellulose acetate overhead transparency before (images on the left) and after (images on the right) photobleaching a "+" pattern into the center of the imaged area. All images were collected with an integration time of 10 ms per pixel. The circular nature of these images arises from the field of view of the objective, as the images were obtained by raster-scanning the angle of the laser as it enters the back of the objective (while remaining centered in the objective).

samples. This approach can be extended to systems with a fluorescence/Raman intensity ratios greater than 20:1, but would likely require turning down the laser intensity (to avoid detector saturation) and using much longer integration times. Thus, the presented OB-CD strategy is expected to be most useful in applications requiring fast analysis of liquid and solid samples whose fluorescence does not overwhelm the underlying Raman chemical fingerprints. This is consistent with previous results [10], which indicated that the trade-off between higher read-noise and higher spectral information content of full-spectral CCD measurements relative to the OB-CD detection strategy would indicate that OB-CD is most advantageous (relative to CCD measurements) in fast (low-signal) applications that unaccessible to CCD-based measurements.

6. Appendix

Fig. 10 shows the training spectra and calculated OB-CD filters for several of the samples in this paper.



(a) (b) (c) (d) (e)

Fig. 10. Area-normalized spectra plotted against Raman shift frequency (wavenumber) shown here pertain to results obtained using a 785 nm and 514 nm excitation laser as noted above each subfigure. Note that the 785 nm spectra were measured with a resolution of 30 cm^{-1} and the 514 nm spectra were measured with a resolution of 12 cm^{-1}. Unless otherwise noted, the spectral component arising from the NIR objective and the Bernstein polynomials were always considered nuisance spectra. (a) Spectra and the resulting OB-CD filters for (in order from top down): n-hexane, methylcyclohexane, the spectral component arising from the NIR objective and the four degree-three Bernstein polynomials. The fraction of the total measurement time that each filter was collecting was 0.385, 0.219, 0.055, 0.034, 0.268, 0.031, and 0.009, respectively. (b) Spectra and the resulting OB-CD filters for (in order from top down): n-hexane, methylcyclohexane, and the four degree-three Bernstein polynomials. The fraction of the total measurement time that each filter was collecting was 0.248, 0.388, 0.010, 0.032, 0.277 and 0.045, respectively. (c) Spectra and the resulting OB-CD filters for (in order from top down): aniline, toluene, spectral component arising from the NIR objective, and the four degree-three Bernstein polynomials. The fraction of the total measurement time that each filter was collecting was 0.484, 0.265, 0.025, 0.003, 0.154, 0.059 and 0.010, respectively. (d) Spectra and the resulting OB-CD filters for (in order from top down): ethanol, water, and the four degree-three Bernstein polynomials. The fraction of the total measurement time that each filter was collecting was 0.276, 0.261, 0.027, 0.116, 0.232, and 0.088 , respectively. (e) Spectra and the resulting OB-CD filters for (in order from top down): Raman features of the plastic film, the spectral component arising from the NIR objective, and the four degree-three Bernstein polynomials. The fraction of the total measurement time that each filter was collecting was 0.099, 0.210, 0.099, 0.271, 0.212, and 0.109, respectively. No components were considered nuisance spectra, as we wanted to accurately estimate the intensity of the fluorescence before and after photobleaching.

Acknowledgments

The work was supported in part by the Office of Naval Research (Contract N0001413-0394 to DBA, GTB, and BJL) and the Simons Foundation (Award #209418 to BJL).