Computational and Statistical Approaches to Optical Spectroscopy

by

Ningren Han

S.M., Massachusetts Institute of Technology (2013)
B.Eng., National University of Singapore (2011)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2018

Signature redacted

Author .. ..
Department of Electrical Engineering and Computer Science
August 31, 2018

Signature redacted

Certified by ....
Rajeev J. Ram
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by ... .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Computational and Statistical Approaches to Optical Spectroscopy

by

Ningren Han

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Compact and smart optical sensors have had a major impact on people's lives over the last decade. Although the spatial information provided by optical imaging systems has already had a major impact, there is untapped potential in the spectroscopic domain. By transforming molecular information into wavelength-domain data, optical spectroscopy techniques have become some of the most popular scientific tools for examining the composition and nature of materials and chemicals in a non-destructive and non-intrusive manner. However, unlike imaging, spectroscopic techniques have not achieved the same level of penetration due to multiple challenges. These challenges have ranged from a lack of sensitive, miniaturized, and low-cost systems, to the general reliance on domain-specific expertise for interpreting complex spectral signals.

In this thesis, we aim to address some of these challenges by combining modern computational and statistical techniques with physical domain knowledge. In particular, we focus on three aspects where computational or statistical knowledge have either enabled realization of a new instrument—with a compact form factor yet still maintaining a competitive performance—or deepened statistical insights of analyte detection and quantification in highly mixed or heterogeneous environments. In the first part, we utilize the non-paraxial Talbot effect to build compact and high-performance spectrometers and wavemeters that use computational processing for spectral information retrieval without the need for a full-spectrum calibration process. In the second part, we develop an analyte quantification algorithm for Raman spectroscopy based on spectral shaping modeling. It uses a hierarchical Bayesian inference model and reversible-jump Markov chain Monte Carlo (RJMCMC) computation with a minimum training sample size requirement. In the last part, we numerically investigate the spectral characteristics and signal requirements for universal and predictive non-invasive glucose estimation with Raman spectroscopy, using an *in vivo* skin Raman spectroscopy dataset. These results provide valuable advancements and insights in bringing forth smart compact optical spectroscopic solutions to real-world applications.

Thesis Supervisor: Rajeev J. Ram
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

Looking back at my years at MIT, there is a long list of people that I would like to thank. First and foremost, Professor Rajeev Ram has been an incredibly helpful advisor along the way. His passion for pursuing research with a meaningful purpose, his curiosity towards new areas and disciplines, and his insights and knowledge of science and technology in general have all impacted me and my research in profound ways. There have been several key defining moments in my Ph.D. career where it was his words or actions that pushed me in the right direction. For that I am forever indebted. I was told that every Ph.D. student inherits his or her advisor's traits in subtle or non-subtle ways. I do not know whether it is true or not, but I certainly hope it is.

I would also like to thank Professor George Verghese and Professor George Barbastathis for being on my thesis committee and providing helpful insights and feedback on my research. They have been incredibly responsive and patient through the busy seasons. I could not find anyone better than them to be the ultimate gatekeepers for my thesis research.

There have been so many memories working in the POE Raman team. I would like to thank everyone in it for his or her support. Dr. Gajendra Singh was the one who introduced Raman spectroscopy to me. The POE Raman team as I know started with him and has now grown to a substantial size. His presence changed my trajectory in this group. I am very grateful for all the help he had provided. Dr. William Herrington worked closely with me on multiple projects. His critical thinking has always promoted me to think harder about various aspects of spectroscopic techniques. He was always the first person that I would consult to when things did not work. Dr. Amir Atabaki's experience and expertise have granted him a unique place in the Raman team. His kind, calm, and hardworking spirit has been a strong supporting force in the group. I hope he succeeds in his next endeavor and I will be there if he needs any help. Neerja Aggarwal has been a positive and energetic figure since she joined the group. Her impact on the group culture extends well beyond the small lab space in many different ways. Erika Ye has impressed me with her tenacity and ability to work through long hours. She has done a great job for her master's study, and I hope she has a fruitful research career in the future. I would also like to express my gratitude to our colleagues at Samsung for their financial support as well as their valuable feedback. In particular, I thank Dr. Seong-Ho Cho, Dr. Ho-Jun Chang, Dr. June-Young Lee, and Dr. Yun-Sang Park. Outside of the Raman team, I have been immersed in the intelligent and energetic environment from the POE group for the past five years. I thank everyone in the group for his or her contributions to the wonderful group environment. My technical horizon has widened significantly over my time in the POE group due to everyone's efforts and hard work. I also want to thank Shayne Fernandes for being an amazing assistant to the group.

Outside of the POE group, Dr. David Burghoff has been an awesome friend and colleague throughout all these years. He and I share way too many interests and ideas in different aspects of life. I hope he will have a long and fruitful career in Notre Dame, and I am looking forward to seeing him continue being the intellectual fountain of his field.

Dr. Matthew Brand offered me a joyful and memorable summer internship experience at MERL. He is no doubt the coolest academic I have ever seen or worked with. I am looking forward to seeing more of his fantastic lenses, holograms, and optics art on exhibition and display.

Dr. Kevin Lee and Dr. Harry Lee from Pharyx Inc. and Dr. Pablo Perez-Pinera from UIUC provided a great amount of help over my early years at the POE group. Although the work we performed did not end up in this thesis, it will always be a part of the fun memories that I cherish from my time at MIT.

I thank all my friends at MIT for their company through these years. Some of them helped me significantly in my job preparation and search over the past year. Some of them have landed enormous support in helping us through difficult times. I hope our friendships will last forever.

I would like to thank my parents and parents-in-law for their love and care. I could not have been here without them. I would also like to thank my wife, Qiyue, for being the perfect one for me. Last but not least, I would like to thank my daughter, Zoe, for bringing tremendous happiness to me and my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Overview

1.1 Optical Spectroscopy As a Sensing Modality

Compact and smart optical sensors have had a major impact on people's lives over the last decade. For example, low-cost imaging sensors have transformed numerous social paradigms by putting a recording device in everyone's pocket. Although the spatial information provided by imaging systems has already had a major impact, there is untapped potential in the spectroscopic domain. As illustrated in Figure 1-1, by transforming molecular information into wavelength-domain data, optical spectroscopy techniques have become one of the most popular scientific tools for examining the composition and nature of materials and chemicals in a non-destructive and non-intrusive manner [5]. Spectroscopic sensing is also one of the few techniques available that can perform simultaneous multi-chemical or multi-component analysis in a true label-free fashion. These characteristics, together with the fact that optical sensing is typically performed in a non-contact or remote mode, have led to the ever-increasing adoption and integration of spectroscopic sensing technologies in the industrial world.

However, in comparison to imaging, spectroscopic techniques have not achieved the same level of penetration despite of their huge potential and demonstrated success in many sensing areas. This is due to challenges ranging from a lack of sensitive, miniaturized, and low-cost systems to the general reliance on domain-specific expertise for interpreting complex spectral signals. As a result, there has been considerable re-

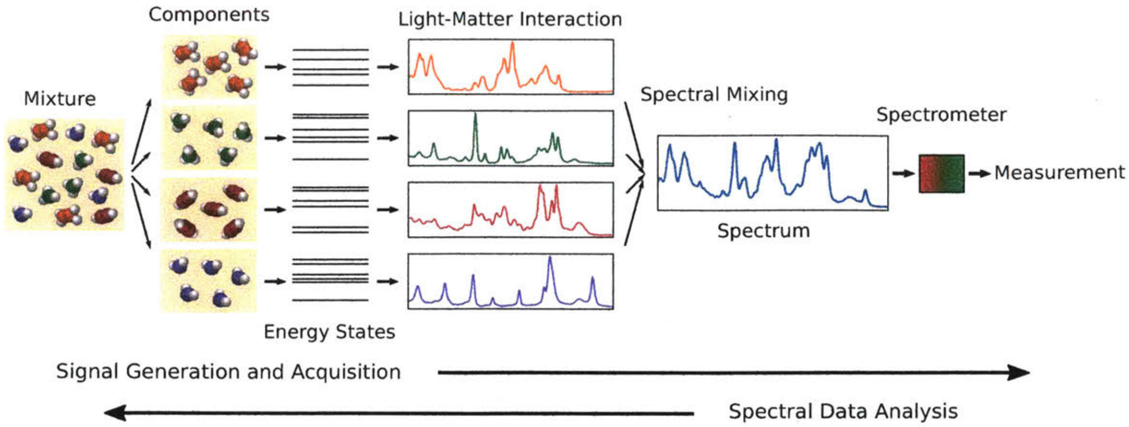


Figure 1-1: Illustration of the signal generation and acquisition process in optical spectroscopy. Molecules in physical mixtures have different energy states governed by quantum mechanics and statistical physics. These energy states can contribute to spectral features during the light-matter interaction in optical spectral sensing. A spectrometer is then used to collected measurements that can be used to retrieve the spectral information. In spectral data analysis, the goal is to infer the molecular information about the mixture under examination from the spectral measurement.

search and development during recent years in bringing forward compact, sensitive and smart spectroscopy solutions with in-the-field analysis only previously achievable from laboratory-grade benchtop instruments with specially-trained spectroscopy experts. This includes instrumentation and computational improvements to the signal acquisition process and application-oriented statistical algorithm development for spectral data analysis. This thesis aims to address some of the contemporary challenges faced in optical spectroscopy by combining modern computational and statistical disciplines with physical domain knowledge and design. As introduction and overview, this chapter will provide the background discussions to the problems that will be addressed in this thesis. Section 1.2 discusses the various approaches for signal acquisition and transformation as illustrated in Figure 1-1. Conventional techniques, as well as the more recent computational approaches, for realizing spectrometers and wavemeters are presented. Next, Section 1.3 introduces algorithms for spectral data analysis with a focus on mixture study. Various algorithmic approaches including explicit mixture modeling, supervised and unsupervised learning, as well as spectral shape modeling are presented. Afterwards, an overview on skin Raman and autofluorescence spectroscopy and its applications in disease diagnostics and monitoring, followed up with a detailed discussion on the background and technology development for skin Raman spectroscopy for non-invasive glucose estimation, are provided in Section 1.4. At last, Section 1.5 provides a structural overview for the content of this thesis.

1.2 Signal Acquisition and Transformation

While many forms of optical spectroscopy exist with various signal generation and collection approaches, transforming the raw optical wave into spectral domain signals for data analysis is the core process shared by all techniques. This puts spectrometer design and optimization at the center stage for many spectroscopy applications. In general, with bright signals that do not have fine spectral features, selecting a general-purpose spectrometer that can properly display the spectrum is a straightforward task. However, this can become progressively more involved and intriguing once more requirements on spectrometer sensitivity, resolution, bandwidth, and form factor are in place. For example, spontaneous Raman spectroscopy is one of the most demanding spectroscopy areas in terms of sensitivity requirement due to the weak Raman scattering process. No satisfactory instrument solutions exist for miniaturized Raman spectrometers (with a footprint smaller than $\approx$ 5 cm × 5 cm × 5 cm) that have a comparable performance to a typical benchtop system. This desire for more compact and low-cost yet high-performing spectrometer systems is behind the many innovations proposed and realized over the past decades [6, 7, 8, 9].

Other than spectrometers that are designed for general spectroscopy including the case with broadband and non-monochromatic light sources, more specialized wavelength meters, or wavemeters, that are specifically for pulsed and continuous-wave (CW) coherent laser beams also play an important role in application areas such as the spectroscopy of atomic systems. A wavemeter usually has accuracy requirement much higher than that of a typical spectrometer and is generally built through an interferometric geometry. Similar to the case of spectrometers, a continuous push for compact, high-accuracy, and broad-bandwidth wavemeters has resulted in a surge of novel concepts for realizing next-generation devices and instruments [10, 11, 12].

Mathematically, for any spectral measurement, assume that the source light has a power spectral density of $S(\lambda)$. For a diffuse light source with $S(\lambda)$, the input light can have many spatial modes that are mutually incoherent. The spectral measurement essentially performs a mapping as

$$I(\mathbf{l}) = \int S(\lambda) H(\mathbf{l}, \lambda)\, d\lambda, \tag{1.1}$$

where $\mathbf{l}$ is the measurement location, $I(\mathbf{l})$ is the field intensity at $\mathbf{l}$, and $H(\mathbf{l}, \lambda)$ is the averaged instrument mapping function across all the input spatial modes. For $H(\mathbf{l}, \lambda)$, $H(\mathbf{l}, \lambda) = \sum_i H_i(\mathbf{l}, \lambda)$, assuming that $H_i(\mathbf{l}, \lambda)$ is the instrument mapping function for spatial mode i. With the proper sensor sampling function, Equation 1.1 can also be represented in the discrete form as

$$\mathbf{I} = \mathbf{HS}, \tag{1.2}$$

assuming that there are M discrete measurements from the sensor and N discrete spectral components from the light source to recover, $\mathbf{I} \in \mathbb{R}^M$, $\mathbf{H} \in \mathbb{R}^{M \times N}$, and $\mathbf{S} \in \mathbb{R}^N$. In general, experimental realization of a spectral measurement system aims at achieving a well-conditioned transformation $\mathbf{H}$, such that the spectral signal $\mathbf{S}$ can be reconstructed from instrument measurement $\mathbf{I}$ through inversion or pseudo-inversion of $\mathbf{H}$ with robust performance against measurement noise. This task can be very challenging for diffused light sources with many spatial modes, as the spatio-spectral ambiguity can result in ill-defined or ill-conditioned $\mathbf{H}$. Due to the conservation of étendue [13, 14], which states that the étendue of light cannot decrease through lossless propagation, one of the only solutions to this problem is to spatially filter the source light through a pinhole or slit, thereby restricting the input spatial modal profile. This means that most spectroscopy solutions are not photon-efficient with diffuse light sources. Brady [15] provided a more fundamental discussion on this topic, linking the sensor multiplexing capacity with the constant radiance theorem (which is a closely related statement to the conservation of étendue).

1.2.1 Commercial Spectrometers and Wavemeters

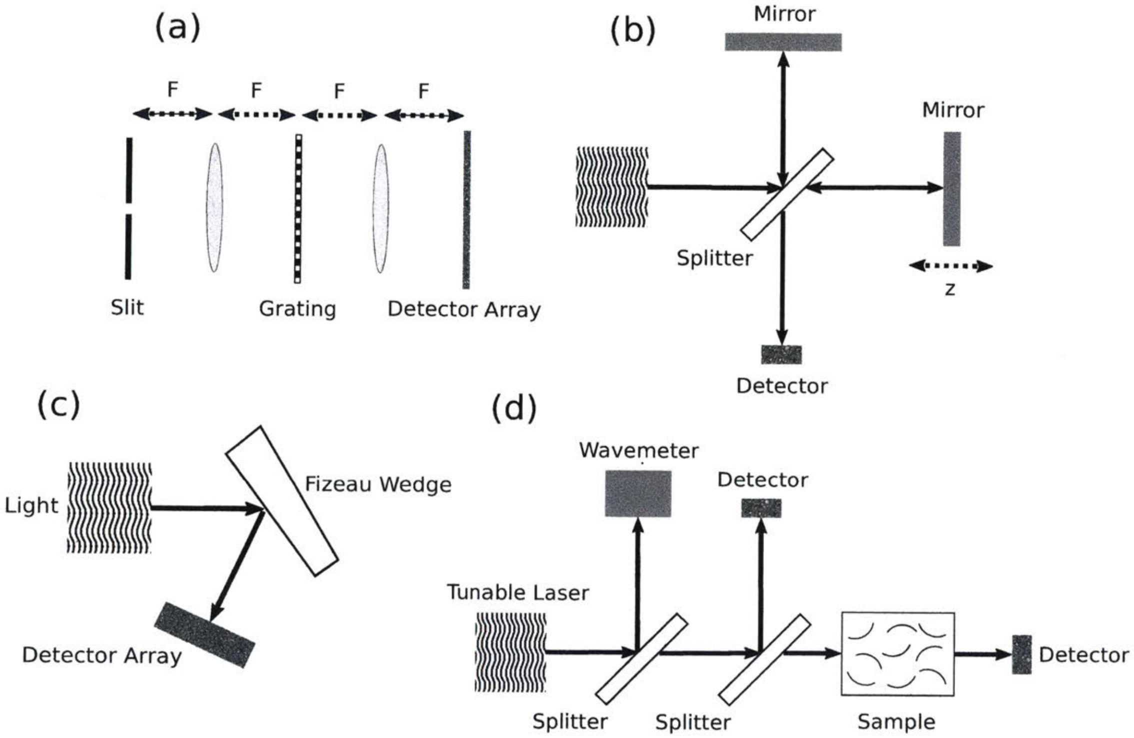


Figure 1-2: Schematic setups for (a) dispersive diffraction grating spectrometer; (b) scanning Michaelson interferometer-based Fourier spectrometer; (c) wavemeter based on the Fizeau interferometer; (d) tunable diode laser absorption spectroscopy (TDLAS).

The design choices for commercial spectrometers have been dominated by two options, the diffraction grating spectrometers and the Michelson interferometer-based Fourier spectrometers. The setups for these spectrometers are shown in Figure 1-2 (a) and (b). The diffraction grating spectrometers utilize the spatially-dispersive responses of periodic diffractive structures and use detector arrays such as CCD or CMOS for signal readout. In addition, a slit is used to reduce the spatio-spectral ambiguity associated with the optical signal transformation introduced by the diffraction grating. With detector line binning, this can achieve near identity matrix transformation as **H**, subject to imaging distortions and aberrations. Fourier spectrometers employ a scanning stage to record the interferogram of the light source through self-interference at various path length differences. In this case, the corresponding **H** is the inverse

Fourier transformation matrix. Both spectrometer approaches achieve near orthogonal transformation for **H**. For visible-to-near infrared applications where there is an abundance of sensitive and low-cost detector arrays, dispersive spectrometers are more favored due to their static nature and strong balance between sensitivity and instrument size. On the other hand, for long wavelength applications from mid-infrared to terahertz frequencies where there is a dearth of large-area and sensitive detector arrays, Fourier spectrometers are the predominant choices.

For wavemeters, there are also two main approaches. They are the scanning Michelson interferometer (same as the Fourier spectrometer) and the static Fizeau interferometer. Figure 1-2 (c) shows an example setup for the Fizeau interferometer [16]. The Fizeau wedge reflects incoming light wave into to two beams with slightly off propagation directions. A detector array is used to record the interference pattern of these two beams. The wavelength information can then be obtained from the interference pattern through signal process and analysis [17]. For commercial wavemeters, the highest accuracy wavemeters are typically based on the Fizeau geometry. In addition, it has better performances against power fluctuations and side modes due to the static nature. Similar to the case of spectrometers, Michelson interferometer-based wavemeters can cover longer wavelength regimes due to the fact that only a single detection element is required [18].

It is worth mentioning that for some spectroscopy applications, a strict "spectrometer" in the conventional sense is not needed. Figure 1-2 (d) shows an example setup for tunable diode laser absorption spectroscopy (TDLAS) [19]. In this case, a diode laser is tuned and two detectors are used for constructing the absorption profile of the sample in the wavelength sweeping domain. A wavemeter is also used to provide the necessary reference wavelength information of the measurement. This is an example where a wavemeter can be a critical component for spectrum acquisition.

Apart from these conventional solutions, there is a myriad of compact spectrometer and wavemeter design approaches that has been proposed and realized. These approaches include integrated optics [7, 20], micro-optics [21], and filter array [9] for spectrometers, and Fabry-Perot interferometer [22] and diffraction grating [10] for wavemeters. However, despite these interesting demonstrations, the commercial landscape for spectrometers and wavemeters are still mostly dominated by the more conventional techniques discussed in this part.

1.2.2 Computational Instrumentation

The Michaelson interferometer is an example where computational inversion is used to retrieve the spectral signal from the instrument measurement. Since the early 2000s, spectrometers and wavemeters that incorporate more sophisticated computational elements have been proposed and experimentally realized. Compared to the conventional approaches, these instruments have shown competitive advantages in areas such as light throughput, form factor, and resolving power amongst others. By utilizing physical modeling and digital computation, spectral information can be retrieved with unconventional optical elements in these designs.

The most prominent example is the coded aperture spectrometer [6, 23]. This is shown in Figure 1-3 (a). Instead of a slit, the coded aperture spectrometer uses a coded mask as the input entrance in a grating spectrometer configuration for throughput improvement. The coded mask employs orthogonal spatial coding such as the Hadamard code. This preserves the well-conditioned nature for $\mathbf{H}$ in coded aperture spectrometers. Deconvolution-like algorithms can then be used for spectral reconstruction. In addition to spectrometer design, the coded aperture configuration has also been demonstrated for high-performance hyper-spectral imaging applications [28]. It is worth noting that the coded aperture spectrometer is one of the few spectrometer approaches that have demonstrated definitive light throughput improvement over the traditional approaches for diffused light sources [6]. Such throughput improvement relies on the fact that for diffused light sources, the light intensity can be assumed to be close to uniform on the entrance mask.

The rest of the computational approaches mostly rely on explicitly measuring the transformation matrix $\mathbf{H}$ with various engineered spatio-spectral responses through a full-spectrum calibration process. Examples of such approaches are shown in Figure 1-2 (b) to Figure 1-2 (h). Xu et al. [24] used a 3-D disordered photonic crystal structure

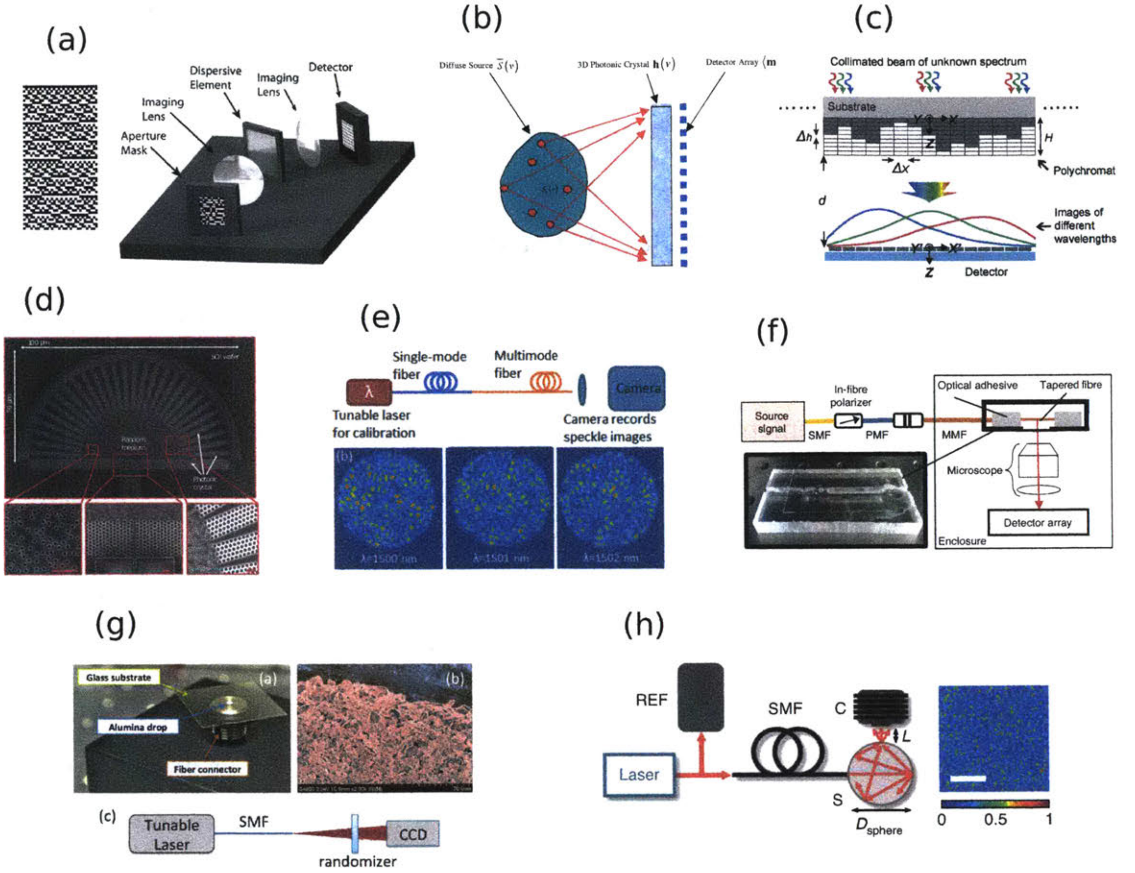


Figure 1-3: Illustrations for a number of spectrometer and wavemeter designs with computational elements in literature. (a) The Hadamard mask and the coded aperture spectrometer [6, 23]. (b) 3-D disordered photonic crystal spectrometer for multi-modal input [24]. (c) Spectrometer with a broadband diffractive structure [25]. (d) Waveguide-based spectrometer with disordered scattering media [8]. (e) Multimode fiber interference-based spectrometer [26]. (f) Multimode waveguide interference-based spectrometer [27]. (g) Wavemeter with a think film diffuser [11]. (h) Wavemeter with an integrating sphere [12].

to construct a spectrometer with multimodal diffused light (Figure 1-3 (b)). Wang and Menon [25] used broadband diffractive structures to reconstruct spectra for light sources including laser, LED, and broadband sources (Figure 1-3 (c)). For waveguide and fiber-based approaches, Redding et al. [8] used scattering through disordered media to realize spectrometers in an ultra-compact footprint (Figure 1-3 (d)). Redding and Cao [26] and Wan et al. [27] used multi-mode interference pattern for spectral reconstruction with light input from single-mode fiber and waveguide (Figure 1-3 (e) and Figure 1-3 (f)). For wavemeter-specific applications, Mazilu et al. [11] used scattering from a thin film diffuser as the dispersive element (Figure 1-3 (g)), achieving picometer level accuracy. Metzger et al. [12] used a fiber-coupled integrating sphere to achieve sub-femtometer accuracy (Figure 1-3 (h)).

With **H** obtained from the full-spectrum calibration process, Equation 1.2 becomes a standard mathematical inverse problem to evaluate the estimated $\hat{\mathbf{S}}$ given any detector measurement **I**. Direct pseudo-inversion of **H** can be easily corrupted due to noise in the measurement [29]. To overcome this, most of these approaches used techniques such as regularization [29] for spectral estimation. While these calibration-based computational spectroscopy approaches have demonstrated impressive resolution (sub-femtometer at 780 nm in Metzger et al. [12]), bandwidth (from 500 to 1600 nm in Wan et al. [27] and from 300 to 2500 nm with simulation in Wang and Menon [25]) and footprint (25 μm × 50 μm in Redding et al. [8]) characteristics, it is worth pointing out that other than Xu et al. [24] and possibly Wang and Menon [25], the rest of them assumed high spatial purity for the input light through spatial filter from single-mode fiber and waveguide. This is fine for coherent light sources such as lasers but does not match the requirement for incoherent applications such as fluorescence measurement. In addition, even with techniques such as regularization, the measurement to obtain the transformation matrix **H** contains noise and errors, which will be amplified and injected in the inversion process [29]. As a result, it is not expected that these spectrometers and wavemeters will have any sensitivity advantage in comparison to conventional methods with orthogonal transformation embedded in the physical process. Moreover, experimentally, the full-spectrum spectral calibration process to retrieve **H** requires tunable monochromatic light sources, which currently do not have simple and compact solutions with high accuracy and broad bandwidth operation. As these instruments might need frequent recalibration due to environmental factors such as response drifts caused by mechanical and temperature influences [30], this aspect could severely hinder their practical use and adoption.

Despite the several potential issues described above, overall, there has been significant innovation and progress in the search for more compact and high-performance spectroscopy solutions. It is imaginable that spectroscopy systems with new elements

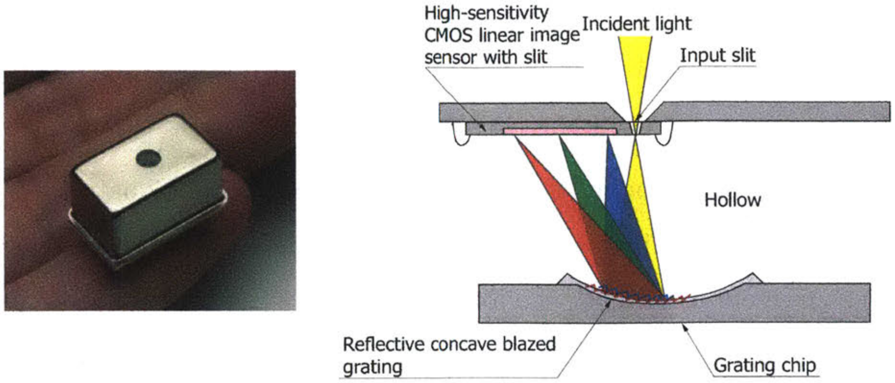

High-sensitivity CMOS linear image sensor with slit
Incident light
Input slit
Hollow
Reflective concave blazed grating
Grating chip

Figure 1-4: The Hamatsu C12666MA mini-spectrometer and its build structure [31].

introduced in this part are on the horizon for wider adoption in the near future. For example, a coded aperture design with existing compact spectrometer technology such as the Hamamatsu C12666MA micro-spectrometer [31], shown in Figure 1-4, can be a viable solution for small footprint and high throughput spectrometer for diffused light sources, although the current resolution achievable with Hamamatsu C12666MA is relatively coarse (≈ 15 nm) at the moment. Computational spectrometers and wavemeters can be used in situations where extremely high resolution might be required and recalibration is allowed. Alternatively, high-precision closed-loop temperature control and stable mechanical housing can be incorporated to mitigate the influences from environmental factors on the calibration matrix. However, the quest for compact spectroscopy solutions with high performance metrics without the need to perform any full-spectrum calibration is still far from completion.

1.3 Algorithms for Spectral Data Analysis

With the many novel concepts in realizing compact and high-performance spectrometers and wavemeters discussed in the previous section, we turn our discussions to the other end of applied spectroscopy problems. This is spectral data processing or chemometric analysis. Common to the core of any optical spectroscopy techniques, chemometric analysis aims at developing robust and statistically-sound methods to extract quantitative information related to the material or substance of interest from the spectral signals. As shown in Figure 1-1, the information flow of spectral data analysis is opposite to that of the signal generation and acquisition process. We start our discussions on this topic by introducing mixture problems in spectroscopy in the following part.

1.3.1 Mixture Problems in Spectroscopy

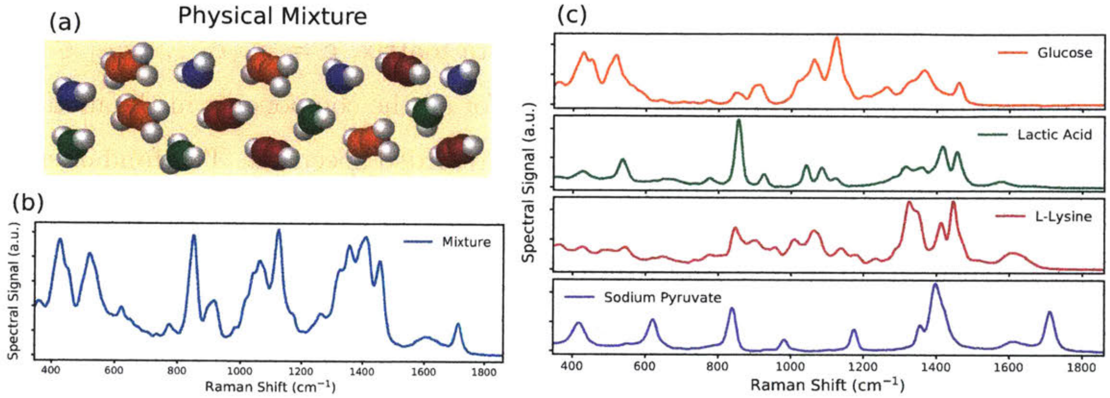


Figure 1-5: (a) The Raman spectrum contains quantitative information from the mixture molecules under examination. (b) An example Raman spectrum for a physical mixture containing four compositions in (c) at a molar concentration ratio of 1 : 0.47 : 0.66 : 0.35. (c) Raman spectra for the four composition materials in the mixture in (b) measured independently at equal concentrations.

Most chemometric problems involve analyzing spectral signals from a physical mixture of chemicals and materials. Figure 1-5 shows example spontaneous Raman spectra of a physical mixture as well as its four composition materials, namely glucose, lactic acid, L-lysine, and sodium pyruvate. Concentration or quantity estimation for any or all the composition materials from a set of mixture spectra can be performed with or without any prior knowledge on the Raman spectra of the composition materials. Alternatively, decision making based on the composition materials can be the direct outcome without the need to explicitly estimate their concentrations or quantities.

Broadly speaking, linear mixture models are the most fundamental and predominantly used modeling assumptions in chemometric research. Assuming that a discrete spectral signal $\mathbf{y}$ is collected with N spectral data points for a mixture substance with K components, we can model $\mathbf{y} = [y_1, y_2, \ldots, y_N] \in \mathbb{R}^N$ as

$$\mathbf{y} = \sum_{i=1}^{K} c_i \mathbf{a}_i + \boldsymbol{\epsilon} = \begin{bmatrix} | & | & \cdots & | & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \ldots & \mathbf{a}_{K-1} & \mathbf{a}_K \\ | & | & \cdots & | & | \end{bmatrix} \mathbf{c} + \boldsymbol{\epsilon} = \mathbf{A}\mathbf{c} + \boldsymbol{\epsilon},$$

where $\mathbf{a_i} = [a_{i,1}, a_{i,2}, \ldots, a_{i,N}] \in \mathbb{R}^N$ is the spectral signal for the i-th component at unit concentration, $\mathbf{A}$ is the component spectra matrix, $\mathbf{c} = [c_1, c_2, \ldots, c_K] \in \mathbb{R}^K$ correspond to the concentrations or quantities for all the components in the mixture, and $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \ldots, \epsilon_N] \in \mathbb{R}^N$ is the noise in the recorded spectrum. The fundamental assumption for this model is that signal strength from each component is linear to its abundance in the mixture and the overall signal is a linear addition from all the constitute component signals. For a set of mixture measurements $\mathbf{Y}$, where each column is a spectral measurement from an independent mixture sample, we have

$$\mathbf{Y} = \begin{bmatrix} | & | & \cdots & | & | \\ \mathbf{y}^{(1)} & \mathbf{y}^{(2)} & \ldots & \mathbf{y}^{(M-1)} & \mathbf{y}^{(M)} \\ | & | & \cdots & | & | \end{bmatrix} = \mathbf{AC} + \mathbf{E}. \tag{1.3}$$

Here, M is the number of mixture samples, $\mathbf{Y} \in \mathbb{R}^{N \times M}$, $\mathbf{A} \in \mathbb{R}^{N \times K}$, $\mathbf{C} \in \mathbb{R}^{K \times M}$, and $\mathbf{E} \in \mathbb{R}^{N \times M}$. For optical spectroscopy problems, Equation 1.3 is in general an over-determined problem with $N \gg K$, meaning that the number of spectral data points (typically $\gtrapprox$ 100) is much greater than the resolvable number of chemical components in the mixtures.

Given $\mathbf{Y}$ and possibly information about $\mathbf{A}$ and $\mathbf{C}$ through physical domain knowledge or reference measurements, for spectroscopy applications, the goal is to infer either the complete $\mathbf{A}$ or $\mathbf{C}$, partial entries in $\mathbf{A}$ or $\mathbf{C}$, or some functional mappings of $\mathbf{A}$ or $\mathbf{C}$. Although our modeling is based on linear assumptions, it is standard to extend the functionality of the associated algorithms into the nonlinear regime through techniques such as the kernel method [32]. In the following parts, we briefly discuss several canonical approaches for dealing with various facets involving Equation 1.3.

1.3.2 Explicit Modeling through Reference Measurements

For mixtures with fixed and known compositions, **A** in Equation 1.3 may be explicitly measured through independent experiments. This can be performed by measuring the components in pure form at high concentrations or quantities, and construct the **A** matrix accordingly. Similar to estimating **S** from **I** in Equation 1.2 in the computational spectrometer and wavemeter approaches, the problem then becomes what is typically encountered in mathematical inverse problems, and many techniques exist that are able to solve it efficiently and accurately [29]. For example, the classical least squares (CLS) estimation can be used to directly calculate the pseudo-inverse of the **A** matrix such as the ones described in Lee et al. [33], Miller and Miller [34], Feng et al. [4]. For measurements with considerable noise that may corrupt the pseudo-inversion, techniques based on singular value decomposition (SVD) or ridge regression can be performed to regularize the solution.

In order for explicit modeling to perform well with high accuracy, it is desirable to obtain the **A** matrix with the exact components in the mixture. In CLS, this is typically obtained by selecting library spectra in **A** carefully with domain knowledge until the point where the residue from the pseudo-inversion reconstruction does not contain any distinguishable spectral shapes. This process can be labor-intensive, and may be prone to human judgment bias. A slightly different approach can be to construct a large library spectra set that covers all the component spectra, plus additional spectra from substances that may not present in the mixture, and use regularizers with sparsity constraints for component selection. Popular choices such as the l_1 regularization can be solved with extremely efficient algorithms for these types of problems [35].

In general, other than instances where the mixture components are known with high confidence and are easy to measure in pure form, explicit modeling is not used as often as some of the supervised learning algorithms as will be discussed next.

However, the process of explicit modeling with CLS-like algorithms can lead to mixture understanding from a physical level, which is often missed with the alternative approaches.

1.3.3 Supervised and Unsupervised Learning Algorithms

Supervised Learning

The linear mixture model assumed above implies that the concentration or quantity information pertaining to any mixture component can be obtained with a linear operation for a given mixture spectrum $\mathbf{y}$. This means that a multivariate calibration model can be constructed without the need to measure the spectrum from any constitute component in the mixture. This type of approach falls under the supervised learning catalog. In regression settings, assuming that there are M independent mixture samples, we have

$$\mathbf{c} = \begin{bmatrix} c^{(1)} \\ c^{(2)} \\ \vdots \\ c^{(M)} \end{bmatrix} = \begin{bmatrix} — & \mathbf{y}^{(1)\mathsf{T}} & — \\ — & \mathbf{y}^{(2)\mathsf{T}} & — \\ \vdots & \vdots & \vdots \\ — & \mathbf{y}^{(M)\mathsf{T}} & — \end{bmatrix} \mathbf{b} + \boldsymbol{\epsilon}' = \mathbf{Y}\mathbf{b} + \boldsymbol{\epsilon}', \tag{1.4}$$

where $\mathbf{c} \in \mathbb{R}^M$ represent the concentration or quantity information for the component of interest in the M mixture samples, $\mathbf{Y} \in \mathbb{R}^{M \times N}$ is the mixture spectra matrix where each row corresponds to spectrum from a mixture sample (a constant term can be appended to each row for offset modeling, which we omit here for simplicity), $\mathbf{b} \in \mathbb{R}^N$ is the regression vector, and $\boldsymbol{\epsilon}' \in \mathbb{R}^M$ is the estimation noise. Here, $\mathbf{c}$ are typically obtained through separate reference measurements with standard chemical assays such as high performance liquid chromatography (HPLC). For models involving quantification for multiple components in the mixture, assuming that there are L

components of interest, we have

$$\mathbf{C} = \begin{bmatrix} - & \mathbf{c}^{(1)\intercal} & - \\ - & \mathbf{c}^{(2)\intercal} & - \\ \vdots & \vdots & \vdots \\ - & \mathbf{c}^{(M)\intercal} & - \end{bmatrix} = \mathbf{YB} + \mathbf{E}',$$

where $\mathbf{C} \in \mathbb{R}^{M\times L}$, $\mathbf{B} \in \mathbb{R}^{N\times L}$, and $\mathbf{E}' \in \mathbb{R}^{M\times L}$. Rows in $\mathbf{C}$ correspond to the concentrations or quantities for all the components of interest in the mixture sample. Meanwhile, columns in $\mathbf{B}$ are the regression vectors for the corresponding components of interest.

Depending on the application scenario, the dimensionality of the mixture spectra matrix can be drastically different. For example, for hyper-spectral image datasets, it can often be the case that $M \gtrapprox N$, meaning that the number of independent measurements can be greater than the number of spectral data points in each measurement. On the other hand, for bio-medical spectroscopy applications, where spectra are taken at high resolutions and spectral samples and reference measurements can be labor-intensive or expensive to acquire, $M \lessapprox N$, or even $M \ll N$, meaning that the number of independent measurements is usually smaller than the number of spectral data points. However, regardless of the dimensionality of the mixture spectra matrix, as indicated in Equation 1.4, $\mathbf{Y}$ is often of much lower rank than either M or N. This means that the number of resolvable chemical components in mixtures is much less than the number of independent measurements or the number of spectral data points. As a result, the multicollinearity issue in multivariate regression analysis is extremely common amongst spectroscopy datasets.

There have been many algorithms and their variants being developed and perfected over the years for multivariate regression analysis, some of which have been widely used with spectroscopy datasets. Examples include partial least squares regression (PLSR) [36], principle component regression (PCR) [37], artificial neural networks (ANNs) [38], support vector regression (SVR) [39] amongst others [34]. With all these algorithms, a training process, which is called “calibration” in the chemometrics community, is first carried out for model construction and hyperparameter selection. The resulting model is subsequently evaluated with cross-validation, bootstrap, or independent test sets. For these algorithms, the model construction and selection process essentially tries to search the optimal subspace that represents the mixture spectra training data, and performs predictions based on projections onto the trained subspace. For supervised classification tasks, the general treatment is similar to that of the regression methods, which we do not discuss in further details.

Unsupervised Learning

Another family of algorithms that has found wide usage in spectral data analysis, especially in the hyper-spectral unmixing community, falls into the unsupervised learning catalog. In these algorithms, **A** and **C** from Equation 1.3 are estimated directly from the observed mixture data **Y**. Afterwards, the estimated spectral bases in **A** are often compared to a spectral library for component (which are called "endmembers" in the hyper-spectral unmixing community) identification [40]. These unsupervised learning techniques are often of the form of matrix factorization, where the individual algorithms are the results of the different factorization criteria. For example, independent component analysis (ICA) is based on assumptions of the statistical properties of the underlying subcomponents [41], whereas the popular vertex component analysis (VCA) for hyper-spectral unmixing comes as a result of geometric constraints [42].

The successful application and convenience of these unsupervised learning algorithms for spectral data unmixing come with relatively strong requirements on the spectral dataset. As an example, many geometric constraint-based unmixing algorithms assume that the simplex formed by the spectra from the pure mixture components has the minimum volume amongst all the possible simplexes that enclose the spectral data cloud [43]. A natural premise for the validity of this assumption is that there needs to be significant variations for the the component abundances in the mixtures, which is often the case for hyper-spectral geological remote sensing, but may fail in other spectroscopy applications, where spectral signals from certain components only have small variations across the mixtures. As a result, compared to the supervised learning algorithms discussed in the previous part, the application of unsupervised learning algorithms for spectral data unmixing is generally more limited. However, many of the canonical unsupervised learning algorithms such as the principle component analysis (PCA) and non-negative matrix factorization (NMF) can be used for dimensionality reduction purpose in conjunction with other supervised regression or classification algorithm. In addition, these techniques can also be applied for other purposes such as spectral artifact removal [44]. Therefore, it is not uncommon to see them playing a significant role in a spectral processing pipeline for any spectroscopy field.

1.3.4 Spectral Shape Modeling

Despite the general popularity of the above-mentioned training-based multivariate learning algorithms for spectral data analysis, there are limitations that prevent them from being effective or optimal for certain applications. The dependence on the training process means that sufficient mixture spectral data together with the ground truth measurement need to be collected first, possibly in large volume, before a reliable model is built. The process of training data collection itself could be prohibitively expensive or labor-intensive. For example, when using Raman spectroscopy as an on-line tool for monitoring the nutrient and metabolite concentrations in biopharmaceutical processes, the performance of PLSR improves significantly with more training samples at the expense of running the process multiple times [45]. In addition, one might need to rerun the training data collection process if certain aspects of the experiment is later modified, e.g. if the growth medium composition is changed in the biopharmaceutical process monitoring example. In these situations, it is therefore preferable to have an analytical method that can directly perform analyte quantification from the mixture spectrum without a large training pool to begin with in these situations.

Generally speaking, the optical spectrum exhibits spectral features originating from the underlying physical light-matter interactions as shown in Figure 1-1. These spectral features can often be modeled explicitly using mathematical profile functions.

For example, the Gaussian, Lorentzian, and Voigt (which is the result of convolution from the Gaussian and Lorentzian profile) profile functions have long been used to model the spectral shapes in vibrational spectroscopy [46, 47, 48]. More recently, other forms of line shapes such as exponential [49] and asymmetric [50] curves have also been used to model absorption spectra for various applications. Direct spectral shape modeling connects the spectral data to the energy transitions induced through optical probing and therefore can be an effective way for quantitative analysis and decision making.

For complex spectra that may contain a multitude of spectral lines and peaks with additional background or baseline signals, statistical tools such as Bayesian inference and modeling can be used for accurate spectral shape modeling. Bayesian inference and modeling has been extensively used for curve shape modeling and fitting [51]. Its adaptation in spectral signal analysis has found a wide range of applications in mixture study [52], peak identification and quantification [53, 54], and noise analysis [55] amongst others. For complex spectra modeling, one of the key issues is to correctly identify and assign an unknown number of spectral lines and peaks in the presence of noise. This is an instance where model selection has to be performed in order to optimally assign the peak and line signals. In the literature, approaches such as reversible jump Markov chain Monte Carlo (RJMCMC) [56, 57], exchange Monte Carlo [53] and sequential Monte Carlo [54] coupled with suitable model selection criteria have been used for this purpose.

Spectral shape modeling represents one of the approach methodologies where physical domain knowledge is directly used for spectral data analysis. This is different from the machine learning-like or data-driven approaches, where the underlying physical nature of optical spectroscopy is more or less ignored in the modeling process (with the exception of linear or non-linear signal mixing). With spectral shape modeling, the spectral signal for every spectrum can be modeled independently. This can drastically reduce algorithm dependency on data volume, which can be crucial for resource-intensive applications as mentioned earlier. In addition, prior knowledge can be incorporated in the modeling process, allowing task-specific algorithms to be designed that maximumly utilize available information in a resource-optimal manner. On the other hand, learning-based approaches are generally much easier to apply and can achieve near-optimal performance with large-volume and high-quality datasets.

Similar to many other statistical disciplines, the algorithm choice for quantitative spectral data processing and analysis is a highly task-dependent problem. Although algorithms such as PLSR and PCA have had wide success in chemometrics research and in many cases are the default go-to algorithms for a new spectral dataset, there is still room for algorithm development and experimental design when it comes to a specific problem. As mentioned earlier, taking advantages of the intertwined nature of physical domain knowledge and its implications on experimental design and statistical treatment can often lead to novel processing algorithms with fewer resource requirements and more robust outcomes than the standard processing algorithms. This can sometimes be crucial for either verifying an experimental idea or demonstrating statistical robustness for a certain spectroscopic application. On the other hand, even with standard processing algorithms such as PLSR, the success of designing a generalizable processing pipeline and validation scheme for a particular dataset and application often relies on good understanding of the statistical nature of the dataset as well as the instrument limits. In the age where machine learning algorithms with efficient implementations are extremely accessible, this is especially important as careless usage of these algorithms can easily lead to over-optimistic results due to issues such as overfitting. It is worth noting that unlike many other statistical disciplines, algorithm choice can impact experimental design and data collection in significant ways in applied spectroscopy and chemometric analysis. As a result, the optimal solution to a spectroscopic application often involves heavily interconnected elements from instrument selection, experiment design, and quantitative analysis. This highlights the unique position of chemometrics at the crossroad of physical science and modern statistical disciplines.

1.4 Case Study: Skin Raman Spectroscopy and Its Biomedical Applications

Out of the large optical spectroscopy family, Raman spectroscopy is perhaps one of the most chemically specific and instrumentally demanding techniques. With recent technological advances in compact and high-power single-mode diode lasers, thin-film optical filters with high suppression properties, high-precision holographic gratings with strong diffraction strengths, and sensitive and low-noise CCD and CMOS detector arrays, Raman spectroscopy has experienced a tremendous boost in application proposals and realizations, especially in the biomedical domain [58, 59]. Figure 1-6 shows the energy transition diagrams for the various light-matter interactions occurred in Raman spectroscopy with biological sample and material. Owning to inelastic photon scattering, Raman spectroscopy is able to directly probe the vibrational and rotational states of molecules in the spectral domain. However, the inelastic photon scattering is much less likely to occur than elastic (Rayleigh) photon scattering, resulting in weak spontaneous Raman signal strengths in a typical scenario. With biological tissues and materials, sample autofluorescence due to intrinsic fluorophores such as NADH, flavin and aromatic amino acids [60, 61] is ubiquitous and often accompanies the Raman light [62]. The shot noise associated with the autofluorescence may overshadow the Raman signal and complicates the analysis [63]. This is one of the main issues faced by spontaneous Raman spectroscopy for sensitive measurements. On the positive side, unlike spectroscopic techniques such as near-infrared absorption spectroscopy, the Raman spectrum of a chemical usually exhibits distinct and highly specific sharp spectral peaks in the probing region corresponding to various energy transition levels, which some would refer to as the "Raman fingerprint" of the chemical. Such high specificity enables universal molecular identification and quantification across a wide range of biomarkers of interest in a label-free fashion. In addition, the excitation wavelength for Raman spectroscopy can be chosen to either resonantly enhance the interaction strength from certain molecules of interest, or to achieve high penetration depth by avoiding absorption from water and other tissue materials. These features are perhaps the main reasons for the popularity of Raman spectroscopy despite of its weak signal strength relative to other optical spectroscopy techniques.

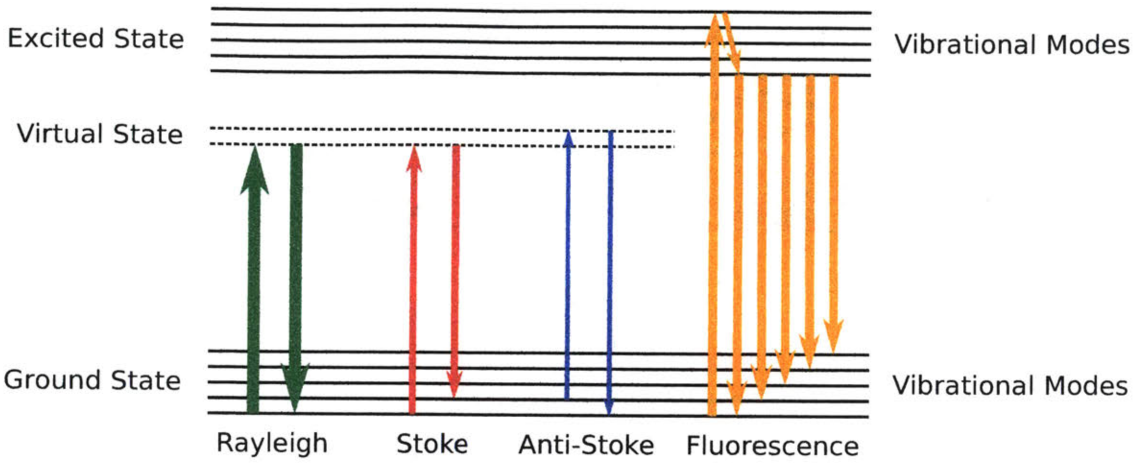


Figure 1-6: Energy transition diagrams for the various light-matter interactions occurred in Raman spectroscopy with biological sample and material. The arrow width for each process corresponds to its relative interaction strength.

1.4.1 Raman and Autofluorescence Signals from Skin

For non-invasive diagnostic and monitoring applications with optical techniques, skin is the most easily accessible human organ for probing. As shown in Figure 1-7 [64], human skin consists of three layers: the epidermis, dermis, and hypodermis (subcutaneous) layer. The outermost epidermis layer is human body's main protection barrier against dehydration and environmental factors such as microbial, chemical, and UV light. It ranges from ≈ 50 μm to 2 mm in thickness for different parts of the body. It is composed of 4 to 5 stratified layers with the outermost being the stratum corneum. The stratum corneum is usually ≈ 10 to 20 μm in thickness and mostly contains dead and flattened cells filled with keratin. The dermis lays beneath the epidermis layer and mainly consists of collagen, elastin, extrafibrillar matrix, as well as cells like fibroblasts, macrophages, and adipocytes. It supports the epidermis layer with both strength and elasticity. At last, the hypodermis, or the subcutaneous layer, supports all the functional skin layers and acts as the main energy store and insulating layer.

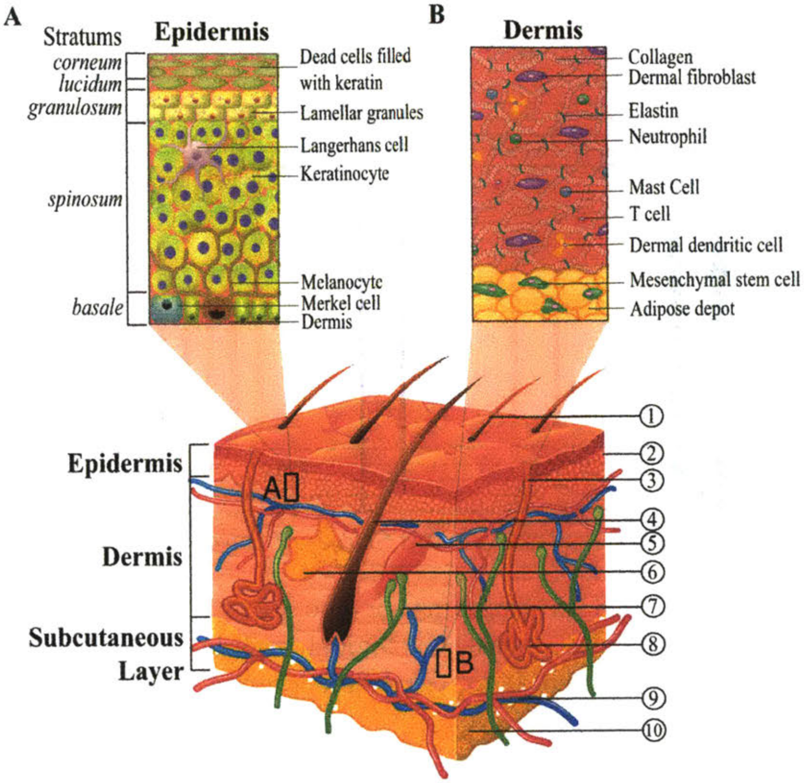


Figure 1-7: Skin layers and structures [64]. Human skin consists of three layers: the epidermis, dermis, and hypodermis (subcutaneous) layer.

Despite the highly heterogeneous composition structure of human skin as discussed above, light transport in these layers across different wavelengths has been well characterized and modeled [65]. For example, multi-layer photon transport models, where absorption and scattering properties for each individual layer can be specified, have been extensively used to study light-tissue interaction [66]. In Raman spectroscopy, photons generated through the inelastic scattering process during the light-tissue interaction contain molecular information of the corresponding interaction region. In the meantime, autofluorescence due to intrinsic fluorophores in the tissue is also present alongside the Raman signal in a collected spectrum. These optical signals encode the underlying physical and biochemical interaction process and can be used to infer the molecular composition and concentration. For example, Figure 1-8 shows *in vitro* and *in vivo* Raman spectra measured from different layers and depths in skin using confocal geometries [67, 68]. An application where water content profile in the stratum corneum is estimated based on the Raman measurement is also shown in the same figure [68]. It is not difficult to imagine that these spectral signals, once equipped with suitable analytical and quantitative tools, can be used to construct a full skin molecular profile in an non-invasive and label-free manner. Several diagnostic applications based on skin Raman and autofluorescence signals will be discussed in the next part.

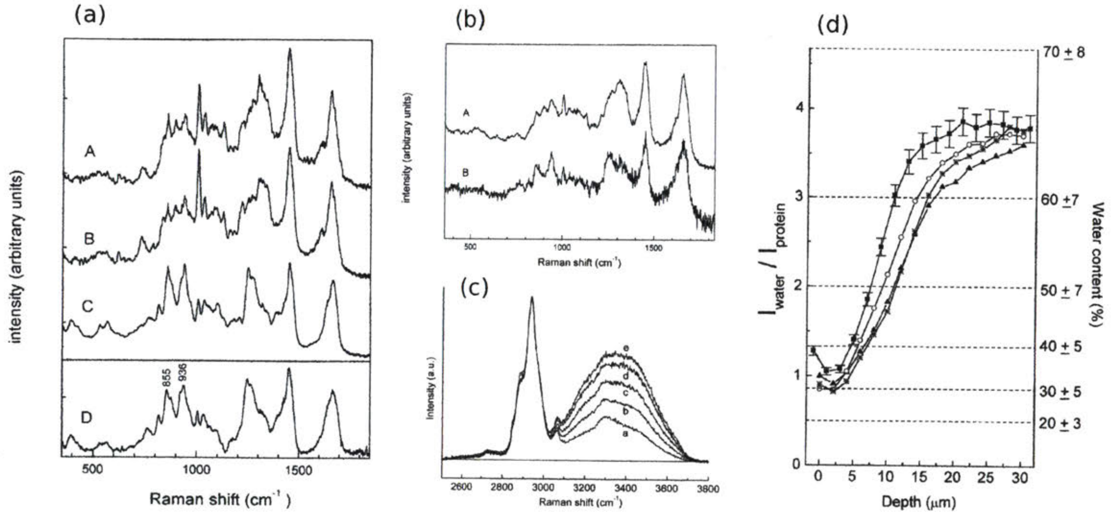


Figure 1-8: *In vitro* and *in vivo* Raman spectra from different layers and depths in skin and their application in water content profiling. (a) *In vitro* Raman spectra for (A) stratum corneum, (B) epidermis, and (C) dermis. (D) is for Type I collagen [67]. (b) *In vivo* Raman spectra for (A) stratum corneum at ≈ 0 μm depth and (B) dermis at ≈ 85 μm depth [67]. (c) *In vivo* Raman spectra of stratum corneum at various depths showing the differences in water content [68]. (d) Water profiles based on the Raman measurements in the stratum corneum [68].

1.4.2 Skin Raman and Autofluorescence as Diagnostic Tools

There have been numerous pursuits trying to utilize the non-invasive and label-free characteristics of skin Raman and autofluorescence spectroscopy for fast disease diagnostics and wellness monitoring applications. One of the most promising examples is to use the Raman signals for skin cancer diagnostics and classification. Zhao et al. [69] developed a real-time Raman spectroscopy system with an off-axis excitation geometry

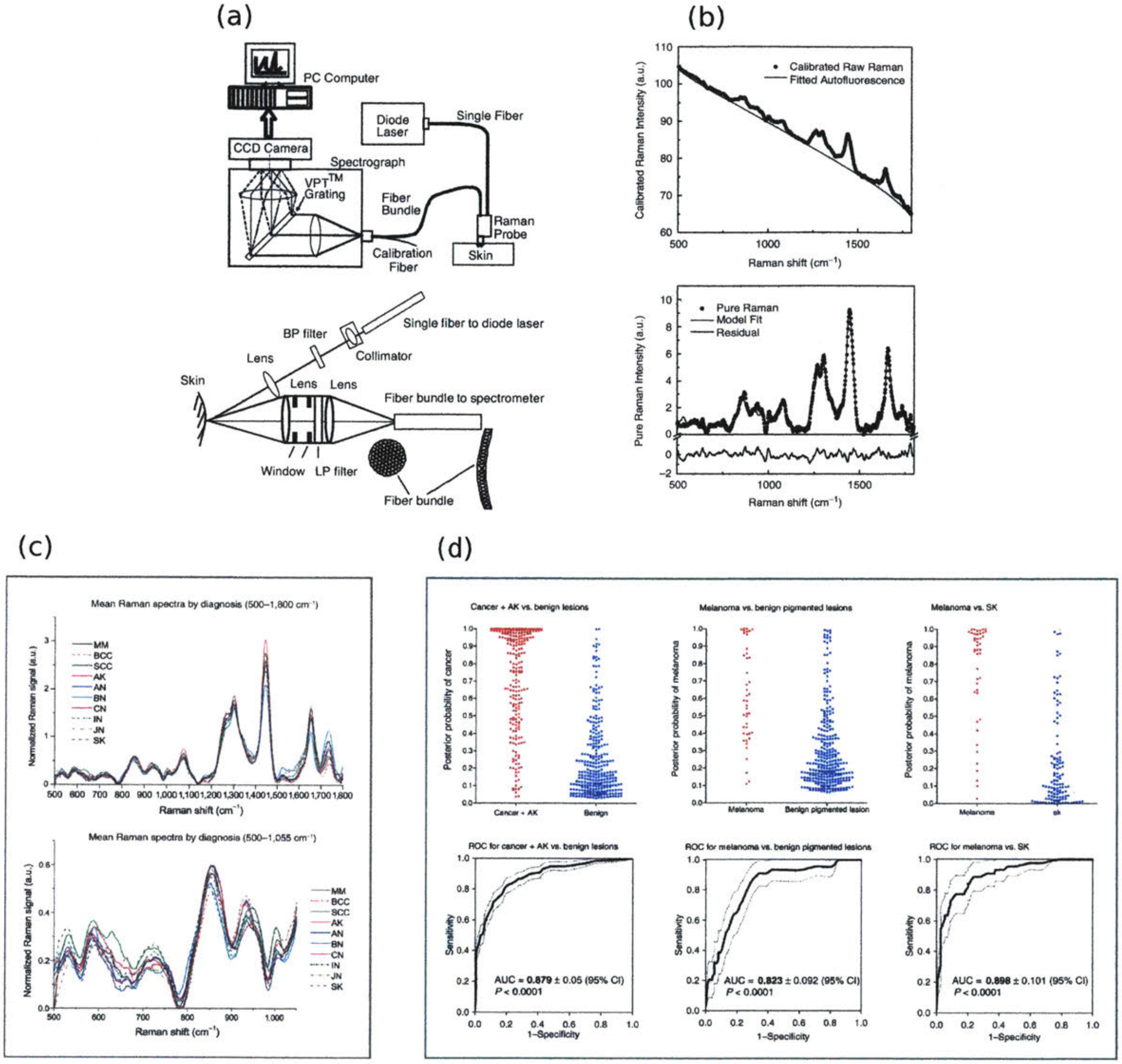


Figure 1-9: A real-time clinical Raman system and its applications in skin analysis and skin cancer diagnostics [69, 70]. (a) Configuration and setup for the real-time clinical Raman system [69]. (b) Raman spectrum from skin and its model fitting with oleic acid, palmitic acid, collagen I, keratin, and hemoglobin [69]. (c) Raman spectra for various skin lesion samples with benign or malignant conditions with the real-time clinical Raman system [70]. (d) Lesion classification posterior probabilities and receiver operating characteristic (ROC) curves based on Raman spectra [70].

for *in vivo* skin diagnostics and analysis applications. In their work, Raman spectra from human volar forearm were modeled with linear mixing from reference Raman measurements of oleic acid, palmitic acid, collagen I, keratin, and hemoglobin. These are shown in Figure 1-9 (a) and (b). In a follow-up study, they used this instrument to demonstrate *in vivo* skin cancer diagnosis with 518 benign and malignant skin lesion samples from 453 patients [70]. Principle component with generalized discriminant analysis (PC-GDA) and PLS were used for classification analysis. Their results show that successful differentiation between (1) malignant/pre-malignant cancers and benign lesions, (2) melanomas and benign pigmented lesions, and (3) melanomas and seborrheic keratoses are achievable with non-invasive and real-time spectral acquisition and analysis (Figure 1-9 (c) and (d)).

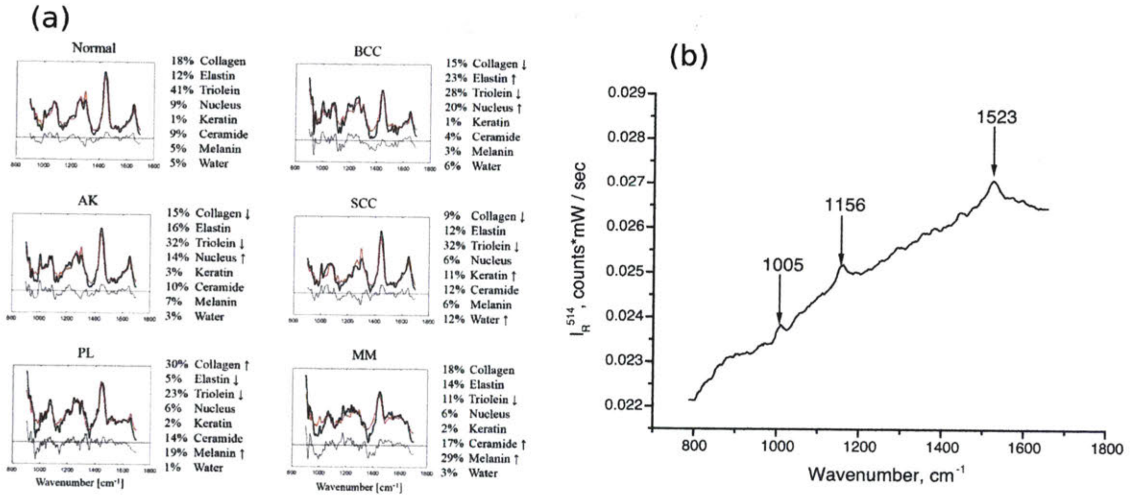


Figure 1-10: (a) A biophysical model based on Raman spectrum from skin and its application in skin cancer Raman spectrum modeling [4]. (b) *In vivo* Raman features from carotenoids beta-carotene and lycopene from human skin with resonance Raman excitation [71].

Through physical modeling of the active components in skin Raman spectra, the composition and its abundance differences can be visualized for various cancer and normal tissue samples. Feng et al. [4] developed a Raman biophysical model based on *in vivo* skin cancer screening data and *in situ* skin constituent measurements. They concluded that collagen, elastin, keratin, cell nucleus, triolein, ceramide, melanin and water were the most important model components of skin Raman spectra. Their relative abundances in skin Raman spectra can be used to help identify normal, benign, and malignant skin tissue samples, owning to their biochemical and structural differences (Figure 1-10 (a)). In addition to the above-mentioned literatures, there are a number of other reports on successful identification and classification of skin cancer with *in vivo* Raman spectroscopy such as Gniadecka et al. [72], Lieber et al. [73].

Other than skin cancer-related applications, resonance Raman spectroscopy in the visible range has been applied to evaluate carotenoids such as beta-carotene and lycopene levels *in vivo* from human subjects (Figure 1-10 (b)) [74, 71]. Carotenoids are believed to play important roles in the anti-oxidant defense system of the skin, and some of them have been proposed to be potential biomarkers for diseases such as prostate cancer [74]. In addition, the carotenoid level in an individual is also a reflection of his/her lifestyle. As a result, non-invasive Raman skin measurement has been proposed as an assistive tool for medical evaluation and therapy control in cosmetic treatments [75].

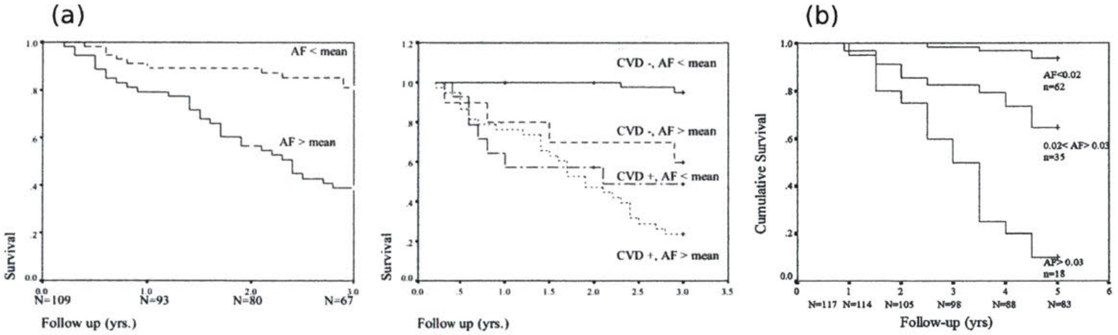


Figure 1-11: Evidence of skin autofluorescence as biomarkers for mortality in hemodialysis and diabetic patients. (a) Mortality of hemodialysis patients as a function of follow-up years with respect to the skin autofluorescence (AF) level and presence of cardiovascular disease (CVD) at baseline [76]. (b) Mortality of diabetic patients as a function of follow-up years with respect to the skin autofluorescence (AF) level [77].

Apart from the Raman signal, the autofluorescence signal from skin has also been proposed as biomarkers for disease monitoring and prediction. Meerwaldt et al. [76] suggested to use skin autofluorescence measurements as biomarkers for cumulative metabolic stress and advanced glycation end products (AGE), which are believed to link to chronic complications in diabetes and end-stage renal disease. They monitored the mortality rate of 29 dialysis patients in a period of 3 years and concluded that the skin autofluorescence can be an independent predictor of mortality in these situations (Figure 1-11 (a)). The same research group also evaluated skin autofluorescence as predictor for cardiac damage and mortality in diabetic patients. In one study, they conducted clinical assessments with 973 Type II diabetic patients and 231 control subjects, and concluded that skin autofluorescence is a biomarker for vascular damage in Type II diabetic patients [78]. In another work, they studied 48 Type I and 69 Type II diabetic patients and 43 control subjects, and followed up in a period of 5 years. They concluded that strong associations can be found between skin autofluorescence and cardiac mortality in diabetic patients (Figure 1-11 (b)) [77]. They linked skin autofluorescence with cumulative metabolic burden and suggested to use it for risk assessment and control.

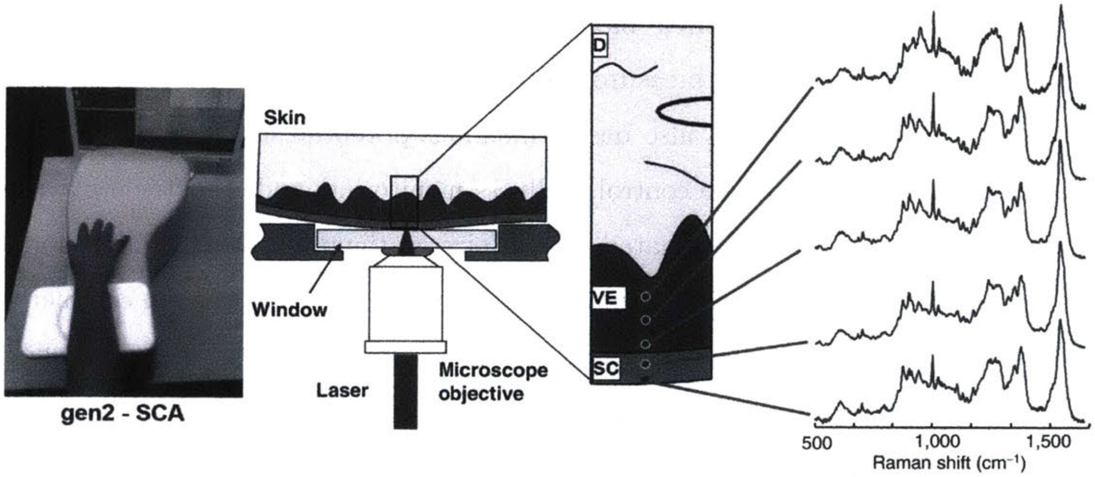


Figure 1-12: The gen2-SCA system from RiverD International B.V. for skin analysis and diagnostics [79].

The above examples just represent a small portion of the overall landscape for the biomedical applications for with Raman and autofluorescence spectroscopy. Commercially, the gen2-SCA system from RiverD International B.V., shown in Figure 1-12, is a confocal Raman system specifically for *in vivo* skin diagnostics and analysis [1]. More broadly, the application for Raman and autofluorescence spectroscopy in biomedical research in general has been rapidly expanding. Ellis et al. [58], Kong et al. [59] are excellent recent review articles on this general topic.

[1] Company link: https://www.riverd.com/

1.4.3 Skin Raman Spectroscopy for Non-Invasive Glucose Estimation

With the several skin condition monitoring and diagnostic applications discussed in the previous part, we turn our attention to the area of skin Raman spectroscopy for non-invasive glucose estimation in this part. Non-invasive and continuous glucose measurement is perhaps one of the most heavily sought-after non-invasive health monitoring domains in the modern era [80, 81]. The ability to non-invasively measure blood glucose level not only has transformative impact on diabetic patients that need to constantly monitor their blood glucose level through minimally invasive techniques such as fingerstick measurements or completely invasive techniques such as transplantable devices, but also has tremendous potential with the non-diabetic community in terms of lifestyle control, wellness monitoring, and diabetes prevention. However, despite having an estimated market size on the order of ten billion dollars and at least millions of potential consumer pool, there is still no successful commercial device in the market [82]. This is not due to a lack of innovative and ingenious efforts from the research community, but rather reflects the highly complex nature of this problem. Smith [82] provides a critical review in this area that highlights the many challenges associated with this problem. As many proposed sensing techniques require multivariate calibration and analysis for glucose concentration estimation, it is no surprise that statistical processing and validation schemes play crucial roles alongside the detection scheme itself in these approaches.

Being one of the most promising non-intrusive optical techniques with high specificity, it is expected that the problem of non-invasive glucose estimation has attracted considerable attentions from the Raman spectroscopy community. The earliest report of *in vivo* non-invasive glucose measurement with Raman spectroscopy was by Enejder et al. [1] in 2005. In this pioneering work, Raman signals with 830 nm excitation were collected from the forearms of 20 healthy volunteers in oral glucose tolerance test (OGTT) experiments. 17 of the 20 measurement sessions were used for performance evaluation. PLSR was used for leave-one-out cross validation. Individual-specific as

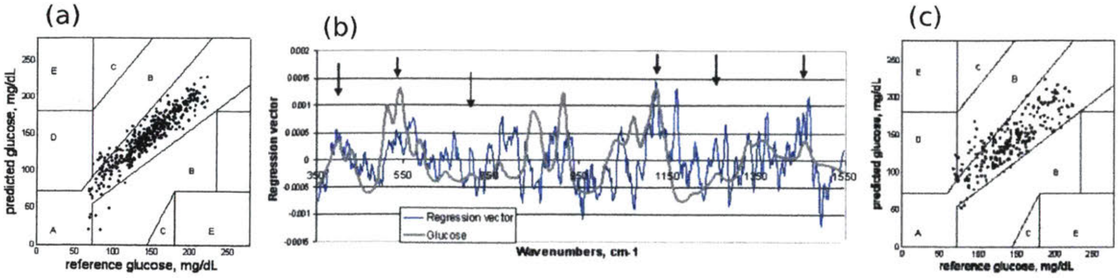


Figure 1-13: Non-invasive Raman glucose estimation results from Enejder et al. [1]. (a) Clarke Error Grid Analysis plot for leave-one-out cross validation and individually calibrated models. (b) Glucose Raman spectrum and regression vector for the individual session with the highest validation accuracy in (a). (c) Clarke Error Grid Analysis plot for leave-one-out cross validation and a globally calibrated model with 9 selected volunteers.

well as universal calibration models were both reported. Overall, the validation results showed promising signs (Figure 1-13). However, in a truly predictive setting, data from an entire session should be left out for model testing to minimize spurious correlations and to reduce the chance of overfitting. Their model performance under this setting was not reported. In addition, their regression vector, shown in Figure 1-13 (b), did not show convincing features from glucose Raman spectrum. After this work, the group followed up with a number of incremental improvements with superb correlation results from the same dataset in the next decade, yet no satisfactory performance in a truly predictive setting was ever reported. The dataset was also not made publicly available for any third-party to explore this area. It is also worth pointing out that most literature in non-invasive glucose measurement with Raman spectroscopy were from members in the same group on the same dataset.

In 2009, C8 MediSensors, a then startup company in Los Gatos, California, published a study on non-invasive glucose measurement with Raman spectroscopy [2]. Shifted laser excitation was used for fluorescence rejection and water Raman signal with a visible laser at 670 nm was used for normalization (Figure 1-14 (a)). Instead of the OGTT experiments, they performed 58 glucose clamp studies with 30 diabetic patients. This resulted in much wider glucose variations as compared to the case in Enejder et al. [1]. A linear analytical model with the inclusion of a single-pole delay element was used for data analysis. In their test scheme, a true predictive setting

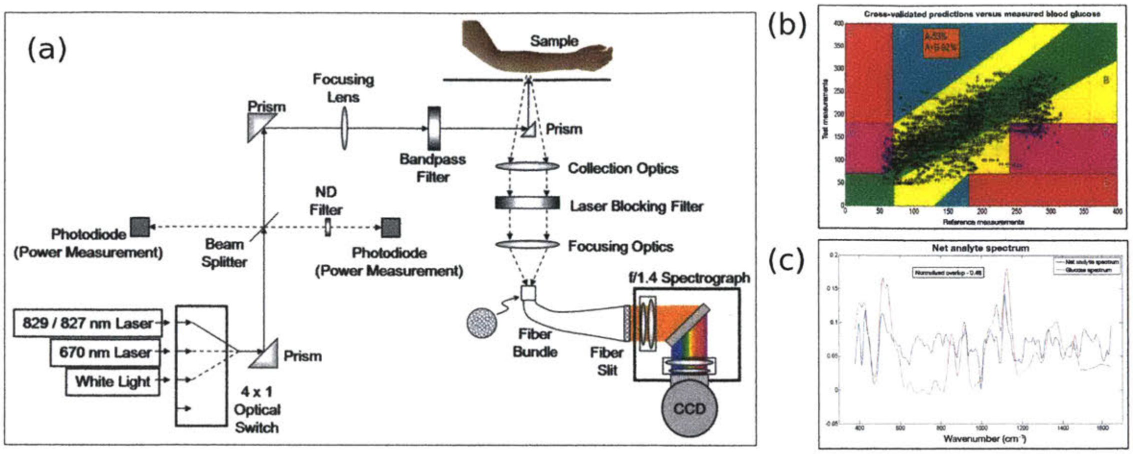


Figure 1-14: Non-invasive Raman glucose estimation experimental setup and results from Lipson et al. [2]. (a) Schematic diagram for the experimental setup. (b) Clarke Error Grid Analysis plot with leave-one-session-out cross validation and a globally calibrated model. (d) The net analyte spectrum and glucose Raman spectrum.

was used where in each test iteration, data from an entire experimental session was left out for performance test. Their predictive results are shown in Figure 1-14 (b) with a median absolute error of ≈ 1.7 mM. While a clear correlation can be observed with their dataset, it is still considered too low by medical professionals in terms of estimation accuracy [82]. Unfortunately, the company filed for bankruptcy in 2013 after a series of misfortunes and departures from leading roles [82], and is no longer in existence.

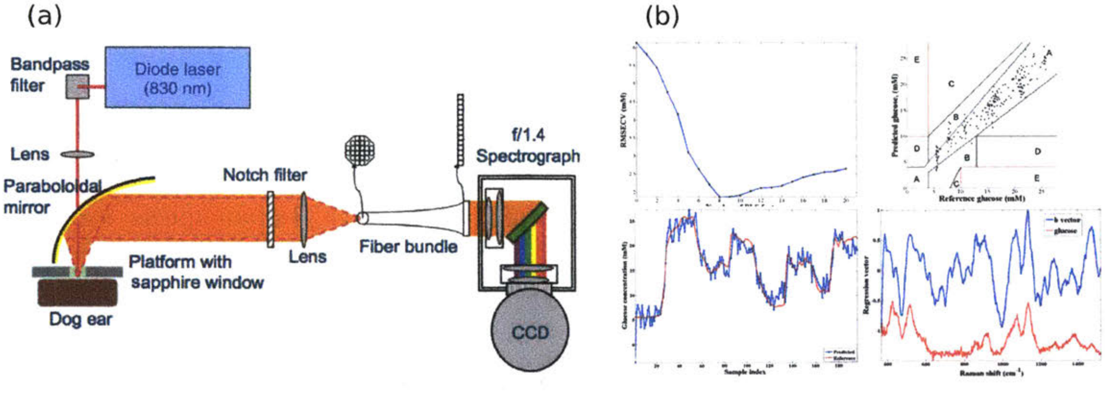


Figure 1-15: Non-invasive Raman glucose estimation experimental setup and results with a dog model from Shih et al. [3]. (a) Schematic diagram for the experimental setup. (b) Validation results with leave-one-level-out cross validation and the regression vector plot.

In addition to human experiments, *in vivo* experiments have also been carried out on animal models. In a recent experiment [3], a beagle was put in anesthesia and its blood glucose concentrations were clamped at various levels in the range from 5.6 to 25.6 mM. The experimental setup was similar to Enejder et al. [1] and is shown in Figure 1-15 (a). The laser power in this study was undisclosed. The CCD integration time per frame was 1.8 s. This was much shorter than those used in similar studies. Leave-one-level-out cross validation was carried out with PLSR and the results are shown in Figure 1-15 (b). Overall, a validation error on the order of $\approx$ 1.5 to 2 mM was obtained. The regression vector plotted in Figure 1-15 (b) shows distinct glucose Raman features. This is in strong contrast compared with the one shown in Figure 1-13 (c). The reported estimation error, however, was worse in this study than those in Enejder et al. [1] and its follow-ups (though they were under slightly different test schemes).

Overall, while there is evidence that it is possible to detect glucose signals from transcutaneous *in vivo* Raman measurements in the literature, the overall quantitative results and their operating conditions reported by different authors are mixed. On the quantitative data analysis side, the details on the training and calibration process, such as how the cross validation process was carried out, were left vague by some authors, making it difficult to evaluate. The closed source nature of the existing datasets in the community significantly hinders collective understanding and exploration of the problem. This is different from many modern data analysis-intensive fields, where data and algorithm sharing have become standard procedures. Similar observations have also been noted by authors in related Raman areas [83]. In addition, few authors have associated estimation and quantification results with key system performance metrics such as glucose signal-to-noise ratio (SNR). This makes result reproduction and verification difficult. Non-invasive glucose estimation with Raman spectroscopy is a problem where in addition to multivariate statistical analysis, there are interconnected elements from optics, device and instrument engineering, biochemistry, and physiology. With such a highly complex nature, it is especially important for the community to address related issues in a more quantifiable and verifiable manner.

1.5 Thesis Overview

In this thesis, we aim to address some of the challenges mentioned in this chapter by combining modern computational and statistical techniques with physical domain knowledge. In particular, we focus on three aspects where computational or statistical knowledge have either enabled realization of a new instrument (with a compact form factor yet still maintaining a competitive performance) or deepened statistical insights on analyte detection and quantification in highly mixed or heterogeneous environments. In Chapter 2 of the thesis, we utilize the non-paraxial Talbot effect to build compact and high performance spectrometers and wavemeters that use computational processing for spectral information retrieval without the need of a full-spectrum calibration process. In Chapter 3, we develop an analyte quantification algorithm for Raman spectroscopy based on spectral shaping modeling using a hierarchical Bayesian inference model and RJMCMC computation with minimum training sample size requirement. In Chapter 4, we numerically investigate the spectral characteristics and signal requirements for universal and predictive non-invasive glucose estimation using an *in vivo* skin Raman spectroscopy dataset. At last, concluding remarks and recommendations for future work are provided in Chapter 5.

1.5.1 Compact and High Performance Computational Spectrometers and Wavemeters Using the Talbot Effect

Starting with spectroscopy instrumentation, a compact and high performance computational spectrometer and wavemeter configuration based on the non-paraxial Talbot effect is proposed and realized in Chapter 2. The Talbot effect refers to the self-imaging phenomenon observed with coherent light after passing through period structures such as diffraction gratings [84]. Owning to its interferometric nature with invertible spatio-spectral response, it has been previously proposed as the building component for realizing spectrometers [85, 86]. However, only coarse resolutions have been reported ($\approx$ 42 nm in Kung et al. [85] and $\approx$ 20 nm in De Nicola et al. [86]) with non-compact geometries. As shown in Figure 1-16, the Talbot effect uses a diffraction grating to discriminate the spectral components – but as opposed to dispersive grating spectrometers that sample the diffraction pattern in the far field, the Talbot spectrometers sample in the near/mid field where the diffraction orders are not spatially separated. Computational inversion is therefore required for spectrum retrieval. With recent technological advances in CCD and CMOS image sensors, the sensor pixel sizes have reached to a point where direct sampling of the Talbot pattern is achievable without any external imaging optics like the ones in previous studies [85]. In addition, operating the spectrometer configuration in the non-paraxial regime with a tilted image sensor allows full-frame interferogram capture without any moving parts, as strong diffractions in the non-paraxial regime limit the Talbot region to be within several millimeters after the diffraction grating. Compared with the recent compact and computational spectrometer solutions discussed in Section 1.2.2, the Talbot spectrometer and wavemeter do not require any full-spectrum calibration, which is a significant advantage for practical adoption. There are two different applications for the compact Talbot spectroscopy that will be discussed. The first one considers spectroscopy in the context of broadband light sources that can be temporally and/or spatially incoherent. The second one considers spectroscopy for the precise determination of the wavelength from a coherent signal, i.e. in the case of a wavemeter. Again, this wavemeter is useful for multiple spectroscopy applications where the laser wavelength is intentionally swept – e.g. for tunable diode laser absorption spectroscopy (TDLAS) discussed in Section 1.2.1.

There are several computational problems that naturally present themselves in the miniaturized Talbot spectroscopy solution. First and foremost, understanding the Talbot image formation through the scalar diffraction theory provides the physics-level foundation. This is provided in Section 2.1, where the generalized Talbot image formation under different incidence situations is investigated in depth through both analytical derivations as well as numerical simulations. The interferometric nature of the Talbot effect is also discussed. Next, direct sampling of the Talbot pattern for high-resolution discrimination of the spectral components not only is a computational processing problem but also dictates the choice of the sensor and its pixel dimension,

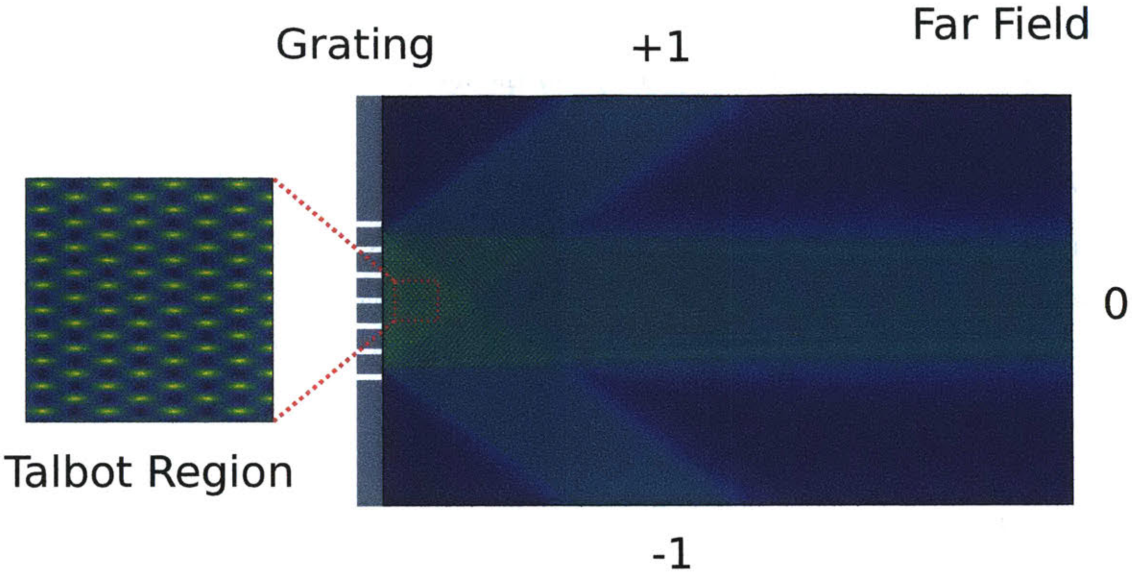


Figure 1-16: Rayleigh–Sommerfeld diffraction solution of a sinusoidal phase grating illustrating the near/mid field Talbot image formation as well as the far field diffraction.

the grating pitch, as well as the orientation of the sensor with respect to the grating. Section 2.2 presents related discussions as a hardware design problem coupled with computational processing of the sampled pattern. Afterwards, calculations and simulations showing the spectrometer performance under temporally incoherent light and incidence angular spread are presented respectively. Section 2.3 shows the experiments for Talbot spectrometer realization and the characterization results. Investigations into resolution tuning, response span, as well as spectrum reconstruction under different inputs are explored and discussed. The result represents the best resolution (sub-nanometer) seen with the Talbot spectrometer under the most compact form factor [87, 88]. Section 2.4 discusses using the Talbot effect specifically for compact and high performance wavemeter applications. Both theoretical performance estimation bounds as well as experimental results are presented and discussed. Here, tone parameter extraction algorithms are used to accurately retrieve the frequency of the periodic signal obtained in the Talbot interferogram. We experimentally demonstrate a compact and high performance wavemeter with below ≈ 10 pm estimation uncertainty with the 1-σ criterion. The chapter ends with conclusions in Section 2.5.

1.5.2 Bayesian Modeling and Computation for Analyte Quantification in Complex Mixtures Using Raman Spectroscopy

Motivated by the shortcomings of training-based multivariate regression algorithms due to their requirements on training data volume mentioned in Section 1.3.4, in Chapter 3 we propose an alternative technique to these algorithms for analyte quantification in complex mixtures using Raman spectroscopy with a Bayesian modeling and computation approach. More specifically, given *a priori* the Raman spectrum measurement of an analyte of interest, which we term as the target analyte in our text, our goal is to quantify its concentration or quantity in a complex mixture spectrum without the need of acquiring additional mixture training data, a scenario that frequently arises in various applications. In addition, the Bayesian approach allows us to simultaneously estimate both the peak and baseline signals. This is different from previous approaches in automatic baseline estimation and correction [89, 90, 91, 92, 93, 94], where baseline estimations were performed without jointly estimating the peak signals. As mentioned in Moores et al. [54], the isolated baseline estimation in these approaches may bring in potential risks of introducing bias and errors due to the fact that the actual Raman signals were ignored during the estimation and correction process.

There exist several publications aiming at bringing the Bayesian modeling framework to spectral data analysis. Razul et al. [56], Fischer and Dose [52], Wang et al. [57], Nagata et al. [53], Tokuda et al. [55] used Bayesian modeling combined with computational methods such as RJMCMC or the exchange Monte Carlo method for accurate spectrum variable estimation in various areas such as nuclear emission spectroscopy and mass spectrometry. For Raman spectral data analysis, Zhong et al. [95] used the Bayesian framework and a combined Gibbs and RJMCMC sampler to infer mixture information from a set of multiplexed surface-enhanced Raman spectroscopy (SERS) measurements. Moores et al. [54] used a sequential Monte Carlo sampler for optimal baseline correction and low-concentration analyte quantification. While building upon the common Bayesian modeling and computation principles, our work differs from these prior work due to the fact that our algorithm employs a two-stage processing for quantifying target analyte concentrations in complex mixtures as an alternative to multivariate regression methods such as PLSR with no requirement on pre-existing mixture training data [96]. In Section 3.1, we provide the hierarchical Bayesian modeling framework and RJMCMC computation procedure for our two-stage algorithm, where the first stage is used to learn the peak information for the pure target analyte spectrum and the second stage is for quantifying its concentrations in mixtures. In Section 3.2, we demonstrate the utility of this algorithm by testing its performance on a wide range of numerically generated datasets and compare its results with several multivariate regression algorithms. The advantages of our algorithm over conventional multivariate regression algorithms are established under the small training sample size regime. In Section 3.3, we report its estimation results on two experimental Raman spectroscopy datasets. The first one is a four-component aqueous mixture study. The second one is for glucose concentration estimation in biopharmaceutical process with Chinese hamster ovary (CHO) cells, which are the most widely used expression systems for industry production of recombinant protein therapeutics such as monoclonal antibodies used in cancer therapy. The chapter is then concluded in Section 3.4.

1.5.3 Numerical Investigations of Non-invasive Glucose Estimation with Raman Spectroscopy

As discussed in details in Section 1.4.3, there are many issues, either technical or non-technical, surrounding the problem of non-invasive glucose estimation with Raman spectroscopy. While the resources required for a complete and comprehensive solution to the problem are beyond what is typically available to small-to-medium research groups and startup companies [82], important questions can still be answered that can significantly contribute to the collective understanding of the problem in the research community. Our lab has developed a portable clinical Raman spectroscopy system, similar to Zhao et al. [69], for skin analysis and diagnostics applications. An *in vivo* skin Raman spectroscopy dataset from healthy volunteers in OGTT experiments was collected with the goal of non-invasive glucose detection and estimation. While no correlation could be detected from the skin Raman spectra to the corresponding fingerstick glucose reference measurements, we use this dataset as a testbed for numerical investigations into the signal requirements for universal and predictive glucose estimation models, by manually adding glucose Raman signals at different strengths to the skin Raman spectra. With this approach, quantitative conclusions can be obtained that can serve as important references for future technology development and experimental design.

In Section 4.1, the portable clinical Raman instrument and the data collection experiments are introduced. In addition, the characteristics of skin Raman spectra, including the variations observed across different individuals, the autofluorescence background and its photo-bleaching, and practical issues such as measurement movement artifacts and ambient light leakage, are presented and discussed in details. In Section 4.2, spectral signal analysis and processing methodologies are introduced and discussed. The signal generation process and variants of the training, validation, and testing schemes under our universal and predictive spectral processing pipeline are presented. Section 4.3 presents the main results on the estimation performance of our universal processing pipeline under different signal generation conditions, and their implications on signal requirements for universal prediction with skin Raman spectroscopy data. Several spectral processing variants are also discussed with recommendations provided for future clinical data analysis. At last, Section 4.4 concludes the chapter with discussions on the implications of our numerical investigations. Through our investigations, we focus on truly predictive modeling in the presence of both cross-individual variations as well as internal spectral correlation structures observed in our dataset. It is worth mentioning that while our investigation focus is on glucose estimation, our signal generation process, methodology and processing recommendations translate naturally to any potential biomarkers measurable from skin Raman and autofluorescence spectroscopy. Given the general spectral characteristics and variabilities observed from skin Raman and autofluorescence spectra across population, these information can be extremely valuable for universal and predictive non-invasive biomarker detection and estimation with skin Raman spectroscopy on a broad scale.

Chapter 2

Compact and High Performance Computational Spectrometers and Wavemeters Using the Talbot Effect

This chapter presents a comprehensive study on utilizing the Talbot effect under non-paraxial situations for building compact and high performance spectrometers and wavemeters. Section 2.1 starts with a brief review of the scalar diffraction theory, and then introduces the Talbot effect and its generalizations under tilted incidence angles. Its interferometric nature is also discussed. Section 2.2 first connects Talbot spectroscopy with Fourier spectrometers. It then presents the Talbot interferogram sampling task as a hardware selection and optimization problem with computational processing and correction for spectrum retrieval. Importantly, the effects of source temporal incoherence and angular spread incidence are numerically investigated with recommendations provided for general spectroscopy. Section 2.3 provides the experimental details for realizing and characterizing the Talbot spectroscopy system. System performances for resolution determination, geometry optimization, and response characterization under different light sources are investigated experimentally. At last, Section 2.4 discusses using the Talbot effect specifically for realizing compact and high precision wavemeters. The Cramér-Rao lower bound (CRLB) for frequency estimation with interferogram-like signals is provided. The estimation results with two different processing algorithms are presented and discussed. The chapter is finally concluded in Section 2.5.

2.1 The Talbot Effect – Theory and Simulations

2.1.1 Brief Review of the Scalar Diffraction Theory

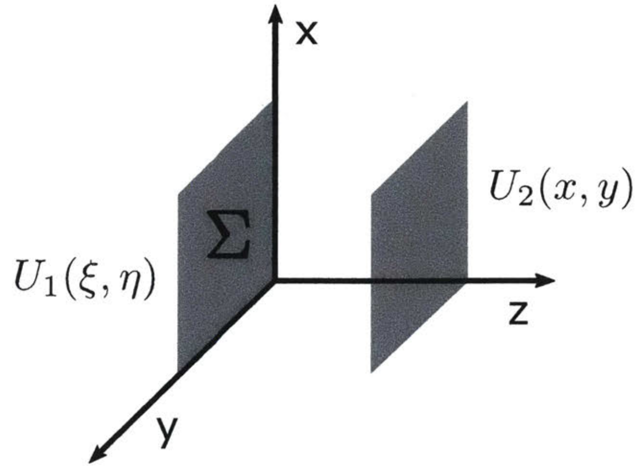


Figure 2-1: The input and output plane for scalar diffraction study.

We start our discussions with a brief review of scalar diffraction theory. For optical waves propagating through space, diffraction occurs when the light wave encounters lateral confinement or disturbance. While the exact solution of diffraction is fundamentally governed by the Maxwell's equations, accurate solutions can be obtained with the much more tractable scalar diffraction theory under mild conditions pertaining to many free-space optical systems [97]. Under the scalar diffraction theory, the propagation of optical wave is described through a scalar optical field $U(x, y, z)$ in three-dimensional space. Amongst several diffraction formulations as described in more details in Goodman [98], the Rayleigh-Sommerfeld diffraction solution states that

$$U_2(x, y) = \frac{z}{j\lambda} \iint\limits_{\Sigma} U_1(\xi, \eta) \frac{e^{jkr}}{r^2}\, d\xi\, d\eta, \tag{2.1}$$

where as shown in Figure 2-1, $U_1(x, y) = U(x, y, z)|_{z=0}$ is the two-dimensional input field, $U_2(x, y) = U(x, y, z)|_{z=z}$ is the two-dimensional output field, Σ defines the input field domain, $r = \sqrt{(x - \xi)^2 + (y - \eta)^2 + z^2}$, λ is the wavelength, and k is the wave vector. The Rayleigh-Sommerfeld diffraction in Equation 2.1 is an embodiment of the Huygens-Fresnel principle, which states that every point on the input plane acts like a spherical point source and the observation is a superposition of all the contributions from these point sources. This principle is at the heart of the wave propagation treatment employed in Fourier optics. The optical impulse response $h(x, y)$ and the optical transfer function $H(f_X, f_Y)$ for the Rayleigh-Sommerfeld diffraction are

$$h(x, y) = \frac{z}{j\lambda} \frac{e^{jkr}}{r^2},$$

and

$$H(f_X, f_Y) = \exp\left[jkz\sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2}\right]. \tag{2.2}$$

Two assumptions are inherent to the Rayleigh-Sommerfeld diffraction solution. The first one is the scalar diffraction treatment and the second one is that $r \gg \lambda$ [98]. An underlying assumption to propagating wave in Equation 2.2 is that $(\lambda f_X)^2 + (\lambda f_Y)^2 < 1$, otherwise the wave becomes evanescent in the z direction.

With modern computing resources, the Rayleigh-Sommerfeld diffraction solution can be readily applied to many scenarios without any problems. For analytical tractability and historical reasons, we introduce the Fresnel approximation to the Rayleigh-Sommerfeld diffraction solution. With Newton's generalized binomial expansion, under the paraxial approximation, r in Equation 2.1 can be approximated as

$$r \approx z\left[1+\frac{1}{2}\left(\frac{x-\xi}{z}\right)^2+\frac{1}{2}\left(\frac{y-\eta}{z}\right)^2\right].$$

By keeping the quadratic expansion terms in the exponent and dropping them in the denominator in Equation 2.1, we reach to the Fresnel approximation as

$$U_2(x,y) = \frac{e^{jkz}}{j\lambda z}\iint\limits_{\Sigma} U_1(\xi,\eta)\exp\left[j\frac{k}{2z}\left[(x-\xi)^2+(y-\eta)^2\right]\right] d\xi\, d\eta. \tag{2.3}$$

The Fresnel approximation in Equation 2.3 essentially replaces the spherical wavefronts of the Huygens-Fresnel principle with quadratic wavefronts. The optical impulse response $h(x,y)$ and the corresponding transfer function $H(f_X, f_Y)$ for the Fresnel approximation are

$$h(x,y) = \frac{e^{jkz}}{j\lambda z}\exp\left[j\frac{k}{2z}\left(x^2+y^2\right)\right], \tag{2.4}$$

and

$$H(f_X, f_Y) = e^{jkz}\exp\left[-j\pi\lambda z\left(f_X^2+f_Y^2\right)\right].$$

As a rough criterion for approximation accuracy, the Fresnel number is defined as

$$N_F = \frac{s^2}{\lambda z},$$

with s being the characteristic size (such as the half width or radius) of the input aperture. It is generally accepted that with a uniform plane wave input, Fresnel approximation can provide reasonable results when N_F is less than ≈ 1.

When the propagation distance is very long such that diffraction is observed in the so-called far field or Fraunhofer regime, further simplification to the Fresnel approximation in Equation 2.3 can be made as the Fraunhofer approximation, which states that

$$U_2(x,y) = \frac{e^{jkz}}{j\lambda z}\exp\left[j\frac{k}{2z}\left(x^2+y^2\right)\right]\iint\limits_{\Sigma} U_1(\xi,\eta)\exp\left[-j\frac{2\pi}{\lambda z}\left(x\xi+y\eta\right)\right] d\xi\, d\eta. \tag{2.5}$$

This is obtained by expanding the quadratic exponent terms in Equation 2.3 and approximating the exponential terms with quadratic exponents in ξ and η inside the integral as unity. The first half of Equation 2.5 is the same as the optical impulse response for the Fresnel diffraction as shown in Equation 2.4, whereas the second half is essentially the Fourier transform of the input filed evaluated at $f_X = \frac{x}{\lambda z}$ and $f_Y = \frac{y}{\lambda z}$. The Fraunhofer approximation typically requires $N_F \ll 1$.

2.1.2 Mid/Near Field Grating Diffraction – The Talbot Effect

Periodic diffraction grating plays a central role in spatially dispersive spectroscopy as well as the computational Talbot spectrometer work in this thesis. Theoretical analysis on wave propagation through a sinusoidal phase grating in the far field regime with Fraunhofer diffraction is provided in Appendix A. This is the foundation for the spatially dispersive grating spectroscopy as discussed in Section 1.2.1. Here, we analyze wave propagation in the near/mid field regime after the diffraction grating, where the diffracted beams are not yet spatially separated.

Assuming that the grating area is large enough such that diffraction due to the grating aperture is negligible, for a sinusoidal phase grating, the input field $U_1(\xi,\eta)$ with uniform plane wave incidence can be modeled as the grating transmission function $t_G(\xi,\eta)$:

$$U_1(\xi,\eta) = t_G(\xi,\eta) = \exp\left[ja\sin\left(2\pi\frac{\xi}{P}\right)\right],$$

where a is the phase modulation amplitude and P is the grating period. Its Fourier transform is shown in Equation A.2. Under the Rayleigh-Sommerfeld diffraction solution, the diffraction transfer function is shown in Equation 2.2, which is $H(f_X, f_Y) = \exp\left[jkz\sqrt{1-(\lambda f_X)^2-(\lambda f_Y)^2}\right]$. This means that the Fourier transform of the optical field at the observation x-y plane with location z can be written as

$$\mathcal{F}\{U_2(x,y)\} = \mathcal{F}\{U_1(x,y)\}H(f_X,f_Y) = \sum_{q=-\infty}^{\infty} J_q(a)\exp\left[jk\sqrt{1-\left(q\frac{\lambda}{P}\right)^2}z\right]\delta\left(f_X - \frac{q}{P}, f_Y\right). \tag{2.6}$$

Here, $J_q(\cdot)$ is the Bessel function of the first kind and order q, originating from the Fourier transform of the sinusoidal phase term as shown in Equation A.1. Subsequently, the optical field is

$$U_2(x,y) = \sum_{q=-\infty}^{\infty} J_q(a)\exp\left[jk\sqrt{1-\left(q\frac{\lambda}{P}\right)^2}z\right]\exp\left(j2\pi q\frac{x}{P}\right). \tag{2.7}$$

The exponential term in Equation 2.6 and Equation 2.7 suggests that propagating waves exist only when $q\lambda < P$. This means that in reality only a finite number of diffraction orders exists after the grating. For simplicity of analysis, we assume that a strongly diffractive grating is used where only the -1, 0, and $+1$ diffractive orders exist. Using the relation that $J_{-q}(x) = (-1)^q J_q(x)$, we have

$$U_2(x,y) = J_0(a)\exp(jkz) + j2J_1(a)\sin\left(2\pi\frac{x}{P}\right)\exp\left[jk\sqrt{1-\left(\frac{\lambda}{P}\right)^2}z\right].$$

The intensity $I_2(x,y)$ can then be obtained as

$$\begin{aligned} I_2(x,y) = J_0^2(a) + 4J_1^2(a)\sin^2\left(2\pi\frac{x}{P}\right) + \\ 4J_0(a)J_1(a)\sin\left(2\pi\frac{x}{P}\right)\sin\left\{k\left[1-\sqrt{1-\left(\frac{\lambda}{P}\right)^2}\right]z\right\}. \end{aligned} \tag{2.8}$$

At z locations where $k\left[1-\sqrt{1-\left(\frac{\lambda}{P}\right)^2}\right]z = 2m\pi + \frac{1}{4}$ with m being an integer, we obtain that

$$I_2(x,y) = \left[J_0(a) + 2J_1(a)\sin\left(2\pi\frac{x}{P}\right)\right]^2,$$

which can be regarded as the self images of the grating function. On the other hand, at z locations where $k\left[1-\sqrt{1-\left(\frac{\lambda}{P}\right)^2}\right]z = 2m\pi + \frac{3}{4}$, we have

$$I_2(x,y) = \left[J_0\left(a\right) - 2J_1\left(a\right)\sin\left(2\pi\frac{x}{P}\right)\right]^2.$$

This still represents the self images of the grating function, except for that now there is a π phase shift compared to the previous plane. Further sub-images can exist on various planes, which we do not discuss in more details [99]. The extension of the analysis to amplitude gratings or gratings with non-sinusoidal phase/amplitude modulation is straightforward. A plot showing a simulated Talbot pattern after a diffraction grating with 0 and ±1 diffractive orders is shown in Figure 2-2.

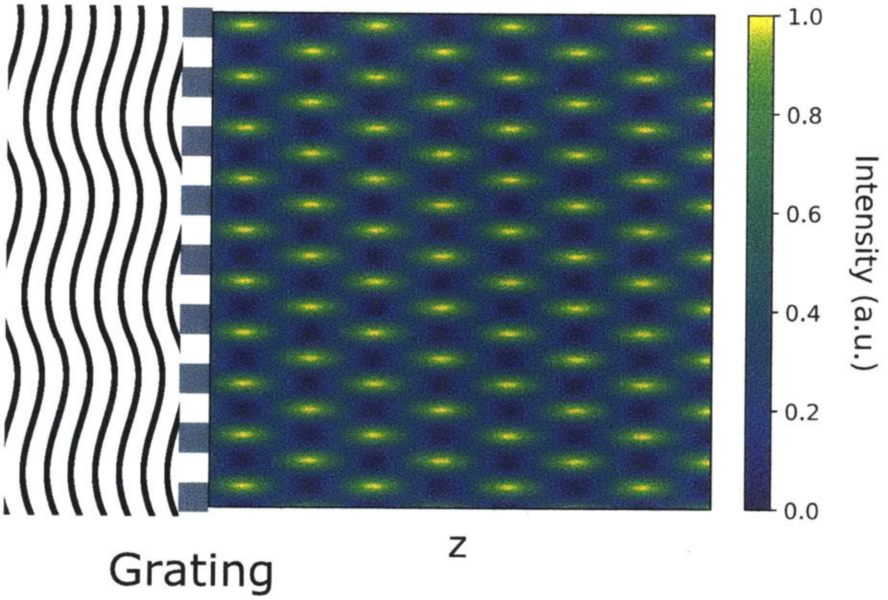


Figure 2-2: Simulated Talbot pattern with a sinusoidal phase grating for a grating pitch of 1.035 µm and incidence wavelength of 700 nm. The power diffraction efficiency is assumed to be 16.7% for the ±1 diffractive orders.

The above grating self images are historically referred to as the Talbot effect after its discovery in 1836 by Talbot [84]. Nearly half a century later, the longitudinal distance over which the self images repeat was provided by Lord Rayleigh [100]. This distance is called the Talbot distance. Lord Rayleigh suggested that the Talbot distance is given as

$$z_T = \frac{\lambda}{1 - \sqrt{1 - \left(\frac{\lambda}{P}\right)^2}}, \tag{2.9}$$

which we re-derived above with the sinusoidal phase grating. A popular approximation to the Talbot distance in Equation 2.9 is

$$z_T = \frac{2P^2}{\lambda}.$$

This is derived under the paraxial approximation with the Fresnel diffraction solution in Equation 2.3.

2.1.3 The Talbot Effect Under Tilted Incidence Angles

Now we investigate the Talbot image formation under tilted incidence angles. This will be important for analysis of the Talbot spectrometer performance under spread angular incidence later, as incidence angular spread can be represented as the ensemble of tilted incidences. Two tilted incidence situations are considered analytically. The first one is tilted incidence in the y-z plane as shown in Figure 2-3, and the second one is tilted incidence in the x-z plane as shown in Figure 2-4. The Talbot image formation leads to different formulations for these two tilted incidence cases. Afterwards, the general tilted incidence formula is given for numerical simulations. The derivations for these results are shown in Appendix B.

Tilted Incidence in y-z Plane

We first consider tilted plane wave incidence in the y-z plane as shown in Figure 2-3. Assume that the tilt angle is θ with respect to the z axis, the transmission function due to the angle tilt is

$$t_T(\xi, \eta) = \exp\left[jk\sin(\theta)\eta\right].$$

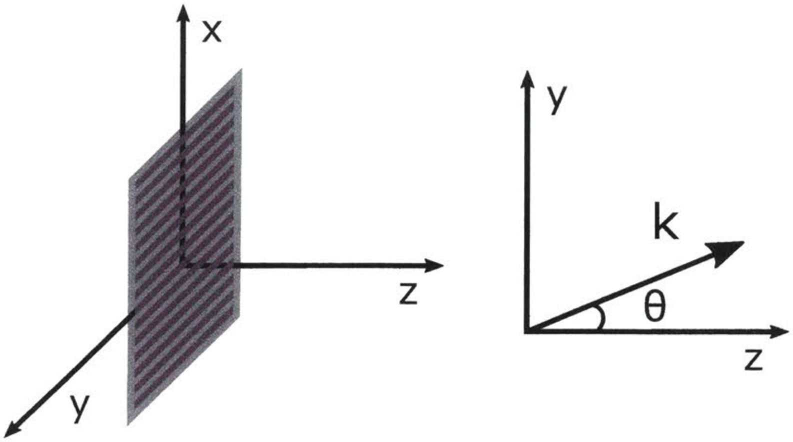


Figure 2-3: Illustration for tilted incidence angle θ in the y-z plane.

With the sinusoidal phase grating, considering only the 0 and ±1 diffractive orders under strongly diffractive cases, the intensity at the observation plane is

$$I_2(x,y) = J_0^2\left(a\right) + 4J_1^2\left(a\right)\sin^2\left(2\pi\frac{x}{P}\right) + \\ 4J_0\left(a\right)J_1\left(a\right)\sin\left(2\pi\frac{x}{P}\right)\sin\left\{k\left[\cos(\theta) - \sqrt{\cos^2(\theta) - \left(\frac{\lambda}{P}\right)^2}\right]z\right\}. \tag{2.10}$$

Equation 2.10 is very similar to Equation 2.8, except for that the Talbot distance here is modified as

$$z_T = \frac{\lambda}{\cos(\theta) - \sqrt{\cos^2(\theta) - \left(\frac{\lambda}{P}\right)^2}}. \tag{2.11}$$

Tilted Incidence in x-z Plane

Next we consider the case where the tilted plane wave incidence is in the x-z plane as shown in Figure 2-4. In this case, the diffraction solution is more complicated than that from the previous section. The transmission function for tilted plane wave incidence with tilt angle ϕ with respect to the z axis is

$$t_T(\xi,\eta) = \exp\left[jk\sin(\phi)\xi\right].$$

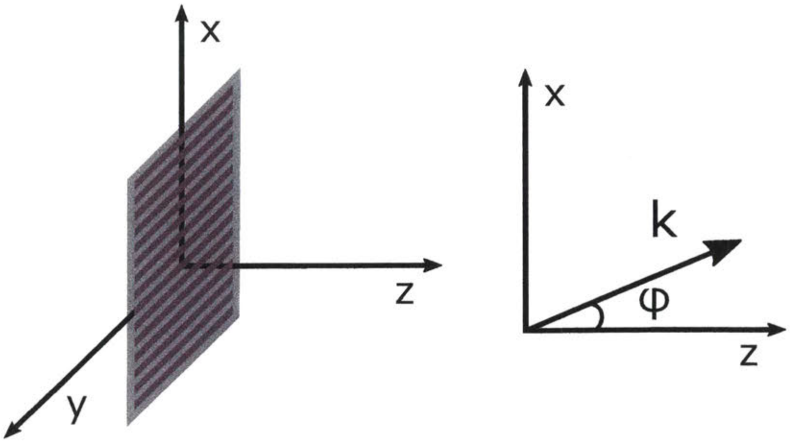


Figure 2-4: Illustration for tilted incidence angle ϕ in the x-z plane.

Correspondingly, the intensity at the observation plane is

$$\begin{gathered}
I_2(x,y) = J_0^2\left(a\right) + 2J_1^2\left(a\right) + \\
2J_0\left(a\right)J_1\left(a\right)\cos\left\{k\left\{\cos(\phi) - \sqrt{1-\left[\sin(\phi)+\frac{\lambda}{P}\right]^2}\right\}z - 2\pi\frac{x}{P}\right\} - \\
2J_0\left(a\right)J_1\left(a\right)\cos\left\{k\left\{\cos(\phi) - \sqrt{1-\left[\sin(\phi)-\frac{\lambda}{P}\right]^2}\right\}z + 2\pi\frac{x}{P}\right\} - \\
2J_1^2\left(a\right)\cos\left\{k\left\{\sqrt{1-\left[\sin(\phi)-\frac{\lambda}{P}\right]^2} - \sqrt{1-\left[\sin(\phi)+\frac{\lambda}{P}\right]^2}\right\}z - 4\pi\frac{x}{P}\right\}.
\end{gathered} \tag{2.12}$$

The above result is slightly more convoluted than the ones from Equation 2.8 and Equation 2.11. Three Talbot distances exist in this solution, with them being

$$\begin{aligned}
z_{T,+1} &= \frac{\lambda}{\cos(\phi) - \sqrt{1-\left[\sin(\phi)+\frac{\lambda}{P}\right]^2}}, \\
z_{T,-1} &= \frac{\lambda}{\cos(\phi) - \sqrt{1-\left[\sin(\phi)-\frac{\lambda}{P}\right]^2}}, \\
z_{T,\pm1} &= \frac{\lambda}{\sqrt{1-\left[\sin(\phi)-\frac{\lambda}{P}\right]^2} - \sqrt{1-\left[\sin(\phi)+\frac{\lambda}{P}\right]^2}}.
\end{aligned} \tag{2.13}$$

Figure 2-5 shows the simulated Talbot patterns for normal incidence as well as tilted incidences with respect to both θ and ϕ. With incidence angle tilt in θ, a slight change in the Talbot distance is observed in (b), whereas an incidence angle tilt in ϕ results in more dramatic change in the pattern formation in (c).

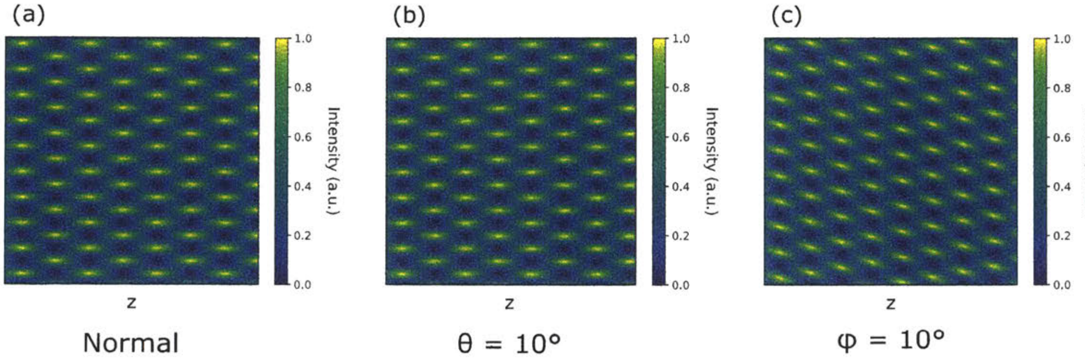


Figure 2-5: Simulated Talbot patterns for (a) normal incidence, (b) tilted incidence with $\theta = 10°$, and (c) tilted incidence with $\phi = 10°$ with a sinusoidal phase grating for a grating pitch of 1.035 µm and incidence wavelength of 700 nm. The power diffraction efficiency is assumed to be 16.7% for the ±1 diffractive orders.

General Tilt

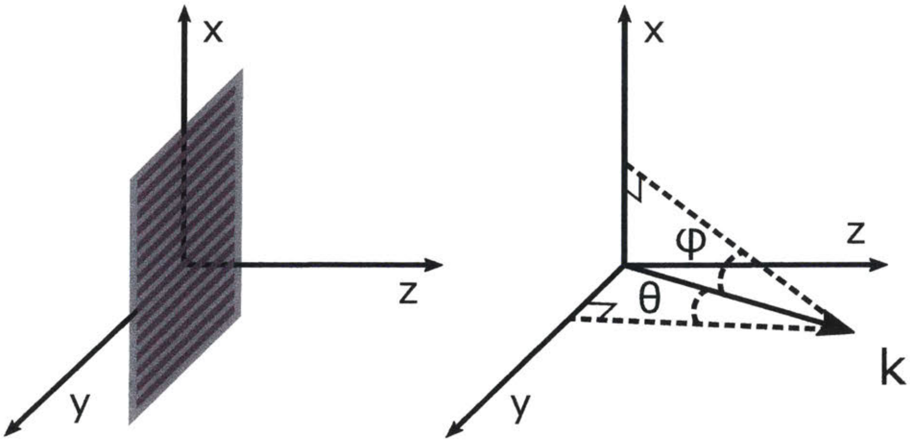


Figure 2-6: Illustration for general tilted incidence with both θ and ϕ.

The individual results for tilt in the y-z and x-z plane demonstrate the different pattern formation response towards incidence angle tilt for the Talbot effect. For the case of general tilt with both θ and ϕ, which is illustrated in Figure 2-6, we have

$$t_T(\xi,\eta) = \exp\left[jk\sin(\phi)\xi + jk\sin(\theta)\eta\right].$$

Similar as before, the intensity at the observation plane is

$$\begin{aligned} I_2(x,y) = \Bigg| & J_0(a)\exp\left[jk\sqrt{\cos^2(\theta)-\sin^2(\phi)}z\right] + \\ & J_1(a)\exp\left\{jk\sqrt{\cos^2(\theta)-\left[\sin(\phi)+\frac{\lambda}{P}\right]^2}z\right\}\exp\left(j2\pi\frac{x}{P}\right) - \\ & J_1(a)\exp\left\{jk\sqrt{\cos^2(\theta)-\left[\sin(\phi)-\frac{\lambda}{P}\right]^2}z\right\}\exp\left(j2\pi\frac{x}{P}\right)\Bigg|^2. \end{aligned} \tag{2.14}$$

Although further analytical simplifications exist for the above equation, we resort to direct numerical evaluation for simulations involving general tilt or spread angle incidences.

2.1.4 The Talbot Effect as Interferometry

Behind the mathematics of image formation for the Talbot effect, a much simpler interpretation is to treat the Talbot effect from an interferometric perspective. For normal plane wave incidence, the wave vectors for the 0 and ±1 diffractive orders are shown in Figure 2-7. For the ±1 diffractive orders, the grating essentially adds a lateral wave vector $k_g = \frac{2\pi}{\lambda}$ to the diffracted beams. The Talbot distance z_T shown in Equation 2.9 results from the interference between the ±1 and the 0 diffracted beams, and corresponds exactly to the longitudinal spatial periodicity for k_z in Figure 2-7.

For tilted incidence plane wave with θ and ϕ from Section 2.1.3, similar analysis can also lead to the exact Talbot distance calculation. For incidence angle of θ, the wave vector along the y direction is essentially unaltered when transmitting through the grating. One can therefore project the original k onto the x-z plane and calculate the diffracted beams as shown in Figure 2-8. The longitudinal wave interference results

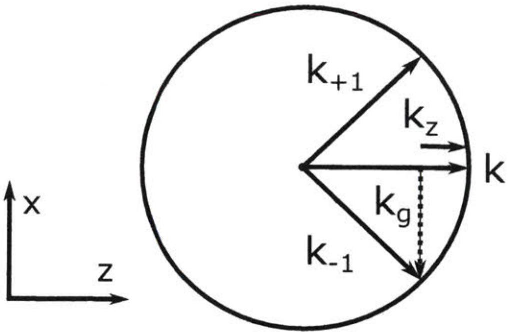


Figure 2-7: k-space diagram showing the diffracted beams with respect to the incidence beam under the normal incidence. The grating adds a lateral wave vector k_g to the diffracted beams.

in k_z, which corresponds to the spatial periodicity in the z direction exactly as the Talbot distance calculated in Equation 2.11.

For incidence angle of ϕ shown in Figure 2-8, k_z from the ±1 diffractive orders no longer equal to each other as in previous cases. As a result, 3 longitudinally interfering terms appear in the solution. These three terms can be calculated from the geometry implied in Figure 2-8 and they match with the three Talbot distances calculated in Equation 2.13.

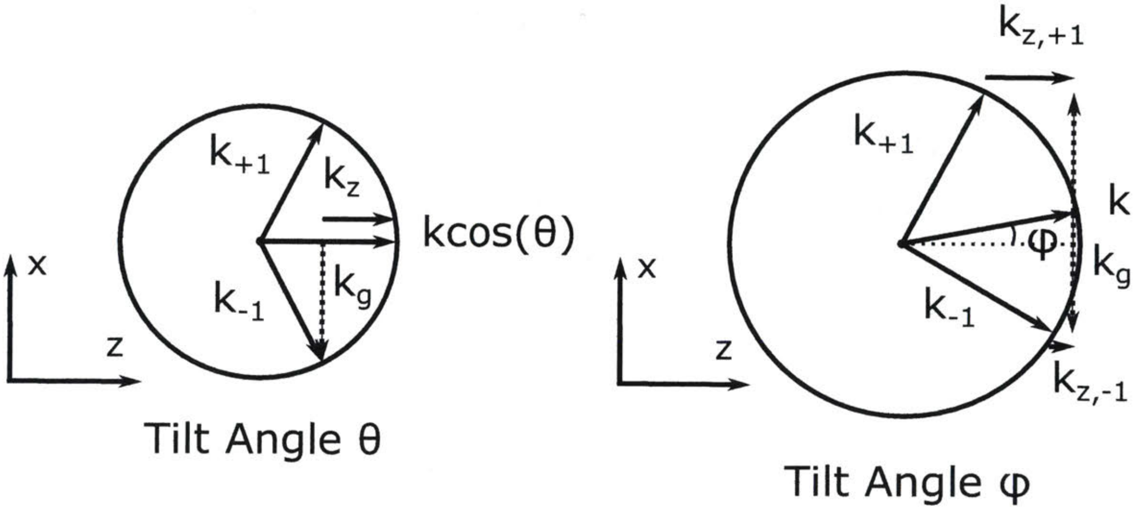


Figure 2-8: k-space diagrams showing the diffracted beams with respect to the incidence beam under tilted incidences for (left) tilt in θ and (right) tilt in ϕ. The grating adds a lateral wave vector k_g to the diffracted beams.

Treating the Talbot effect as the result from interference from the various diffracted beams greatly simplifies the analysis on the Talbot distance. Extension to multiple diffractive orders is straightforward. The periodic self images and sub-images of the diffraction grating in the x direction can also be interpreted as the interference amongst various diffraction orders in the x dimension in Figure 2-7 and Figure 2-8.

2.2 Spectrometer Design and Simulations with the Talbot Effect

With the theoretical foundations for the Talbot effect described in the previous section, we explore the design space for building spectrometers with the Talbot effect with simple 1-D diffraction gratings and modern image sensors in this section. Various considerations such as grating and sensor selection, geometrical configuration, and design trade-offs are discussed in details. Simulations of temporally incoherent sources and the effect of incidence angle spread on spectral resolution are also explored in depth in this section.

2.2.1 The Talbot Effect for Fourier Spectrometers

We first return to the intensity distribution for normal plane wave incidence in Equation 2.8. This has a high resemblance with the Michelson interferometer intensity response with normal plane wave incidence [101], which means that the power spectral density $S(\lambda)$ for the optical signal can be retrieved from the Talbot interferogram along the z direction with a Fourier transform. Diving into more details, we first express the intensity as a function of the Talbot wave vector k_T as

$$\begin{aligned} I(k_T; x, y, z) &= D_0^2 + 4D_1^2 \sin^2\left(2\pi \frac{x}{P}\right) + 4D_0D_1 \sin\left(2\pi \frac{x}{P}\right) \sin\left(k_T z\right) \\ &= A_0(x) + A_1(x) \sin\left(k_T z\right), \end{aligned}$$

where we use D_0 and D_1 to represent the Fourier coefficients in the expansion for the grating transmission function, and $A_0(x)$ and $A_1(x)$ to represent the constant terms with respect to k_T and z in Equation 2.8. Assume that the power spectral density with respect to the transformed Talbot wave vector is $S(k_T)$, we can then express the overall intensity $I(x, y, z)$ as

$$I(x,y,z) = \int_{0}^{+\infty} S(k_T) I(k_T; x, y, z)\, dk_T = A_0(x) \int_{0}^{+\infty} S(k_T)\, dk_T +$$
$$A_1(x) \int_{0}^{+\infty} S(k_T) \sin(k_T z)\, dk_T = A_0(x) \int_{0}^{+\infty} S(k_T)\, dk_T + \frac{A_1(x)}{2j} \int_{-\infty}^{+\infty} S(k_T) e^{jk_T z}\, dk_T, \tag{2.15}$$

where we have assumed that $S(k_T)$ is anti-symmetric around $k_T = 0$. Denoting

$$I'(z) = \frac{2j}{\sqrt{2\pi} A_1(x)} \left[I(x,y,z) - A_0(x) \int_{0}^{+\infty} S(k_T)\, dk_T \right],$$

this means that $I'(z) = \mathcal{F}^{-1}\{S(k_T)\}$ and $S(k_T) = \mathcal{F}\{I'(z)\}$, with the assumption that $I'(z)$ is anti-symmetric around $z = 0$.

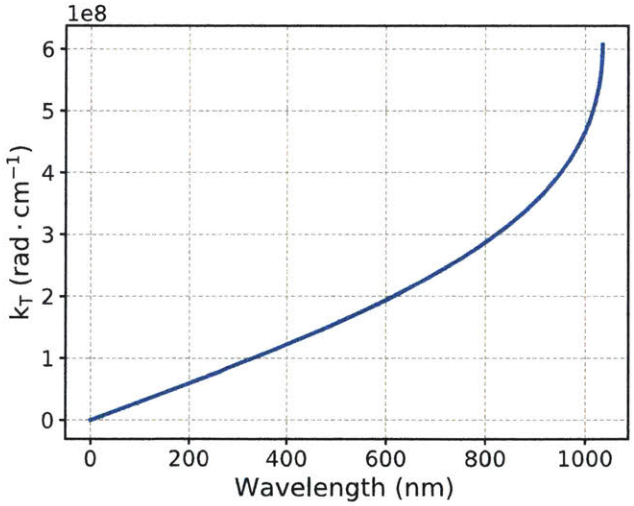


Figure 2-9: k_T as a function of λ for the 0 and ±1 diffractive orders with a grating pitch of $P = 1.035$ µm.

Based on the fact that the Talbot wave vector k_T relates to the input wavelength as

$$k_T = \frac{2\pi}{z_T} = 2\pi\frac{1-\sqrt{1-\left(\frac{\lambda}{P}\right)^2}}{\lambda}, \tag{2.16}$$

and

$$\lambda = \frac{4\pi k_T P^2}{4\pi^2 + k_T^2 P^2},$$

we can get $S(\lambda)$ from $S(k_T)$ with the following transformation

$$S(\lambda) = S(k_T)\frac{dk_T}{d\lambda} = S(k_T)\frac{2\pi\left[1-\sqrt{1-\left(\frac{\lambda}{P}\right)^2}\right]}{\lambda^2\sqrt{1-\left(\frac{\lambda}{P}\right)^2}}.$$

Figure 2-9 shows k_T as a function of λ for wavelengths smaller than the grating period for the 0 and ±1 diffractive orders at $P = 1.035$ µm. For shorter wavelengths, higher diffractive orders may exist, which can result in additional Talbot periods that introduce wavelength ambiguity. We will cover design considerations on this subject in the next part.

2.2.2 Talbot Interferogram Sampling – Design Considerations

With recent technological advances in CCD and CMOS image sensors, the sensor pixel sizes have reached a point where direct sampling of the Talbot interferogram is achievable without any external imaging optics like the ones in previous studies [85]. In addition, operating the spectrometer configuration in the non-paraxial regime with a tilted image sensor allows full-frame interferogram capture without any moving parts, as strong diffractions in the non-paraxial regime limit the Talbot region to be within several millimeters after the diffraction grating. To sample the Talbot interferogram intensity along the z direction as shown in Equation 2.15, we use a tilted 2D image sensor in close proximity to the grating as shown in Figure 2-10. A sample simulated 1-D Talbot interferogram for a single-frequency laser source at 700 nm wavelength captured by an image sensor with 1.67 µm pixel size is shown in Figure 2-11.

Discussions on component selection and system configuration are provided in this section for optimizing the Talbot interferogram sampling. We note here that just as in the design for conventional grating spectrometers, the actual component selection is highly dependent on the target application at hand – wavelength span, resolution requirement, light throughput, dynamic range, form factor, and cost considerations all play important roles in building a system for a particular application. The versatility of design choices of the Talbot spectrometer also highlights its potential broad appealing in terms of application domains.

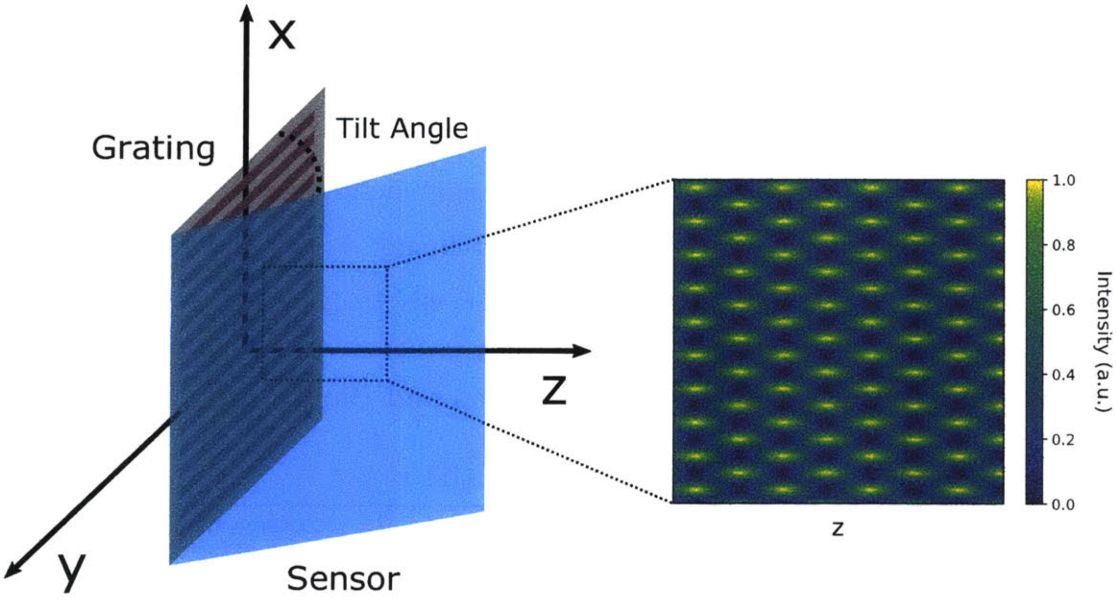


Figure 2-10: Illustration of Talbot interferogram sampling with a tilted image sensor in close proximity to the grating.

Tilt Axis

Two ways exist to sample the Talbot interferogram across the z axis with a tilted image sensor, the first one is to tilt the image sensor around the x axis and the second one is to tilt around the y axis. Tilting around the y axis is more complicated as x cannot be kept as a constant with respect to z in this situation, resulting in non-constant $A_0(x)$ and $A_1(x)$ in Equation 2.15. The situation is further exacerbated with tilted ϕ incidence as in Equation 2.12, where x appears inside the cosine terms

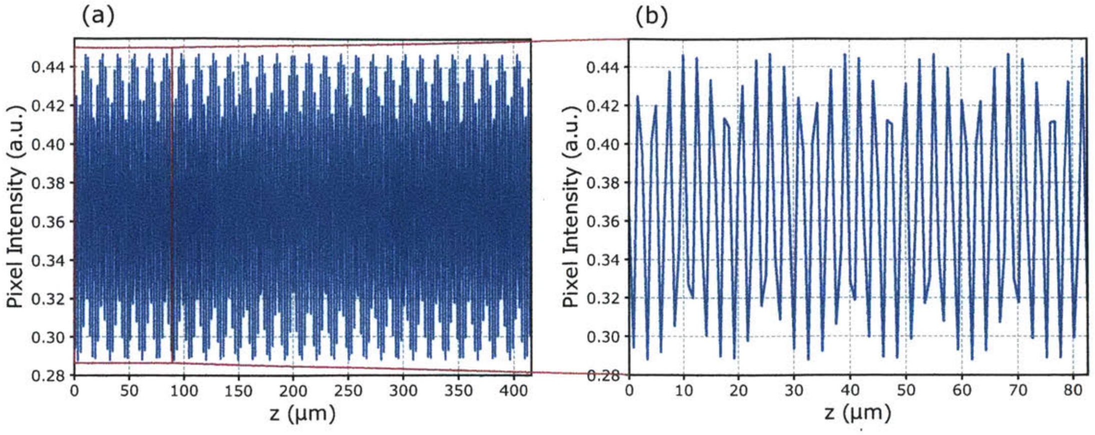


Figure 2-11: Simulation of a sampled interferogram for a single-frequency laser source at 700 nm captured by an image sensor with 1.67 μm pixel size tilted at 30°. (b) is a closer look at the interferogram pattern in (a) within 80 μm distance.

for all the three periodic terms. Therefore, it is more desirable to tilt around the x axis such that x can be kept as constant across each interferogram recording row. As a result, this is the system geometry we choose for this study.

Tilt Angle – Tuning the Spectral Resolution and Bandwidth

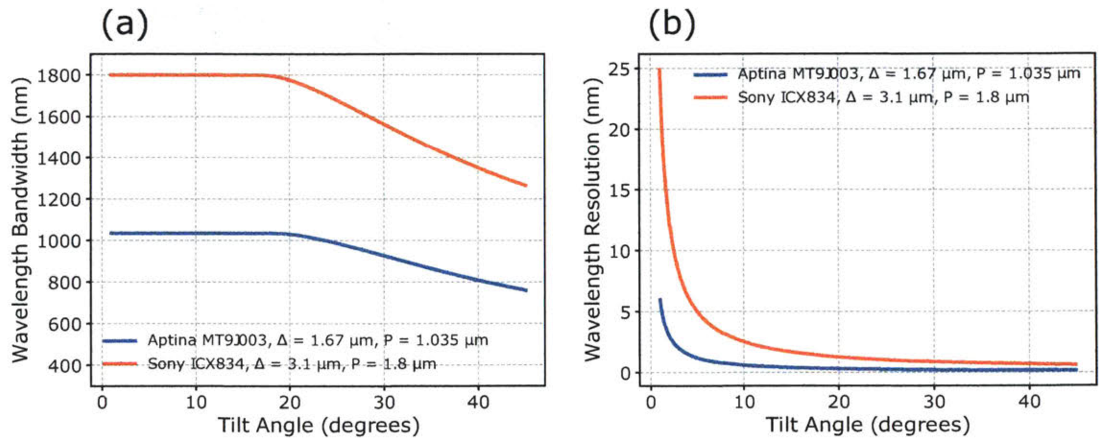


Figure 2-12: Plots showing the wavelength bandwidth and resolution as a function of image sensor tilt angle α for two image sensor and grating combinations.

An image sensor has a fixed sampling budget determined by the number of spatial samples one can obtain along the sampling direction. This imposes an inherent space-bandwidth product limit to our optical system. Assume that the length of the image sensor is L, the pixel pitch size is Δ, the number of pixels along L is $N = \frac{L}{\Delta}$, and the tilt angle is α, the total 1-D sampling depth is $L\sin(\alpha)$ and the sampling periodicity is $\Delta\sin(\alpha)$. According to the Nyquist-Shannon sampling theorem, this means that in the reciprocal k space, the sampling spacing δ_k and half-bandwidth B_k are

$$\delta_k = \frac{2\pi}{L\sin(\alpha)},$$

and

$$B_k = \frac{\pi}{\Delta\sin(\alpha)}. \tag{2.17}$$

Subsequently, the resolution δ_λ and bandwidth B_λ in the wavelength domain can be obtained through Equation 2.16 or from the plot in Figure 2-9. The above equations mean that one can change the operating resolution and bandwidth by changing the image sensor tilt angle α. Two plots showing the bandwidth and resolution change as a function of the tilt angle α for two image sensors in Table 2.1 are shown in Figure 2-12. The corresponding grating pitch P is indicated in the plot legend. As the operating wavelength cannot exceed the grating pitch P, the wavelength bandwidth is constant as P for smaller tilt angles as shown in Figure 2-12 (a). In this region, tilting the image sensor more results in higher resolution without sacrificing the operating bandwidth. After this region, a trade-off between resolution and bandwidth with respect to the tilt angle exists, which means that one might need to select an optimal tilt angle based on the design requirement, especially for the smaller pixel sensor as the bandwidth limitation starts to impact optical wavelengths with higher tilt angles.

Image Sensor Selection – Balancing the Visibility and Sensitivity

The self imaging nature of the Talbot phenomenon implies that pattern sampling has to occur on the grating period scale. For optical frequencies of range around 300 to 1000 nm, this means that the sampling unit, namely the pixel size, has to be on the order of micrometers. As a result, modern CMOS and CCD image sensors come as

Sensor	Aptina MT9P031	Aptina MT9J003	Samsung S.LSI	Sony ICX834
Type	CMOS	CMOS	CMOS	CCD
Pixel Size (µm)	2.2	1.67	1.12	3.1
Pixel Number	2592 × 1944	3856 × 2764	4208 × 3120	4250 × 2838
Dimensions (mm)	5.70 × 4.28	6.440 × 4.616	4.713 × 3.494	13.2 × 8.8
Dynamic Range (dB)	70.1	65.2	61.9	75

Table 2.1: Specifications for several image sensors for consideration with the Talbot spectrometer.

natural choices for capturing the Talbot interferogram.

Smartphone and compact camera sensors typically have an optical format of below $\approx 1/1.8$ inches, and a pixel size of ≈ 1 to 2.5 µm [102]. These sensors are attractive because of their small pixel size and form factor, ubiquitous usage, and low cost. Although a majority of them employ a back-illuminated structure to enhance light collection [103], their major disadvantage is low sensitivity due to small light collection area. However, many applications involving relatively strong light signals do not have stringent requirement on sensor sensitivity. For example, portable white light and fluorescence microscopes have been realized with smartphone cameras and are capable of disease diagnostics applications [104, 105, 106]. Therefore, there is still a large application domain where portability and low cost are the major factors driving the popularity for these types of sensors.

Moving up the frame size ladder, image sensors of around ≈ 1 inch optical format up till full-frame size for premium compact cameras and digital LSR cameras can have pixel sizes of ≈ 2.4 to 8 µm [102]. The much improved sensitivity for these sensors coupled with potential cooling systems can be adequate for scientific imaging and astrophotography. Consequently, these sensors are more suitable for low light spectroscopy applications such as Raman spectroscopy. Large pixel sizes in these sensors may impose limitations on interferogram visibility and grating selection, which will be discussed later in this section. Specifications for several image sensors are shown in Table 2.1, out of which three of them (the Aptina and Samsung sensors) have been used to realize Talbot spectrometer systems in this study. The Sony ICX834 sensor is attractive since commercial scientific-grade cameras have been built with this sensor with TEC cooling for low-light applications, yet it has a relatively small pixel size compared to other image sensors with similar functionalities, which is important for building a Talbot spectrometer for reasons that will be discussed in the following part. It is therefore a potential candidate for realizing a Talbot spectrometer system for sensitive measurements.

For any interferometric measurements, the fringe visibility, which is defined as

$$v = \frac{I_{max} - I_{min}}{I_{max} + I_{min}},$$

is useful as it is related to the sensor dynamic range required for retrieving the spectral signal. Assume that a sensor pixel size of Δ is used and the sensor sampling function is a square function with a pixel fill factor of 100%, we denote the sensor dimensions as ξ and η, where ξ is along the same axis as x and η is the dimension for the interferogram row. The sampled pixel reading at discrete pixel locations m and n can be written as

$$I(m,n) = \int\limits_{\xi_m - \frac{\Delta}{2}}^{\xi_m + \frac{\Delta}{2}} \int\limits_{\eta_n - \frac{\Delta}{2}}^{\eta_n + \frac{\Delta}{2}} I[x = \xi, y = \eta\cos(\alpha), z = \eta\sin(\alpha)]\, d\xi\, d\eta =$$

$$\int\limits_{\xi_m - \frac{\Delta}{2}}^{\xi_m + \frac{\Delta}{2}} \int\limits_{\eta_n - \frac{\Delta}{2}}^{\eta_n + \frac{\Delta}{2}} D_0^2 + 4D_1^2 \sin^2\left(2\pi\frac{\xi}{P}\right) + 4D_0D_1\sin\left(2\pi\frac{\xi}{P}\right)\sin\left[k_T\sin(\alpha)\eta\right]\, d\xi\, d\eta =$$

$$D_0^2\Delta^2 + 4D_1^2 \int\limits_{\xi_m - \frac{\Delta}{2}}^{\xi_m + \frac{\Delta}{2}} \sin^2\left(2\pi\frac{\xi}{P}\right) d\xi + 4D_0D_1 \int\limits_{\xi_m - \frac{\Delta}{2}}^{\xi_m + \frac{\Delta}{2}} \sin\left(2\pi\frac{\xi}{P}\right) d\xi \int\limits_{\eta_n - \frac{\Delta}{2}}^{\eta_n + \frac{\Delta}{2}} \sin\left[k_T\sin(\alpha)\eta\right] d\eta. \tag{2.18}$$

The middle term is constant with respect to η, and can be simplified as

$$4D_1^2 \int\limits_{\xi_m - \frac{\Delta}{2}}^{\xi_m + \frac{\Delta}{2}} \sin^2\left(2\pi\frac{\xi}{P}\right) d\xi = 2D_1^2\Delta + \frac{PD_1^2}{\pi}\cos\left(\frac{4\pi\xi_m}{P}\right)\sin\left(\frac{2\pi\Delta}{P}\right).$$

This is a term that has a linear-like increasing trend with Δ. The last term in Equation 2.18 depends on sensor readings from the η dimension, which we further simplify as follows

$$\begin{aligned} 4D_0D_1 \int_{\xi_m-\frac{\Delta}{2}}^{\xi_m+\frac{\Delta}{2}} \sin\left(2\pi\frac{\xi}{P}\right) d\xi \int_{\eta_n-\frac{\Delta}{2}}^{\eta_n+\frac{\Delta}{2}} \sin\left[k_T \sin(\alpha)\eta\right] d\eta = \\ \frac{8D_0D_1P}{k_T\sin(\alpha)\pi} \sin\left(\frac{2\pi\xi_m}{P}\right) \sin\left[\frac{k_T\sin(\alpha)\Delta}{2}\right] \sin\left(\frac{\pi\Delta}{P}\right) \sin[k_T\sin(\alpha)\eta_n] = \\ \frac{4D_0D_1P\Delta}{\pi} \sin\left(\frac{2\pi\xi_m}{P}\right) \mathrm{sinc_u}\left(\frac{\pi k_T}{2B_k}\right) \sin\left(\frac{\pi\Delta}{P}\right) \sin[k_T\sin(\alpha)\eta_n] \end{aligned} \tag{2.19}$$

This is a periodic term in η_n, which we expect as the interferogram signal. Several terms constant with respect to η_n appear as the amplitude for the interferogram signal. The $\sin\left(\frac{\pi\Delta}{P}\right)$ term indicates that the interferogram signal can have zero amplitude when $\Delta = NP$, where N is a positive integer. Further, the interferogram visibility has a local maximum when

$$\Delta = \left(N + \frac{1}{2}\right) P \quad \text{for } N = 1, 2, ... \tag{2.20}$$

Smaller N leads to lower overall DC bias term in $I(m, n)$, which results in higher interferogram visibility v. Figure 2-13 (a) shows how the fringe visibility changes as a function of Δ/P for a grating with $P = 1.035$ µm and power diffraction efficiency for the ±1 diffractive orders as 16.7%. Figure 2-13 (b) shows the corresponding sampled interferogram along the z direction. As mentioned earlier, image sensors with a larger pixel size in general have better sensitivity. However, as revealed in Figure 2-13 (a), this can lead to small fringe visibility. Therefore, balancing sensitivity and visibility is one of the major design challenges in Talbot spectrometer system realization.

Another term in Equation 2.19, which is $\sin\left[\frac{k_T\sin(\alpha)\Delta}{2}\right]$, also deserves some attentions. With the bandwidth definition for B_k in Equation 2.17, we get $\sin\left[\frac{k_T\sin(\alpha)\Delta}{2}\right] = \sin\left(\frac{\pi k_T}{2B_k}\right)$, and can transform this into an unnormalized sinc function as shown in the equation. This means that with a fixed Δ, the visibility for different wavelengths are different due to the sampling process. Therefore calibration is needed for accurate spectrum reconstruction. Figure 2-14 shows the effect of calibrating the FFT-inverted

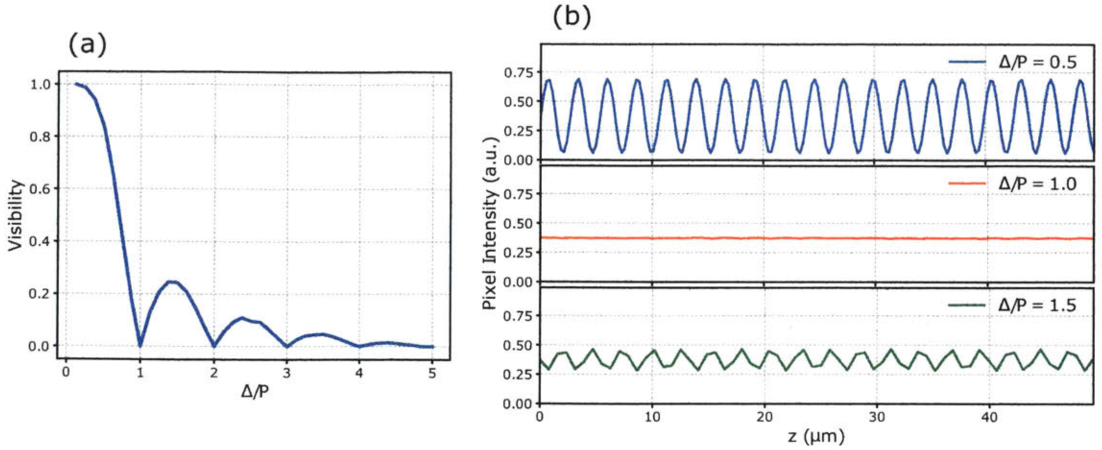


Figure 2-13: (a) Simulated results showing the interferogram fringe visibility change as a function of pixel size to grating pitch ratio. (b) Simulated Talbot interferograms in z direction for image sensors with various pixel sizes. The wavelength for the simulation is 700 nm.

spectrum by point-wise division with $\mathrm{sinc_u}\left(\frac{\pi k_T}{2B_k}\right)$, which results in a flat spectral response as desired.

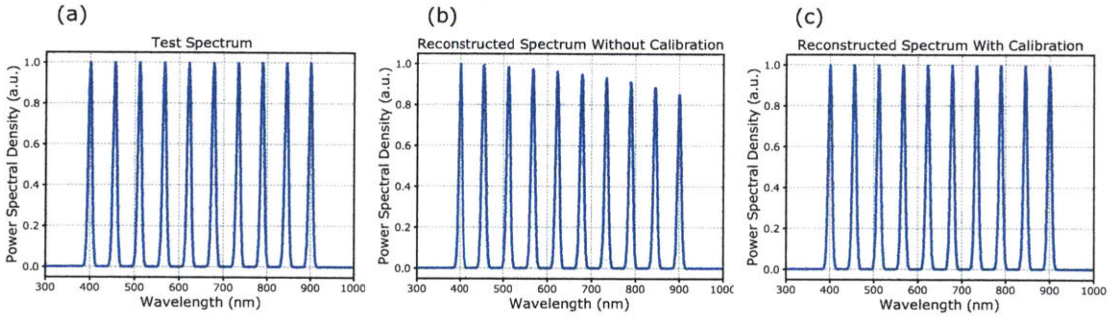


Figure 2-14: Illustration of spectral calibration for a test spectrum with flat-amplitude peaks across a wide spectral range.

As revealed in Figure 2-13 (a), overall small pixel sensors should be preferred for higher visibility, with some caution to avoid the zero visibility spots in theory. For smartphone and compact camera sensors with a pixel size of $\Delta \approx 1$ to 2.5 μm, this means that we can use a grating pitch size of $P \approx \frac{2}{3}\Delta \approx 0.6$ to 1.7 μm with $N = 1$, which can provide efficient ±1 order diffraction for optical wavelengths. For more sensitive image sensors having a pixel size of $\Delta \gtrapprox 3$ μm, two directions exist. The first one is to still use a grating with a pitch size of ≈ 0.6 to 1.7 μm. However, since now $N \geq 2$ from Equation 2.20, fringe visibility suffers. The second one is to use a larger grating pitch size with $N = 1$ from Equation 2.20, which means that now the grating pitch $P \gtrapprox 2$ µm. Higher diffractive orders may exist for such grating pitch for optical wavelengths, which may introduce wavelength ambiguity due to Talbot patterns formed from higher diffractive orders. Grating selection will be discussed in more details in the next part.

We note here that for small pixel sensors, the actual fringe visibility in experiment can deviate from what's theoretically depicted in Figure 2-13 (a). Practical issues such as grating-sensor alignment, pixel fill factor, effective pixel sampling function, and response function of micro-lens array under extreme oblique incidence angles can change the overall response function significantly as compared to that in Figure 2-13 (a). However, the overall trend of reduced visibility with larger pixel sizes is an important consideration when selecting the image sensor. In general, there are more design constraints for image sensors with pixel sizes $\gtrapprox 3$ µm for Talbot spectrometer design with optical wavelengths. In addition, one of the the major advantages of the Talbot spectrometer, which is its compact form factor for competitive performance metrics in resolution and bandwidth, starts to wither with large image sensors and the associated optical components. We leave the Talbot spectrometer design with large pixel image sensors as a potential future research direction and instead focus on compact systems with smartphone and compact camera sensors in this study.

Grating Selection – Operating Wavelength Considerations

Once the image sensor is decided based on the light condition, a grating can be selected accordingly with the visibility constraints in mind. Design flexibility can still exist at this stage due to the balancing of interferogram visibility and operating wavelength considerations.

Transmission gratings are the most straightforward types of gratings to use for the Talbot spectrometer configuration, which is what we focus on in our discussions. In general, a phase grating is preferred over an amplitude grating due to higher light throughput as well as higher diffraction efficiency in general.

Our theoretical analysis uses sinusoidal phase gratings, which are relatively easy to fabricate with interference lithography techniques. For sinusoidal phase gratings with a grating pitch of P, the operating wavelength range where only 0 and ±1 diffractive orders exist for normal incidence is $\left[\frac{P}{2}, P\right]$. For wavelengths below $\frac{P}{2}$, wavelength ambiguity may exist due to Talbot pattern formed with higher diffractive orders. An alternative is to use square-wave phase gratings, where even diffractive orders do not exist [107]. Therefore, they have a wider higher-diffractive-order-free wavelength range of $\left[\frac{P}{3}, P\right]$. For image sensors with a pixel size of ≈ 3 µm with a grating pitch $P \approx 1.8$ µm, this means that only one pair of higher diffractive order can exist above ≈ 600 nm for normal incidence. For narrowband spectroscopy applications such as Raman spectroscopy at 632 nm excitation wavelength, a compact Talbot spectroscopy system is possible with the Sony ICX834 sensor at its peak quantum efficiency range above 60% from ≈ 600 to 700 nm. On the other hand, for smartphone sensors with a pixel pitch of ≈ 1.5 µm and a grating pitch of ≈ 1 µm, the operating wavelength range covers almost the entire visible spectrum, which will be useful for broadband spectroscopy applications.

Even with the existence of higher diffractive orders, they are typically much weaker than that of the 0 and ±1 diffractive orders. For example, with square-wave phase gratings, the power diffraction efficiency is $DE_{m=odd} = \frac{1}{m^2} DE_{\pm1}$ [107]. Therefore, the third diffractive order efficiency is around an order of magnitude lower than that of the first diffractive order. As a result, spectral artifacts due to the higher diffractive orders can be small compared to the main signal, or may be easy to remove for narrowband applications. Other types of gratings such as blazed gratings, which can be optimized for a narrow spectral range with high diffraction efficiency, can also be used for building the Talbot spectrometer at oblique incidence angles. This can be advantageous over the current configuration with normal incidence due to potentially cleaner interfering pattern with less interfering beams.

Table 2.2 provides a summary of image sensors and gratings either used in this study or suggested for possible future work. The gratings for the Aptina sensors are chosen such that the grating pitches are roughly 2/3 of the pixel sizes of the image sensors. The grating for the Samsung S.LSI sensor is chosen to achieve high resolution with availability constraint from the vendor. These three systems have all been experimentally realized in this study. The grating for the Sony ICX834 sensor is chosen aiming at sensitive measurements with Raman spectroscopy at 632 nm excitation wavelength. This is a potential system for future research.

Sensor	Pixel Size	Dimension	Grating Size
Aptina MT9P031	2.2 µm	5.70 mm × 4.28 mm	1.608 µm
Aptina MT9J003	1.67 µm	6.440 mm × 4.616 mm	1.035 µm
Samsung S.LSI	1.12 µm	4.713 mm × 3.494 mm	1.035 µm
Sony ICX834	3.1 µm	13.2 mm × 8.8 mm	1.900 µm

Table 2.2: Summary for several image sensors and gratings either used in this study or suggested for possible future work.

2.2.3 Temporal Incoherence and Angular Spread Incidence Study

In this section, we study the Talbot spectrometer response with temporally incoherent light sources as well as sources with incidence angular spread. These aspects have important implications for realizing the system experimentally as well as revealing the performance expectations for practical diffuse light sources. Simulations in this section are performed based on the Rayleigh-Sommerfeld diffraction solutions for sinusoidal phase gratings as discussed in Section 2.1.3.

Temporal Incoherence

We proceed with partial temporal coherence simulations to characterize the Talbot spectrometer with broadband spectra. Our simulation approach for partial temporal coherence is to discretize the power spectral density $S(\lambda)$ and sum over the spectral intensity responses from the spectral components as [97]

$$I(x, y, z) \approx \sum_{i=1}^{I} S(\lambda_i) I(\lambda_i; x, y, z) \delta\lambda.$$

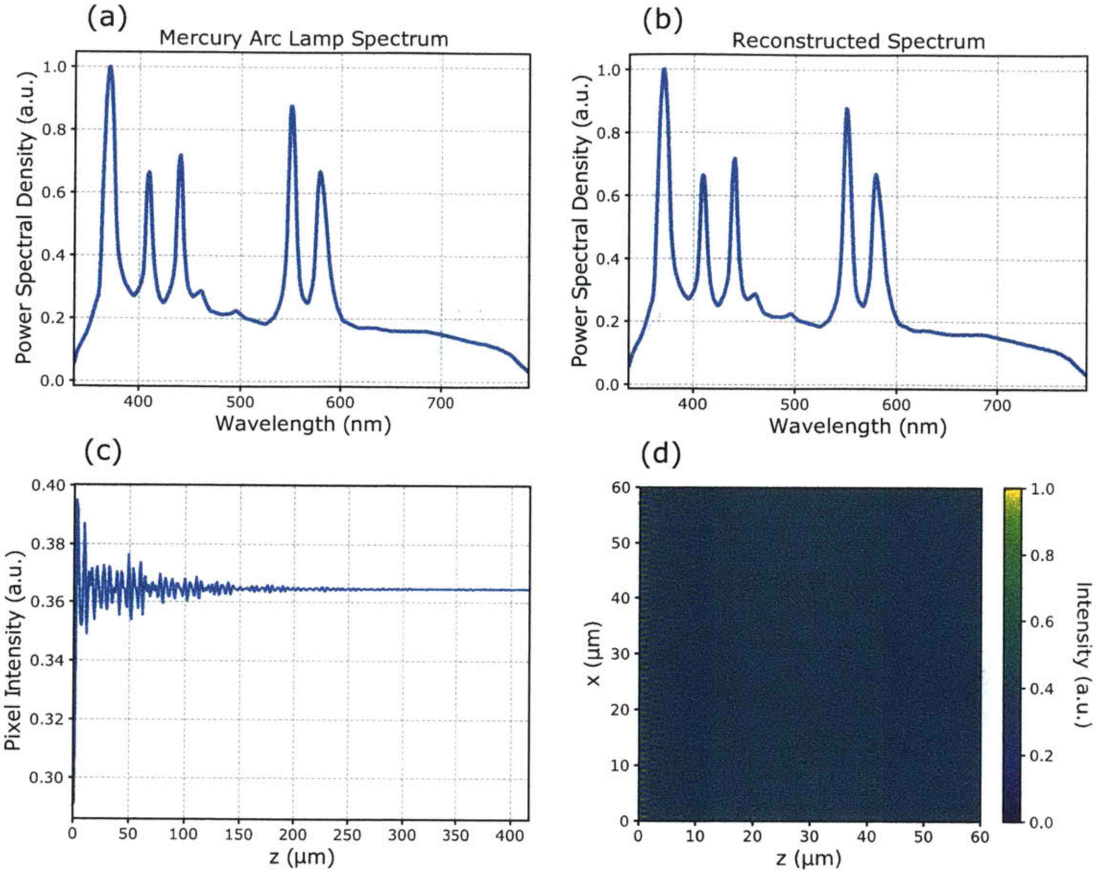


Figure 2-15: Simulation illustrations for spectrum reconstruction with the sampled Talbot interferogram for a broadband mercury arc lamp source. The simulated image sensor is Aptina MT9J003 as in Table 2.2. (a) The input light spectrum for the mercury arc lamp source spanning across ≈ 450 nm in the visible spectral range. (b) Reconstructed spectrum from the sampled interferogram with the sensor tilting at 30°. (c) Simulated 1-D interferogram sampled by the image sensor. (d) Simulated 2-D Talbot pattern after the grating.

This approach effectively ignores any spectral cross-correlations amongst the various spectral components, which is reasonable for many light sources such as thermal, fluorescence, and Raman sources. Figure 2-15 shows the simulation results where a broadband Mercury arc lamp spectrum is used as the input light source, and the Aptina MT9J003 system as in Table 2.2 is used in the simulated system. The Mercury arc lamp source covers over ≈ 450 nm in the visible spectral range and the spectrum is obtained digitally from Thorlabs [108]. The image sensor tilt angle is assumed to be 30° in the simulation and the sensor is assumed to be in contact with the grating on

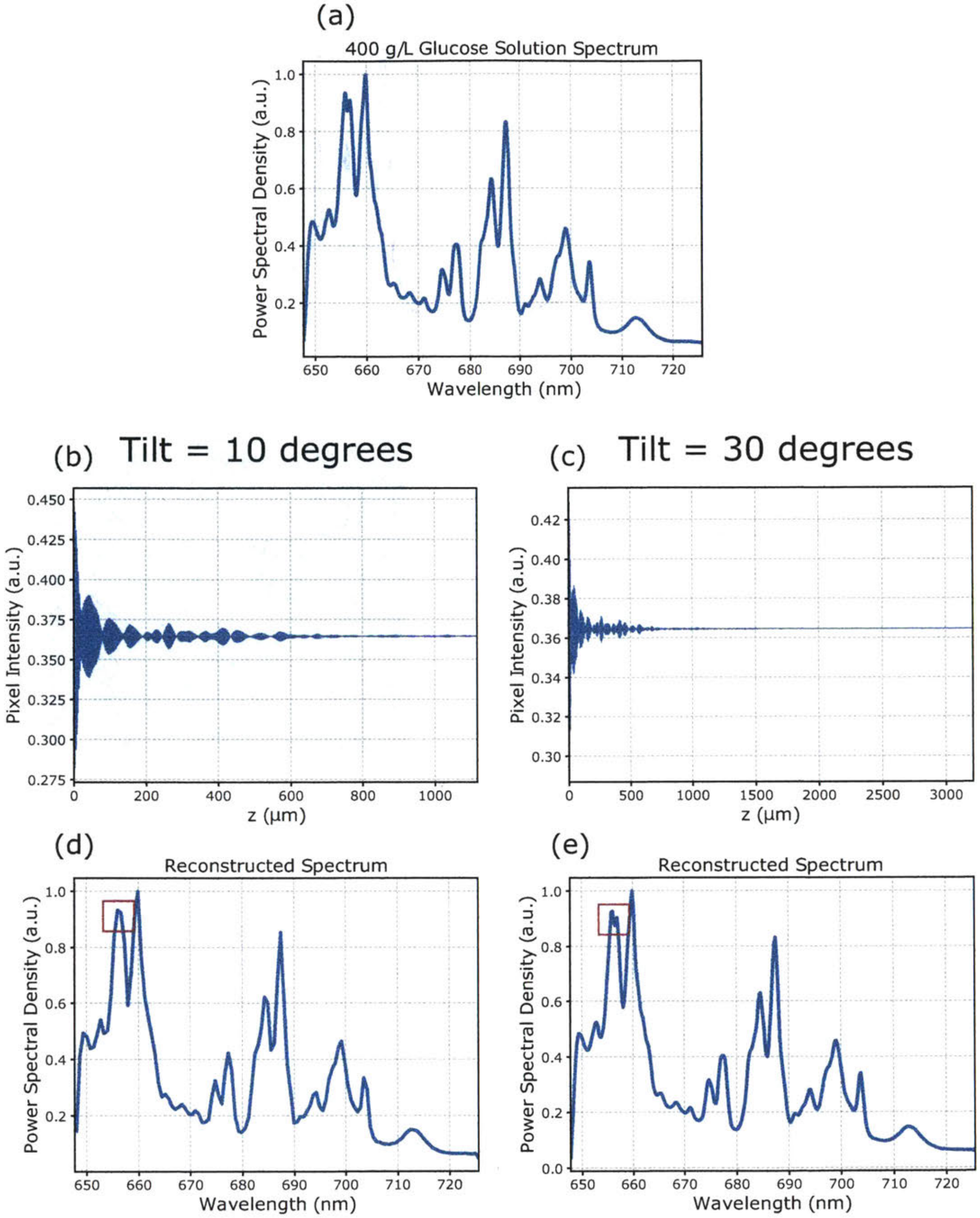


Figure 2-16: Simulation illustrations for spectrum reconstruction with the sampled Talbot interferogram for Raman spectrum of glucose solution at 400 g/L concentration. The simulated image sensor is Aptina MT9J003 as in Table 2.2. (a) The input light spectrum as the Raman spectrum of glucose solution at 400 g/L concentration. (b) and (c) The 1-D Talbot interferogram sampled by the image sensor tilting at 10° and 30°. (d) and (e) The reconstructed spectra for the two tilt angles. The resolution enhancement with a larger tilt angle results in resolving the doublet peak at ≈ 656 nm in (e).

the pivot axis such that the Talbot interferogram in the near-field is sampled. As can be seen from Figure 2-15 (b), almost exact spectral reconstruction can be obtained in this case. The sampled 1-D interferogram and the 2-D Talbot pattern are also shown in the figure.

We further simulate the system response with a relatively narrowband spectral source with fine spectral features, which is the Raman spectrum of glucose solution at 400 g/L concentration collected in our lab. The excitation wavelength is at 632 nm and the Raman spectrum covers around 70 nm in the visible range. The plot is shown in Figure 2-16. The sensor and grating are the same as in the previous simulation. We simulate the collected Talbot interferograms and reconstructed spectra with two camera tilt angles at 10° and 30°. The spectrometer resolution at 10° tilt angle is worse than 1 nm and should not be able to resolve the doublet peak at $\approx$ 656nm. This is what we can observe in the reconstructed spectra as shown in Figure 2-16 (d) and (e). This further corroborates our design guidelines for the relationship between image sensor tilt angle and spectral resolution.

For any light source with finite temporal coherence, fringes in the interferogram start to wash out after the coherence length of the light source. This is illustrated in Figure 2-17, where spectral sources with a Gaussian line shape at various Full width at half maximums (FWHMs) are used. The coherence length can be roughly characterized as

$$l_c = \frac{\lambda^2}{\delta\lambda},$$

with λ being the center wavelength and $\delta\lambda$ being the FWHM. l_c for the various input spectra plotted in Figure 2-17 correspond well with the length scales of the interferogram fringes in the plots. This is expected due to the interferometric nature of the Talbot effect.

For Fourier spectrometers like the Michelson interferometer, it is crucial to sample around the zero path-length delay region, especially for broadband sources. The same is true for the Talbot spectrometer, which means that interferogram needs to be collected starting right after the grating. From a practical system-building

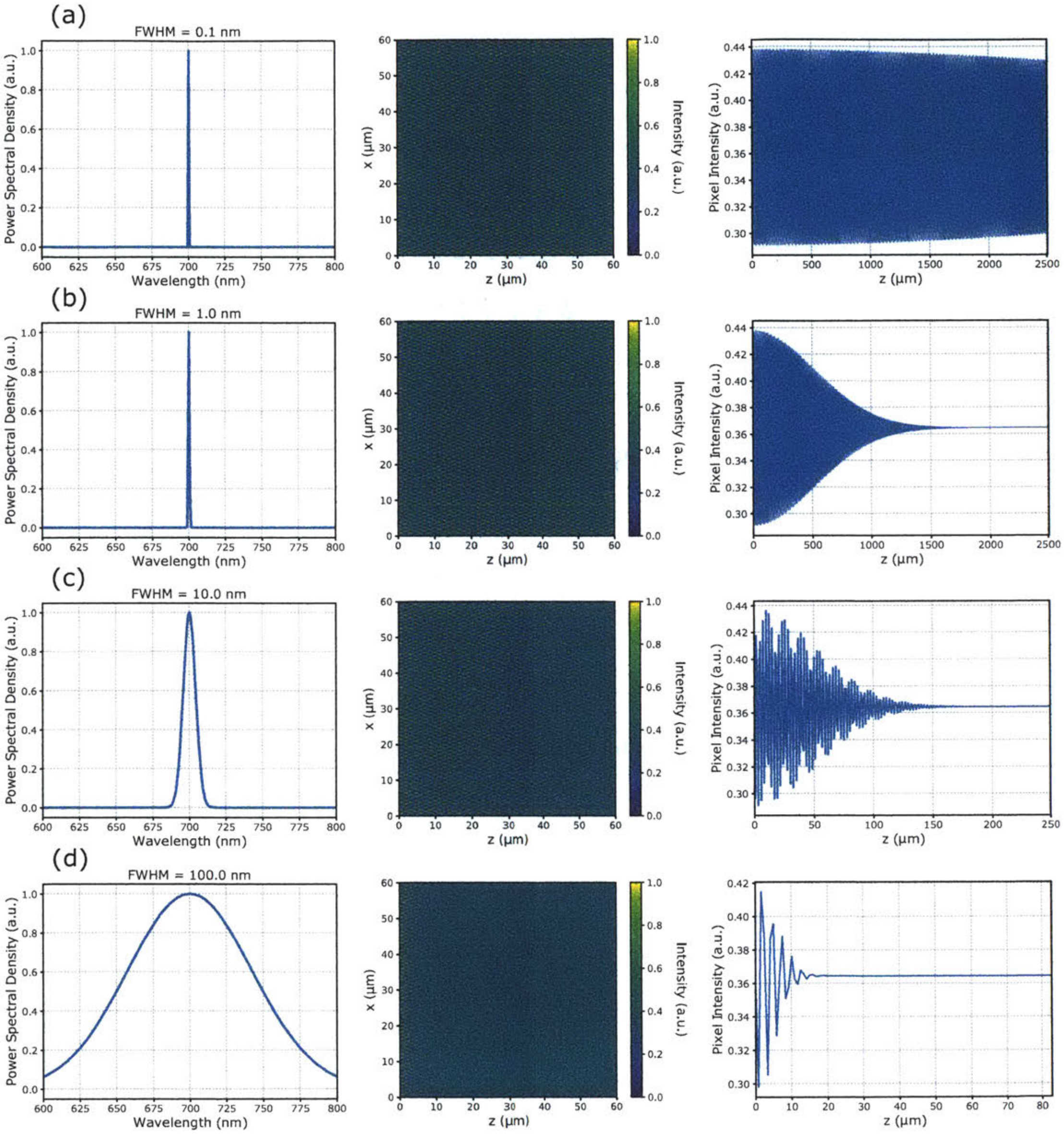


Figure 2-17: Illustrations for the Talbot pattern and interferogram change as the temporal coherence of the input light source varies. The extent of the fringes in the Talbot interferogram reduces as the FWHM of the input spectrum increases.

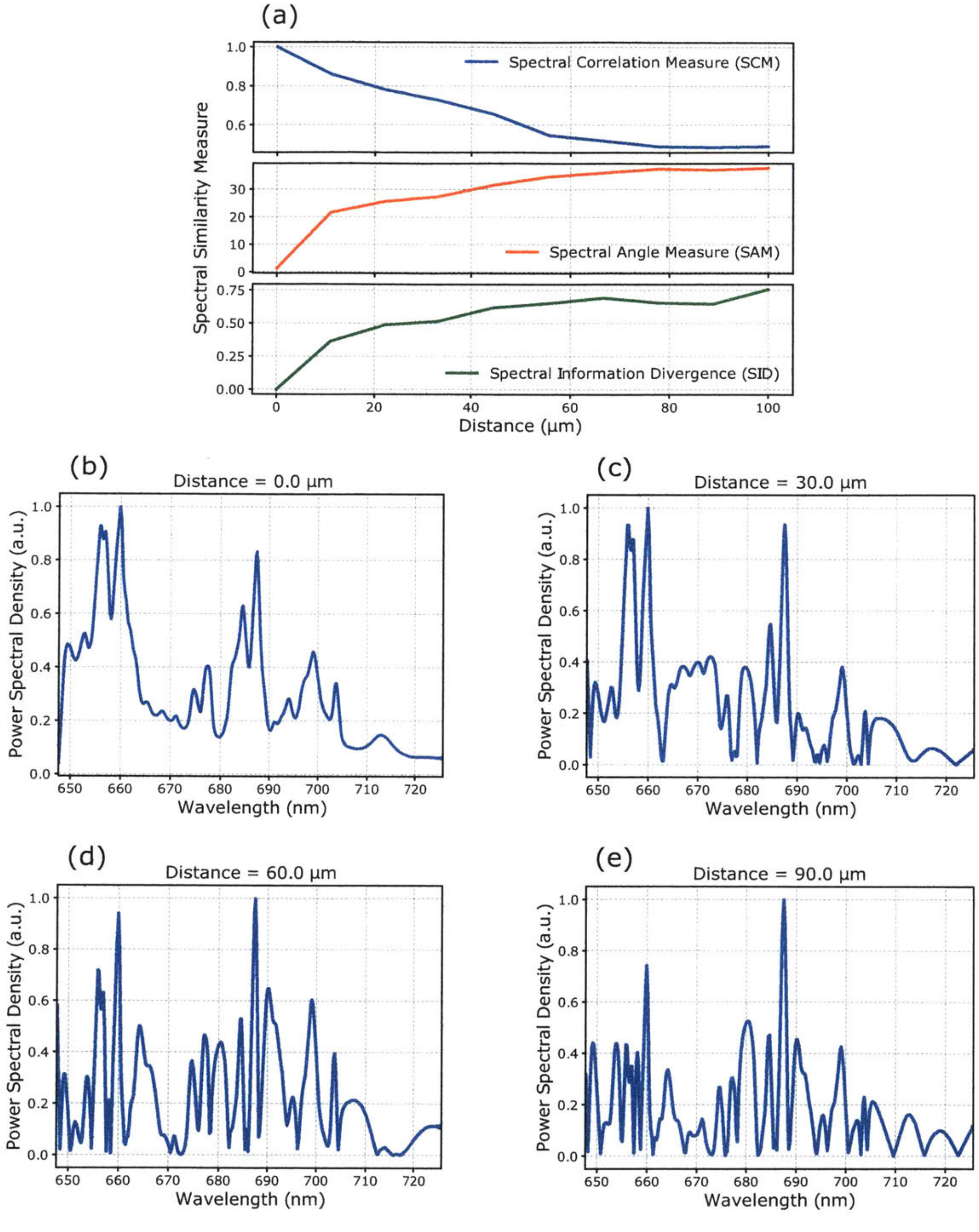


Figure 2-18: Illustrations for the grating-sensor distance requirement for spectral reconstruction with temporally incoherent light sources. (a) SCM, SAM and SID change as a function of the grating-sensor distance. (b) – (e) Reconstructed spectra as the grating-sensor distance varies from 0 to 90 µm.

perspective, this means that the grating has to touch the surface of the image sensor. We illustrate this point by investigating how the reconstructed spectrum evolves as the grating-sensor distance changes, which is defined as the closest distance in the z direction between the grating and the sensor surface. The results are plotted in Figure 2-18. Several popular spectral similarity measures [109] are used to quantify the distances between the reference spectrum, which is the glucose Raman spectrum that we choose as the input light source, and the reconstructed spectra under different grating-sensor distances. Assume that there are N spectral points for the reference spectrum $\mathbf{p}$ and the reconstructed spectrum $\mathbf{q}$, the spectral correlation measure (SCM) is defined as

$$\mathrm{SCM} = \frac{N\sum_{i=1}^{N} p_i q_i - \left(\sum_{i=1}^{N} p_i\right)\left(\sum_{i=1}^{N} q_i\right)}{\sqrt{N\sum_{i=1}^{N} p_i^2 - \left(\sum_{i=1}^{N} p_i\right)^2}\sqrt{N\sum_{i=1}^{N} q_i^2 - \left(\sum_{i=1}^{N} q_i\right)^2}}.$$

The spectral angular measure (SAM) is defined as

$$\mathrm{SAM} = \cos^{-1}\left[\frac{\sum_{i=1}^{N} p_i q_i}{\sqrt{\left(\sum_{i=1}^{N} p_i^2\right)\left(\sum_{i=1}^{N} q_i^2\right)}}\right].$$

Lastly, the spectral information divergence (SID) is defined as

$$\mathrm{SID} = D_{KL}(\hat{\mathbf{p}}||\hat{\mathbf{q}}) + D_{KL}(\hat{\mathbf{q}}||\hat{\mathbf{p}}),$$

where $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ are the normalized spectral signals and the Kullback-Leibler (KL) divergence $D_{KL}(\hat{\mathbf{p}}||\hat{\mathbf{q}})$ from $\hat{\mathbf{p}}$ to $\hat{\mathbf{q}}$ is

$$D_{KL}(\hat{\mathbf{p}}||\hat{\mathbf{q}}) = \sum_{i=1}^{N} \hat{p}_i \log\left(\frac{\hat{p}_i}{\hat{q}_i}\right).$$

The plot showing how these spectral measure changes as a function of grating-sensor distance is shown in Figure 2-18 (a), and the corresponding reconstructed spectra are shown in Figure 2-18 (b) – (e). As can be seen from the plot, almost-perfect reconstruction is only possible when the sampling happens right after the grating surface at the zero distance. As the grating-sensor distance increases, more spectral reconstruction distortion is observed, especially for the slowly varying content in the original spectrum. This is expected as the sharp spectral peaks in the input Raman signal have longer coherence lengths and are more resistant towards sampling loss closer to the zero path-length difference region. All the spectral similarity measures become worse as the grating-sensor distance increases, with a strong decrease in similarity measure even when the grating-sensor distance is around ≈ 10 µm. This highlights the importance of minimizing the grating-sensor distance for temporally incoherent signals.

Incidence with Angular Spread

An important specification for any spectrometer is its light throughput, which can be characterized with its response function towards diffuse light with an incidence angle spread. In this part we investigate the system response of the Talbot spectrometer under spread incidences. An illustration for incidence angle spread on the grating surface is shown in Figure 2-20. We consider incidence beam angular spread over two directions, namely θ and ϕ, which are both defined in the previous sections.

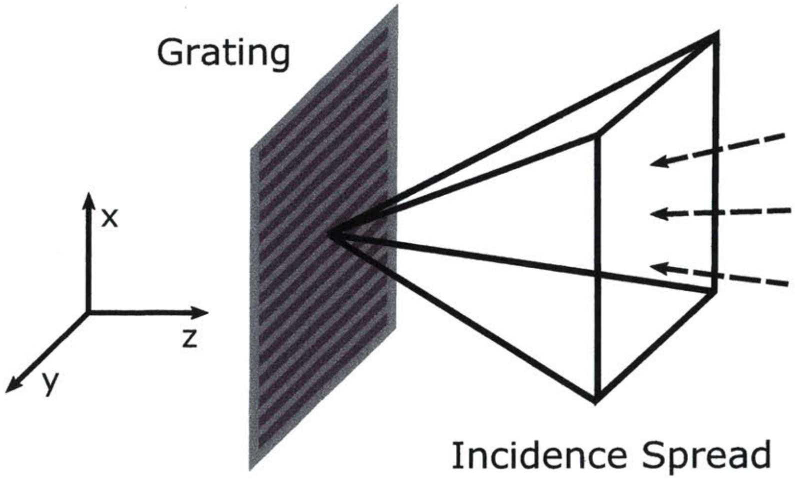


Figure 2-19: Illustration for incidence angle spread on the grating surface.

To quantify and visualize the effect of angular spread on spectral reconstruction,

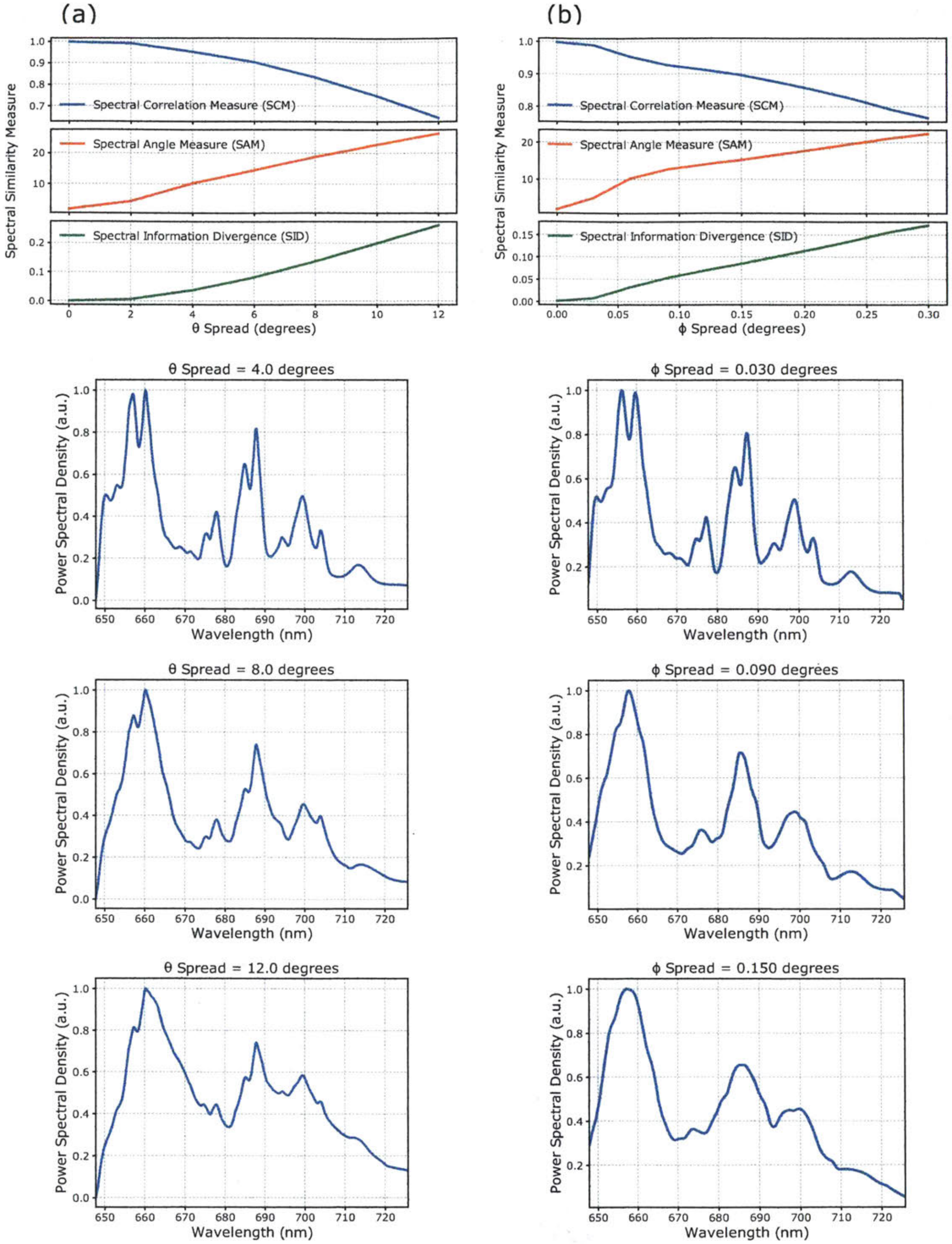


Figure 2-20: Simulations illustrating the effect of incidence angle spread in θ or ϕ on the reconstructed spectra for the Aptina MT9J003 system as in Table 2.2. Column (a) SCM, SAM and SID as spread in θ changes and some sample reconstructed spectra. Column (b) SCM, SAM and SID as spread in ϕ changes and some sample reconstructed spectra. The tilt angle is 30° in the simulation.

we perform simulations similar to the previous part, where an input spectrum of Raman glucose solution is assumed and the reconstructed spectrum is obtained from the sampled interferogram. Equation 2.14 for Talbot pattern calculation under general incidence with tilt in θ and ϕ is used in this simulation. In addition to discretizing the power in the wavelength domain, we also discretize the incidence power for different angular tilt in θ and ϕ within the angular spread and add the contributions from all the discretizations in the intensity domain as

$$I(x, y, z) \approx \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} S(\lambda_i, \theta_j, \phi_k) I(\lambda_i, \theta_j, \phi_k; x, y, z) \, \delta\phi \, \delta\theta \, \delta\lambda.$$

The underlying assumption for this treatment is that the worst-case spatial incoherence is assumed such that all the intensity fields from different incidence angles are added incoherently. This can be a reasonable assumption for diffuse sources such as thermal light but may be violated for other types of light sources. For more tailored simulations, stochastic transmission screen methods can be used for partial spatial coherence studies [97], which we do not explore in this study.

Figure 2-20 shows the results where the reconstructed spectra are obtained through various incidence angle spread in either θ or ϕ. Different spectral similarity measures including the SCM, SAM and SID are plotted for both cases. As can be seen from the plots, angular spread effectively blurs the reconstructed spectra with the worst-case spatial incoherence assumption. However, the tolerance for the two directions are different. θ spread is much more tolerable than ϕ spread in terms of preserving the spectral resolution. This is due to the fact that spectral dispersion is mainly introduced in the x direction, which corresponds to the ϕ tilt. As a result, small disturbance in ϕ can result in significant pattern response change as discussed in Section 2.1.3. This angular spread characteristics is similar to that of conventional grating spectrometers, which use slits to restrict the incidence angle spread onto the grating more in one direction than the other. The similarity should come as no surprise since the same dispersive element is used in both types of spectrometers. The difference for the two spectrometer types are in terms of sample collection region, where the grating spectrometers collect diffraction images in the far field and the Talbot spectrometers capture interferogram patterns in the near/mid field.

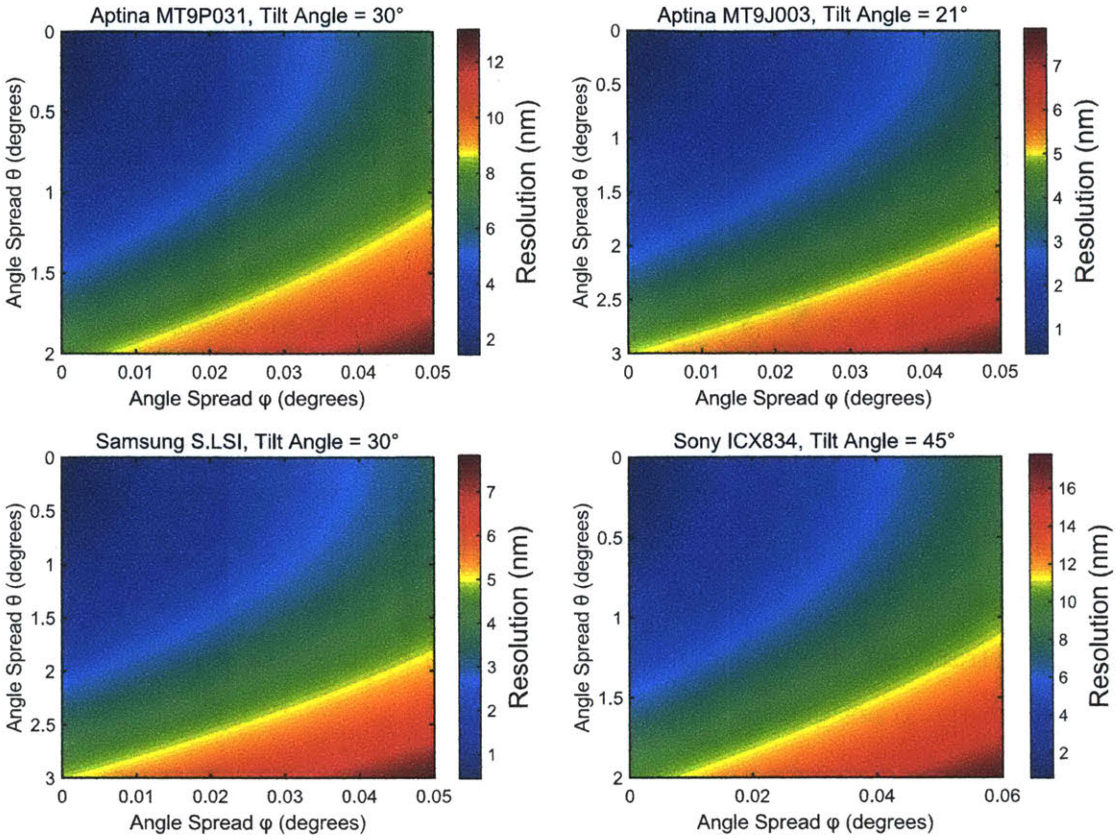


Figure 2-21: Theoretical resolution as a function of incidence angle spread in θ and ϕ for the Talbot spectrometer systems in Table 2.2.

We further theoretically characterize the achievable resolution in terms of incidence angle spread. This is based on the fact that the Talbot wave vector is a result of the interference amongst various diffractive orders as discussed in Section 2.1.4. A tilted incidence results in a shift in the Talbot wave vector and a spread incidence results in a blur in the wave vector domain. Therefore, the effect of incidence angular spread can be characterized by considering the Talbot wave vector blurring with varying spread in both θ and ϕ. Figure 2-21 shows the effective resolution change for all the Talbot systems in Table 2.2 as a function of angular spread in both θ and ϕ. The light throughput, in terms of the etendue, for a set wavelength resolution can also be calculated from the plot. For example, for the Aptina MT9J003 sensor, the calculated etendue for $\approx$ 1 nm resolution is $\approx$ 130 µm^2. This is around half of that from a commercial compact spectrometer like the Ibsen FREEDOM spectrometer with a much larger image sensor. However, as the Talbot spectrometer utilizes near/mid field pattern sampling, it can be of considerably smaller footprint than that of the Ibsen FREEDOM system (which is built upon a dispersive system with far field imaging).

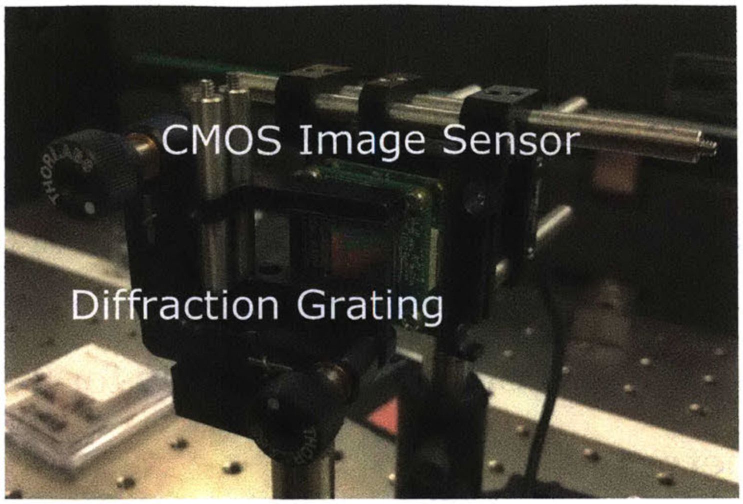
2.3 Experimental Setup and Results for Talbot Spectrometers

2.3.1 Experimental Setup

Over the course of this work, three Talbot spectrometer systems had been built with the Aptina and Samsung sensors as shown in Table 2.2. Characterization had been performed from $\approx$ 780 nm to 950 nm with a tunable Ti:Sapphire laser. All the image sensors did not have any color filters but did have the micro-lens array on top of the pixels. The fused silica surface relief gratings used in the study were produced as holographic masterpieces by Ibsen Photonics.

The experimental setup for characterizing the Talbot spectrometer systems is shown in Figure 2-22. In the top plot, the Aptina CMOS image sensor is shown behind the diffraction grating. In the bottom plot, a complete system with the Samsung S.LSI CMOS sensor is shown. A single-mode optical fiber was used to deliver the input light. Afterwards, a fiber collimation lens and a 10X beam expander were used to expand the beam to fill the entire grating surface with normal incidence. Translational and rotational stages were used to host the image sensor in order to provide the necessary alignment for maximizing the visibility and also to tilt the image sensor for resolution characterization. The Aptina CMOS sensor had an integrated readout board for data acquisition and transfer, whereas the Samsung S.LSI CMOS sensor was controlled by an independent readout board. Both sensors are shown in Figure 2-22.

In additional to the experimental setups we built to characterize the Talbot spectrometer systems, we also built prototypes for portable systems, where the frames



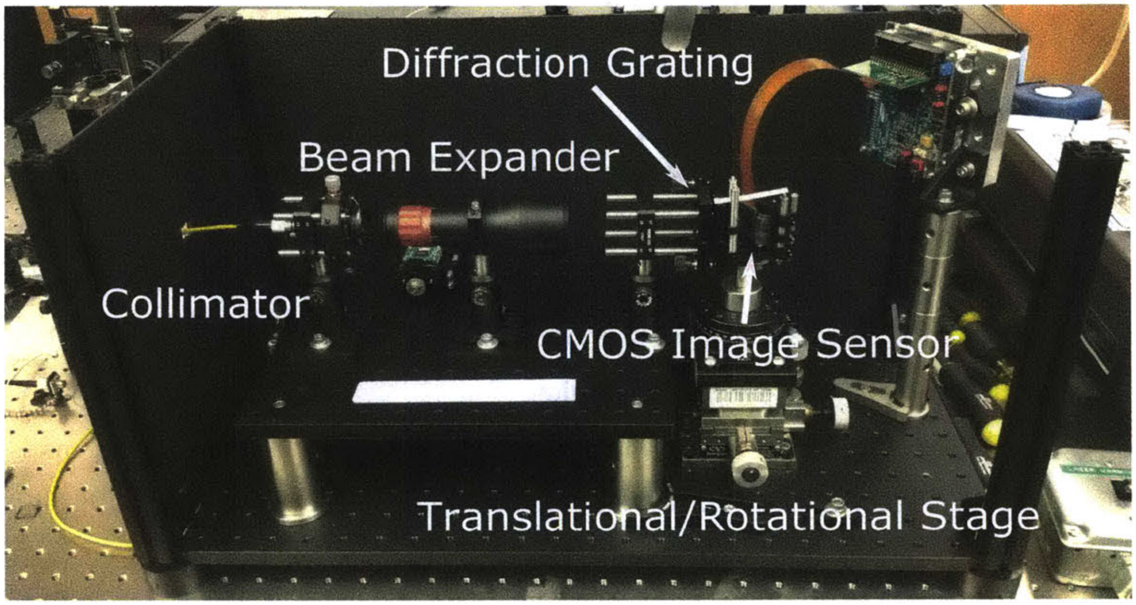


Figure 2-22: Experimental setup for characterizing the Talbot spectrometer systems. (top) The diffraction grating and the Aptina CMOS image sensor. (bottom) The setup used to characterize the relationship between spectral resolution and sensor tilt angle with the Samsung S.LSI sensor.

were built from 3D printed parts to demonstrate the compactness of our systems. Figure 2-23 shows images of the instruments with the Aptina sensors. Some of the experimental work, such as the high resolution wavemeter characterization that will be detailed later, were carried out with the portable systems.

Figure 2-24 shows a series of plots for simulated and experimentally captured

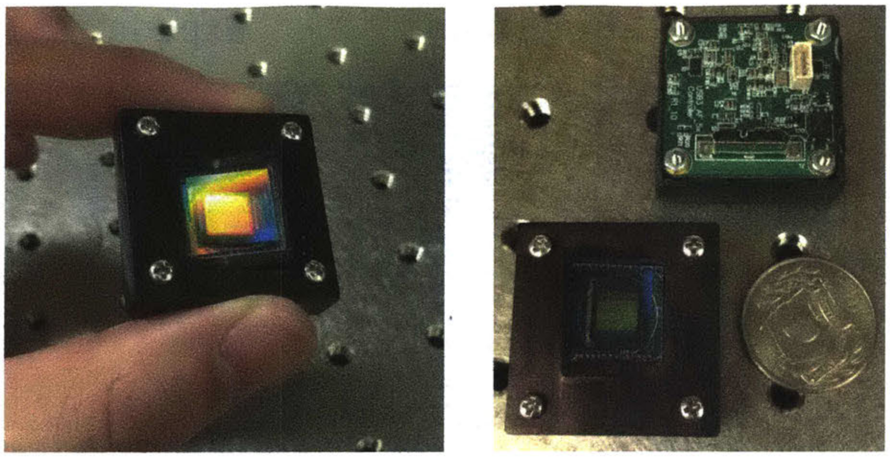


Figure 2-23: Prototypes for the portable Talbot spectrometer systems with the Aptina CMOS image sensors.

Talbot images from the Aptina MT9J003 sensor at various tilt angles with 830 nm laser light input. The images under the experiment column are raw data from the image sensor and are scaled for display. As can be seen from the plots, the experiment agrees very well with the simulation. Tilting the sensor results in periodicity change in the row direction. This means that more periods can be sampled with a larger tilt angle and as a result, a finer resolution can be achieved. After the images are captured, row-wise FFT can be performed to retrieve the spectral information, which will be discussed in the coming parts.

2.3.2 Resolution Characterization with Angle Tilt

An important experimental characterization is to verify the spectral resolution change as a function of image sensor tilt angle. We experimentally verified this with all the three image sensors from Aptina and Samsung. Figure 2-25 shows example plots of FFT-inverted spectra as the tilt angle changes for the Samsung S.LSI sensor with 830 nm laser incidence (Ondax SureLock). We used the Hanning window for apodization prior to the FFT inversion. The FWHM decreases as more tilt is introduced as can be seen from the plot. Since the laser source has $\approx$ 250 MHz linewidth, its spectral shape can be safely regarded as a delta function at 830 nm for our application. This means that the spectral resolution can be characterized approximately as the FWHM

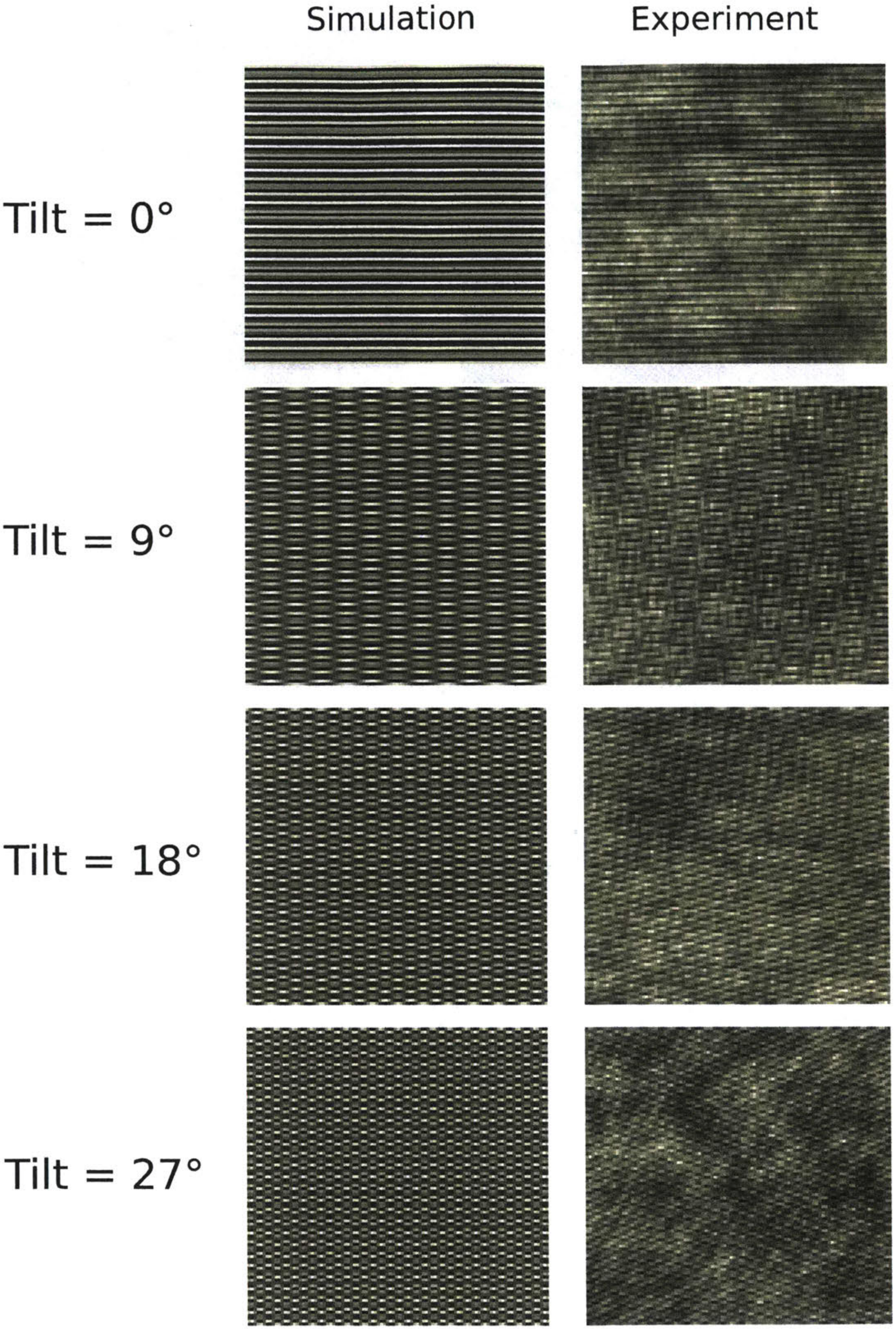


Figure 2-24: Simulated and experimentally captured Talbot images from the Aptina MT9J003 sensor at various tilt angles with 830 nm laser light.

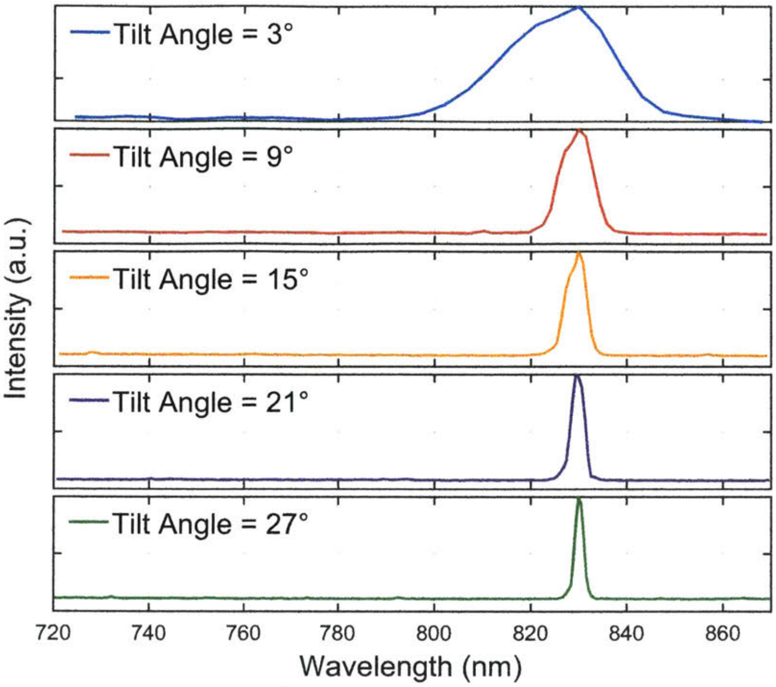


Figure 2-25: FFT inverted spectra for 830 nm laser light at various image sensor tilt angles. The FWHM of the reconstructed laser line decreases as the tilt angle increases. The sensor used in this plot is the Samsung S.LSI sensor.

of the reconstructed peak for the laser line.

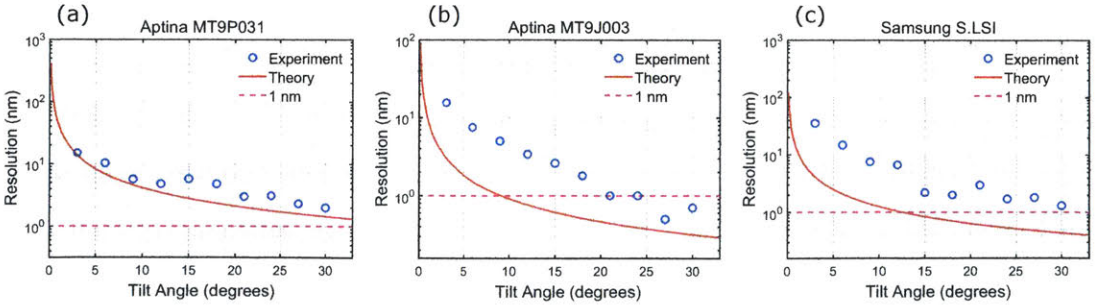


Figure 2-26: Experimental and theoretical plots showing the resolution change as a function of the tilt angle for the three Talbot spectrometer systems depicted in Table 2.2. The resolution is calculated as the FWHM of the reconstructed linewidth for the laser light at 830 nm.

With this in mind, we quantitatively characterized the FWHM of the laser line at various tilt angles from 3° to 30° for the three smartphone and compact camera sensor systems, namely the Aptina MT9P031, the Aptina MT9J003 and the Samsung S.LSI sensors. The results are plotted in Figure 2-26. In addition to the experimentally obtained data, we also plot the theoretical resolution as a function of tilt angle in the same plots. Overall the experiment agrees well with the theory, with the best match for the Aptina MT9P031 sensor system with a 2.2 µm pixel size and a 1.608 µm grating pitch size. The greater deviation from the other two systems can be due to the fact that the gratings used for the Aptina MT9J003 and Samsung S.LSI systems had a smaller pitch than that used with the Aptina MT9P031 system, and therefore can result in more deflection for the ±1 diffractive orders. This means that the incidence beams for the Aptina MT9J003 and Samsung S.LSI systems had more tilted incidence angles, and therefore can have more unpredictable sensor sampling responses. In addition, the micro-lens arrays can also introduce more sampling function distortions with smaller-sized pixels and more incidence angle tilt. As a result, some differences are observed from the experiment compared to the theory for the smaller pixel sensors with smaller grating pitches. However, the overall trend of higher resolution with more tilt angle is well preserved, indicating that our theory is able to provide useful guidelines in terms of designing the spectrometer configuration for target resolution requirements.

We further characterized the spectral response with two mutually-incoherent lasers as the input source. One laser was the Ondax laser at 830 nm as before and the other laser was the tunable Ti:Sapphire laser. The light from these two sources were combined by a fiber coupler and the output was fed into the fiber collimator as in Figure 2-22. We fine-tuned the Ti:Sapphire emission wavelength such that the spectral spacing between the Ti:Sapphire and the Ondax laser was varied, and the corresponding spectra were reconstructed as shown in Figure 2-27. The data was collected with the Aptina MT9J003 sensor tilting at 20° that had a theoretical resolution of $\approx$ 0.5 nm and a measured FWHM of $\approx$ 1 nm. As can be seen from the plot, wavelength separation of down till 0.9 nm can be resolved by the Rayleigh criterion. At a wavelength separation of 0.3 nm, which is below both the theoretical resolution limit and the measured FWHM of the laser linewidth with our system,

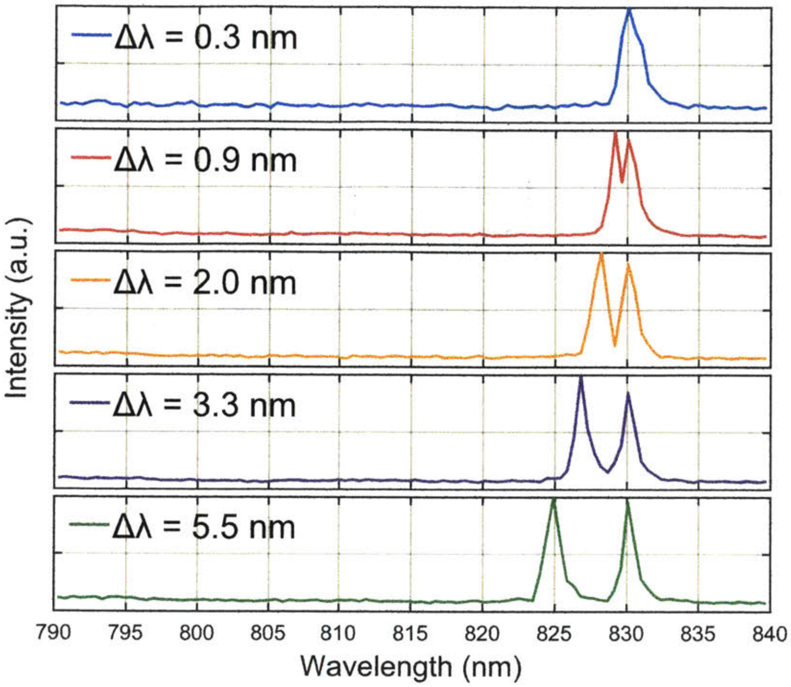


Figure 2-27: Reconstructed spectra for two laser lines separated by various spectral spacings for the Aptina MT9J003 Talbot system with a 20° tilt angle. At $\delta\lambda = 0.3$ nm, which is below the theoretical resolution for this tilt angle, the two laser lines can no longer be resolved.

the two peaks can no longer be resolved. This result further confirms our resolution theory with a different experimental resolution criterion in addition to the FWHM of the laser line that was discussed previously.

2.3.3 Response Span with Tunable Laser Characterizations

Next, we characterized the bandwidth response of the Talbot spectrometer with the broadly tunable Ti:Sapphire laser. We tuned the Ti:Sapphire laser from 780 nm to 950 nm in steps of 10 nm, and recorded the sensor response with the Aptina MT9J003 image sensor system at 20° tilt angle. The reconstructed spectra are shown in Figure 2-28. There are several aspects worth noticing with this characterization. Firstly, the Talbot spectrometer was able to reconstruct spectral signal over a broad

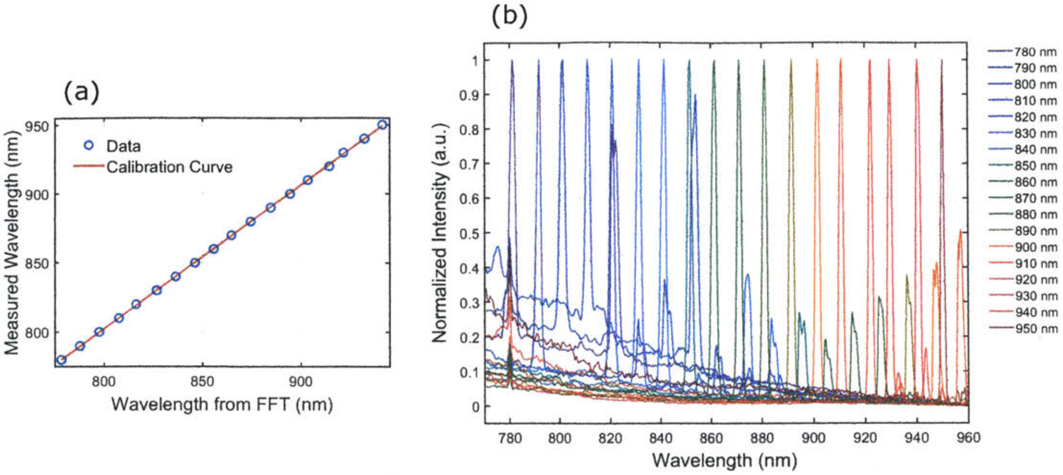


Figure 2-28: (a) Calibration curve used for spectral wavelength estimation. (b) Reconstructed spectra across 780 nm to 950 nm with the tunable Ti:Sapphire laser for the Talbot spectrometer system built with the Aptina MT9J003 image sensor.

spectral range across ≈ 170 nm as shown in Figure 2-28 (b). This should come as no surprise. Our demonstration over even broader spectral range is only limited by the availability and tunability of our source at optical wavelengths. Secondly, due to the fact that the incidence angle cannot be controlled as perfectly normal, two major Talbot peaks are shown in the reconstructed spectra, which is caused by incidence tilt in ϕ as discussed in Section 2.1.3. Thirdly, also due to the imperfect incidence angle control, the reconstructed wavelengths have some offset compared to the reference wavelength. This means that a wavelength calibration needs to be performed for accurate wavelength estimation. Precise wavelength calibration can be performed with Equation 2.8 for general tilt in both θ and ϕ based on the spacing between the two Talbot peaks and their offset to the reference wavelength. Here, we adopted a simpler linear regression approach which is able to provide good calibration results as shown in Figure 2-28 (a).

Problems associated with having more than one peak in the reconstructed spectrum can be solved in two practical ways. The first one is to place the sensor carefully such that it is mostly sampling the Talbot pattern from the interference of the 0 and +1 diffractive orders. The second one is to tilt the incidence angle in the ϕ direction such that only one 1 diffractive order can exist. We demonstrate the first approach

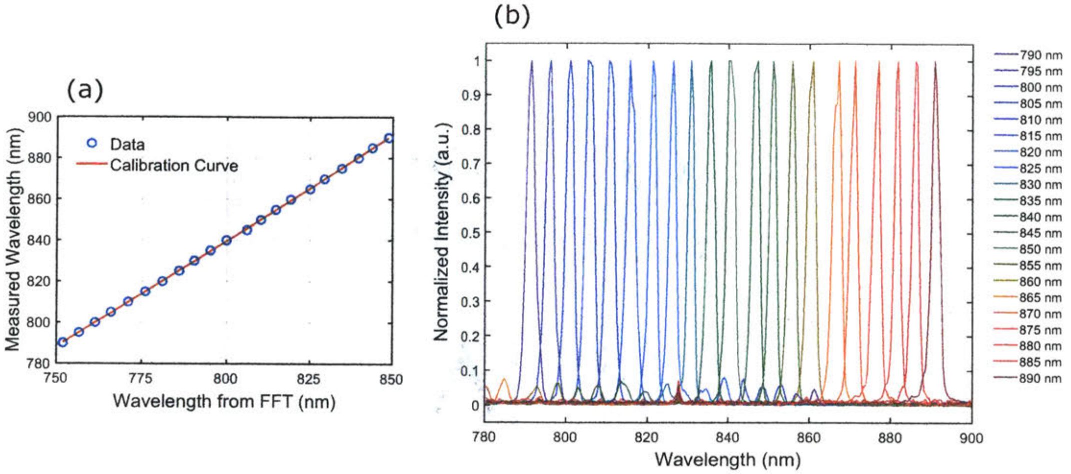


Figure 2-29: (a) Calibration curve used for spectral wavelength estimation. (b) Reconstructed spectra across 790 nm to 890 nm with the tunable Ti:Sapphire laser for the Talbot spectrometer system built with the Samsung S.LSI image sensor.

with Figure 2-29, where we used the translational stage as in Figure 2-22 (b) to only sample the Talbot pattern from two diffractive orders. The image sensor used in this case is the Samsung S.LSI sensor at 30° tilt angle. Figure 2-30 shows the simulated and captured Talbot pattern from this system. As opposed to the chessboard-like patterns obtained in Figure 2-24, the Talbot pattern in Figure 2-30 shows tilted stripe-like patterns due to the two-beam-interference nature. In Figure 2-29, the spectral wavelength calibration curve is provided in (a), and the FFT-inverted spectra spanning across $\approx$ 100 nm in steps of 5 nm are provided in (b). Due to the careful placement of the image sensor, the reconstructed spectra are much cleaner than that from Figure 2-28. The wavelength span is smaller than that from the Aptina sensor due to the fact that the Samsung S.LSI sensor has a worse quantum efficiency above $\approx$ 900 nm than that of the Aptina sensor, and the control board was not able to perform long integrations required to capture images with good SNR. Due to the fact that some part of the image sensor still captured both ± 1 diffractive orders, smaller secondary peaks can be observed in the figure. They can be eliminated by using a larger-area grating that has a bigger two-beam-interference region.

The effective dynamic range of the Samsung S.LSI Talbot spectrometer system can also be characterized. We plot the normalized intensity in terms of dB in Figure 2-31.

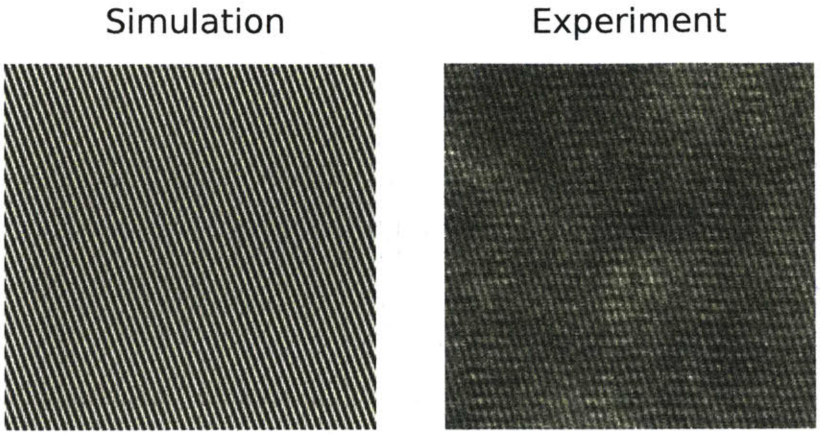


Figure 2-30: Simulated and experimentally captured Talbot images with the Samsung S.LSI sensor. The system geometry is set such that the Talbot pattern from only the 0 and +1 diffractive orders are captured by the image sensor.

The artifact peak at ≈ 828 nm is due to auto-focus pixels on the image sensor. The plot shows ≈ 20 dB dynamic range above the noise floor, which is descent but not in matching terms with conventional spectrometers. This reveals a potential challenge for the Talbot spectrometer for general purpose spectroscopy applications. As implied in the discussions in Section 2.2.2, current image sensors with the available pixel sizes may have a difficult time getting high visibility with the Talbot interferogram sampling. This issue can be further aggravated for practical image sensors due to the fact that they are generally not optimized for extremely tilted incidence angles as in the case for the Talbot spectrometer. Therefore, a large portion of the captured signal does not contain any useful spectral information. In addition, limitations in ADC bit depth on the image sensor can also potentially impact its ability to resolve fine changes in spectral reconstruction. Some of these limitations are shared with recent imaging modalities such as digital holography with smartphone and compact camera sensors [110], and may be solved or partially resolved as the image sensor technology develops. However, given the current sensor technology, these are important considerations that should be scrutinized carefully before realizing systems for practical applications.

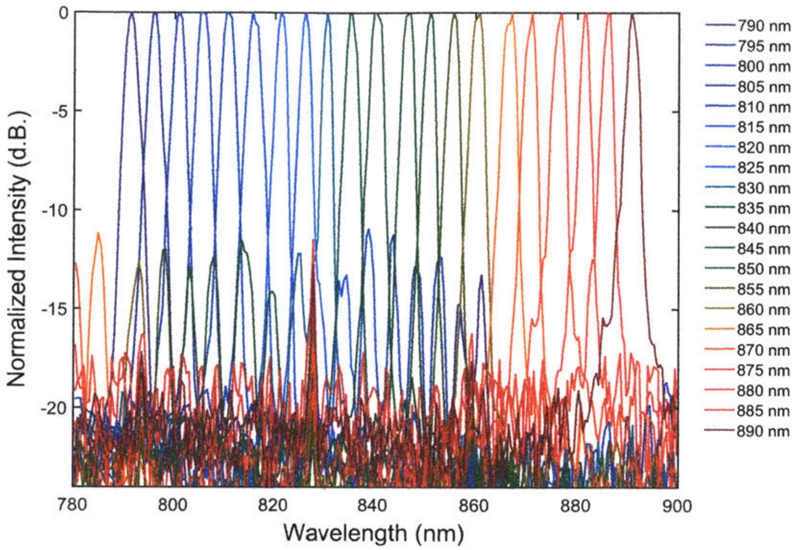


Figure 2-31: Reconstructed spectra in log scale across 790 nm to 890 nm with the tunable Ti:Sapphire laser for the Talbot spectrometer system built with the Samsung S.LSI image sensor.

2.4 Compact and High Precision Wavemeters Using the Talbot Effect and Signal Processing

2.4.1 Frequency Analysis and Estimation with Periodic Signals

For coherent light signals such as a laser source, the sampled interferogram rows (across the depth dimension) with the Talbot spectrometer essentially contain periodic signals where the spatial periodicity corresponds to the laser frequency according to Equation 2.9. Extracting the frequency (and possibly amplitude and phase) information for a periodic signal, which some term as the tone parameter estimation problem [111], can be achieved with precision[1] much higher than that from direct FFT or the spectrogram [111]. This problem has a long history in the signal processing

[1] While "super-resolution" has been used in certain fields for high accuracy peak or signal localization, we here use "high precision" to refer to this operation. In general, "resolution" is used to refer to the FFT-defined frequency separation under the sampling theorem or the Rayleigh resolution criterion in our text.

community with applications ranging from radar and sonar systems [112, 113], audio and acoustics [114, 115], astronomy [116] and many more. As a direct result of its immense presence in engineering and scientific problems, many algorithms such as the maximum likelihood estimation [111], MUSIC (MUltiple SIgnal Classification) [117, 112], and ESPRIT (Estimation of Signal Parameters via Rotational Invariance Technique) [113] amongst others [118] have been proposed and realized to solve it with extremely fine precisions.

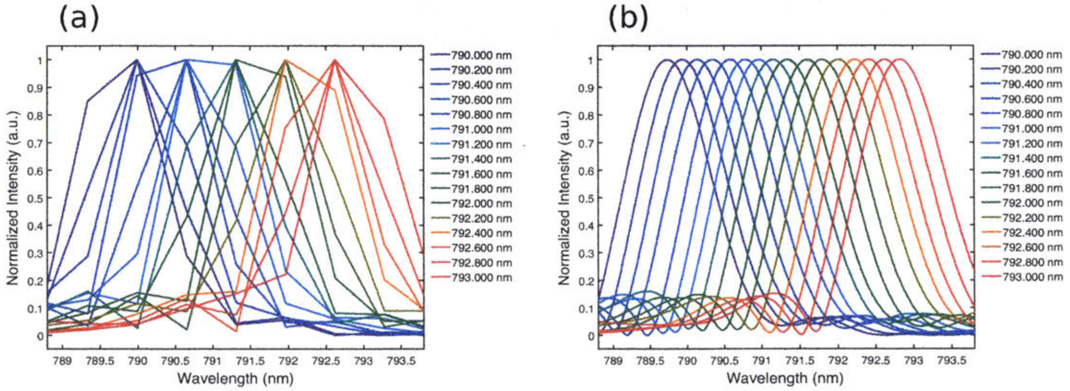


Figure 2-32: (a) Reconstructed spectra using direct FFT for laser wavelengths in steps of $\approx$ 200 pm with a tunable external cavity diode laser. (b) Reconstructed spectra using FFT with prior zero-padding for the spectral sources in (a).

While direct FFT has been used for spectrum retrieval with the Talbot spectrometer, for laser wavelength estimation, similar tone parameter extraction ideas can be applied to achieve much higher laser wavelength estimation precisions. The idea is illustrated in Figure 2-32, which shows example plots for reconstructed spectra for laser wavelengths in steps of $\approx$ 200 pm with a tunable external cavity diode laser from Sacher Lasertechnik. Figure 2-32 (a) uses direct FFT for spectral processing, whereas Figure 2-32 (b) zero-pads the interferogram rows prior to the FFT operation for precision enhancement. As can be seen from the figure, with zero-padding, much finer interpolated spectral shapes can be achieved, resulting in much more accurate center frequency estimation than that from the direct FFT estimation.

For real sinusoidal parameter estimation, assume that the underlying periodic signal $y[n]$ is

$$y[n] = A\cos(2\pi f_0 \Delta n + \phi),$$

for $n = 0, 1, 2, ..., N-1$, where A is the amplitude for the periodic signal, f_0 is its frequency, Δ is the sampling interval, and ϕ is the phase. The observed discrete noisy signal $x[n]$ is

$$x[n] = y[n] + w[n],$$

where $w[n]$ is the white Gaussian noise with variance σ^2. The Cramér-Rao lower bound (CRLB), which is the theoretical lowest error bound achievable with an unbiased estimator, for the various parameters in the model is therefore [119]

$$\mathrm{var}[\hat{A}] \geq \frac{2\sigma^2}{N},$$

$$\mathrm{var}[\hat{f}_0] \geq \frac{6\sigma^2}{\pi^2 A^2 \Delta^2 N(N^2-1)},$$

$$\mathrm{var}[\hat{\phi}] \geq \frac{4(2N-1)\sigma^2}{A^2 N(N+1)}.$$

The most relevant one for our application is $\mathrm{var}[\hat{f}_0]$. The model SNR, which is usually defined as the ratio between the variance of the signal and the variance of the noise, is

$$\mathrm{SNR} = \frac{\mathrm{var}[y]}{\mathrm{var}[w]} = \frac{A^2}{2\sigma^2}.$$

Rewriting the CRLB for frequency estimation, we have

$$\mathrm{var}[\hat{f}_0] \geq \frac{3}{\pi^2\,\mathrm{SNR}\,\Delta^2 N(N^2-1)} = \frac{3N(\delta f)^2}{\pi^2\,\mathrm{SNR}\,(N^2-1)} \approx \frac{3\,(\delta f)^2}{\pi^2\,\mathrm{SNR}\,N}.$$

Here, δf is the FFT-defined frequency domain spacing. The standard deviation for the frequency estimation error is therefore

$$\mathrm{std}[\hat{f}_0] \gtrapprox \frac{\delta f}{\sqrt{3\,\mathrm{SNR}\,N}}.$$

Based on experimentally captured Talbot images from single-frequency laser sources, the SNR from the above definition for the main Talbot peak in a single interferogram row is estimated to be ≈ 0.2 to 1.2 depending on the obtained interferogram visibility. This means that for an image sensor like the Aptina MT9J003, which has a sensor dimension of 3856×2764, the standard deviation of single-row (across the larger dimension) frequency estimation can be $\approx 0.85\%$ to 2.08% of that obtained from the FFT bin size δf. If ensemble estimation such as aggregated mean based on row-wise estimations is performed, a further reduction of $\sqrt{2764} \approx 53$ in terms of standard deviation of the estimation error is possible, assuming negligible effects from issues like the wavefront aberration, which may cause potential systematic estimation bias. This means an estimation error standard deviation of $\approx 0.016\%$ to 0.040% of δf is possible. With δf below 1 nm for most geometries tested under optical wavelengths, this corresponds to sub-picometer precision based on the 1-σ criterion and around picometer precision with the 3-σ criterion.

2.4.2 Experimental Realization, Algorithm, and Result Discussions

Experiments aiming at exploring the performance limits for the Talbot wavemeter with signal processing for wavelength estimation were carried out. Compact Talbot wavemeter setup using 3-D printed parts similar to the one shown in Figure 2-23 was used for device characterization. A tunable Ti:Sapphire laser (SolsTiS, M Squared Lasers) was used as the light source. A high-precision wavemeter (Bristol Instruments) was used to provide the reference wavelength measurements with sub-picometer precisions. Single-mode optical fiber was used for light delivery. A fiber collimation lens and a 10X beam expander were used to fill the image sensor area. In addition, a linear polarizer was used after the fiber collimation lens for polarization clean-up. The Talbot system used was the Aptina MT9J003 sensor with 1.035 µm grating pitch size. The laser source was tuned in steps of ≈ 100 to 200 pm. For each wavelength, consecutive images were obtained for estimation variance analysis.

For estimating the wavelength from the Talbot interferogram image, row-wise wavelength estimation was carried out for all the image rows across which depth samplings were performed. Afterwards, the mean of all the wavelength estimations was used as the final aggregated estimation result. Two algorithms have been extensively tested. The first one is an algorithm based on peak localization with zero-padded FFT [114]. The interferogram rows are first zero-padded to augment the array dimension by one to two orders of magnitude. FFT is then applied on the signal for spectrum retrieval. Afterwards, the maximum of the spectral peak is identified, and a parabolic approximation based on this point and its adjacent points is used for peak maximum localization. The second one is the MUSIC algorithm [117, 112], which is an eigenspace method to identify a known number of sinusoidal signals in the presence of Gaussian white noise. It is considered by many as one of the most promising algorithms for frequency estimation tasks [120]. For this algorithm, we used the MATLAB implementation (rootmusic) for our numerical processing.

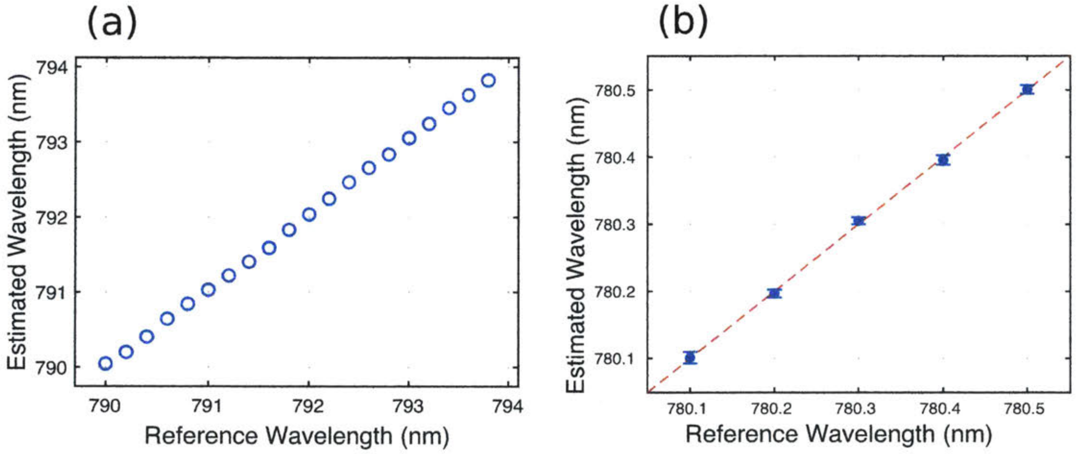


Figure 2-33: (a) Wavelength estimation results for a laser source tuned across ≈ 4 nm with 200 pm step sizes. The algorithm used here is the FFT peak localization algorithm. (b) Wavelength estimation results for a laser source tuned with 100 pm step sizes. The algorithm used here is the MUSIC algorithm. The dots in the plots are the means from 10 consecutive acquisitions. The error bars represent the standard deviations from the 10 measurements.

Overall, both algorithms were able to provide accurate wavelength estimations much better than those obtained through direct FFT inversion. Figure 2-33 shows plots for wavelength estimations across a narrow wavelength span. Figure 2-33 (a) shows wavelength estimations with the FFT peak localization algorithm for steps of 200 pm across $\approx$ 4 nm. In Figure 2-33 (b), the means as well as the standard deviations across 10 consecutive acquisitions are shown for wavelength steps of 100 pm with the MUSIC estimation algorithm. Wavelength calibrations were performed prior to the plot to account for the non-perfect incidence angle control in our setup as mentioned in Section 2.3.3. To quantify the estimation uncertainty, which defines the resolution achievable with our approach, we used 10 wavelength measurements, each with 10 consecutive acquisitions, to calculate the estimator standard deviation. The mean standard deviation for the FFT peak localization algorithm was $\approx$ 8.5 pm, whereas the mean standard deviation for the MUSIC algorithm was $\approx$ 6.6 pm. Both algorithms were able to provide sub-10 picometer estimation standard deviation, with the MUSIC algorithm having a slightly more accurate estimation.

The effect of mean aggregation across different interferogram rows is investigated next. This is shown in Figure 2-34, where we varied the number of averaging rows from 1 to 2701 in steps of 100. Figure 2-34 (a) shows the results for the FFT peak localization algorithm and Figure 2-34 (b) shows the results for the MUSIC algorithm. As can be seen from the plots, mean aggregation across the interferogram rows can enhance the estimation accuracy significantly. For both estimation algorithms, close to an order of magnitude estimation precision improvement can be achieved by averaging all the available rows from the image sensor. The most significant improvement is across the first $\approx$ 1000 rows, with plateaued performance after the first $\approx$ 1000 rows. The plateaued performance improvement is likely caused by the fact that in our experiment, aberrations with the collimation setup as well as the non-ideal pixel sampling either due to the oblique incidence or the micro-lens array can introduce phase errors in our interferogram signal. This causes Fourier-domain spectral peak distortions as well as peak position misalignment across different interferogram rows, weakening the efficacy of mean aggregation in terms of uncertainty reduction. This can likely be improved by better collimation setup for aberration reduction or by using algorithms that can correct systematic phase errors during estimation.

While the MUSIC algorithm yielded better accuracies in terms of estimation

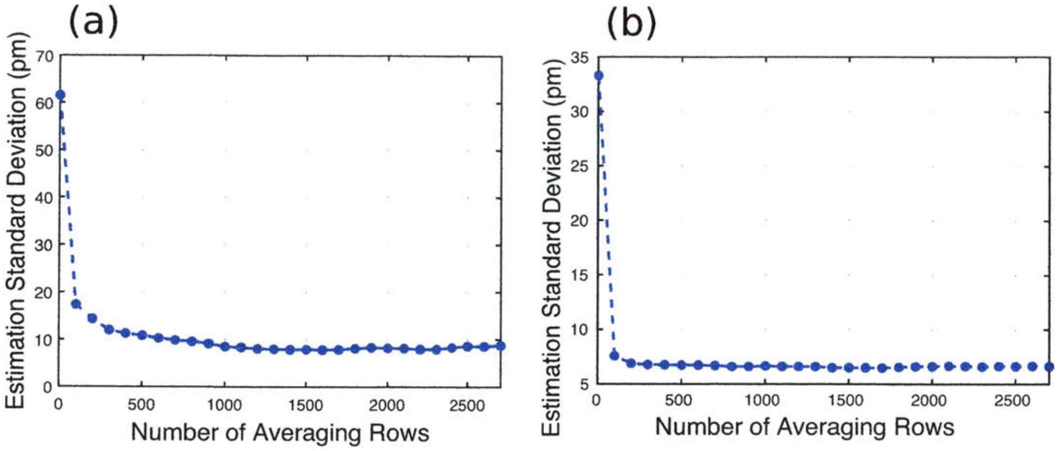


Figure 2-34: Estimation standard deviation as a function of the number of averaging rows for (a) the FFT peak localization algorithm, and (b) the MUSIC algorithm. The standard deviations were calculated from 100 acquisitions with 10 different wavelength points around 780 nm.

consistency, a significant advantage for the FFT peak localization algorithm is its computational speed. While extra memory is needed for storing the zero-padded image, row-wise FFT across a two-dimensional image has vectorized and well-optimized code executions [121]. The remaining operations generally have linear time complexity and are easily vectorized. As discussed earlier, aggregating row-wise estimation results is one of the keys to achieve an accurate estimation, therefore being able to perform fast and parallel frequency estimations across several thousand interferogram rows can be extremely advantageous. In general, our experiments suggested at least ≈ 20 times faster computational speed for the FFT localization algorithm as compared to the MUSIC algorithm. In addition, the MUSIC algorithm requires the number of sinusoidal signals to be known in advance. In our numerical study, this was taken care of by spectrally filtering the wavelength region of interest prior to MUSIC estimation to ensure the estimation robustness. This aspect can be handled more easily with the FFT localization algorithm, as one can incorporate a heuristic approach for peak identification and thresholding within the FFT localization steps in a straightforward manner.

2.5 Conclusions

In conclusion, we have demonstrated using the non-paraxial Talbot effect with modern image sensors to build high performance spectrometers and wavemeters. Due to the dimensions of the relevant elements, sampling and processing the Talbot interferogram for high spectral discrimination not only is a computational problem but also involves hardware selection and optimization. Strategies and recommendations for dealing with source temporal incoherence and angular spread incidence have been provided. By experimentally realizing several Talbot spectroscopy systems with different image sensor and grating selections, we verified our design theory and demonstrated the compact and high performance Talbot spectroscopy solution with nanometer resolution across a broad wavelength bandwidth. With further statistical signal processing, we demonstrated high precision wavemeters using our device with estimation resolution below 10 picometers (using the 1-σ criterion). Unlike the recent advances in computationally-enabled compact spectroscopy solutions, our Talbot spectroscopy solution does not require any full-spectrum calibration process, which is a significant advantage for practical adoptions. While the visibility of the interferogram is suboptimal at the moment, we envision that the general performance of Talbot-based spectroscopy systems will improve as the image sensor technology advances with more densely packed pixel arrays. As discussed in this chapter, using the Talbot effect for general spectroscopy can have stringent requirements on issues like sensor-grating positioning. Significant further effort might be required to demonstrate its performance limits. However, for coherent light sources, Talbot wavemeters based on the existing approach are already having operating bandwidths and precisions close to those offered in the commercial domain. With further engineering and optimization, it is foreseeable that the compact Talbot wavemeter solution can take off in the near future.

Chapter 3

Bayesian Modeling and Computation for Analyte Quantification in Complex Mixtures Using Raman Spectroscopy

This chapter discusses our development in an analyte quantification algorithm for complex mixtures using Raman spectroscopy. We use a Bayesian inference and modeling approach with reversible jump Markov chain Monte Carlo (RJMCMC) computation for spectral shape modeling. This framework is introduced in Section 3.1 with detailed discussions on functional modeling, prior selection, the Bayesian computation process, and our two-stage algorithm. Section 3.2 introduces numerical experiments for performance validation and exploration with our algorithm. The mixture interfering environment and noise condition, as well as algorithm performance comparison with several popular multivariate regression algorithms are explored. Section 3.3 presents the estimation results of our algorithm on two experimental datasets. The first one is a physical mixture dataset and the second one is a dataset for nutrient monitoring with mammalian cell culture processes. At last, Section 3.4 provides the conclusions for this chapter.

3.1 Bayesian Formulation and Monte Carlo Computation for Spectral Data Analysis

3.1.1 Functional Model for Spectral Signal

Raman spectra are typically collected as one-dimensional signals from a CCD or CMOS detector placed after a dispersive element such as a diffraction grating. Assuming that there are N spectral data points, we model the discrete Raman signal as

$$\mathbf{y} = f_P(\boldsymbol{\nu}) + f_B(\boldsymbol{\nu}) + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\mathbf{y} \in \mathbb{R}^N$ represents the spectrum array, $\boldsymbol{\nu} \in \mathbb{R}^N$ represents the corresponding Raman shift in wavenumbers, $f_P(\boldsymbol{\nu})$ and $f_B(\boldsymbol{\nu})$ are the functional arrays describing the shape for the Raman peaks and baseline of the signal, and $\boldsymbol{\epsilon} \in \mathbb{R}^N$ is the noise term. $f_P(\boldsymbol{\nu})$ is modeled as the sum of individual Raman peaks each corresponding to an energy transition level as

$$f_P(\boldsymbol{\nu}) = \sum_{j=1}^{k_P} \beta_{P,j} g(\boldsymbol{\nu}; \theta_{P,j}),$$

where $g(\boldsymbol{\nu}; \theta_{P,j})$ is the functional form for the shape of the j-th peak with $\theta_{P,j}$ containing the shape variables, and $\beta_{P,j}$ is the corresponding amplitude variable. Depending on the relative contributions from the amplitude correlation time and the coherence lifetime to the effective lifetime of the excited energy states, the functional line shape of a Raman peak can be of the Gaussian profile, the Lorentzian profile, or a combination of both, in which case it can be represented by the Voigt profile [48]. As a popular approximation to the computationally-expensive Voigt profile, the pseudo-Voigt profile uses a linear combination of the Gaussian profile and the Lorentzian profile controlled by a weight factor to adjust their relative contributions [122]. This is what we choose to model the line shape of the Raman peaks in our study. With l_j as the centroid location, w_j as the full width at half maximum (FWHM), and ρ_j as the weight factor for the j-th peak, we denote the peak variables for the j-th peak as $\theta_{P,j} = (l_j, w_j, \rho_j)$. This leads to

$$g(\boldsymbol{\nu}; \theta_{P,j}) = \rho_j \exp\left\{-\frac{4\ln 2\,(\boldsymbol{\nu} - l_j)^2}{w_j^2}\right\} + (1-\rho_j)\frac{1}{1+\left[\frac{2(\boldsymbol{\nu}-l_j)}{w_j}\right]^2}. \tag{3.2}$$

Meanwhile, the baseline signal $f_B(\boldsymbol{\nu})$ is modeled with a B-spline function, which can be represented as

$$f_B(\boldsymbol{\nu}) = \sum_{j=1}^{k_B} \beta_{B,j} B_{d,j;t}(\boldsymbol{\nu}).$$

Here, $B_{d,j;t}(\boldsymbol{\nu})$ is the j-th basis function with degree d and knots $\mathbf{t}$, and can be derived from the Cox-de Boor recursive formula [123]. k_B is the number of spline basis functions and $\beta_{B,j}$ is the amplitude coefficient for the j-th basis. In our modeling, the knots $\mathbf{t} \in \mathbb{R}^{k_t}$ are chosen as equally-spaced locations in the wavenumber domain and the number of knots k_t satisfies the constraint that $k_t = k_B + d + 1$. In addition, we choose to have a fixed number of spline basis with $k_B = 4$ and set the degree d of the basis function as 3. For Raman spectroscopy, the noise $\boldsymbol{\epsilon}$ may come from a variety of sources including signal shot noise, detector dark current shot noise, temperature and environment fluctuations, laser instability, and so on. With the contributions from these independent sources, we approximate the noise in the observed signal as independent and identically distributed (i.i.d.) Gaussian random noise across the spectral domain.

With the above formulation, Equation 3.1 can be expressed in a typical Bayesian linear regression form as

$$\mathbf{y} = \mathbf{X}_k(\boldsymbol{\theta}_P)\boldsymbol{\beta}_k + \boldsymbol{\epsilon}, \tag{3.3}$$

with $\mathbf{y} \in \mathbb{R}^N$, $\boldsymbol{\epsilon} \in \mathbb{R}^N$, $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_P, \boldsymbol{\beta}_B) \in \mathbb{R}^k$, $k = k_P + k_B$ as the overall model dimension, and $\mathbf{X}_k(\boldsymbol{\theta}_P) \in \mathbb{R}^{N\times k}$ as

$$\mathbf{X}_k(\boldsymbol{\theta}_P) = \begin{bmatrix} g(\nu_1;\theta_{P,1}) & \dots & g(\nu_1;\theta_{P,k_P}) & B_{d,1;t}(\nu_1) & \dots & B_{d,k_B;t}(\nu_1) \\ g(\nu_2;\theta_{P,1}) & \dots & g(\nu_2;\theta_{P,k_P}) & B_{d,1;t}(\nu_2) & \dots & B_{d,k_B;t}(\nu_2) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ g(\nu_N;\theta_{P,1}) & \dots & g(\nu_N;\theta_{P,k_P}) & B_{d,1;t}(\nu_N) & \dots & B_{d,k_B;t}(\nu_N) \end{bmatrix}.$$

With the Gaussian random noise assumption mentioned above, we have

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_N),$$

where σ^2 is the noise variance and $\mathbf{I}_N$ is the identity matrix with dimension N.

Given an observed Raman spectrum $\mathbf{y}$, we can jointly estimate the signal decomposition matrix $\mathbf{X}_k(\boldsymbol{\theta}_P)$ as well as the corresponding regression coefficients $\boldsymbol{\beta}_k$ in Equation 3.3. As the number of Raman peaks k_P is in general not known ahead of the time, model selection is required. We solve this estimation problem by incorporating a hierarchical Bayesian model and using trans-dimensional MCMC computation for model selection and variable estimation.

3.1.2 Prior Selection

We start solving our model by incorporating Zellner's g-prior [124], which is a popular choice in Bayesian linear regression and variable selection due to its computational efficiency and the convenience of forming the prior covariance structure from the design matrix itself, into our formulation. The prior for $\boldsymbol{\beta}_k$ is

$$\boldsymbol{\beta}_k|\mathbf{X}_k(\boldsymbol{\theta}_P), g, \sigma^2 \sim \mathcal{N}\left(\boldsymbol{\beta}_{k,0}, g\sigma^2\left(\mathbf{X}_k^\intercal(\boldsymbol{\theta}_P)\mathbf{X}_k(\boldsymbol{\theta}_P)\right)^{-1}\right), \tag{3.4}$$

with prior mean $\boldsymbol{\beta}_{k,0}$ and a positive scale variable g. Meanwhile, we impose an improper Jeffery's prior on σ^2 as

$$p(\sigma^2) \propto \sigma^{-2}.$$

Various strategies exist for the modeling of g such as empirical Bayes and fully Bayesian [125], here we put an uninformative diffuse inverse-gamma(ϵ, ϵ) prior to g as

$$g \sim \mathcal{IG}(a_g, b_g),$$

with $a_g, b_g \to 0$ similar to Razul et al. [56]. This allows a convenient Gibbs update for g due to its conditional conjugacy property.

The number of Raman peaks k_P present in the spectrum is modeled with a Poisson distribution with rate or mean variable Λ as [1]

$$k_P|\Lambda \sim \text{Poisson}(\Lambda),$$

and we further model Λ with a weak and uninformative conjugate gamma(ϵ, ϵ) prior as

$$\Lambda \sim \mathcal{G}a(a_\Lambda, b_\Lambda),$$

with $a_\Lambda, b_\Lambda \to 0$.

Given k_P Raman peaks, we assume conditional independence for the prior distributions for the peak variables in $\boldsymbol{\theta}_P$. With the wavenumber region spanning across $[l_{\min}, l_{\max}]$ and $\Delta l = l_{\max} - l_{\min}$, we assign a uniform flat prior to the locations $\mathbf{l} \in [l_{\min}, l_{\max}]^{k_P}$ of the peaks, which leads to

$$\mathbf{l}|k_P \sim \prod_{i=1}^{k_P} \mathcal{U}(l_i; l_{\min}, l_{\max}) = \left(\frac{1}{\Delta l}\right)^{k_P} \prod_{i=1}^{k_P} \mathbb{1}_{[l_{\min}, l_{\max}]}(l_i).$$

For the widths of the peaks $\mathbf{w} \in \mathbb{R}^{k_P}$, it is desirable to obtain prior information in order to design a suitable prior distribution. As will be described in more details in Section 3.2.1, we surveyed around 100 Raman peak widths found in common materials and fitted these samples with an inverse-gamma distribution. To account for the limited sample space that we have surveyed and to adopt a conservative

[1] Although it is more precise to model it as a truncated Poisson distribution due to the finite number of components allowed in our computation, the choice of an untruncated Poisson distribution results in a cleaner conditional posterior distribution without losing much accuracy.

approach [126], we intentionally weaken this prior knowledge by scaling the variance of the inverse-gamma distribution by a factor of 4 while keeping the mode of the distribution fixed. With a_w and b_w denoting the parameters corresponding to the scaled inverse-gamma distribution, we have the prior distribution for $\mathbf{w}$ as

$$\mathbf{w}|k_P \sim \prod_{i=1}^{k_P} \mathcal{IG}(w_i; a_w, b_w) = \left(\frac{b_w^{a_w}}{\Gamma(a_w)}\right)^{k_P} \left(\prod_{i=1}^{k_P} w_i^{-a_w-1}\right) \exp\left\{-b_w \sum_{i=1}^{k_P} \frac{1}{w_i}\right\}.$$

As noted in Bradley [48], the line shape of the Raman peak can depend on the state of matter of the material due to the impact of the environment on the effective lifetime of the excited energy states for the molecules. For example, solids tend to have Gaussian profiles, gases tend to have Lorentzian profiles, and liquids tend to have features of both. It is therefore possible to assign specific priors for the relative weights $\boldsymbol{\rho} \in [0,1]^{k_P}$ between the Gaussian and Lorentzian profile for the Raman peaks based on knowledge of the material. Here for general purpose, we assign another uninformative flat prior in the range of $[\rho_{\min}, \rho_{\max}]$ with $\rho_{\min} = 0$, $\rho_{\max} = 1$, and $\Delta\rho = \rho_{\max} - \rho_{\min}$ for $\boldsymbol{\rho}$, which leads to

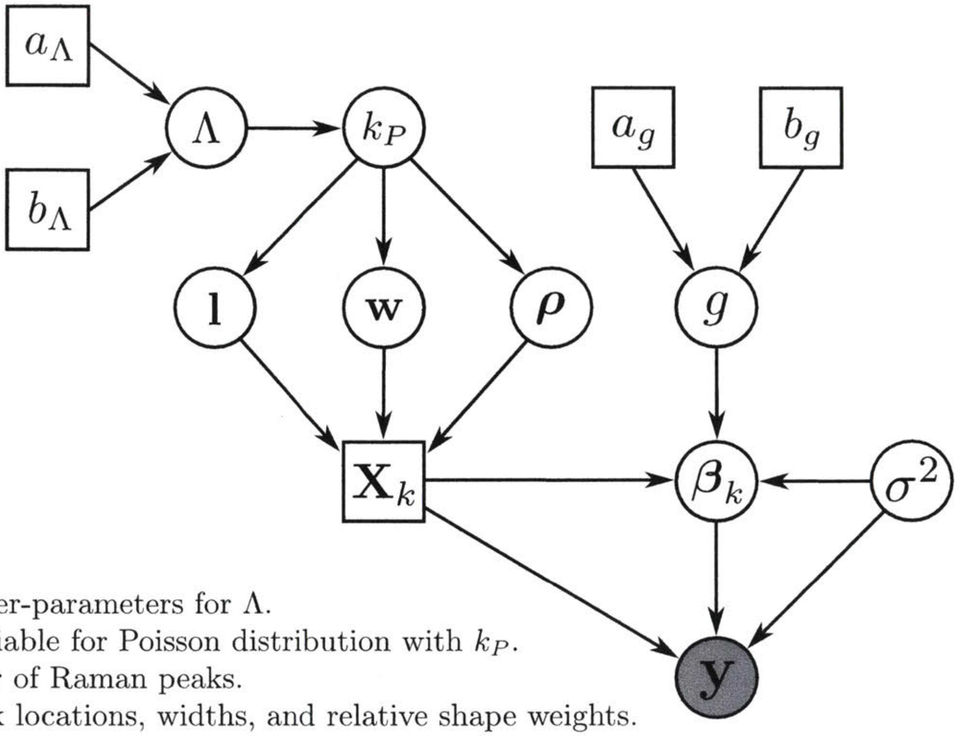
$$\boldsymbol{\rho}|k_P \sim \prod_{i=1}^{k_P} \mathcal{U}(\rho_i; \rho_{\min}, \rho_{\max}) = \left(\frac{1}{\Delta\rho}\right)^{k_P} \prod_{i=1}^{k_P} \mathbb{1}_{[\rho_{\min},\rho_{\max}]}(\rho_i).$$

The graphical model representing the hierarchical Bayesian structure of the spectral signal is shown in Figure 3-1. With the likelihood function of our model as

$$p\left(\mathbf{y}|\mathbf{X}_k(\boldsymbol{\theta}_P), \boldsymbol{\beta}_k, \sigma^2\right) = \frac{1}{\sqrt{|2\pi\sigma^2 \mathbf{I}_N|}} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}_k(\boldsymbol{\theta}_P)\boldsymbol{\beta}_k\right)^\intercal \left(\mathbf{y} - \mathbf{X}_k(\boldsymbol{\theta}_P)\boldsymbol{\beta}_k\right)\right\},$$

we can express the joint posterior distribution for all the variables as

$$p(g, \sigma^2, \Lambda, k_P, \boldsymbol{\theta}_P, \boldsymbol{\beta}_k|\mathbf{y}) \propto$$
$$p(g)p(\sigma^2)p(\Lambda)p(k_P|\Lambda)p(\mathbf{l}|k_P)p(\mathbf{w}|k_P)p(\boldsymbol{\rho}|k_P)p\left(\boldsymbol{\beta}_k|\mathbf{X}_k(\boldsymbol{\theta}_P), g, \sigma^2\right) p\left(\mathbf{y}|\mathbf{X}_k(\boldsymbol{\theta}_P), \boldsymbol{\beta}_k, \sigma^2\right). \tag{3.5}$$



a_Λ, b_Λ: Hyper-parameters for Λ.
Λ: Rate variable for Poisson distribution with k_P.
k_P: Number of Raman peaks.
$\mathbf{l}, \mathbf{w}, \boldsymbol{\rho}$: Peak locations, widths, and relative shape weights.
$\mathbf{X}_k$: Signal decomposition matrix.
a_g, b_g: Hyper-parameters for g.
g: Scale variable for the g-prior.
$\boldsymbol{\beta}_k$: Amplitude variables for the signal decomposition matrix $\mathbf{X}_k$.
σ^2: Noise variance.
$\mathbf{y}$: Observed spectral array.

Figure 3-1: Graphical model for the hierarchical Bayesian structure of the spectral signal.

3.1.3 Bayesian Computation

No closed-form solution exists for evaluating the joint posterior distribution from Equation 3.5, we resort to numerical approximation with statistical sampling. In addition, the number of Raman peaks k_P is generally not known beforehand and affects the dimensionality of the variable space $\mathcal{X}_{k_P}$ for the model. Therefore, model selection across the model space with different k_P is required. A diversity of criteria and methodologies exists for Bayesian model selection [127]. Here, we adopt a unified approach for joint variable estimation and model selection with the RJMCMC technique introduced by Green [128], Richardson and Green [129]. The RJMCMC method samples from the union space $\mathcal{X} = \cup_{k_P \in \mathcal{K}}(\{k_P\} \times \mathcal{X}_{k_P})$ for the potential models, where $\mathcal{K}$ in our case is a countable set containing all the possible Raman peak number in the spectrum, by constructing a reversible Markov chain in the general state space. The trans-dimensional moves across the models in RJMCMC can be incorporated inside the general Metropolis-Hastings paradigm in a straightforward manner. With marginalization, the posterior probability of being in any variable space $\mathcal{X}_{k_P}$ can be obtained, and model selection can be performed accordingly. For more in-depth discussions on the model determination aspects with RJMCMC, Hastie and Green [130] serves as an excellent reference.

With the hierarchical Bayesian structure imposed by our model, several variables can be conveniently updated with Gibbs sampling. In the following text, we use $|\dots$ to denote conditioning on all other random variables. With Gibbs sampling, g can be updated with an inverse-gamma distribution as

$$g|\cdots \sim \mathcal{IG}\left(a_g + \frac{k}{2}, b_g + \frac{1}{2\sigma^2}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k,0})^\intercal \mathbf{X}_k^\intercal(\boldsymbol{\theta}_P)\mathbf{X}_k(\boldsymbol{\theta}_P)(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k,0})\right).$$

Λ can be updated as

$$\Lambda|\cdots \sim \mathcal{G}a(a_\Lambda + k_P, b_\Lambda + 1).$$

Denoting $\hat{\boldsymbol{\beta}}_k = (\mathbf{X}_k^\intercal(\boldsymbol{\theta}_P)\mathbf{X}_k(\boldsymbol{\theta}_P))^{-1}\mathbf{X}_k^\intercal(\boldsymbol{\theta}_P)\mathbf{y}$ as the maximum likelihood (ML) estimation of $\boldsymbol{\beta}_k$ and $s^2 = (\mathbf{y} - \mathbf{X}_k(\boldsymbol{\theta}_P)\hat{\boldsymbol{\beta}}_k)^\intercal(\mathbf{y} - \mathbf{X}_k(\boldsymbol{\theta}_P)\hat{\boldsymbol{\beta}}_k)$ as the squared residue of the ML estimation, we define

$$\tilde{b}_{\sigma^2} = \frac{s^2}{2} + \frac{1}{2(g+1)}(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{k,0})^\intercal \mathbf{X}_k^\intercal(\boldsymbol{\theta}_P)\mathbf{X}_k(\boldsymbol{\theta}_P)(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{k,0}).$$

With this definition, our posterior for σ^2 and $\boldsymbol{\beta}_k$ can be updated as

$$\sigma^2|\cdots \sim \mathcal{IG}\left(\frac{N}{2}, \tilde{b}_{\sigma^2}\right),$$

and

$$\boldsymbol{\beta}_k|\sigma^2, \cdots \sim \mathcal{N}\left(\frac{1}{g+1}\left(\boldsymbol{\beta}_{k,0} + g\hat{\boldsymbol{\beta}}_k\right), \frac{g}{g+1}\sigma^2\left(\mathbf{X}_k^\intercal(\boldsymbol{\theta}_P)\mathbf{X}_k(\boldsymbol{\theta}_P)\right)^{-1}\right). \tag{3.6}$$

The conditional posterior distribution for the rest of the variables $\boldsymbol{\theta}_P$ and k_P does not admit a tractable form. To sample $\boldsymbol{\theta}_P$ and k_P, we first integrate out σ^2 and $\boldsymbol{\beta}_k$ from the full joint posterior distribution for simplification. This leaves the conditional posterior probability for $\boldsymbol{\theta}_P$ and k_P known to a proportionality as

$$\begin{aligned} p(\boldsymbol{\theta}_P, k_P|\dots) \propto \frac{\Lambda^{k_P}}{k_P!}\tilde{b}_{\sigma^2}^{-\frac{N}{2}} \frac{1}{\sqrt{|(g+1)\mathbf{I}_k|}} \left(\frac{1}{\Delta l}\right)^{k_P} \prod_{i=1}^{k_P} \mathbb{1}_{[l_{\min},l_{\max}]}(l_i) \\ \left(\frac{b_w^{a_w}}{\Gamma(a_w)}\right)^{k_P} \left(\prod_{i=1}^{k_P} w_i^{-a_w-1}\right) \exp\left\{-b_w \sum_{i=1}^{k_P} \frac{1}{w_i}\right\} \left(\frac{1}{\Delta\rho}\right)^{k_P} \prod_{i=1}^{k_P} \mathbb{1}_{[\rho_{\min},\rho_{\max}]}(\rho_i). \end{aligned} \tag{3.7}$$

Samples of $\boldsymbol{\theta}_P$ and k_P can be obtained with the RJMCMC method, which can be viewed as a generalization of the Metropolis-Hastings algorithm with additional trans-dimensional moves. Denoting the current variable state as $\mathbf{x} = (\boldsymbol{\theta}_P, k_P)$, for any proposed variable state $\mathbf{x}' = (\boldsymbol{\theta}'_P, k'_P)$, we can calculate the corresponding Metropolis-Hastings acceptance probability $a(\mathbf{x}, \mathbf{x}') = \min\{1, A(\mathbf{x}, \mathbf{x}')\}$, where $A(\mathbf{x}, \mathbf{x}')$ can be calculated for each move type respectively.

For within-dimensional moves where the dimensionality of the variable space stays the same and $k'_P = k_P$, we can use symmetric random walks in $\boldsymbol{\theta}_P$ to generate $\boldsymbol{\theta}'_P$. This is essentially the Metropolis algorithm and $A(\mathbf{x}, \mathbf{x}')$ is the ratio between the posterior density function for the proposed state $\mathbf{x}'$ and the current state $\mathbf{x}$ as

$$\begin{aligned} A_{\text{within}}(\mathbf{x}, \mathbf{x}') = \left(\frac{\tilde{b}'_{\sigma^2}}{\tilde{b}_{\sigma^2}}\right)^{-\frac{N}{2}} \prod_{i=1}^{k_P} \mathbb{1}_{[l_{\min},l_{\max}]}(l'_i) \\ \left[\prod_{i=1}^{k_P}\left(\frac{w'_i}{w_i}\right)^{-a_w-1}\right] \exp\left\{-b_w \sum_{i=1}^{k_P}\left(\frac{1}{w'_i} - \frac{1}{w_i}\right)\right\} \prod_{i=1}^{k_P} \mathbb{1}_{[\rho_{\min},\rho_{\max}]}(\rho'_i). \end{aligned}$$

For trans-dimensional moves, pairs of reversible moves need to be devised. With proper engineering, by generating assistive random variable $\mathbf{u}$ from proposal density $g(\mathbf{u})$, the proposed state $\mathbf{x}'$ can be constructed by a deterministic function $h(\cdot)$ as $(\mathbf{x}', \mathbf{u}') = h(\mathbf{x}, \mathbf{u})$, where $\mathbf{u}'$ is a random variable that can be generated from proposal density $g'(\mathbf{u}')$ so that one can reversely jump from $\mathbf{x}'$ back to $\mathbf{x}$ based on the inverse of $h(\cdot)$. The transformation $h(\cdot)$ needs to be a diffeomorphism with matching dimensions for $(\mathbf{x}, \mathbf{u})$ and $(\mathbf{x}', \mathbf{u}')$, which means $n_{\mathbf{x}} + n_{\mathbf{u}} = n_{\mathbf{x}'} + n_{\mathbf{u}'}$ with n being the variable dimension. Let m and m' be the indices for reversible move pairs across the dimensions of $\mathbf{x}$ and $\mathbf{x}'$ in set $\mathcal{M}$ containing all the possible moves and $q(m|\mathbf{x})$ be the probability of taking move m at state $\mathbf{x}$, we can calculate $A(\mathbf{x}, \mathbf{x}')$ as

$$A(\mathbf{x}, \mathbf{x}') = \frac{p(\mathbf{x}')q(m'|\mathbf{x}')g'(\mathbf{u}')}{p(\mathbf{x})q(m|\mathbf{x})g(\mathbf{u})} \left| \frac{\partial(\mathbf{x}', \mathbf{u}')}{\partial(\mathbf{x}, \mathbf{u})} \right|,$$

where $|\cdot|$ denotes the determinant of the transformation Jacobian.

Here, we design four trans-dimensional moves to facilitate cross model mixing, where similar strategies can be found in applications such as Bayesian mixture estimation [129]. These moves are:

1. Birth of a new peak.

2. Death of an existing peak.

3. Split of an existing peak.

4. Merge of two adjacent peaks.

For the birth move with $k'_P = k_P + 1$, a peak is generated with $\theta_b = (l_b, w_b, \rho_b)$, where l_b is randomly drawn from $[l_{\min}, l_{\max}]$, w_b is randomly drawn from density function $p_{w_b}(w_b)$, and ρ_b is randomly drawn from $[\rho_{\min}, \rho_{\max}]$. For $p_{w_b}(w_b)$, we choose to use the empirically fitted inverse-gamma distribution that is described in Section 3.2.1. With this, $A_{\text{birth}}(\mathbf{x}, \mathbf{x}')$ can be shown as

$$A_{\text{birth}}(\mathbf{x}, \mathbf{x}') = \left(\frac{\tilde{b}'_{\sigma^2}}{\tilde{b}_{\sigma^2}} \right)^{-\frac{N}{2}} \frac{\Lambda}{k'_P} (g+1)^{-\frac{1}{2}} \frac{b_w^{a_w}}{\Gamma(a_w)} w_b^{-a_w-1} e^{-\frac{b_w}{w_b}} \frac{1}{k'_P} \frac{1}{p_{w_b}(w_b)}.$$

Meanwhile, for the reversed process of randomly killing an existing peak with peak variables $\theta_d = (l_d, w_d, \rho_d)$, we have $k'_P = k_P - 1$ and

$$A_{\text{death}}(\mathbf{x}, \mathbf{x}') = \left(\frac{\tilde{b}'_{\sigma^2}}{\tilde{b}_{\sigma^2}} \right)^{-\frac{N}{2}} \frac{k_P}{\Lambda} (g+1)^{\frac{1}{2}} \frac{\Gamma(a_w)}{b_w^{a_w}} w_d^{a_w+1} e^{\frac{b_w}{w_d}} k_P \, p_{w_b}(w_d).$$

For the split move, a random peak is selected and split into two adjacent peaks. Assume that the selected peak has the peak variables as $\theta_s = (l_s, w_s, \rho_s)$, we split the peak into two peaks with $\theta_s^+ = (l_s^+, w_s^+, \rho_s^+)$ and $\theta_s^- = (l_s^-, w_s^-, \rho_s^-)$ as

$$l_s^+ = l_s + \delta_l u_l, \quad l_s^- = l_s - \delta_l u_l,$$

$$w_s^+ = w_s + \delta_w u_w, \quad w_s^- = w_s - \delta_w u_w,$$

$$\rho_s^+, \rho_s^- \sim \mathcal{U}(0,1),$$

where δ_l and δ_w are the hyper-parameters to specify the variable split ranges, $u_l \sim \mathcal{U}(0,1)$, and $u_w \sim \mathcal{U}(-1,1)$. The corresponding $A_{\text{split}}(\mathbf{x}, \mathbf{x}')$ with $k_P' = k_P + 1$ is

$$A_{\text{split}}(\mathbf{x}, \mathbf{x}') = \left(\frac{\tilde{b}'_{\sigma^2}}{\tilde{b}_{\sigma^2}}\right)^{-\frac{N}{2}} \frac{\Lambda}{k_P'}(g+1)^{-\frac{1}{2}} \frac{1}{\Delta_l} \frac{b_w^{a_w}}{\Gamma(a_w)} \left(\frac{w_s^+ w_s^-}{w_s}\right)^{-a_w - 1} e^{-b_w\left(\frac{1}{w_s^+} + \frac{1}{w_s^-} - \frac{1}{w_s}\right)} 8\delta_l\delta_w.$$

For the reversed merge move, the peak variables $\theta_m^+ = (l_m^+, w_m^+, \rho_m^+)$ and $\theta_m^- = (l_m^-, w_m^-, \rho_m^-)$ from the two selected adjacent peaks are merged into $\theta_m = (l_m, w_m, \rho_m)$ as

$$l_m = \frac{1}{2}\left(l_m^+ + l_m^-\right), \quad w_m = \frac{1}{2}\left(w_m^+ + w_m^-\right), \quad \rho_m \sim \mathcal{U}(0,1).$$

With $k_P' = k_P - 1$, $A_{\text{merge}}(\mathbf{x}, \mathbf{x}')$ can be calculated as

$$A_{\text{merge}}(\mathbf{x}, \mathbf{x}') = \left(\frac{\tilde{b}'_{\sigma^2}}{\tilde{b}_{\sigma^2}}\right)^{-\frac{N}{2}} \frac{k_P}{\Lambda}(g+1)^{\frac{1}{2}} \Delta_l \frac{\Gamma(a_w)}{b_w^{a_w}} \left(\frac{w_m^+ w_m^-}{w_m}\right)^{a_w + 1} e^{b_w\left(\frac{1}{w_m^+} + \frac{1}{w_m^-} - \frac{1}{w_m}\right)} \frac{1}{8\delta_l\delta_w}.$$

For the trans-dimensional move pairs, we make sure that the moves within each pair are reversible. For the split and merge move pair, this means that if a split move creates two peaks that are not adjacent in the current spectrum, or if the selected adjacent peaks in the merge move have larger variable differences in l and w than those that are allowed in the split move, we would discard the proposal to ensure reversibility.

With the hybrid Gibbs and RJMCMC sampling schedules described above, we can describe an algorithm for joint peak variable and baseline estimation with a Raman spectrum. At each sampling iteration, the RJMCMC move for this iteration is first determined with a categorical random variable m with the support corresponding to the indices of the available moves in $\mathcal{M}$. $\boldsymbol{\theta}_P$ and k_P are updated subsequently based on the move type m. This essentially creates a combined mixture MCMC transition kernel for the update of $\boldsymbol{\theta}_P$ and k_P. Afterwards, $(g, \Lambda, \sigma^2, \boldsymbol{\beta}_k)$ are updated with Gibbs sampling. Once the Markov chain is fully mixed, model selection based on k_P can be performed. For example, the *maximum a posteriori* (MAP) estimation for k_P can be obtained as

$$\hat{k}_P = \arg\max\nolimits_{k_P} p(k_P|\mathbf{y}).$$

For spectra having visually distinct and well-spaced peaks, the above Bayesian sampling schedule works well with a fixed and equally-likely move proposal distribution for m. However, for more complex spectra having a large number of peaks that may have tightly spaced or partially overlapping peaks, we notice that frequent label switching caused by trans-dimensional moves in steady state can become problematic for variable estimation [131]. In addition, in these situations, during the early inter-state mixing iterations, negative amplitudes can be assigned to some peaks (while still maintaining an overall spectral signal match with the observed spectrum). These peaks with negative amplitudes may stay throughout the iterations, which create unphysical decomposition results. We address these two problems with the following approaches.

As a solution to the first problem, we employ a heuristic approach by gradually decreasing the probability of taking trans-dimensional moves throughout the iterations. At iteration i, the move type is determined from $m^{(i)}$ sampling from the categorical proposal distribution $p_m^{(i)}(m)$ with probability mass as $(p_w^{(i)}, p_b^{(i)}, p_d^{(i)}, p_s^{(i)}, p_m^{(i)})$ for each move – $p_w^{(i)}$ corresponds to the within move, $p_b^{(i)}$ and $p_d^{(i)}$ correspond to the birth and death moves, and $p_s^{(i)}$ and $p_m^{(i)}$ correspond to the split and merge moves. Using $p_b^{(i)}$ as an example, we adjust its value in each iteration as

$$p_b^{(i)} = \left(p_b^{(0)}\right)^{1/T^{(i)}},$$

with $p_w^{(0)} = p_b^{(0)} = p_d^{(0)} = p_s^{(0)} = p_m^{(0)}$, $T^{(0)} = 1$, $\lim_{i \to \infty} T^{(i)} = 0$, and a linearly decreasing cooling schedule for $T^{(i)}$. We perform this adjustment for all the trans-dimensional moves. Meanwhile, we increase the within-dimensional move probability accordingly with the constraint that $p_w^{(i)} + p_b^{(i)} + p_d^{(i)} + p_s^{(i)} + p_m^{(i)} = 1$. This treatment is similar to simulated annealing, and effectively creates a non-homogeneous Markov chain in the general state space [132]. Convergence results can be obtained with simulated annealing-like algorithms [133], which we do not pursue in this work. Once the steady state is reached, all samples are effectively drawn from the same model with $\hat{k}_P$. Therefore, variable values can be estimated without explicitly performing model selection.

For the second problem, we enforce a non-negativity constraint on the amplitude coefficients $\boldsymbol{\beta}_P$ for the peaks. During the sampling iterations, if any peak with a negative amplitude is generated, we discard the sample and restart the current iteration step until all peaks are of non-negative values. This emulates rejection sampling and effectively adds a non-negative support constraint on $\boldsymbol{\beta}_P$ for the prior and posterior probability distributions in Equation 3.4 and Equation 3.6. We note here that even without the annealing schedule described above, this re-sampling operation is only mostly required during early iterations where the computation is rapidly converging in the model domain. Once the model dimension is reasonably converged, negative peak amplitude generation seldomly occurs.

Denoting $\boldsymbol{\phi}_{\hat{k}_P} = (g, \Lambda, \sigma^2, \boldsymbol{\beta}_k, \boldsymbol{\theta}_P)$ for the variables associated with $\hat{k}_P$, we can estimate the conditional posterior expected values for $\boldsymbol{\phi}_{\hat{k}_P}$ as

$$\mathbb{E}_{p(\boldsymbol{\phi}_{\hat{k}_P} | \mathbf{y}, \hat{k}_P)}[\boldsymbol{\phi}_{\hat{k}_P}] \approx \frac{1}{M} \sum_{i=1}^{M \to \infty} \boldsymbol{\phi}_{\hat{k}_P}^{(i)} \tag{3.8}$$

with $\boldsymbol{\phi}_{\hat{k}_P}^{(i)}$ being the i-th sample drawn from the conditional posterior distribution $p(\boldsymbol{\phi}_{\hat{k}_P} | \mathbf{y}, \hat{k}_P)$ and M being the total number of samples. Alternative estimation criterion such as the MAP estimator can also be used here.

3.1.4 Two-Stage Algorithm for Analyte Quantification

The above Bayesian formulation and computation provide a framework to simultaneously estimate the peak and baseline signal in a Raman spectrum. In order to further it into a quantification algorithm applicable to practical scenarios, we propose a two-stage algorithm built upon this framework.

In many analyte quantification tasks involving Raman spectroscopy, the goal is to quantify one or more target analyte in mixture spectra. For simplicity of presentation we restrict our attentions to one target analyte but note that the extension to multiple target analytes is straightforward. We also assume an aqueous mixture environment in our analysis. While not required by many multivariate regression techniques such as PLSR, the actual spectrum of the target analyte is often easy to acquire through a separate reference measurement. Our algorithm takes advantage of this aspect by first learning the peak representation for the target analyte at a known concentration. This can be achieved through the Bayesian computation process described above working on the reference target analyte spectrum. Once the peak variables for the target analyte are learned in this first stage, we move on to the second stage with the mixture spectrum where the concentration for the target analyte in the mixture needs to be determined. In this stage, the processing is slightly modified relative to the first stage to take into account of the learned representation of the target analyte. The modifications are described as follows.

With the peak and baseline decomposition for the reference target analyte spectrum in the first stage as in Equation 3.1, and $\hat{\boldsymbol{\theta}}_P$ and $\hat{\boldsymbol{\beta}}_P$ corresponding to the estimated peak variables for the target analyte according to Equation 3.8, we define the Raman peak signal at unit concentration for the target analyte as

$$\tilde{f}_P(\boldsymbol{\nu}) = \frac{f_P(\boldsymbol{\nu})}{c_{\text{pure}}} = \sum_{j=1}^{k_P} \frac{\hat{\beta}_{P,j}}{c_{\text{pure}}} g(\boldsymbol{\nu}; \hat{\theta}_{P,j}), \tag{3.9}$$

where c_{pure} is the target analyte concentration in the reference measurement.

In the second stage, the observed signal in the mixture spectrum now can be

```
First stage
Input   : Reference spectrum of the target analyte
Initialize $T^{(0)}$, $(p_w^{(0)}, p_b^{(0)}, p_d^{(0)}, p_s^{(0)}, p_m^{(0)})$, and the spectrum variables
for $i \leftarrow 1$ to $I$ do
    Determine move type $m^{(i)}$ with $(p_w^{(i)}, p_b^{(i)}, p_d^{(i)}, p_s^{(i)}, p_m^{(i)})$
    Based on move type $m^{(i)}$, sample $\boldsymbol{\theta}_P$ with RJMCMC and update $k_P$
    Sample $g$, $\Lambda$, $\sigma^2$, $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_P, \boldsymbol{\beta}_B)$ with Gibbs sampling
    if $\exists \beta_{P,j} \in \boldsymbol{\beta}_P$ s.t. $\beta_{P,j} < 0$ then
        Discard current samples and restart current iteration
    end
    Update $i, T^{(i)}$, $(p_w^{(i)}, p_b^{(i)}, p_d^{(i)}, p_s^{(i)}, p_m^{(i)})$
end
Estimate $\hat{\boldsymbol{\theta}}_P$, $\hat{\boldsymbol{\beta}}_P$, and calculate $\tilde{f}_P(\boldsymbol{\nu})$
Second stage
Input   : Mixture spectrum
Initialize $T^{(0)}$, $(p_w^{(0)}, p_b^{(0)}, p_d^{(0)}, p_s^{(0)}, p_m^{(0)})$, and the spectrum variables
for $i \leftarrow 1$ to $I$ do
    Determine move type $m^{(i)}$ with $(p_w^{(i)}, p_b^{(i)}, p_d^{(i)}, p_s^{(i)}, p_m^{(i)})$
    Based on move type $m^{(i)}$, sample $\boldsymbol{\theta}_I$ with RJMCMC and update $k_I$
    Sample $g$, $\Lambda$, $\sigma^2$, $\boldsymbol{\beta}_k = (c_{\text{mix}}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_B)$ with Gibbs sampling
    if $\exists \beta_{I,j} \in \boldsymbol{\beta}_I$ s.t. $\beta_{I,j} < 0$ then
        Discard current samples and restart current iteration
    end
    Update $i, T^{(i)}$, $(p_w^{(i)}, p_b^{(i)}, p_d^{(i)}, p_s^{(i)}, p_m^{(i)})$
end
Estimate $\hat{c}_{\text{mix}}$
```

Algorithm 1: The two-stage algorithm for analyte quantification in mixture spectrum with Bayesian modeling and computation.

modeled as

$$\mathbf{y} = f_T(\boldsymbol{\nu}) + f_I(\boldsymbol{\nu}) + f_B(\boldsymbol{\nu}) + \boldsymbol{\epsilon},$$

where here $f_T(\boldsymbol{\nu})$ represents peaks originating from the target analyte, $f_I(\boldsymbol{\nu})$ represents peaks from the other analytes in the mixture, which we call the interfering analytes, and the rest follows as previous. The target analyte signal $f_T(\boldsymbol{\nu})$ is related to its concentration in the mixture c_{mix} as

$$f_T(\boldsymbol{\nu}) = c_{\text{mix}} \tilde{f}_P(\boldsymbol{\nu}).$$

In order to solve for c_{mix}, a similar Bayesian computation process relative to the first stage can be carried out except for that $\tilde{f}_P(\boldsymbol{\nu})$ is kept as a fixed basis with estimated $\hat{\boldsymbol{\theta}}_P$ and $\hat{\boldsymbol{\beta}}_P$ as in Equation 3.9. With $\boldsymbol{\theta}_I$ as the peak variables for the interfering analytes, this means that $\mathbf{X}_k(\boldsymbol{\theta}_I) \in \mathbb{R}^{N \times k}$, now depends on $\boldsymbol{\theta}_I$, is

$$\mathbf{X}_k(\boldsymbol{\theta}_I) = \begin{bmatrix} \tilde{f}_P(\nu_1) & g(\nu_1;\theta_{I,1}) & \dots & g(\nu_1;\theta_{I,k_I}) & B_{d,1;t}(\nu_1) & \dots & B_{d,k_B;t}(\nu_1) \\ \tilde{f}_P(\nu_2) & g(\nu_2;\theta_{I,1}) & \dots & g(\nu_2;\theta_{I,k_I}) & B_{d,1;t}(\nu_2) & \dots & B_{d,k_B;t}(\nu_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \tilde{f}_P(\nu_N) & g(\nu_N;\theta_{I,1}) & \dots & g(\nu_N;\theta_{I,k_I}) & B_{d,1;t}(\nu_N) & \dots & B_{d,k_B;t}(\nu_N) \end{bmatrix}$$

to take into consideration of the target analyte spectrum. Correspondingly, $\boldsymbol{\beta}_k = (c_{\text{mix}}, \boldsymbol{\beta}_I, \boldsymbol{\beta}_B) \in \mathbb{R}^k$, k_I represents the number of peaks coming from the interfering analytes in the mixture, and $k = k_I + k_B + 1$. Afterwards, all the variable sampling schedule and estimation procedure from the previous section can be directly applied to estimate $\hat{c}_{\text{mix}}$ and the rest of the variables. The overall algorithm is shown in Algorithm 1.

3.2 Simulation Study

3.2.1 Numerical Experiment Setup

We first set up the numerical experiment environment for exploring the performance of our algorithm under various situations. For all the simulated spectra, the Raman shift wavenumber range spanned across 400 cm^{-1} to 1600 cm^{-1} and contained 300 equally-spaced spectral points. We simulated our studies under the same use case as in actual practice where a reference measurement of the Raman spectrum for the target analyte at a known concentration was first given and our goal was to quantify its concentration in mixture measurements in the presence of other interfering analytes at unknown concentrations.

For any simulated analyte including the target analyte, we modeled its Raman spectrum at unit concentration by explicitly generating the Raman peaks. The

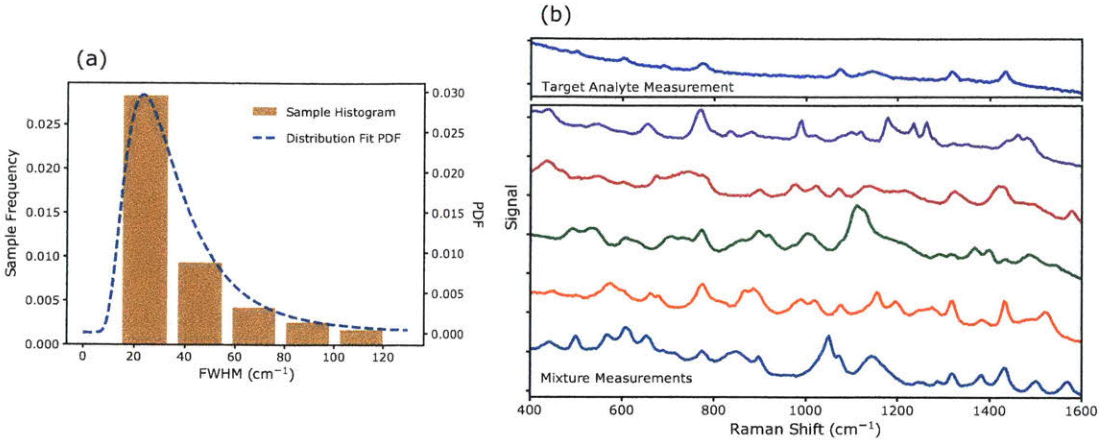


Figure 3-2: (a) Histogram for the widths of around 100 Raman peaks surveyed in common materials and the PDF plot for an inverse-gamma probability distribution fit. This PDF was used to generate the simulated Raman spectra in our studies. (b) Examples of the simulated target analyte spectrum and 5 mixture spectra each with 5 randomly generated interfering analytes. $\sigma = 1$ for all these simulated measurements.

number of Raman peaks k_P for the analyte was first determined. Afterwards, we generated the corresponding peak random variables $\boldsymbol{\theta}_P = (\mathbf{l}, \mathbf{w}, \boldsymbol{\rho}, \mathbf{a})$ from predefined probability distributions. Here, $\mathbf{l}$, $\mathbf{w}$, and $\boldsymbol{\rho}$ are defined in the previous section and $\mathbf{a} \in [0,1]^{k_P}$ are the peak amplitudes at unit concentration. For all the peaks, $\mathbf{l}$ were independently and uniformly generated across the available spectrum span, and $\boldsymbol{\rho}$ and $\mathbf{a}$ were both independently and uniformly generated from range $[0, 1]$. For $\mathbf{w}$ to be representable to actual Raman peaks encountered in practice, we surveyed around 100 Raman peaks from 11 common materials including phenylalanine, tryptophan, tyrosine, alanine, glycine, glucose, lactic acid, acetic acid, succinic acid, ethanol and water. We extracted the Raman peaks and their widths and fitted these width samples with an inverse-gamma distribution. The histogram for the peak width samples and the fitted probability density function (PDF) are shown in Figure 3-2 (a). We denote this PDF as $p_g(\mathbf{w})$ and sampled $\mathbf{w}$ independently from this probability distribution in our simulations. Once k_P and $\boldsymbol{\theta}_P$ were both determined, the Raman spectrum at unit concentration could be represented as

$$\tilde{f}_P(\boldsymbol{\nu}) = \sum_{j=1}^{k_P} a_j g(\boldsymbol{\nu}; \theta_{P,j})$$

with $g(\boldsymbol{\nu}; \theta_{P,j})$ defined in Equation 3.2. Given $\tilde{f}_P(\boldsymbol{\nu})$ for each analyte, we could generate any mixture spectrum by adding together the spectral signals from all the constituent analytes adjusted linearly by their respective concentrations in the mixture. In addition, we also added the baseline signal represented by a low-order polynomial curve, where for each baseline curve, small random perturbations were added to the fitting points used to generate the polynomial curve to ensure a varying baseline across all the spectra. At last, we added independent and additive Gaussian random noise with standard deviation σ across the array to generate the final spectra. For the reference target analyte spectrum, no spectral signal from any other analyte was added, but we still included the baseline signal and the noise to resemble an actual measurement.

In the following numerical experiments, we fixed the target analyte spectrum with $k_P = 10$ randomly generated Raman peaks across all the mixture studies. For each mixture, the number of co-existing interfering analytes is denoted as N_I. The number of Raman peaks was set as 10 for each interfering analyte similar to that of the target analyte. The concentration for each analyte in the mixture including the target analyte was uniformly and randomly generated from range $[0, 60]$. For the reference target analyte spectrum generation, we set its concentration at 30 and noise scale σ at 1. The mixture spectra were all randomly and independently generated from the process described above. Sample plots for the target analyte spectrum as well as 5 mixture spectra each with 5 random interfering analytes are shown in Figure 3-2 (b). $\sigma = 1$ for all the spectra in the figure.

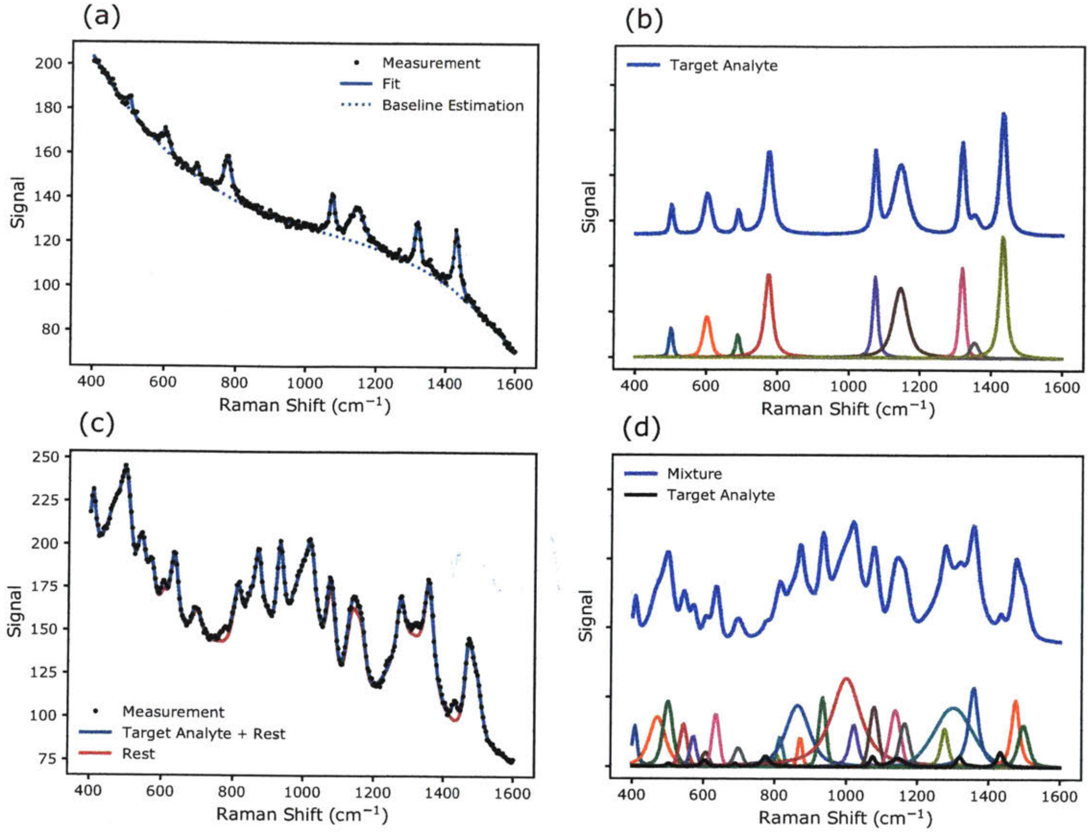


Figure 3-3: (a) and (b) Example plots for a simulated target analyte spectrum, and its baseline estimation and peak decomposition results with our algorithm. (c) and (d) Example plots for analyte assignment and peak decomposition for a mixture spectrum based on peak variables obtained from (a) and (b) for the target analyte. The resulting spectrum amplitude for the target analyte was subsequently used to estimate its concentration in the mixture. In this simulation, $\sigma = 1$ for the target analyte spectrum, $N_I = 5$ and $\sigma = 1$ for the mixture spectrum. The concentration for the target analyte in the mixture was 5 and the estimated concentration from our algorithm was 4.6

3.2.2 Mixture Environment Study

With the above settings, we were able to validate our algorithm with simulated target analyte and mixture spectra. As an illustrative example, we show an estimation result in Figure 3-3. Figure 3-3 (a) and (b) show the baseline estimation and peak decomposition results for a simulated reference target analyte spectrum. The estimated peak variables obtained in this step were further used to quantify the target analyte concentrations in mixtures, as shown in Figure 3-3 (c) and (d). With the mixture

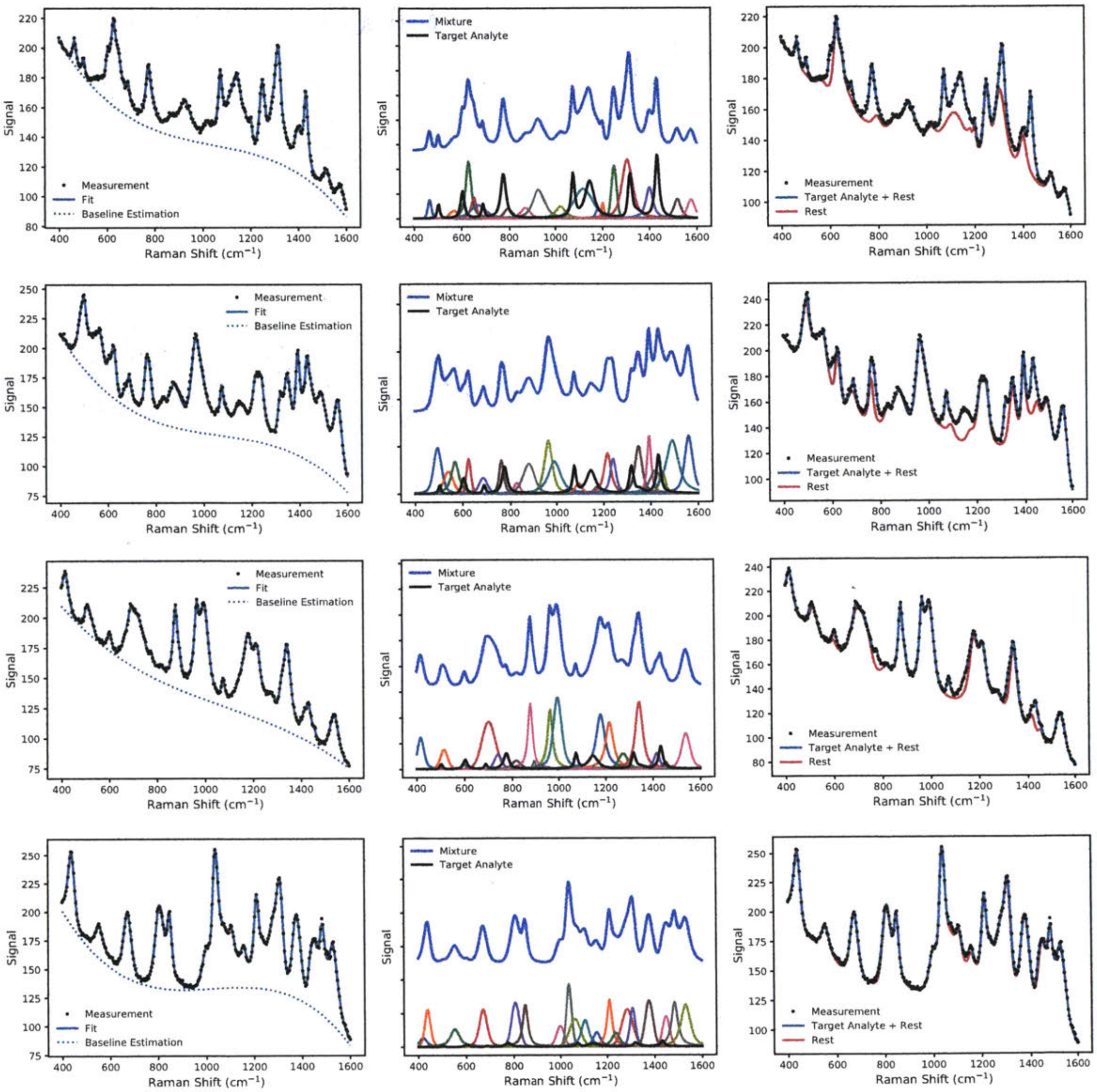


Figure 3-4: Example plots showing the baseline estimation and analyte assignment in mixture spectra for various target analyte concentrations. The concentrations for the target analyte in the mixtures decrease as we move from the top row to the bottom row. In all plots, $N_I = 5$ and $\sigma = 1$.

spectrum, other than the amplitude coefficient of the learned target analyte, a variable number of Raman peaks were also fitted with the RJMCMC computation to explain the peak signals from the rest of the interfering analytes. This ensured that all the peaks appearing in the mixture spectrum were properly assigned to either the target analyte or the interfering analytes, resulting in the most likely amplitude estimation for the target analyte peaks in the mixture spectrum. This in turn corresponded to the concentration of the target analyte in the mixture. In the simulation as shown in Figure 3-3, $\sigma = 1$ for the target analyte spectrum, $N_I = 5$ and $\sigma = 1$ for the mixture spectrum. The concentration for the target analyte in the mixture was 5 and the estimated concentration from our algorithm was 4.6.

In addition to Figure 3-3, Figure 3-4 shows more mixture spectra decomposition examples. The concentrations for the target analyte in the mixtures decrease as we move from the top row to the bottom row. In all plots, $N_I = 5$ and $\sigma = 1$. While 60 Raman peaks were being randomly generated for each plot, many of them overlapped as can be seen from the plots. Due to the fact that the spectrum for the target analyte was already learned in the first stage, the algorithm was able to properly decompose wider peaks into narrower partially overlapping peaks that were associated with the target analyte, despite of heavy peak degeneracies in the mixture spectra. A closer examination of the plots also reveals that the success of our algorithm in quantification tasks relies heavily on the number of Raman peaks in the target analyte spectrum. For a target analyte with $\gtrsim 10$ Raman peaks, even in a highly mixed environment like the ones generated in Figure 3-4, there is a high probability that at least several target analyte Raman peaks are not clouded by any interfering peaks. As a result, this would likely result in proper analyte assignment and in turn, accurate target analyte concentration estimation. In practice, many analytes of interest in analytical chemistry and biology have a respectable number of distinct Raman peaks like the ones shown in Figure 1-5. This validates the general usage of our algorithm.

Plots of the sampling process for all the modeling variables for a simulation run are shown in Figure 3-5 through Figure 3-8. In this simulation, $N_I = 5$, and $\sigma = 1$ for both the target analyte spectrum and the mixture spectrum. Figure 3-5 and Figure 3-6

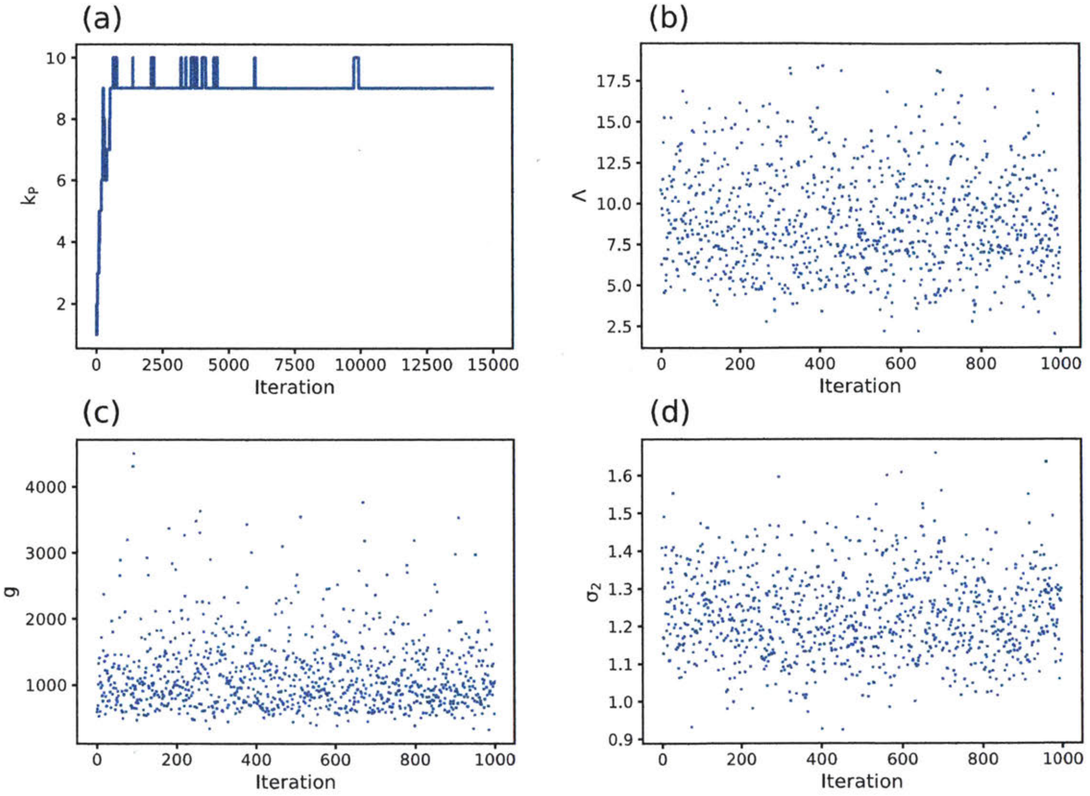


Figure 3-5: Variable samples for k_P, Λ, g, and σ^2 for an example target analyte spectrum. k_P is plotted for the entire iteration whereas the rest are plotted for the last 1000 iterations.

show the simulation variables for the first stage with the target analyte spectrum. The evolution of the number of Raman peaks k_P, which dictates the variable dimension of the model, throughout the entire iteration process is shown in Figure 3-5 (a). Our annealing schedule ensured that the algorithm converged to a single posterior model space as can be seen from this plot. All the peak variable evolutions are shown in Figure 3-6 for the last 1000 iterations. Due to our annealing schedule, no label switching is observed, which greatly facilitates the variable estimation step. After this stage, all variables can be estimated through the samples shown in Figure 3-5 and Figure 3-6.

Figure 3-7 and Figure 3-8 show similar plots for a mixture spectrum. In Figure 3-8 (f), $\beta_{T,\text{mix}}$ is defined as $\hat{c}_{\text{mix}}/c_{\text{pure}}$ in our computation. The estimated value of this variable was used to compute the target analyte concentration in this mixture spectrum.

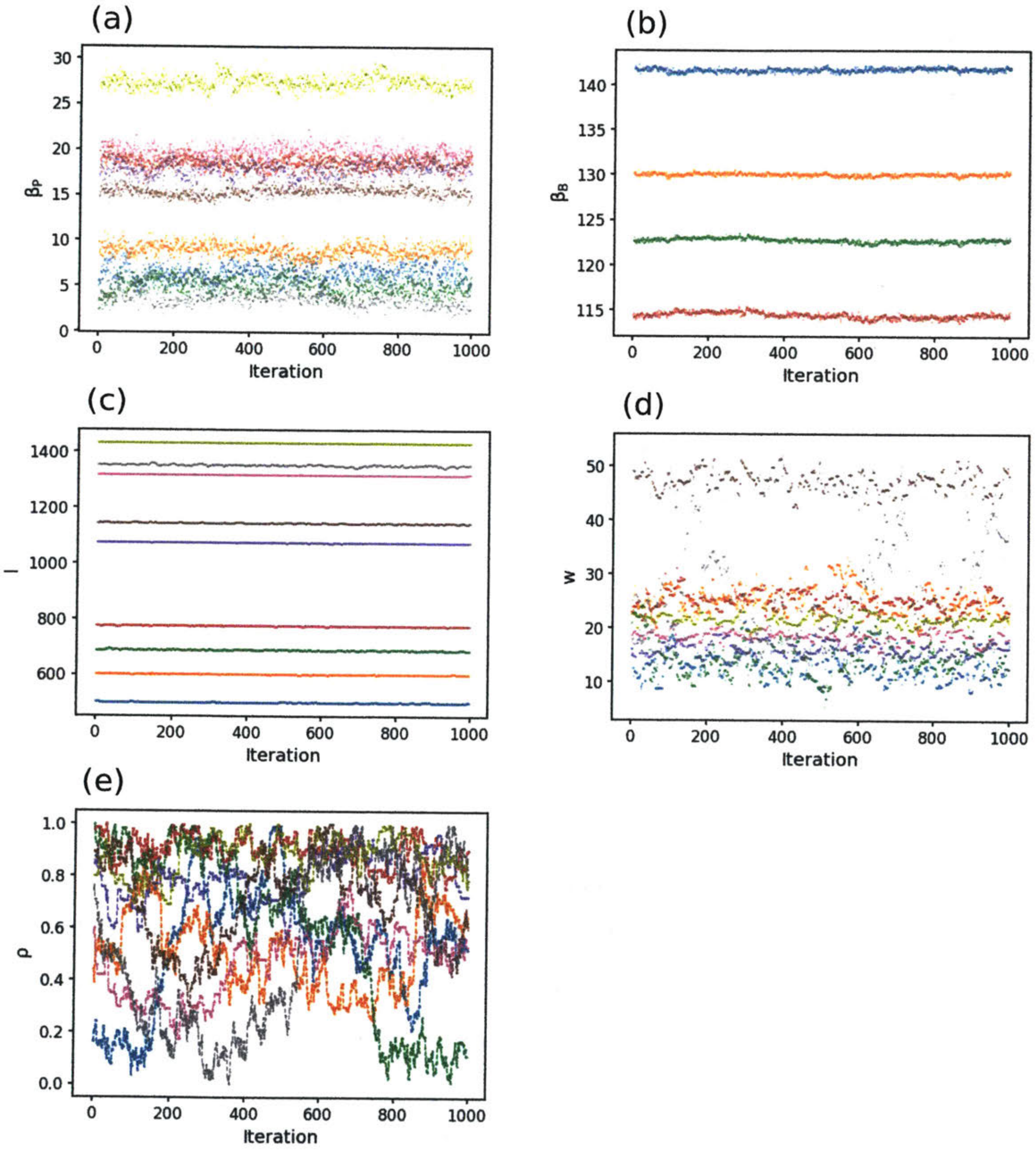


Figure 3-6: Variable samples for $\boldsymbol{\beta}_P$, $\boldsymbol{\beta}_B$, $\mathbf{l}$, $\mathbf{w}$, and $\boldsymbol{\rho}$ for an example target analyte spectrum. All variables are plotted for the last 1000 iterations.

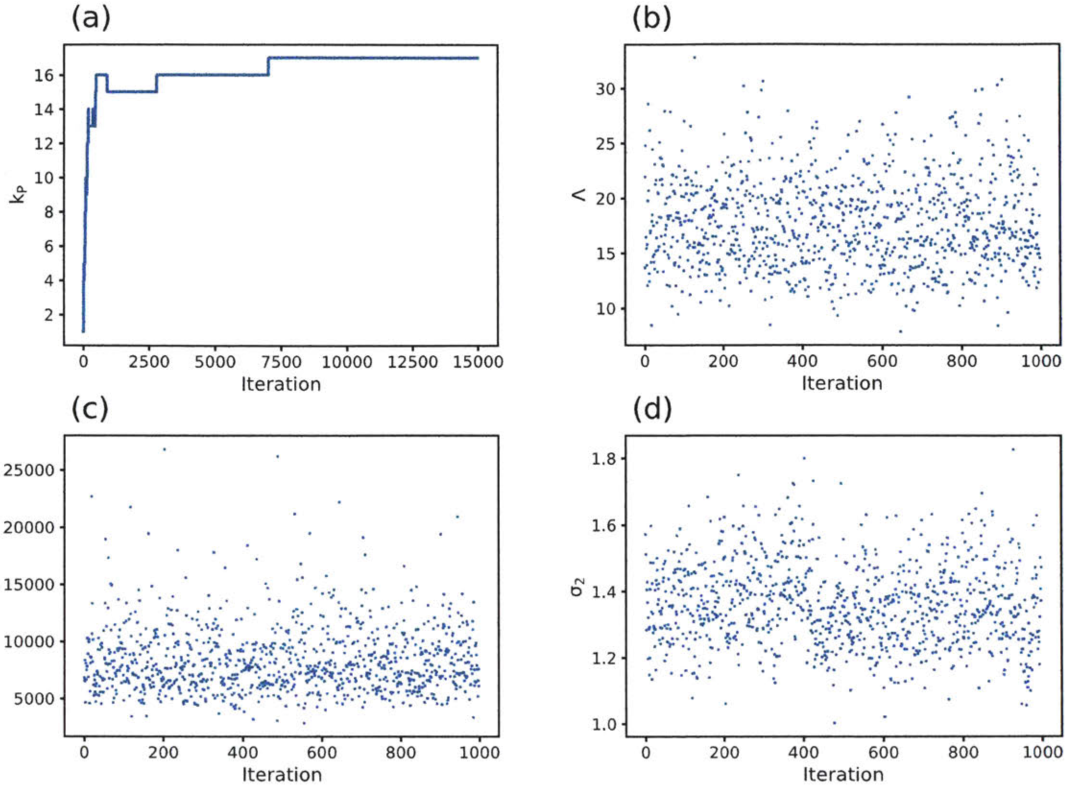


Figure 3-7: Variable samples for k_P, Λ, g, and σ^2 for an example mixture spectrum. k_P is plotted for the entire iteration whereas the rest are plotted for the last 1000 iterations.

From Figure 3-8 (a), it can be seen that the model converged to $k_P = 17$, which was below the number of interfering Raman peaks (50) in this simulation. This was due to the peak overlapping phenomenon mentioned earlier – a significant portion of Raman peaks from the composition analytes overlapped, and without any prior information such as the ones obtained from the target analyte spectrum decomposition, it would be extremely difficult to perform finer peak assignments. However, once the target analyte spectrum representation is given as $\tilde{f}_P(\boldsymbol{\nu})$, the breakdown and decomposition of broader mixture peaks can become more likely, if they overlap with peaks from the target analyte spectrum. As a result, peak decomposition and analyte assignment in this stage are more physically rooted. For an extreme case, if the spectra from all the constituent analytes are provided as prior information, the mixture spectrum decomposition will most likely result in perfect analyte assignment and becomes more

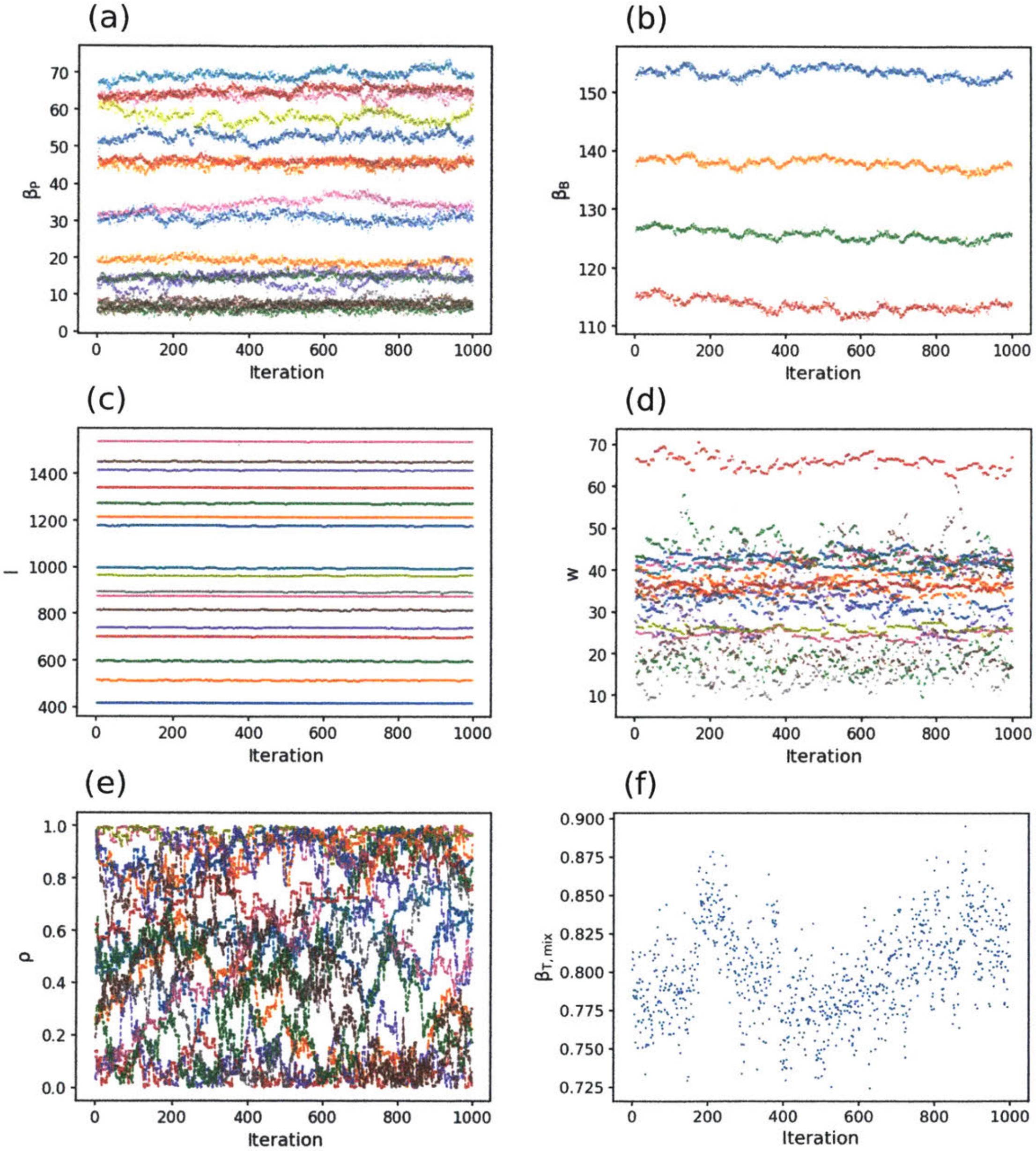


Figure 3-8: Variable samples for $\boldsymbol{\beta}_P$, $\boldsymbol{\beta}_B$, $\mathbf{l}$, $\mathbf{w}$, $\boldsymbol{\rho}$, and $\beta_{T,\text{mix}}$ for an example mixture spectrum. All variables are plotted for the last 1000 iterations. $\beta_{T,\text{mix}}$ is defined as $\hat{c}_{\text{mix}}/c_{\text{pure}}$ in our computation.

like the case in CLS as discussed in Section 1.3.2.

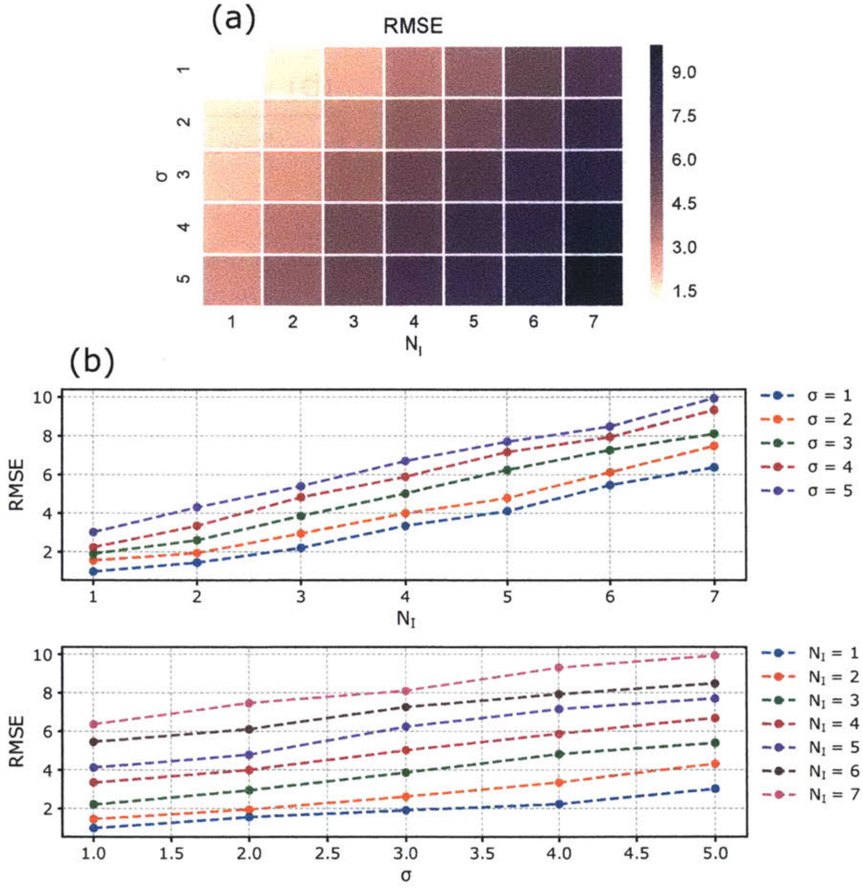


Figure 3-9: (a) RMSE heatmap plot with varying N_I and σ. (b) 1D plots showing the RMSE change with fixed σ (above) and fixed N_I (below). N_I is the number of co-existing interfering analytes in the mixture and σ is the standard deviation of the additive Gaussian random noise. All RMSEs were calculated based on 1000 independently generated random mixtures.

Next, we evaluated the generalized performance of our algorithm when the number of co-existing interfering analytes N_I and the additive Gaussian noise scale σ varied. The noise scale σ was kept as 1 for the reference target analyte spectrum. For each mixture test set with a fixed N_I and σ, we generated 1000 mixture spectra with randomly generated interfering analytes and randomly generated concentrations for all the constituent analytes as described previously. We used the root mean squared error (RMSE) between our estimations and the ground truth values across all the 1000 measurements as the performance metric. In total, we generated 35 test sets where we

N_I	1	2	3	4	5	6	7
$\sigma = 1$	1.0	1.4	2.2	3.3	4.1	5.5	6.4
$\sigma = 2$	1.6	1.9	2.9	4.0	4.8	6.1	7.5
$\sigma = 3$	1.9	2.6	3.9	5.0	6.2	7.3	8.1
$\sigma = 4$	2.2	3.4	4.8	5.9	7.2	7.9	9.3
$\sigma = 5$	3.0	4.3	5.4	6.7	7.7	8.5	9.9

Table 3.1: RMSE values with varying N_I and σ (also plotted in Figure 3-9). N_I is the number of co-existing interfering analytes in the mixture and σ is the standard deviation of the additive Gaussian random noise. All RMSEs were calculated based on 1000 independently generated random mixtures.

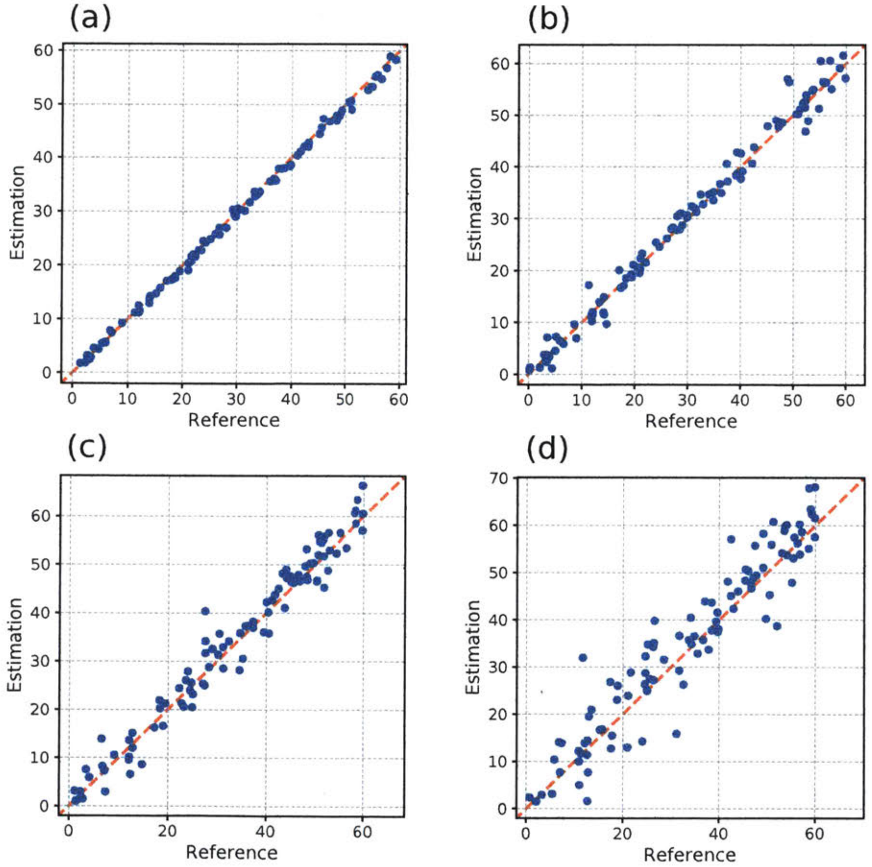


Figure 3-10: Estimation scatter plots for simulations with $\sigma = 1$. (a) $N_I = 1$. (b) $N_I = 3$. (c) $N_I = 5$. (a) $N_I = 7$.

varied N_I from 1 to 7 and σ from 1 to 5, both in steps of 1. The resulting RMSE across the 1000 test spectra for each set is shown in Table 3.1 and the corresponding 2D heatmap is shown in Figure 3-2 (a). In addition, 1D plots showing how RMSE changes

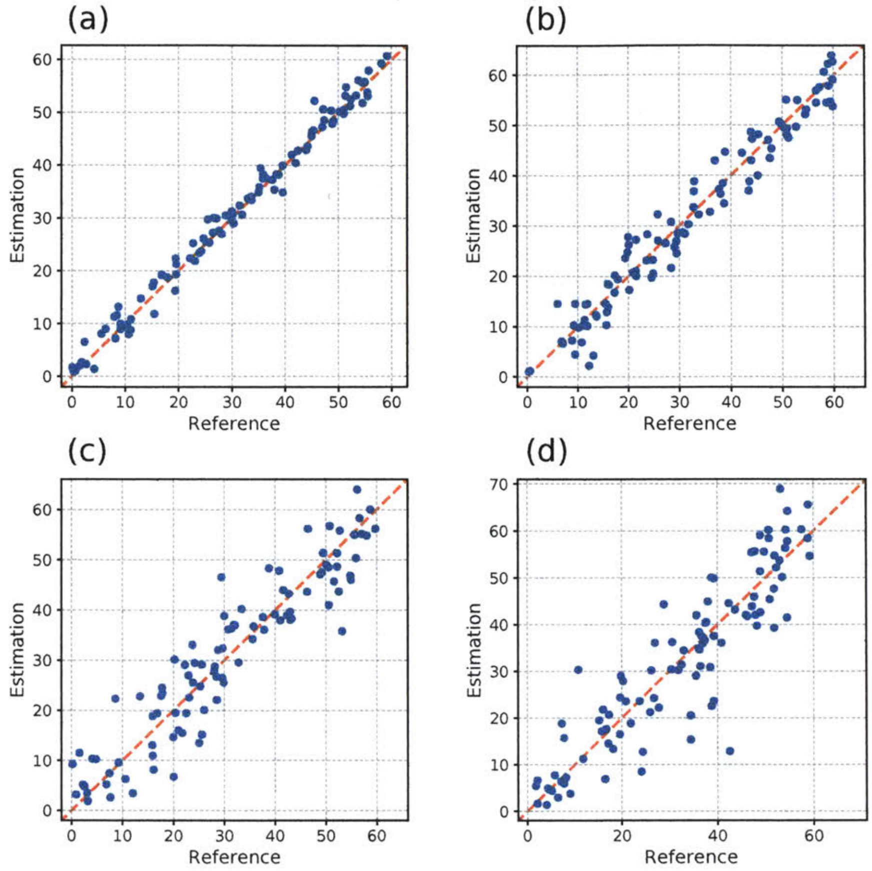


Figure 3-11: Estimation scatter plots for simulations with $\sigma = 3$. (a) $N_I = 1$. (b) $N_I = 3$. (c) $N_I = 5$. (a) $N_I = 7$.

with N_I under fixed σ (and vice versa) are shown in Figure 3-2 (b). Scatter plots with 100 samples for visualizing the estimation variabilities under different interfering analyte settings are shown in Figure 3-10 through Figure 3-12. $\sigma = 1$ for Figure 3-10, $\sigma = 3$ for Figure 3-11, and $\sigma = 5$ for Figure 3-12.

As seen from these results, overall our algorithm was able to provide a consistent and reliable estimation for the target analyte concentration over a wide range of test conditions. The RMSEs under all test cases were below $\approx 17\%$ of the concentration variation from 0 to 60 for the target analyte. As the measurement became noisier or more interfering analytes were added to the mixture, more estimation error was observed. This is expected as the effect of more noise or more interfering analytes increases the uncertainty of correct analyte peak assignment and peak amplitude

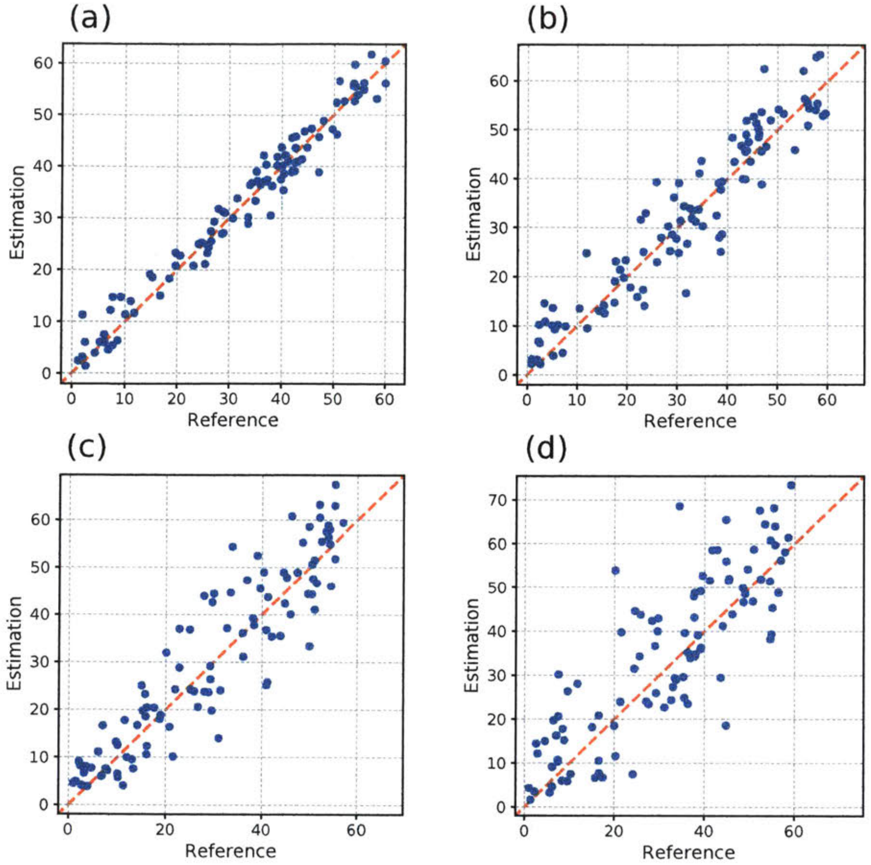


Figure 3-12: Estimation scatter plots for simulations with $\sigma = 5$. (a) $N_I = 1$. (b) $N_I = 3$. (c) $N_I = 5$. (a) $N_I = 7$.

estimation. A closer examination of error change with fixed N_I or fixed σ in Figure 3-9 (b) indicates a linear increase of error with the other variable. This suggests that a simple linear model with RMSE being the dependent variable, and N_I and σ being the independent variables may be able to predict our algorithm's performance in a more generalized scenario. However, further research is needed to rigorously analyze the error bound of our algorithm under these situations.

3.2.3 Comparison Study

We further compared the performance of our algorithm against several popular multivariate regression quantification algorithms in chemometrics. Three algorithms including partial least squares regression (PLSR), principle component regression

(PCR) and ridge regression (RR) were selected for this comparison study. The implementations for these algorithms were from the scikit-learn 0.19.1 package with Python 3.6. An important distinction between these multivariate regression algorithms and our algorithm is that they are typically built based on a mixture training set with cross validation for model selection. This requires pre-existing mixture spectra as well as the corresponding ground truth reference measurements for the target analyte in the mixtures. In practice, the ground truth reference measurements are usually obtained through a separate chemical assay such as high performance liquid chromatography (HPLC). On the contrary, our algorithm only requires the reference spectrum of the target analyte at a known concentration as prior information before the actual estimation.

Our focus in this study is to investigate how the quantification results compare across the different algorithms as the number of mixture training data varies. This comparison study is relevant for practical applications as mixture training data itself is often labor or resource intensive to acquire in large volume since in addition to spectral data collection, special sample handling or additional analytical measurement like HPLC is usually required. In contrast, the reference spectrum of the target analyte required by our algorithm is much easier to prepare in practice. For the three multivariate regression algorithms, we created training and test sets with a randomized process similar to previously described, except for that now we kept the same fixed mixture components across each training/test set with randomized concentrations. This is to ensure the proper settings for the multivariate regression algorithms. For each dataset, we generated 100 samples in the test set and varied the sample size in the training set from 6 to 24 in steps of 3. During model training, we first performed 3-fold cross validation and parameter grid search within the training set to choose the optimal hyper-parameters for each algorithm (i.e., number of loading vectors in PLS, number of retaining principle components in PCR, and the regularization parameter value in RR respectively). We then refitted the model with the optimal hyper-parameter on the entire training set and applied the resulting model on the test set to calculate the RMSE for the dataset. Since these multivariate regression

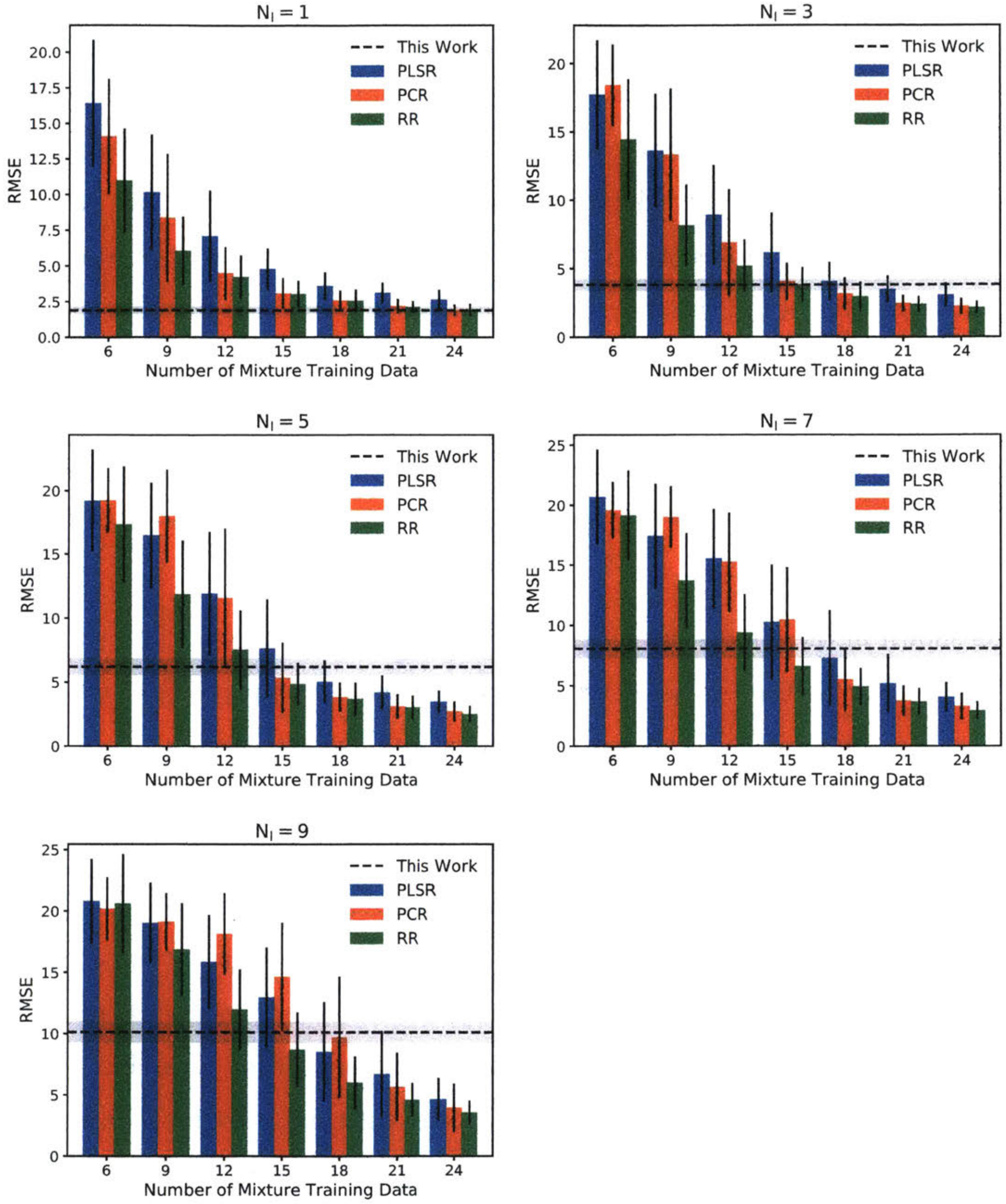


Figure 3-13: Error plots for our algorithm and three other multivariate regression algorithms with various mixture training data sizes for different N_I with $\sigma = 3$. The error bars in the bar plots indicate the standard deviation of RMSE for the multivariate regression algorithms across 100 independently simulated datasets each consisting of 100 test spectra. The horizontal shaded area around the dotted line indicates the standard deviation of RMSE for our algorithm across 10 independently simulated datasets each consisting of 100 test spectra. N_I is the number of co-existing interfering analytes in the mixture in our simulations.

algorithms have a high variance under the small training sample size regime, we repeated this process a number of times on independently generated training and test sets and report its performance based on aggregated statistics across these independent runs. In Figure 3-13, we show the average RMSE across 100 independently simulated datasets for the three multivariate regression algorithms as the number of mixture training data varies for different N_I at $\sigma = 3$. The error bars in the plots represent the standard deviation of RMSE across the 100 independent runs. As comparison, we also show the average RMSE for our algorithm across 10 test sets each consisting of 100 mixture spectra in the same plots. The shaded area around the error line indicates the standard deviation of RMSE across the 10 test sets. Since our algorithm does not use the mixture data for training, the error line stays horizontal across the axis for mixture training sample size.

As shown in these plots, for all the three multivariate regression algorithms, the prediction error for the test set decreases monotonically as a function of the number of mixture training data. This is expected as with more training data, it is more likely for these algorithms to effectively capture the sample subspace of the mixture data during the training process, thereby increasing the regression accuracy. Under small training sample size regime with less than ≈ 15 training spectra, there is a clear advantage of our algorithm in terms of both estimation accuracy and consistency under all testing situations. On the contrary, once there are enough training data to accurately construct the regression models, our algorithm is unable to match their performance and the accuracy gap widens with more training data. Under low interfering conditions, our algorithm remains competitive across the range of the training sample size change. This is due to the fact that with low interfering conditions, it is more likely to unambiguously resolve analyte peak assignment and accurately estimate peak amplitudes, resulting in near-optimal quantification.

Although a non-negligible accuracy gap is present for the high interfering conditions in Figure 3-13, we note here that with perfect linearity and fixed mixture components with concentrations generated from uniform distributions in both the training and test sets, our numerical experiment was constructed such that the conventional multivariate

Training Data Size	6	9	12	15	18	21	24
$N_I = 1$							
This Work				1.9 ± 0.3			
PLSR	16.4 ± 4.4	10.2 ± 4.0	7.1 ± 3.2	4.8 ± 1.5	3.6 ± 1.0	3.1 ± 0.7	2.6 ± 0.7
PCR	14.1 ± 4.0	8.4 ± 4.5	4.5 ± 1.8	3.1 ± 1.1	2.6 ± 0.7	2.2 ± 0.5	1.9 ± 0.4
RR	11.0 ± 3.6	6.1 ± 2.4	4.2 ± 1.5	3.0 ± 0.9	2.5 ± 0.8	2.1 ± 0.4	1.9 ± 0.4
$N_I = 3$							
This Work				3.8 ± 0.4			
PLSR	17.7 ± 4.0	13.7 ± 4.1	8.9 ± 3.7	6.2 ± 2.9	4.1 ± 1.4	3.5 ± 1.0	3.1 ± 0.9
PCR	18.4 ± 2.9	13.4 ± 4.8	6.9 ± 3.9	4.1 ± 1.3	3.2 ± 1.2	2.5 ± 0.6	2.3 ± 0.6
RR	14.5 ± 4.4	8.2 ± 3.0	5.2 ± 1.9	3.8 ± 1.3	3.0 ± 1.0	2.4 ± 0.6	2.1 ± 0.5
$N_I = 5$							
This Work				6.2 ± 0.7			
PLSR	19.2 ± 4.0	16.5 ± 4.2	11.9 ± 4.8	7.6 ± 3.8	5.0 ± 1.7	4.2 ± 1.3	3.5 ± 0.9
PCR	19.2 ± 2.5	18.0 ± 3.6	11.6 ± 5.4	5.4 ± 2.7	3.8 ± 1.1	3.1 ± 1.0	2.7 ± 0.8
RR	17.3 ± 4.6	11.9 ± 4.2	7.5 ± 3.1	4.9 ± 1.7	3.7 ± 1.3	3.0 ± 0.9	2.5 ± 0.6
$N_I = 7$							
This Work				8.1 ± 0.8			
PLSR	20.7 ± 3.9	17.4 ± 4.4	15.5 ± 4.1	10.3 ± 4.7	7.3 ± 4.0	5.2 ± 2.4	4.1 ± 1.2
PCR	19.6 ± 2.4	19.0 ± 2.6	15.3 ± 4.1	10.5 ± 4.3	5.5 ± 2.6	3.8 ± 1.2	3.3 ± 1.1
RR	19.1 ± 3.7	13.7 ± 3.9	9.4 ± 3.2	6.6 ± 2.4	4.9 ± 1.5	3.7 ± 1.1	2.9 ± 0.7
$N_I = 9$							
This Work				10.1 ± 0.9			
PLSR	20.8 ± 3.4	19.0 ± 3.3	15.8 ± 3.8	13.0 ± 4.1	8.5 ± 4.1	6.7 ± 3.5	4.6 ± 1.7
PCR	20.2 ± 2.6	19.1 ± 2.3	18.1 ± 3.3	14.6 ± 4.4	9.7 ± 5.0	5.7 ± 2.8	4.0 ± 2.0
RR	20.6 ± 4.0	16.9 ± 3.8	12.0 ± 3.3	8.7 ± 3.0	6.0 ± 2.1	4.6 ± 1.4	3.6 ± 1.0

Table 3.2: RMSE values for our algorithm and three other multivariate regression algorithms with various mixture training data sizes for different N_I (also plotted in Figure 3-13). The number before/after the ± sign indicates the mean/standard deviation of RMSE across independent runs. N_I is the number of co-existing interfering analytes in the mixture in our simulations.

regression algorithms were set to achieve near-optimal performance under situations with moderate-to-large training data volume. In practice, apart from experimentation and instrumentation-related issues as described in Wolthuis et al. [134], the performance of these multivariate regression algorithms are more critically dependent on the quality of the training data, including the size of the training data as well as the measurement conditions for the training and test data. In general it is desirable to select training data that are most representable to the mixture conditions of the test data with sufficient concentration variabilities for critical analytes [45]. These requirements can be difficult to satisfy without substantial resources being allocated to the training data collection and verification process. In addition, mixture environment may introduce undesirable effect to the spectral signal. For example, it is known that the Raman peak may shift with environmental pH for many chemicals such as certain amino acids [135]. Peak shifts introduce non-linearity into the spectral basis and may reduce estimation accuracy for linear regression algorithms. It is therefore desirable to maintain the pH of the environment for both the training and test data during the measurement for PLSR-like linear regression algorithms [33]. In contrast, our algorithm is less sensitive to these various requirements so long as the target analyte spectrum stays the same in the mixture comparing to its reference measurement in native form. Therefore, in practice our algorithm may still be comparable or even outperform these multivariate regression algorithms with larger training data volume depending on the nature of the measurement.

3.3 Experimental Data Study

3.3.1 Physical Mixture Raman Spectroscopy Data

Following the numerical experiments, we further tested our algorithm on experimental Raman spectroscopy datasets collected in our lab. The first dataset was a four-component aqueous mixture study with Raman spectroscopy. The four components were glucose, lactic acid, L-lysine, and sodium pyruvate. Four mixture solutions were made where the concentrations for these components varied across the mixtures. These concentrations are shown in Table 3.3. In addition, the spectra for all the pure components were measured at 500 mM concentrations. For each set of the Raman measurement, 5 repeated spectra were collected in sequence. As spectral preprocessing, we first took the median across the 5 measurements for each spectral data point for cosmic ray removal and noise reduction. Afterwards, a 21-point Savitzky-Golay filter with a polynomial order of 3 was applied across the spectral dimension to further enhance the spectral signal-to-noise ratio (SNR). Finally, a direct subtraction of Raman spectrum measured with water was carried out to remove background Raman signals from water as well as the optical components along the light path. The Raman spectra for all the four mixtures as well as all the pure components after the pre-processings are shown in Figure 3-14.

	Glucose (mM)			Lactic Acid (mM)			L-Lysine (mM)			Sodium Pyruvate (mM)		
	True	Estimation	Error	True	Estimation	Error	True	Estimation	Error	True	Estimation	Error
Mixture 1	71.4	78.6	7.2	147.1	156.6	9.5	145.8	120.5	25.3	111.1	136.4	25.3
Mixture 2	142.9	139.9	3.0	205.9	211.2	5.3	104.2	72.5	31.7	83.3	85.1	1.8
Mixture 3	178.6	176.4	2.2	117.6	140.9	23.3	62.5	42.8	19.7	83.3	110.2	26.9
Mixture 4	107.1	107.7	0.6	29.4	28.3	1.1	187.5	133.9	53.6	222.2	240.7	18.5

Table 3.3: Constituent compositions and their respective concentrations for all the mixtures in this study. The estimated concentrations with our algorithm are also shown alongside their true values.

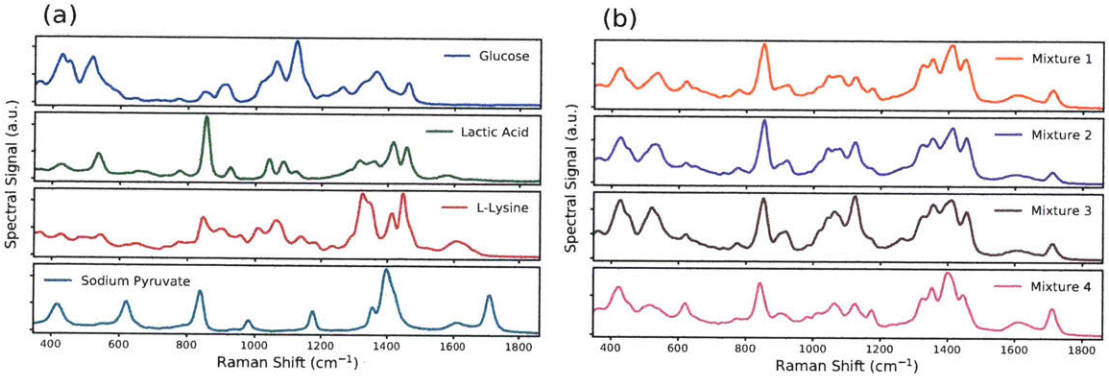


Figure 3-14: (a) Raman spectra for the four composition materials in pure form measured at equal concentrations at 500 mM. (b) Raman spectra for the four mixture samples used in this study. Preprocessings described in the text have been applied to the raw spectra data prior to the plots.

We ran our algorithm on each one of the four components, and estimated its concentrations in all the four mixtures in turn. The estimation results, which are the average of 10 independent computation runs, are shown in Table 3.3. In addition, an estimation scatter plot is shown in Figure 3-15. Overall, despite the fact that only the reference spectrum for the target analyte was provided, our algorithm was able to perform accurate estimations for most of the test analytes, with the estimation for glucose closest to the reference values. The largest estimation errors were for L-lysine. Figure 3-15 suggests that the estimated concentrations for L-lysine seem to be off by some constant offset values from the reference concentrations. It is therefore possible to perform a post-estimation calibration with linear regression to further improve the estimation accuracy. The Raman spectrum for L-lysine in Figure 3-14 (a) indicates that among the four composition materials, L-lysine has the most peak overlaps with the rest of the composition materials, which is perhaps why its estimation results were the worst among the four materials.

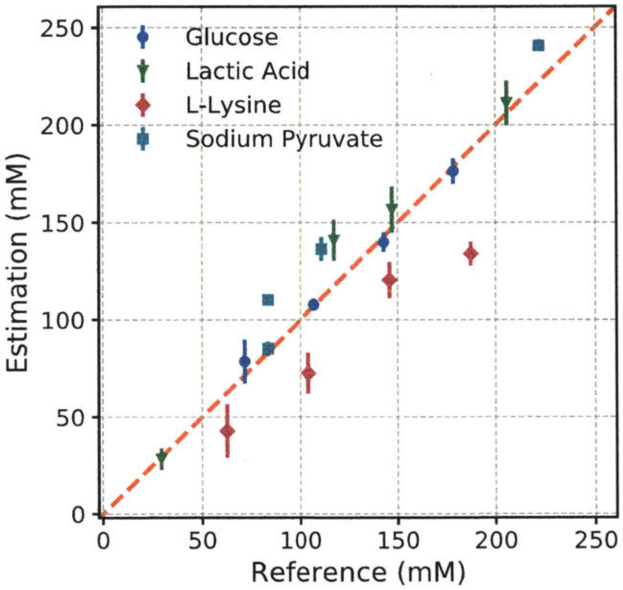


Figure 3-15: Estimation scatter plot for all the constituent materials in the four mixture samples used in this study. The reference and estimated concentrations are shown in Table 3.3. The estimated values were obtained as the mean of 10 independent algorithm runs. The error bars indicate the standard error of the mean (SEM) for the estimations across the runs.

It is worth pointing out that the experimental measurements for this dataset were not taken to maximumly optimize the estimation performance of our algorithm. The target analyte reference spectra measurements were all performed at 500 mM concentrations, which were much larger than the actual expected concentrations in the mixtures. In practice, we suggest to measure the target analyte spectrum at concentrations around the same neighborhood of its concentrations in mixtures, to minimize potential errors introduced by signal non-linearity in terms of analyte concentration and its spectral peak strength (with the exception for low SNR situations). In addition, the measurements were performed where the liquid samples were directly on top of a cover slip without any specialized sample holders. This may introduce extra measurement variance from our experience. Nonetheless, our estimations shown in Figure 3-15 are promising with only small relative errors compared to the reference measurements for all the four composition materials.

3.3.2 Raman Spectroscopy Data for CHO Cell Culture

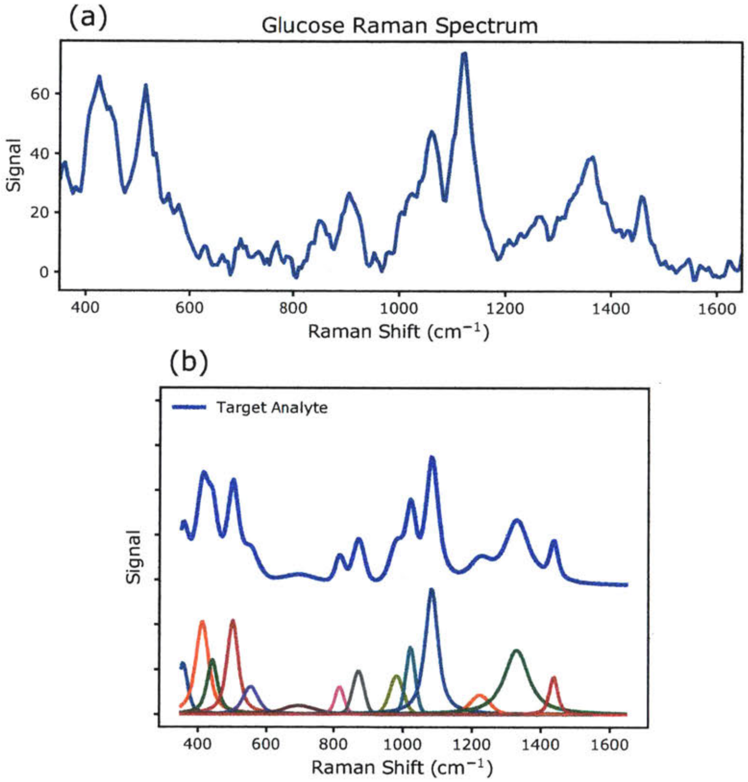


Figure 3-16: (a) Glucose Raman spectrum measured at 40 mM with our system. (b) Peak decomposition for the glucose Raman spectrum in (a).

Next, we further tested our algorithm on experimental Raman spectroscopy data collected for biopharmaceutical applications. Raman spectra were collected to monitor the concentration of the main carbon source, glucose, in the growth environment during the fermentation process of Chinese hamster ovary (CHO) cells, which are the most widely used expression systems for industry production of recombinant protein therapeutics. The initial CHO growth medium included all the nutrients required by

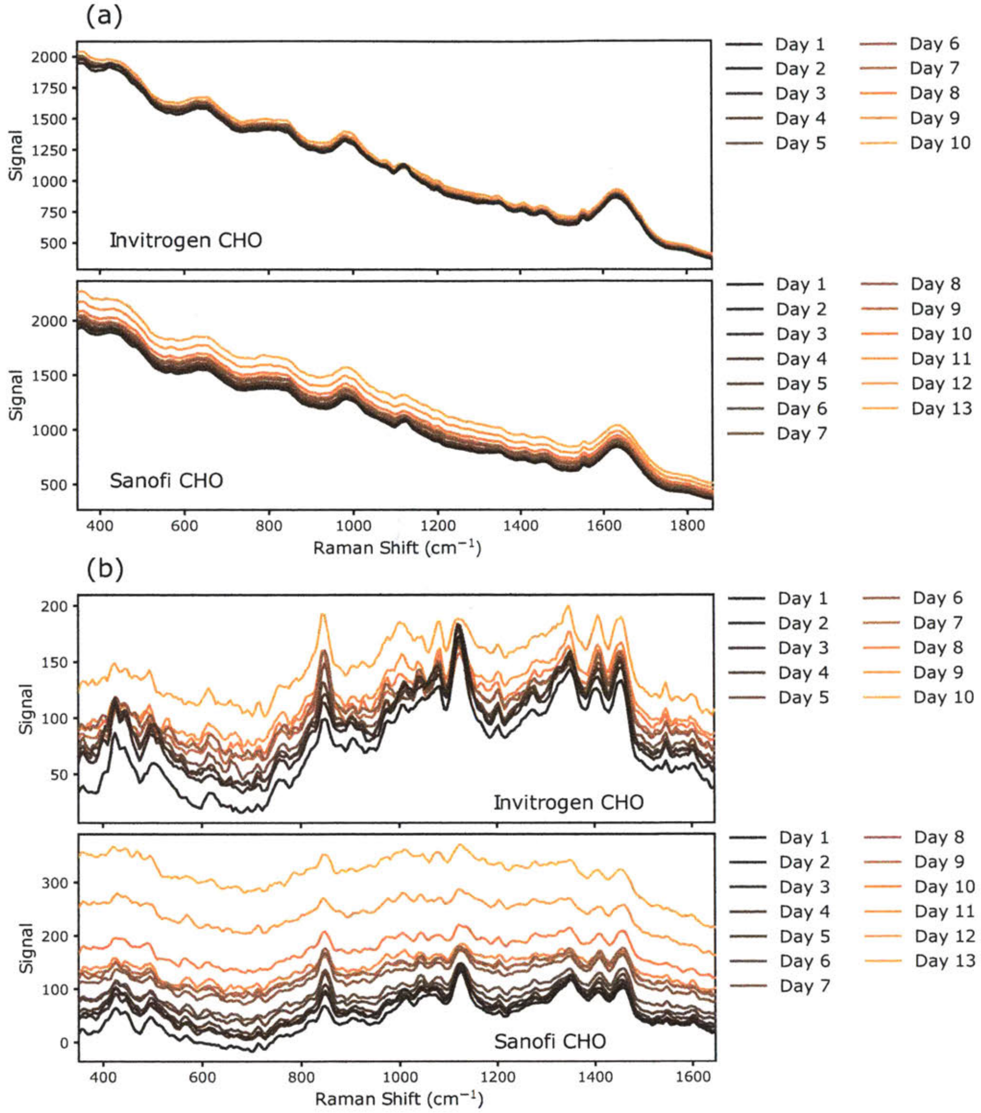


Figure 3-17: Raman spectra of the CHO culture supernatant for the Invitrogen and Sanofi CHO cell lines. (a) Spectra without any background subtraction. (b) Processed spectra after background subtraction with the water Raman spectrum.

the cells such as the necessary carbon sources, nitrogen sources, salts and trace elements. As the fermentation advanced, nutrients were consumed by the cells and metabolites were being produced and released into the growth environment. Therefore, the culture environment represented a complex aqueous mixture and was changing constantly over the fermentation process. Knowledge of key nutrients such as glucose during the fermentation process through an on-line measurement such as Raman spectroscopy can help regulate the fermentation condition for better yield or reproducibility [45]. During our experiment, supernatant from the culture material was collected on a daily basis. Raman spectra of the collected supernatant were immediately measured with a confocal Raman spectroscopy system at 830 nm excitation wavelength. Meanwhile, HPLC measurement was used to obtain the reference concentrations for glucose in the supernatant samples. The instrumentation and experimental setup are described in more details in Singh et al. [136]. Two independent experiments with different CHO cell lines from Invitrogen Inc. and Sanofi-Aventis Deutschland GmbH Inc. were carried out respectively. In addition to Raman spectra from the supernatant samples, Raman spectrum for pure glucose dissolved in solution was also collected with the same instrument.

The fermentation experiments lasted a total of 10 days for the Invitrogen CHO cell line and 13 days for the Sanofi CHO cell line. Therefore, 10 and 13 supernatant Raman measurements were collected for these two cell lines respectively. The reference Raman spectrum for pure glucose solution was taken at 40 mM concentration. For each set of the Raman measurement, 10 repeated spectra were collected in sequence. As spectral preprocessing, we first took the median across the 10 measurements for each spectral data point for cosmic ray removal and noise reduction. Afterwards, a 21-point Savitzky-Golay filter with a polynomial order of 3 was applied across the spectral dimension to further enhance the spectral signal-to-noise ratio (SNR). A spectral window from 350 cm^{-1} to 1650 cm^{-1}, which covered all the major Raman peaks in glucose and CHO Raman spectra, was selected for further processing. Finally, a direct subtraction of Raman spectrum measured with water was carried out to remove background Raman signals from water as well as the optical components along the light path. The processed spectra for glucose and its peak decomposition from the first stage of our algorithm are shown in Figure 3-16. The Raman spectra from the CHO cell line measurements are shown in Figure 3-17, where spectra without any water background subtraction are shown in in Figure 3-17 (a) and spectra with the water background subtraction are shown in Figure 3-17 (b). As can be seen from the figure, after the water background subtraction, Raman peaks from the underlying composition materials start to show up on the relevant scale. The mixture environments for the two cell lines were different due to the fact that the growth media for these two cell lines had different compositions. This results in the differences in the corresponding Raman spectra shown in Figure 3-17. The overall baseline drifts over days for each cell line were likely caused by the changing refractive index of the supernatant due to its composition change over the course of the fermentation process.

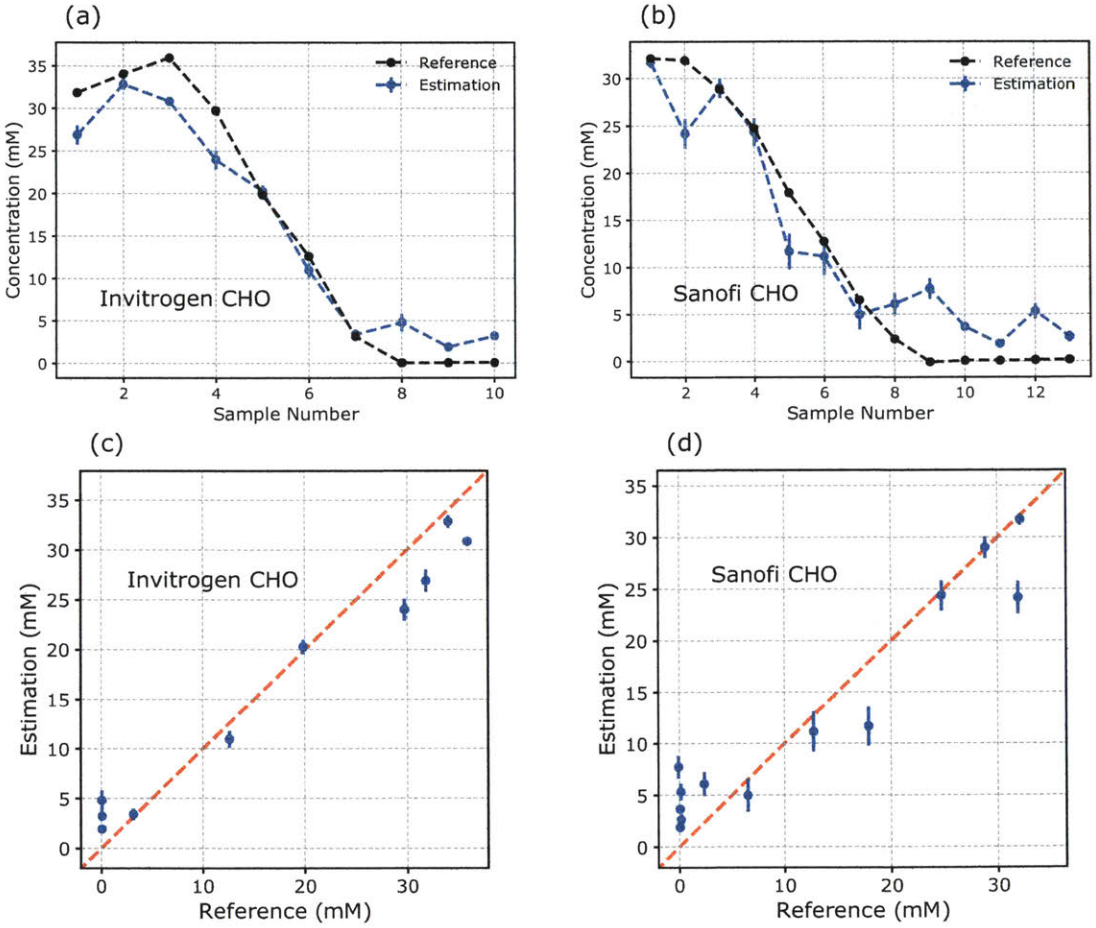


Figure 3-18: Plots for glucose estimation with Raman spectroscopy for the Invitrogen (left) and Sanofi (right) CHO cell line measurements. The estimated values were obtained as the mean of 10 independent algorithm runs. The error bars indicate the standard error of the mean (SEM) for the estimations across the runs. The reference values were obtained with independent HPLC measurements.

We applied our algorithm with the same modeling parameter settings as with the previous tests on the measured glucose and CHO Raman spectra. The average

Cell Line	RMSE (mM)	MAE (mM)	R^2
Invitrogen CHO	3.5	2.9	0.94
Sanofi CHO	4.2	3.3	0.89

Table 3.4: Estimation results for our algorithm with Raman spectroscopy for the Invitrogen and Sanofi CHO cell line measurements.

of 10 algorithm runs is plotted in Figure 3-18 together with the HPLC reference measurements for both cell lines. The error bars in the plots indicate the standard error of the mean (SEM) of the estimation runs. The HPLC measurements were estimated to have ±0.5 mM accuracy. Overall our algorithm shows a consistent and reliable estimation of glucose, as shown in Table 3.4, with RMSE of 3.5 mM, mean absolute error (MAE) of 2.9 mM, and R^2 of 0.94 for the Invitrogen CHO measurement, and RMSE of 4.2 mM, MAE of 3.3 mM, and R^2 of 0.89 for the Sanofi CHO measurement. The error is comparable with the 3-σ limit of detection for pure glucose solution with our measurement system, which was ≈ 6 mM based on peak-SNR estimation. Comparing with conventional PLSR-like multivariate regression algorithms, our algorithm only requires additional measurements of pure glucose solution and water. Otherwise additional experimental runs need to be planned in order to accumulate enough data for model training and validation. With resource-intensive applications like industrial fermentation, a sample-efficient approach like our algorithm can significantly reduce the research cost and development cycle. It is worth noting that it is also possible to explicitly measure spectra for all constituent analytes in the mixture and use classical least squares (CLS) regression to quantify analyte concentrations without acquiring additional mixture training data [33, 136]. However, the library spectra collection process can be labor-intensive. In addition, it is usually difficult to know all the mixture constituents ahead of time in a general setting. Therefore, in practice, our algorithm has advantages in terms of both performance competitiveness as well as looser requirement on training or additional measurements.

3.4 Conclusions

In summary, we have developed a two-stage quantification algorithm with the Bayesian modeling framework and the RJMCMC computation. We tested our algorithm on both simulated as well as experimentally collected Raman spectroscopy datasets to validate its usage. The successful quantification of glucose concentration in a complex aqueous cell culture environment without any mixture training data suggests its promising potential for applications involving Raman spectral analysis.

In practice, collecting high quality Raman spectroscopy training datasets with reference measurements in sufficient volume for multivariate regression algorithms can be a long, challenging, and labor/resource-intensive process for many application disciplines. In addition, due to the intertwined nature of statistical data analysis and experimental design in chemometrics, timely quantitative feedback can often impact aspects of experimental design in a significant way. An analyte quantification algorithm without any requirement on mixture training data such as the one developed in this work is therefore desirable in many scenarios. From this aspect, we envision our algorithm to be an important complementary tool to the multivariate regression algorithms for quantification analysis with Raman spectroscopy datasets.

Chapter 4

Numerical Investigations of Non-invasive Glucose Estimation with Raman Spectroscopy

In this chapter, numerical investigations of the signal requirement for non-invasive glucose estimation with Raman spectroscopy based on an experimentally collected Raman dataset are provided. Section 4.1 provides the experimental background, the data collection process, as well as a comprehensive presentation on the acquired spectral dataset from our experiment. Many practical issues with medical spectroscopy datasets are discussed in details. Section 4.2 first discusses the spectral noise analysis and its implications for the collected dataset. It then provides our signal generation methodology, the validation and test scheme, and the algorithm processing pipeline. Afterwards, Section 4.3 discusses our investigation results for the signal requirement of universal and predictive glucose estimation based on Raman spectroscopy. Various assessment approaches are presented for both the overall model performance as well as the model performances from individual sessions. Different processing variations and their results are also presented and discussed. At last, the chapter is concluded with Section 4.4.

4.1 Experiment and Data Collection

4.1.1 Raman Instrument

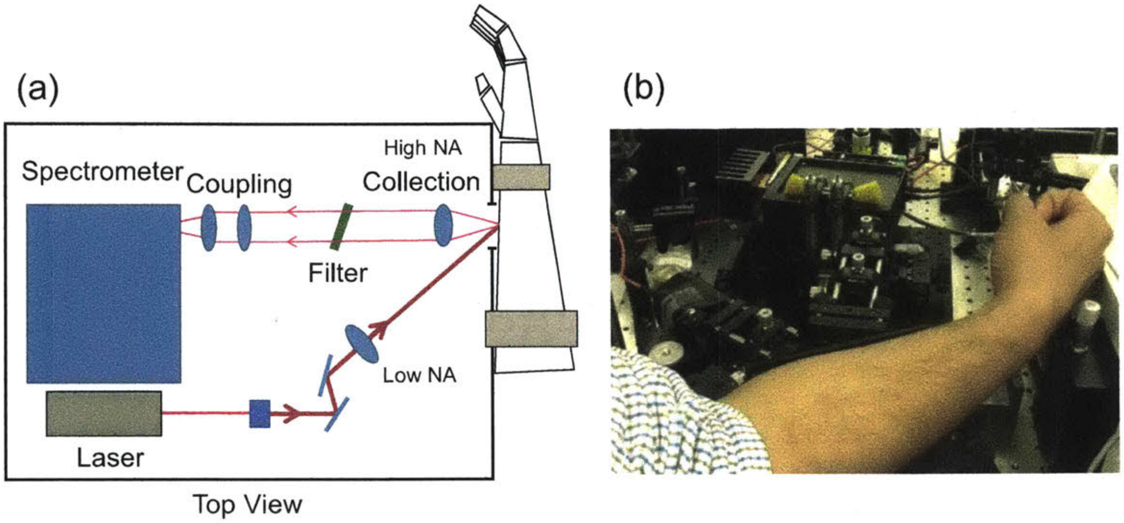


Figure 4-1: (a) A schematic diagram showing the off-line illumination Raman spectroscopy system used for *in vivo* volunteer study. (b) A picture showing the actual Raman spectroscopy system beside a volunteer forearm.

All experimental data for this study were obtained from a portable Raman spectroscopy system built in-house designed for clinical skin analysis and diagnostics applications similar to Zhao et al. [69]. Figure 4-1 shows the schematic diagram for the off-line illumination geometry of the system and a picture for the actual Raman instrument. The laser light from a 785 nm laser (SureLock, Ondax) was first expanded and then passed through a bandpass filter (MaxLine, Semrock). A low NA lens (40 mm focal length, 12.5 mm diameter, Thorlabs) was then used to focus the laser light onto the sample with an excitation spot area of $\approx$ 1 mm^2. The sample was kept behind an anti-reflection coated 0.5 mm thick quartz window (Polysciences). For all the clinical studies, the excitation laser power was set as 60 mW. This gives an excitation power density of $\approx$ 6000 mW/cm^2, which is above the ANSI maximum permissible level (MPE) of skin to 785 nm laser exposure at $\approx$ 295 mW/cm^2 [137], but lower than previous studies ($\approx$ 30000 mW/cm^2 in Enejder et al. [1] and $\approx$ 22000 mW/cm^2 in Lipson et al. [2]). Light scattered in the sample and a high NA lens (16 mm focal length, 25 mm diameter, Thorlabs) was then used to collect the Raman signal. It then passed through a long pass edge filter (RazorEdge, Semrock), which blocked the the laser light. Finally, a 40 mm focal length cylindrical lens (Thorlabs) and another 12.5 mm focal length cylindrical lens (Thorlabs) were used to focus the collected light into the 2.5 mm long and 25 µm wide slit of the spectrometer (EAGLE, Ibsen). The collected light through the spectrometer slit was then diffracted with a fused silica-based transmission grating and was detected on the one-stage TEC cooled 1024 × 122 pixels CCD detector (S7031-1007S, Hamamatsu) with 24 µm pixel sizes. The TEC cooling can reach down to $-10°$C. A flip mirror was also used to help obtain the bright-field image of the sample area before Raman data acquisition. At last, the entire Raman instrument was optically enclosed. Figure 4-2 shows the final clinical study-ready instrument.

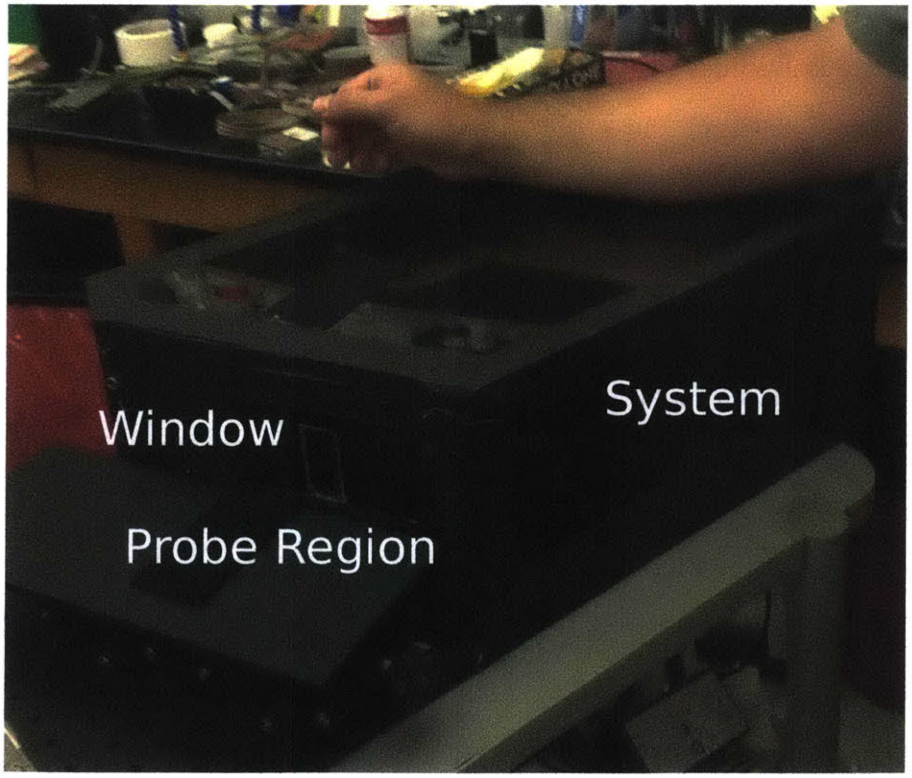


Figure 4-2: A picture showing the portable Raman instrument with optical enclosure.

4.1.2 *In vivo* Skin Raman Spectrum Collection

With the portable clinical Raman spectroscopy system described above, experiments aiming at investigating the feasibility of non-invasive glucose estimation with skin Raman spectroscopy for healthy volunteers were carried out. After approval from the Institutional Review Board at MIT (COUHES), a series of oral glucose tolerance test (OGTT) experiments were performed on 15 healthy volunteers at the MIT/IMES Clinical Research Center on MIT Cambridge Campus. For the OGTT, the fasting volunteers drank a solution containing 75 g of D-glucose over a short period of time, typically less than 10 minutes. In a healthy volunteer, consuming the glucose drink will cause his/her blood glucose level to rise and then fall due to the natural body insulin response. This happens over a period of about two hours. Throughout the clinical session, the subject's blood glucose levels were measured using a Nova biomedical Stat Strip glucose meter. This method of glucose measurement required a small drop of capillary blood obtained by fingerstick. The fingerstick blood draws and Stat Strip glucose measurements were performed by trained healthcare professionals. Before the start of each day's measurements, the meters used were put through a quality control check using high and low concentration glucose solutions. Additionally, the meters used are checked for calibration by MIT Medical department twice a year. Raman measurements from the subject's volar forearm were collected continuously during the OGTT experiment, with a 30 s integration time per spectrum frame. Meanwhile, the fingerstick glucose measurement occurred approximately every 10 mins since before the volunteer consumed the glucose beverage. The full experiment required volunteers to return for a second OGTT experiment. A small subset of the volunteers repeated the experiment consuming a water-only drink of the same volume as the glucose beverage to provide a long period of near constant glucose level for comparison study.

In total, 29 distinct clinical sessions were carried out over the period of three months. Out of these 29 distinct visits, spectra from one session exhibited excessive ambient light leakage and poor spectral quality due to excessive arm adjustment and movement. As a result, data from this session were removed from any future analysis.

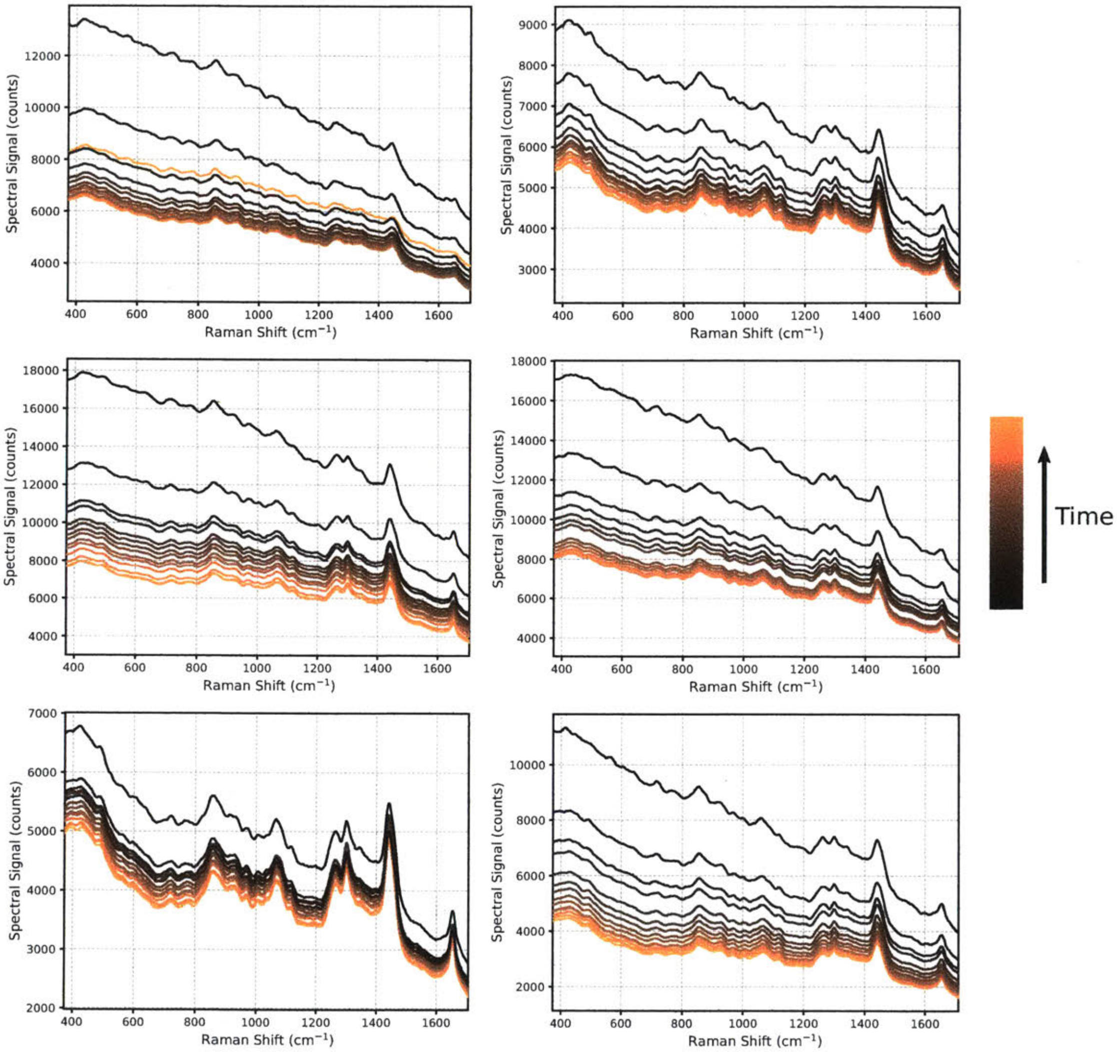


Figure 4-3: Example skin Raman spectra collected from 6 different clinical visits with 6 distinct volunteers. Spectra in each plot are color-coded in terms of collection sequence. Adjacent spectra in display were collected ≈ 10 mins apart, each corresponding to a fingerstick glucose measurement.

No other session data were taken out despite some visual SNR differences observed from visit to visit and some minor ambient light contamination for certain visits. A very small amount of spectra were unusable due to volunteers re-adjusting their positions, these spectra were treated as missing data and were substituted with the mean of the previous and next uncontaminated spectra available in the same clinical session.

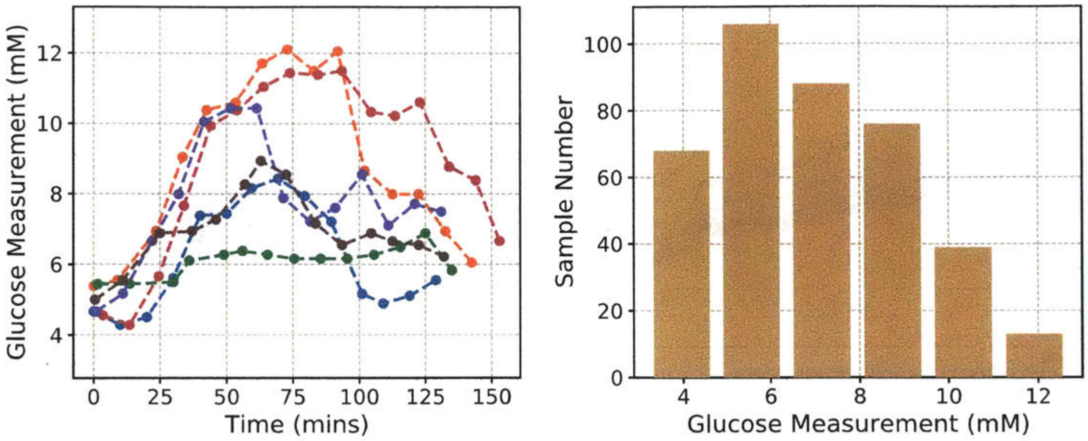


Figure 4-4: (a) Glucose fingerstick measurement profiles for 6 distinct volunteers showing the diverse body insulin responses volunteers had due to glucose consumption. (b) Histogram of the 390 fingerstick glucose measurements collected from 28 clinical visits.

Figure 4-3 shows example skin Raman spectra from 6 different clinical visits for 6 distinct volunteers. For spectral preprocessing, we first took the median across 10 adjacent acquisition frames (where each acquisition had an integration time of 30 s) for each spectral data point for cosmic ray removal and noise reduction. Afterwards, an 11-point Savitzky-Golay filter with a polynomial order of 2 was applied across the spectral dimension to further enhance the spectral SNR. In Figure 4-3, spectra in each plot are color-coded in terms of collection sequence. Adjacent spectra in display were collected $\approx$ 10 mins apart, each corresponding to a fingerstick glucose measurement. The overall reduction in signal strength is due to autofluorescence photobleaching, which will be discussed in more detail later.

For the fingerstick measurements, Figure 4-4 (a) shows example OGTT fingerstick glucose measurement profiles for 6 distinct volunteers. The diverse response profiles shown in the plot reflect the different body insulin responses volunteers had due to glucose consumption. Figure 4-4 (b) shows the histogram for all the fingerstick measurements across the 28 visits. In total there were 390 measurements, with glucose concentrations ranging from $\approx$ 3.2 to 12.7 mM. Some main sample statistics for the fingerstick glucose measurements are shown in Table 4.1.

Summary (Measurement Number = 390)	Min	Max	Mean	Median	Standard Deviation
Value (mM)	3.2	12.7	6.9	6.8	2.1

Table 4.1: Sample statistics for the 390 fingerstick glucose measurements from 28 clinical sessions.

4.1.3 Spectral Characteristics

As can be observed from Figure 4-3, the skin Raman spectra collected from volunteers' forearms exhibit distinct Raman features on top of varying background signals. While the spectral signals contain rich information regarding the tissue composition and environment under examination thanks to the high specificity, label-free, and multi-analyte nature of Raman spectroscopy, nonidealities exist with practical *in vivo* spectral datasets like the one collected in this study. In this part, we discuss in details the spectral characteristics observed in our dataset. Certain aspects of the data acquisition process may be improved experimentally in future trials to alleviate some of the issues seen here, whereas others are more fundamental and have implications for signal detection in general for biological and medical Raman spectroscopy.

Autofluorescence Background

One of the most significant obstacles to many for non-invasive skin Raman analysis and diagnostics, and more broadly for biological and medical Raman spectroscopy, is the strong background signal in Raman spectrum. This has been one of the longest-standing problems in the community where numerous techniques, either instrumentally or computationally, have been proposed yet there is no consensus solution in terms of satisfactory performance and ease of adoption up to date. Instrumental solutions include shifted excitation [138], time-domain gating [139], frequency-domain methods [140], amongst others that are discussed in Wei et al. [141]. Examples of computational removal include Lieber and Mahadevan-Jansen [90], Zhao et al. [91] and the technique discussed in Chapter 3. The implications of the strong background signal add complications in two ways. Firstly, it is extremely difficult to precisely determine the boundary between the background and the Raman peak signals only from information available in the raw spectral data, especially when broad Raman peaks are present. Therefore, any computational evaluation algorithm for the background signal would most likely either underfit or overfit the true underlying signal in certain ways, thereby introducing bias into any following estimation algorithms. Secondly and more importantly, the shot noise from the dominant background can dwarf the already-weak Raman signal from potential biological analyte of interest. This imposes a more fundamental limit on the detectability of analytes in biological tissues and materials. Instrumentation modifications such as time or frequency-domain techniques have shown promises of enhanced SNR with fluorescence suppression [63]. However, a full comparison with state-of-the-art CCD-based Raman spectroscopy systems needs to be performed in order to formally establish their SNR performance, counting factors such as photon detection efficiency and time integration. For the foreseeable future however, the problem caused by the dominant background signal from biological tissues and materials in Rama spectroscopy is likely to persist. As a result, the optimal background rejection or removal technique will continue to be problem or application-specific like the case in the past decade [62].

For skin Raman spectrum, it is generally believed that the origin of the background signal comes from the autofluorescence from intrinsic fluorophores such as NADH, collagen, flavins, and melanin amongst others, which are ubiquitous in biological skins and tissues [60, 61, 142]. Even with near-infrared laser excitation, considerable autofluorescence still exists. With prolonged light exposure, the autofluorescence signal typically experiences an overall reduction in intensity as shown in Figure 4-3. This is usually referred to as autofluorescence photobleaching, which adds additional dynamics to the overall signal on top of any potential changes from physiological attributes. The origin of autofluorescence photobleaching with near-infrared excitation is not yet fully understood and research efforts have been carried out for endogenous fluorophore assignment [142, 143]. It is worth noting that debate still exists on the exact physical origin of the background signal from skin tissue Raman spectra with near-infrared excitation. Bonnier et al. [144] suggested that as opposed to the photochemical root, the background signal from skin tissue Raman spectrum under near-infrared excitation

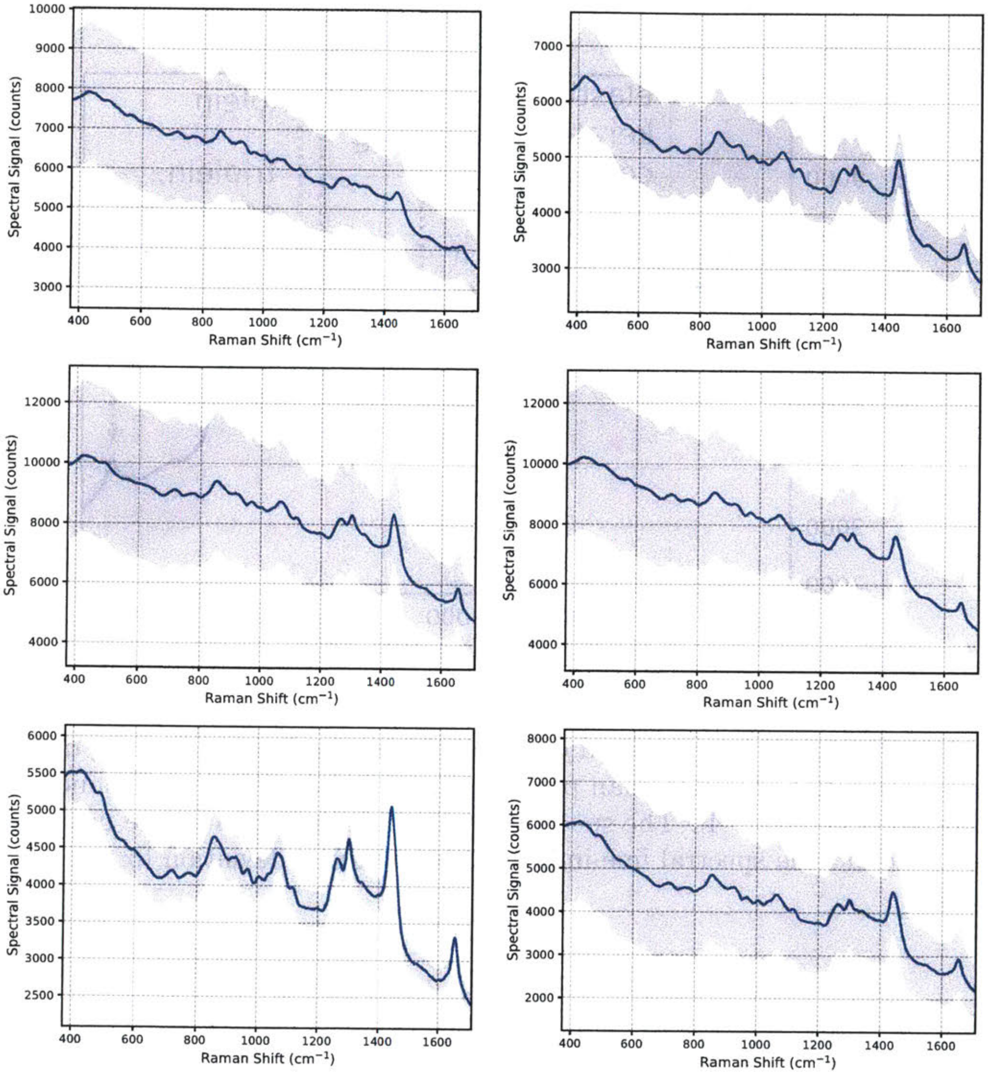


Figure 4-5: The means and standard deviations for the Raman spectral data plotted in Figure 4-3. The spectral line in each subplot indicates the mean and the grey area around the spectral line indicates the standard deviation for each spectral data point.

is caused by scattering due to morphology. Despite the true underlying physical origin, their results highlight the complications associated with the chemical and physical inhomogeneities in biological tissues and materials. In our work, we refer to the background signal in the Raman spectrum as autofluorescence in line with the mainstream opinion in the research community, noting that this assumption of physical origin does not affect our statistical signal analysis and model construction framework in any particular way.

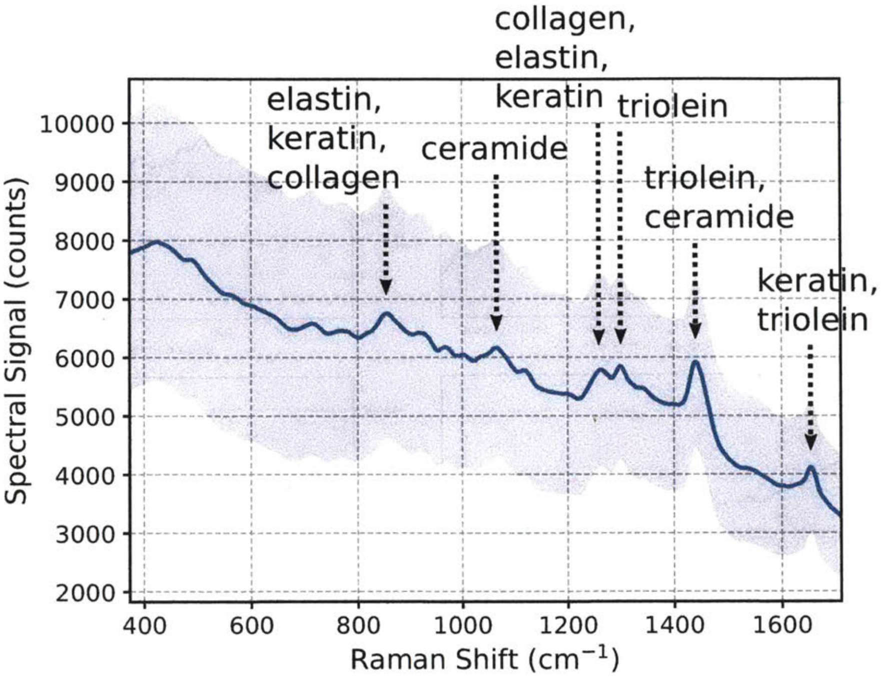


Figure 4-6: The mean and standard deviation for all the Raman spectral data across 28 clinical sessions. Major Raman peaks are associated with their chemical origins according to Feng et al. [4]. The spectral line in the plot indicates the mean and the grey area around the spectral line indicates the standard deviation for each spectral data point.

Figure 4-5 replots data in Figure 4-3 such that the spectral line in each subplot indicates the mean of the collected spectral data and the grey area around the spectral line indicates the standard deviation. The plots show that the degree of the photobleaching effect can vary significantly amongst different volunteers and visits. This is likely due to the disparities for the abundances of the endogenous fluorophores inside the skin tissues for different volunteers or different probing spots. In addition, we plot the total spectral mean and standard deviation across all the 28 sessions in a similar fashion in Figure 4-6. The spectral line in display is a general representation of skin Raman spectrum collected with our instrument. In general, despite the variations associated with the autofluorescence background within each collection session, the Raman peaks in the spectra are visually stable, as can be seen from Figure 4-3. This should be expected as the majority of the Raman peaks come from skin tissue constituents such as collagen, elastin, triolein, nucleus, keratin, and ceramide amongst others [4], which are not expected to change during the two-hour OGTT session. Some of these major Raman peaks are labeled with their chemical origins in Figure 4-6 according to Feng et al. [4]. Although the variations in glucose level are higher in the OGTT experiments compared to normal glucose variation range in such a time window, direct change in the glucose signal is expected to be very small in comparison to the visually discernible Raman peaks in our plots. Related discussions will be presented in more details in Section 4.2.2.

Movement Artifacts

With long-duration *in vivo* experiments, the background signal and its photobleaching can also exhibit time-dependent dynamic behavior due to body movement. This is illustrated in Figure 4-7, where we define spectral signal mean (SSM) as the mean of spectral data points in an acquisition frame. The plots in the left column are examples where the arms of the volunteers were relatively stationary throughout the two-hour sessions. As a result, no major movement artifacts are observed in these plots. On the contrary, the plots in the right column have significant overall background changes due to arm adjustment and movement. The arm movement essentially introduced a new spot under light exposure. Subsequently, a sudden increase in the overall background occurred, followed by restarted photobleaching. During all the clinical visits, our Raman instrument was standing vertical such that volunteers can rest their arms comfortably directly on top of the probe window. Volunteers' arms were also strapped to the instrument to minimize major arm movement. At the beginning of the sessions, volunteers were told to try their best not move their arms during the OGTT experiments. However, minor arm movement and adjustment, which were the cause for these movement artifacts, can be inevitable for certain volunteers depending on their comfortable levels during the two-hour-long experiments.

Experimental improvements can be made to potentially resolve this problem in future trials. A specially designed Raman probe [145] with stable mechanical strap can be used to minimize problems associated with probing spot mis-alignment due

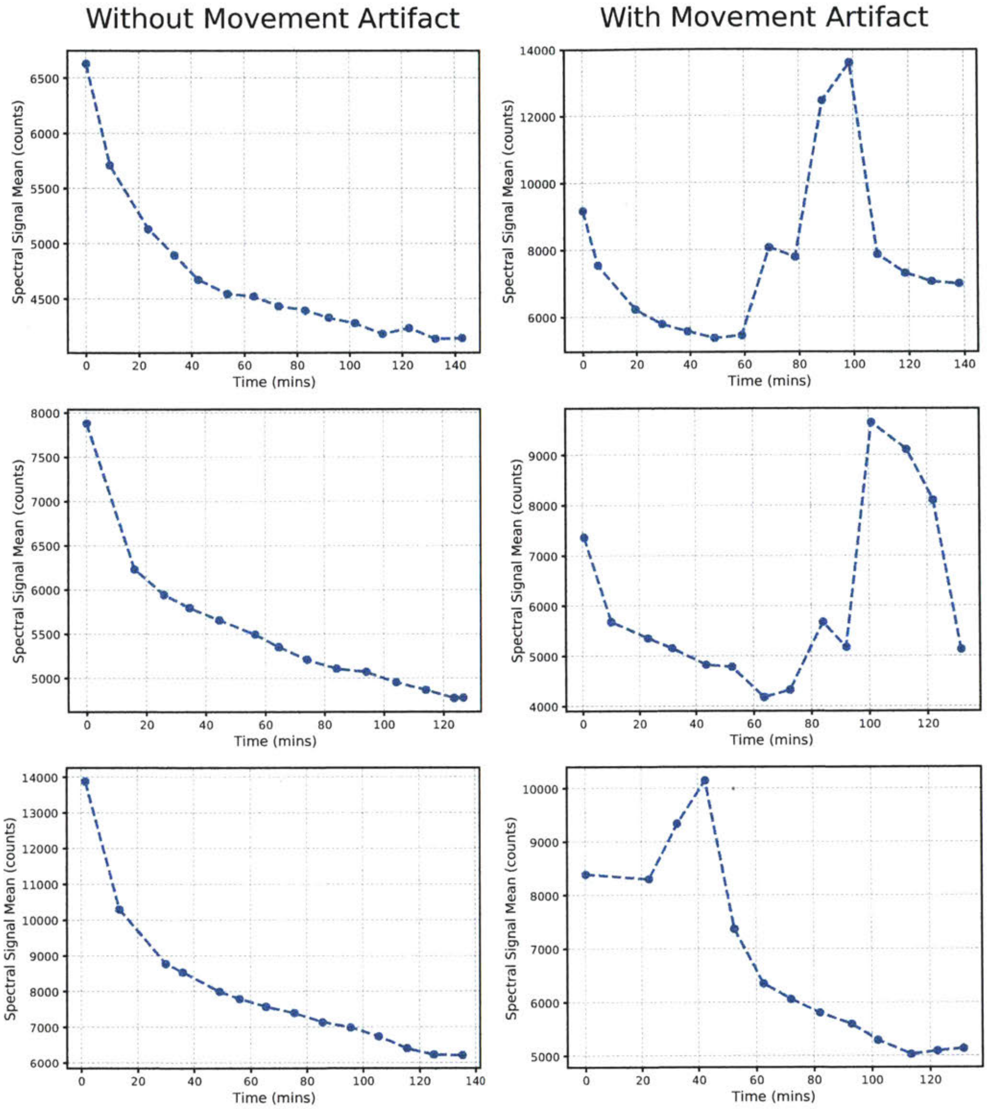


Figure 4-7: The time evolution of spectral signal mean (SSM) for selected clinical sessions with ≈ 10 mins spectral acquisition spacings. (left column) Example data plots showing the autofluorescence photobleaching with no arm movement disruptions. (right column) Example data plots showing the SSM evolution with arm movement artifacts.

to body movement. In addition, extensive light exposure over a large area has been proposed to pre-bleach the skin to improve spectral quality [142, 146]. This has the added benefits of further increasing the spectral SNR with lower overall background shot noise. One potential future experimental direction is therefore to establish a safe, robust, and repeatable skin pre-bleach protocol for future *in vivo* clinical studies.

Ambient Light Leakage

One other issue that is worth bringing up is the ambient light leakage problem during the two-hour-long experiment. In addition to the optical enclosure as shown in Figure 4-2, volunteers' arms were covered with a black optical cloth for ambient light blockage during the experiments. However, for certain visits, spectral bands from the fluorescence light bulb in the room were shown in the collected Raman spectra, roughly on the same intensity order as the main Raman features in the spectra. An example is shown in Figure 4-8. Examinations into the system suggested that this was most likely caused by the non-perfect contact between volunteer's arm and the probe window in Figure 4-2. Volunteers with smaller arm circumferences were more likely to have ambient light leakage with their recording sessions compared to those with larger arm circumferences, which had better contact with the probe window. Overall, out of the 28 clinical sessions, a small fractions of them (around 8) exhibited some degree of ambient light leakage in a subset of the recorded spectra.

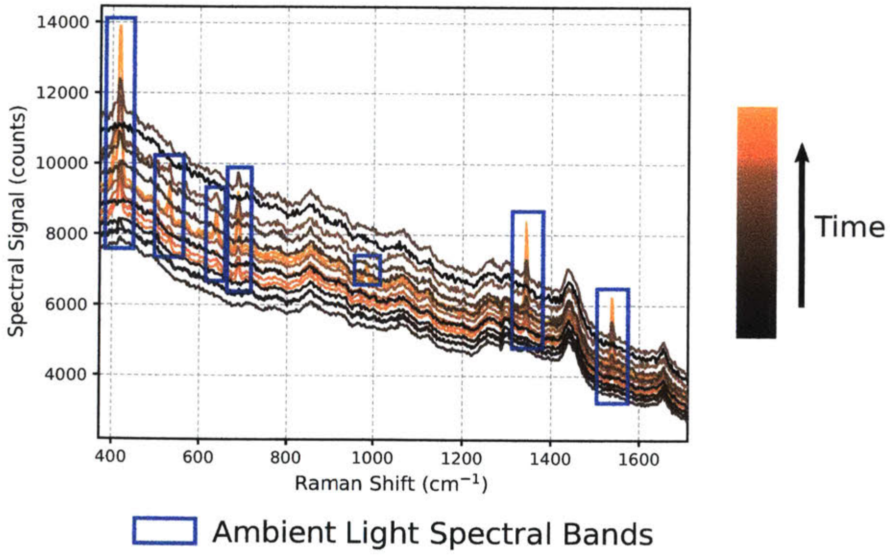


Figure 4-8: Raman spectra from a clinical session where some ambient light leakages are observed in the recorded spectra. The Raman spectra in display were not smoothed with the Savitzky-Golay filter.

In our experimental settings, the ambient light exhibited narrow discrete spectral bands that do not impose much interference to the glucose Raman spectrum (shown in Figure 4-11 (b)). As a result, no special processing was implemented to account for these spectral artifacts. When there is significant ambient light presence from various light sources that may interfere with the analytical results, algorithmic processings can be implemented to improve detection performance [147]. Experimentally, future improvements can be carried out for the optical interface between the Raman instrument and the probing region on volunteer to eliminate the ambient light leakage problem. For example, a smaller probing window or special Raman probe that is tightly strapped onto volunteer's forearm are potential solutions.

4.2 Signal Analysis and Processing Methodology

4.2.1 Spectral Means and Variations

Session	1	2	3	4	5	6	7
Mean (counts)	5101	5658	5309	5159	5630	4035	6072
Standard Deviation (counts)	506	1685	1406	384	859	269	1710
Coefficient of Variation	9.9%	29.8%	26.5%	7.4%	15.3%	6.7%	28.2%
Session	8	9	10	11	12	13	14
Mean (counts)	4732	7904	7537	7765	6389	10547	6805
Standard Deviation (counts)	657	2504	2124	1756	1103	2856	1688
Coefficient of Variation	13.9%	31.7%	28.2%	22.6%	17.3%	27.1%	24.8%
Session	15	16	17	18	19	20	21
Mean (counts)	5433	6270	5152	5950	5454	5662	8032
Standard Deviation (counts)	1082	720	1102	1312	590	896	1954
Coefficient of Variation	19.9%	11.5%	21.4%	22.1%	10.8%	15.8%	24.3%
Session	22	23	24	25	26	27	28
Mean (counts)	7796	5553	4647	4153	5798	5735	4581
Standard Deviation (counts)	2330	786	829	1257	864	1304	465
Coefficient of Variation	29.9%	14.1%	17.8%	30.3%	14.9%	22.7%	10.2%

Table 4.2: The mean, standard deviation, and coefficient of variation for the spectral signal mean (SSM) within each of the 28 clinical visits. The mean of SSM ranges from ≈ 4000 to 10600 counts. The standard deviation of SSM ranges from ≈ 270 to 2860 counts. The coefficient of variation ranges from $\approx 7\%$ to 32%.

With the sources of background signal variations discussed in Section 4.1.3, we first calculated certain statistical quantities associated with the spectral dataset for signal characterization. The calculations are based on Raman spectra taken ≈ 10

mins apart in each session, same as the ones displayed in Figure 4-3. The overall mean signal level for spectra from a clinical session is important as it is directly related to the Poisson shot noise in the spectra, which is the dominant noise source in our measurements. The mean of the spectral signal mean (SSM) for the spectra taken in a clinical session is used to represent this quantity. Meanwhile, the effect of autofluorescence photobleaching and arm movement artifact can be roughly estimated by calculating the overall SSM variation in a session. This can be represented by statistical quantities such as the standard deviation of SSM within the session. Lastly, the relative SSM variation can be characterized by the coefficient of variation of SSM, which is defined as the ratio between its standard deviation and mean. A summary table showing the mean, standard deviation, and coefficient of variation for the SSM within each of the 28 clinical sessions is shown in Table 4.2. Meanwhile, Figure 4-9 shows the sample histograms for these three quantities. Within our dataset, the mean of SSM ranges from ≈ 4000 to 10600 counts. The standard deviation of SSM ranges from ≈ 270 to 2860 counts. The coefficient of variation ranges from $\approx 7\%$ to 32%. These quantities indicate that a substantial change in the overall signal level, mostly due to the background autofluorescence signal change, should be expected from a cohort of volunteers or test subjects with skin Raman spectroscopy.

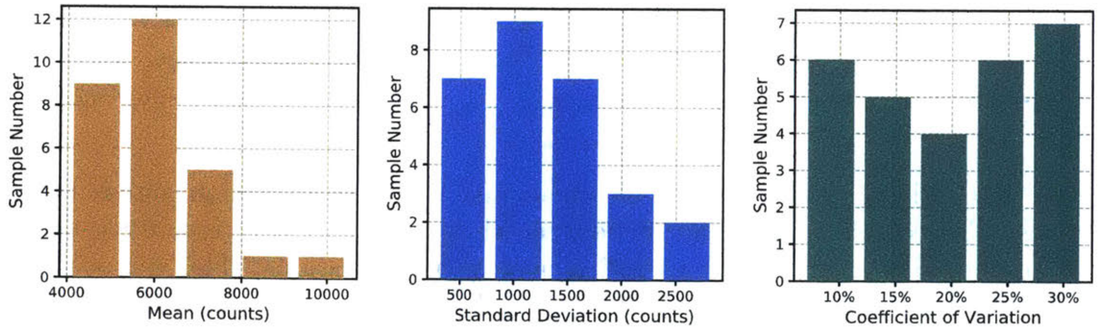


Figure 4-9: Sample histograms for the mean (left), standard deviation (middle), and coefficient of variation (right) of the SSM from all the 28 sessions.

In general, it can be observed that spectral data with higher overall intensity levels experience more photobleaching (which also means more movement artifacts if the arm is not stable throughout the session). This can be visualized in Figure 4-10, where in the left subplot, the SSM means and standard deviations are plotted for all the 28 sessions. The correlation coefficient for these two quantities is ≈ 0.85. In the right subplot of Figure 4-10, the SSM coefficient of variation for all the sessions is plotted. This quantity has a correlation coefficient of ≈ 0.52 with the SSM mean and ≈ 0.88 with the SSM standard deviation. The quantities and their relationships shown in this figure indicate that pre-bleach of the probe area should be an effective way to lower down the overall background signal for skin Raman analysis and diagnostics applications. This is especially true for volunteers with high background levels, as they experienced more background drops (and hence higher SSM variations) with light exposure in our experiments. However, more investigation into this topic needs to be carried to establish a safe operating condition for the pre-bleach protocol with robust, repeatable and quantifiable outcomes.

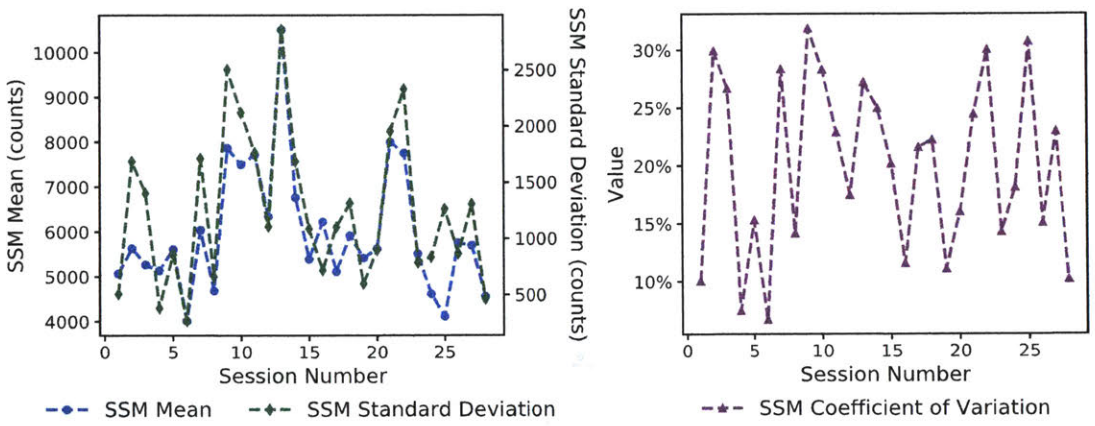


Figure 4-10: (left) Plot for the SSM mean and standard deviation across the 28 sessions. The correlation coefficient between these two quantities is ≈ 0.85. (right) Plot for the SSM coefficient of variation across the 28 sessions. This quantity has a correlation coefficient of ≈ 0.52 with the SSM mean and ≈ 0.88 with the SSM standard deviation.

4.2.2 Noise Analysis and Its Implications

For any optical detection modalities, the fundamental detectability and quantification accuracy depend on the specificity and the SNR of the detection scheme. While the highly specific molecular fingerprint of Raman spectroscopy is often cited for its potential for sensing applications, the signal and noise requirement is usually the bottleneck for Raman sensing due to the fact that often the Raman signal is within orders of the noise limit. For optical detection with CCD or CMOS sensors, various noise sources exist in the signal acquisition process. These include sensor readout noise, signal shot noise, dark current shot noise, sensor fixed pattern noise, and quantization noise amongst others. The electrons generated in the pixels follow Poisson statistics where the variance equals to the expected number of electrons. The electrons are then converted to digital counts through a series of signal amplification, on-chip operations such as binning, and analog-to-digital conversion. In order to estimate and quantify the shot noise from the signal count in the acquired spectra, we first carried out CCD gain calibration for our clinical instrument.

We define the effective gain of the CCD sensor as G, which corresponds to the conversion between electrons generated inside the pixels and the CCD readout in terms of CCD counts. This is in the unit of e^-/count. To obtain the effective gain G, we acquired dark current spectra with various acquisition times. We calculated the mean and variance of the signal counts with these acquisition levels through repeated measurements and generated the mean-variance plot in Figure 4-11 (a). The count mean and variance follow a linear relationship and the slope of the regression curve corresponds to the inverse of the effective gain G [148]. This is so that after converting to electron number, the incremental variance equals to the incremental mean to account for the noise contribution from the shot noise alone, assuming that the rest of the noise sources do not vary with the signal count. From the regression curve, G is calculated as ≈ 7 e^-/count for our CCD sensor [1].

Knowledge of the effective gain G helps translating the CCD count into electron number. This can facilitate noise analysis, as the electron number follows Poisson statistics. From Table 4.2, it can be seen that the mean overall signal level from our clinical Raman spectra ranges from ≈ 4000 to 10600 counts per spectral data point.

[1] Notice that this conversion coefficient accounts for factors such as line binning in our calculation. As a result, it is not directly comparable to the gain obtained from the sensor datasheet.

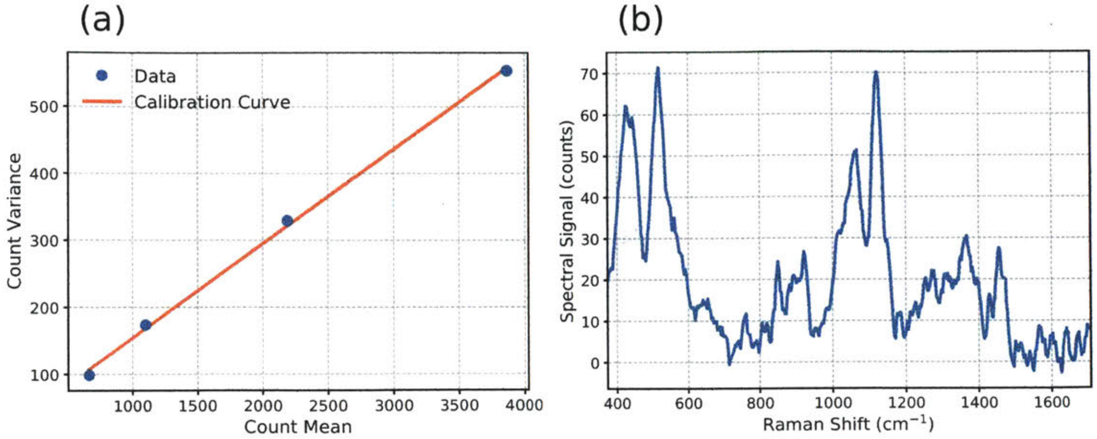


Figure 4-11: (a) Calibration curve for characterizing the effective gain G for electron to digital count conversion coefficient. Counts in this plot were obtained through various integration times with just the dark current. (b) Glucose Raman spectrum measured at 100 mM concentration in water with the clinical instrument. The spectrum in display was obtained through averaging of 50 consecutive spectral acquisitions. Water background was subtracted prior to display. All settings were the same as those in the clinical experiments.

On the other hand, the mean dark current count in our CCD with 30 s integration time is $\approx$ 6300 counts per spectral data point. Counting these two sources as the dominant shot noise origin, this suggests the the shot noise standard deviation ranges from $\approx$ 38 to 49 counts per spectral data point given the effective gain G as calculated above. In the mean time, in Figure 4-11 (b), we plot the Raman spectrum of 100 mM glucose in solution, measured with the same system under the same excitation power and integration time settings. The plot was obtained as the average of 50 consecutive acquisitions with 30 s integration times. The water background was subtracted for display purpose. This gives $\approx$ 70 counts of signal for the main Raman peak of glucose. With our OGTT experiments, the average fingerstick glucose measurement concentration was $\approx$ 7 mM. With a perfect linearity assumption, this corresponds to $\approx$ 5 counts for the main Raman peak, measured in clear solution. Assuming that the *in vivo* Raman measurement is background (which includes both the autofluorescence signal and dark current) shot noise limited, this means that for the average glucose concentration expected to be present in volunteer's body fluid, the expected peak

SNR[2] is at most ≈ 0.10 to 0.13. While peak SNR alone is not enough for determining quantifiability, given the high variabilities observed in skin Raman spectra, this is most likely too low for direct estimation. Averaging consecutive acquisition frames can improve the SNR. However, the benefit is limited due to the fact that SNR only increases proportional to the square root of the overall integration time with shot noise limited measurements. As a result, we do not believe that the current system as it stands is able to perform direct glucose measurement *in vivo* for physiologically relevant glucose levels.

While this Raman instrument is not equipped with the sensitivity required for *in vivo* glucose measurement as indicated by the calculations above, instrumental improvements can be made to increase its light throughput, though this will likely result in a larger overall instrument size with limited portability. For example, spectrometers that are further cooled down with negligible dark currents with larger input slit sizes can be used with modified collection optics for higher light gathering power [149]. Larger slit sizes will require longer optical propagation lengths in order to maintain the resolution target. Alternatively, the coded aperture design discussed in Section 1.2.2 can be used for light throughput improvement without sacrificing instrument size and portability. While potential instrument improvement is outside of the scope for this study, the spectral dataset already collected in the OGTT experiments still presents many opportunities for studying the skin Raman spectral characteristics and the signal requirements and implications for universal and predictive chemometric models for Raman skin analysis and diagnostics applications. This in turn can inform the requirements for future instrument improvement and provide valuable experiment design guidelines for future trials. For many applied chemometric applications, the predictive capability of an algorithm processing pipeline often relies on key factors such as the SNR, training data sample size, interfering conditions, and spectral conditions of the training and test datasets. These factors are often difficult to assess alone without experimental inputs. As a result, the skin Raman spectral dataset with a

[2] Here, the peak SNR definition is different from the SNR defined in Section 2.4.1 for ease of analysis in this context.

moderate sample size collected in this study is highly desirable for such evaluation and analysis tasks. In particular, given the large variabilities, overall signal changes both within-session and cross-session, and internal structures in skin Raman spectra as discussed earlier in this chapter, the glucose signal requirement for a universal and predictive chemometric model remains to be a key open question that needs to be addressed. As a result, this will be the main objective for the statistical data analysis in this chapter.

4.2.3 Signal Generation, Validation and Test Scheme, and Algorithm Pipeline

Signal Generation Approach

To numerically investigate the glucose signal requirement for a universal and predictive chemometric model with skin Raman spectra, we adopted a signal generation procedure where glucose signal was manually added to the raw skin Raman spectra from the clinical studies. With numerical control over signal generation, the signal strength of the glucose spectral signal can be varied such that quantitative conclusions can be drawn in terms of the relationship between prediction accuracy and glucose signal strength amongst other related issues. For signal generation, we first interpolated the fingerstick glucose measurements down to 30 s acquisition spacings, such that all Raman spectra acquired during a session (which all had 30 s integration times and acquisition spacings with adjacent frames) have corresponding glucose measurements in time vicinities of less than 30 s. Afterwards, for each acquired raw Raman spectrum, a glucose Raman signal was determined. The shape of the glucose Raman signal was obtained from the glucose solution measurement shown in Figure 4-11 (b). The amplitude of the glucose Raman signal was determined by the corresponding original or interpolated fingerstick glucose measurement at the same acquisition time and a multiplicative amplification factor. This means that for a fingerstick glucose measurement level of c mM and an amplification factor of A, the added glucose signal strength is equivalent to Ac mM concentration glucose in clear solution, assuming perfect linearity. Before the generated glucose signal was added to the skin Raman spectrum, Poisson noise was added to it for every spectral data point in the electron number domain based on the effective gain G calculated from Section 4.2.2. In addition, quantization into discrete digital counts was performed to resemble the actual signal acquisition levels.

Several intrinsic assumptions were made in the above process. Firstly, perfect linearity was assumed in terms of glucose signal strength and its concentration. This is in general a reasonable assumption in most situations. However, when the concentration variation is large, nonlinear signal changes introduced by factors such as refractive index variations can introduce a non-negligible effect. Secondly, interpolation was used to obtain virtual fingerstick measurements such that all Raman spectra have corresponding glucose concentration labels. This is so that the effect of averaging consecutive acquisition frames can be numerically explored. Lastly and most importantly, light absorption and scattering due to the turbidity of skin tissue were neglected with this treatment. The chemical and physical inhomogeneities inside skin tissues can introduce significant perturbations to photon transport in comparison to the case of clear solution measurements [66]. This may require spectral signal correction and calibration for quantitative analysis, with the help from additional optical measurements such as reflectance measurement or Monte Carlo light transport simulations [150, 151, 152]. While it is possible to take the turbidity-induced photon transport phenomenon from skin tissues into our signal generation process, this would require assumptions of the absorption and scattering properties for the skin tissues from the volunteers, which adds complications to our model and may cloud judgment over our investigation purpose. As a result, this is not pursued in the current work.

With these assumptions, we note that the actual signal requirement for predictive chemometric modeling is likely higher than the results obtained from our numerical experiments. Nonetheless, our investigation should provide a reasonable lower-bound check (in terms of the signal requirement) for predictive analysis. With further experimental modifications and investigations, additional modeling attributes can be incorporated to more fully account for the physical and biochemical conditions and environments of the measurements. For example, tissue phantom with known optical properties can be studied with and without glucose to cross compare the natural signal generation process and the numerical approach presented in this study [152]. This can help determine any potential signal correction function for analysis involving more detailed physical modeling.

Training, Validation, and Test

In supervised machine learning, the process of model training, selection, and evaluation typically requires partition of the available dataset into three subsets, the training set, validation set, and the test set [32]. The training and validation set are used to construct the training model and perform hyperparameter optimization. The test set is generally used to evaluate the model performance. With a finite dataset sample size, techniques such as holdout set, cross-validation (CV), and bootstrapping can be used in model validation and test [153]. While this process is a standard practice in machine learning and applied statistics research, it has not always been followed by practitioners in the applied spectroscopy community [83]. Due to the high dimensionality nature of spectral arrays, generally limited sample sizes, and potential internal correlation structures common in spectral datasets, extra caution must be paid in designing the validation and test scheme for spectral data analysis. Otherwise overfitting or spurious correlations caused by potential confounding factors can lead to overly optimistic results. This is especially true if model training, validation and test are performed with a single CV scheme, where hyperparameter selection and model performance evaluation are both performed within this single CV iteration [154]. We note here that unfortunately this approach has a rather regular appearance in the applied spectroscopy literature.

In our numerical study, all spectral data with actual fingerstick glucose measurements were used in model training, validation and test. This corresponds to 390 spectra from 28 independent sessions. A nested CV scheme has been used for model training and evaluation. The inner CV was used for algorithm pipeline preprocessing, training and hyperparameter selection. This was 10-fold CV in our study. The outer CV was used for comparing Raman-predicted glucose levels and the fingerstick measurements. This was used to obtain the generalized prediction accuracy for our processing pipeline. In our approach, as opposed to a volunteer-specific training schedule [1], spectral data from different sessions or volunteers were used together for model construction. There are several reasons for this choice of training. Firstly, such a universal training approach is based on the fact that skin autofluorescence and Raman signals are from common tissue and cellular constituents shared by all volunteers. As a result, a universal model with enough training data should be able to capture the sample subspace necessary for glucose prediction. In addition, in practical scenarios, measurement schemes that require constant volunteer-specific training or calibration is likely unfavorable in comparison to those that are universally trained or calibrated. Secondly, combining data from different sessions or volunteers allows a much larger pool of training data to be constructed. In general, estimator variance is expected to decrease as the training sample size increases [155]. This allows most training algorithms to generalize better and have higher prediction accuracy. This also enables more efficient usage of the data from a cost perspective, as collecting medical spectroscopy data is generally resource-intensive. Lastly, using all the spectral data across sessions allows session-wise test schemes that minimize confounding factors such as the internal correlation structures within spectral data collected in a single session. Session-wise test scheme is also the only way to measure the true predictive performance of the model. This aspect will be discussed next in details.

Test Scheme

Within the glucose Raman sensing community, leave-one-out-cross-validation (LOOCV) has been the dominant test scheme with OGTT-like experiments. However, there are subtle but important differences under the common terminology. We summarize them as follows.

Local Test Scheme [1]: Spectral calibration models are built completely from spectral data collected in a single clinical session. LOOCV in this case means leaving out one Raman spectrum per test from the current clinical session.

Global Test Scheme [1]: Spectral calibration models are built from spectral data collected from all clinical sessions. LOOCV in this case means leaving out one Raman spectrum per test from the global spectral set containing spectra from all clinical sessions.

Session-Wise Test Scheme [2]: Spectral calibration models are built from spectral data collected from all clinical sessions. LOOCV in this case refers to leaving out Raman spectra from an entire clinical session from the global spectral set containing spectra from all clinical sessions.

For spectral datasets like the one collected in this study, care must be paid in terms of recognizing the internal correlation structure of the spectral signal. For our dataset, this can be observed in Figure 4-3, where Raman spectra from a single session exhibit higher internal spectral correlations compared to spectral correlations amongst data across different sessions. This is due to the fact that other than the autofluorescence background, the vast portion of the spectral signal does not change during a clinical session. For many machine learning and chemometric algorithms, the derivations are based on the assumption that samples are independently and identically drawn from certain underlying probability distributions. When the overall sample size is small as with the case in our study, confounding factors such as the internal spectral correlations within the same session can potentially lead to spurious correlations if spectral samples from the same session end up in both the training/validation set and the test set. As a result, we believe that session-wise test scheme should be the most rigorous test scheme for measuring the predictive capability of any processing algorithms or pipelines in these situations.

In our work, the outer CV scheme, or the test scheme, was performed with leave-one-session-out CV (LOSOCV). In this case, for each outer CV iteration, the spectra and glucose measurements in the test set were from an independent single session that was not used in the training and validation stage. Completion of the outer CV loop means that all sessions were left out for test exactly once. The overall test accuracy performance could then be obtained. This score was subsequently used for comparing various signal conditions and processing variations.

Preprocessing and Algorithm Pipeline

With the model training, validation, and test scheme discussed above, we developed a preprocessing and algorithm pipeline with partial least squares regression (PLSR). PLSR has been the most widely used regression algorithm in the chemometrics community due to its effectiveness, simplicity, and computational efficiency [36]. It is often the default choice for regression applications in spectral data analysis. For spectral preprocessing, a fixed number of acquisition frames was first determined over which a median filter along the time domain was applied for all spectral data points to remove spectral spikes caused by cosmic ray events and also to enhance the SNR of the spectra. Afterwards, an 11-point Savitzky-Golay filter with a polynomial order of 2 was applied across the spectral dimension to further smoothen the spectral signal. As mentioned earlier, only preprocessed Raman spectra that have corresponding experimental fingerstick measurements were used in the model for training, validation and test.

For training and validation, a grid search for the number of optimal loading vectors in the PLSR was performed with the inner 10-fold CV scheme. During this stage, spectral training data were first scaled to have zero mean and unit variance prior to PLS decomposition. The average of the validation mean squared errors for the 10-fold CV loop was used for loading vector number optimization. Subsequently, all data in the training and validation set were used to retrain the PLSR model with the optimal loading vector number. This retrained model was then used to predict the glucose levels in the left out session with the outer LOSOCV scheme. The process continued until all the sessions were tested exactly once.

4.3 Investigations of the Signal Requirement on Universal Predictive Glucose Estimation

With the strategies laid out in Section 4.2.3, we numerically investigate various aspects of the signal requirement on universal predictive glucose estimation accuracy with

Raman spectroscopy for our dataset in this section. Many areas related to the required signal strength for predictive analysis and its implications on future instrumental and experimental design are explored and discussed in details. The results obtained in this section can serve as comprehensive guidelines on the predictive modeling aspect of non-invasive skin Raman sensing for future studies.

4.3.1 Overall Performance of Predictive Modeling with Amplified Glucose Signals

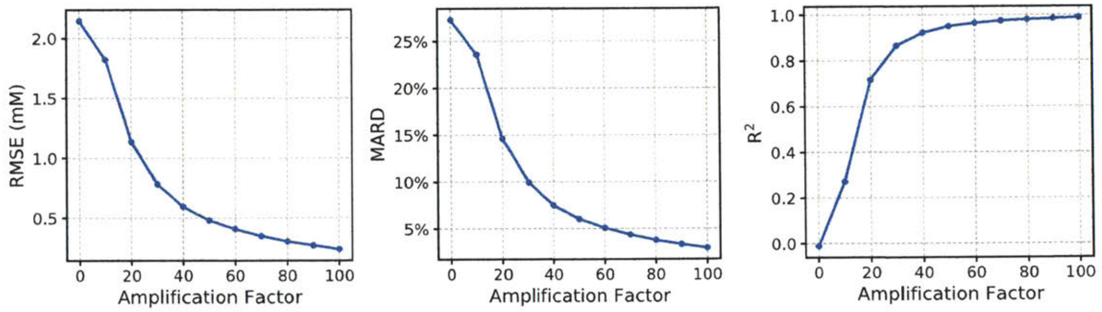


Figure 4-12: RMSE, MARD, and R^2 plots for the overall glucose estimation accuracy with varying glucose signal amplification factors and LOSOCV predictive test scheme.

The effect of glucose signal amplification on prediction accuracy under the LOSOCV test scheme was first investigated. As defined in Section 4.2.3, the amplification factor is the multiplicative coefficient for glucose signal enhancement. The amplification factor was varied from 0 (which corresponded to the original data) to 100 in steps of 10. The size of the median filter for spectral smoothing was chosen as 1 for single acquisition analysis. Figure 4-12 shows the overall root mean squared error (RMSE), mean absolute relative difference (MARD), and R^2 change as a function of the amplification factor under the LOSOCV test scheme. Table 4.3 shows the corresponding values. MARD is defined as the mean of the ratio between the relative estimation errors and the actual measurement values. At an amplification factor of 0, which corresponded to the case for the original dataset, no predictive behavior was observed. This is in line with our noise analysis as discussed in Section 4.2.2. Higher accuracies were obtained with larger amplification factors, with the most accuracy improvement across the ≈ 10 to 20 amplification factor range. The accuracy improvement shows diminishing returns as the amplification factor increases further. Smith [82] suggested that for a non-invasive glucose estimation technique, a MARD of 15% and below would be considered promising[3]. This corresponds to an amplification factor of ≈ 20 with our results. With the noise analysis in Section 4.2.2, an amplification factor of 20 corresponds to an average peak SNR of ≈ 2.0 to 2.6 with our measurements. This should be interpreted as the minimum peak SNR requirement in order for our system to produce a competitive and predictive result with universal modeling under all the assumptions discussed in Section 4.2.3. For our measurements to be comparable to standard strip test systems such as the fingerstick measurements, the MARD needs to be within 5 to 10% [82]. This corresponds to an amplification factor of at least ≈ 30 with an average peak SNR of ≈ 3.0 to 3.9.

Amplification Factor	0	10	20	30	40	50	60	70	80	90	100
RMSE (mM)	2.14	1.82	1.14	0.78	0.59	0.48	0.41	0.35	0.30	0.27	0.24
MARD	27.3%	23.6%	14.6%	9.9%	7.5%	6.0%	5.1%	4.4%	3.8%	3.3%	3.0%
R^2	-0.01	0.27	0.72	0.87	0.92	0.95	0.96	0.97	0.98	0.98	0.99

Table 4.3: RMSE, MARD, and R^2 for the overall glucose estimation accuracy with varying glucose signal amplification factors and LOSOCV predictive test scheme. Data are also plotted in Figure 4-12.

The Clarke Error Grid Analysis was developed in 1987 and has since been used extensively by researchers, medical practitioners, and blood glucose monitoring manufacturers for accuracy evaluation based on different risk levels associated with the instrument estimations [156]. The Clarke Error Grid Analysis plots for amplification factors of 0, 10, 20, and 30 with our processing under the LOSOCV test scheme are shown in Figure 4-13. The corresponding region assignment summary is shown in Table 4.4. While for all the tested cases, more than $\approx 97\%$ of the data points fall inside the A and B regions, and no data fall into the E region, care must be paid in interpreting the results. This is because the Clarke Error Grid Analysis alone can lead

[3] With respect to a robust reference measurement such as the YSI measurement. Here, we make the assumption that our experimental glucose reference measurements are accurate in our signal generation approach for simplicity purpose.

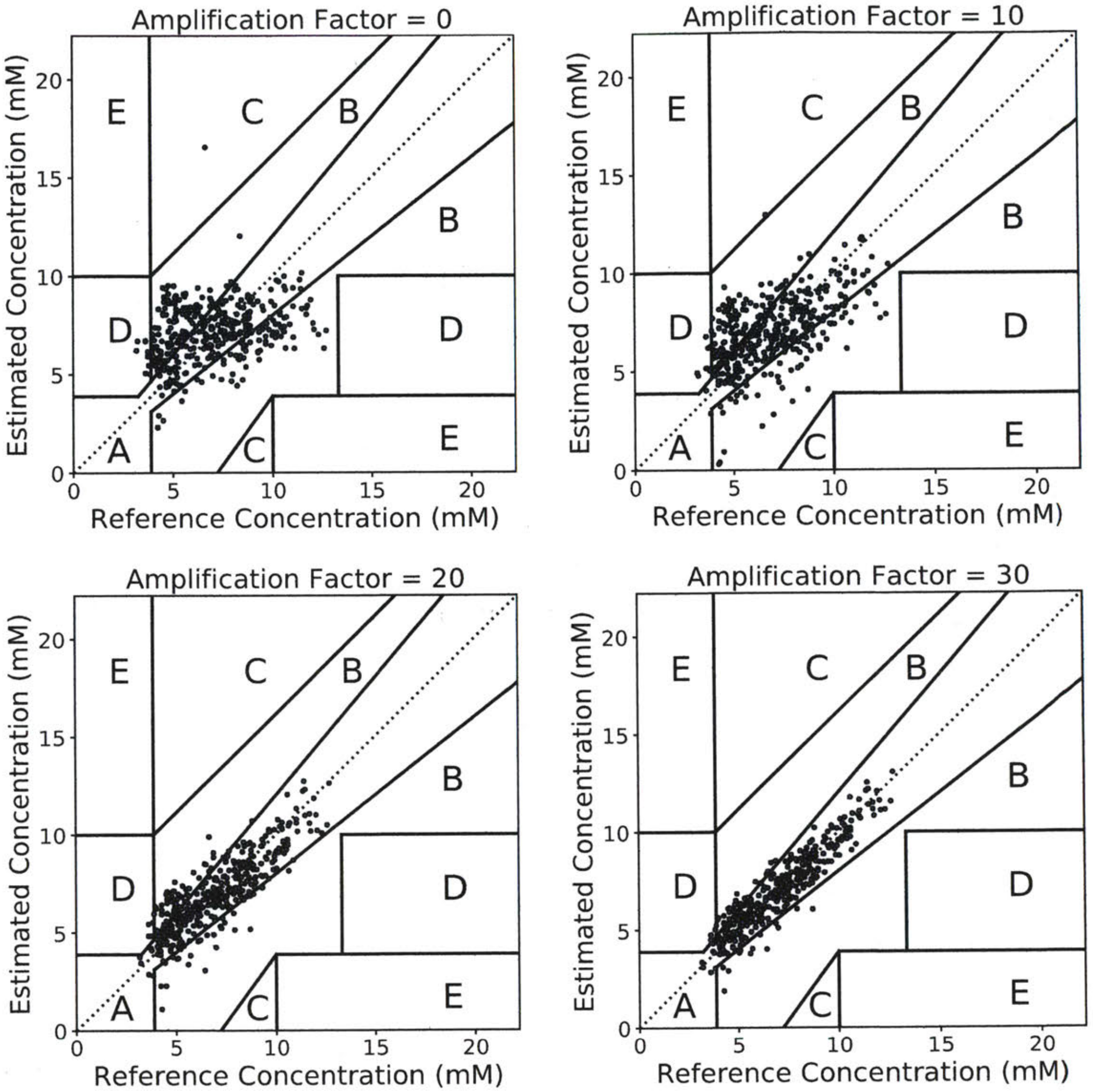


Figure 4-13: Clarke Error Grid Analysis plots for glucose estimations with amplification factors of 0, 10, 20, and 30 under the LOSOCV test scheme. The region assignment for each plot is shown in Table 4.4.

to overly-promising conclusions for cases where the glucose variation range is relatively limited. An example of such situation is the OGTT experiment with healthy volunteers. In general, the target for a traditional glucose meter is to have ≈ 98% data in the A and B regions and to have less than ≈ 0.1% in the E region [82]. This requirement is met with an amplification factor of 20 from the Clarke Error Grid Analysis, and is consistent with the conclusion obtained with the MARD evaluation. An amplification factor of 30 results in A region assignment to be ≈ 88% of the overall data number. This can be considered as a strong estimation performance. Taking a closer look, the estimations are especially accurate with glucose reference measurements on the higher range in our dataset. This is most likely due to the stronger glucose signals in these situations. Other than the Clarke Error Grid Analysis presented here, the Consensus Error Grid Analysis [157] has gained increasing adoptions from blood glucose monitoring manufacturers over the last decade. Similar region assignment and analysis can be performed with the Consensus Error Grid Analysis, which is omitted here.

Amplification Factor	0		10		20		30	
Region	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
A	183	46.9%	217	55.6%	298	76.4%	343	87.9%
B	195	50.0%	163	41.8%	87	22.3%	43	11.0%
C	1	0.3%	1	0.3%	0	0%	0	0%
D	11	2.8%	9	2.3%	5	1.3%	4	1.0%
E	0	0%	0	0%	0	0%	0	0%

Table 4.4: Region assignments for the Clarke Error Grid Analysis plots shown in Figure 4-13 for amplification factors of 0, 10, 20, and 30.

For linear chemometric models like PLSR, many authors have proposed the usage of regression vector check for confirming the correlation source [158]. In linear models, the regression vector is the trained product vector, where the prediction value is typically produced from the inner product of the test spectrum and the regression vector (with possible additional scaling and offset adjustments). In general, the regression vector contains strong signatures from the target component spectrum and therefore can be used to confirm the correlation source. For example, with classical least squares (CLS) models [33], the regression vector is a linear combination of the model bases or mixture component spectra, which contains all the potential target spectral signatures. The regression vectors from our PLSR processing pipeline with amplification factors of 0, 10, 20, and 30 are plotted in Figure 4-14. As can be observed from the figure, other than the zero amplification case, strong glucose signatures are observed in the regression vectors, with more distinct features in more amplified situations. This is achieved despite the fact that our training data consist of highly disparate spectral

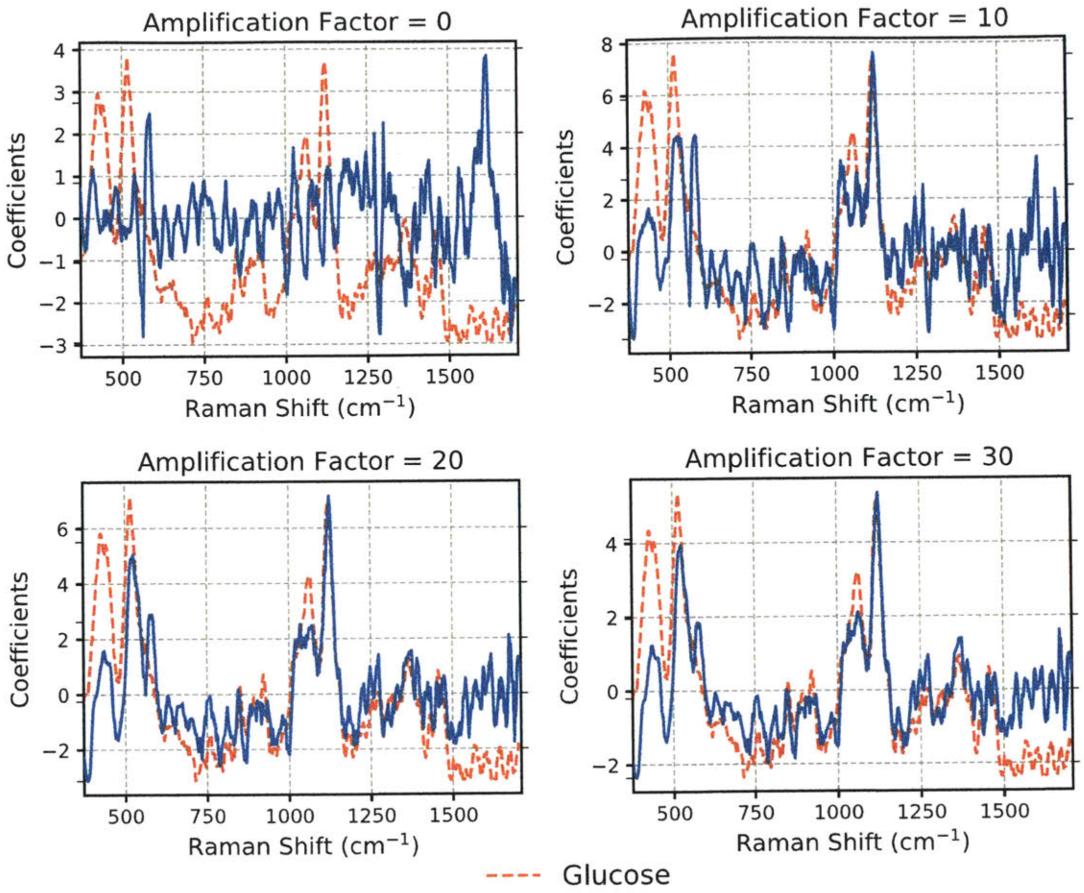


Figure 4-14: Regression vectors obtained from the PLSR algorithm for the universal predictive model with amplification factors of 0, 10, 20, and 30. The glucose Raman spectrum is also shown in the plots. The correlation coefficients between the glucose Raman spectrum and the regression vectors in the plots are shown in Table 4.5.

shapes from different volunteers, as evidenced in Figure 4-3. For a quantitative measure, Table 4.5 shows the correlation coefficients between the glucose Raman spectrum and the regression vectors plotted in Figure 4-14. A substantial increase in correlation coefficient is observed across the amplification factor improvement from 0 to 10. This suggests that the glucose signal is the main source for the improved prediction. This feature is similar to the one reported in Shih et al. [3], where all spectral data were obtained from a single-session *in vivo* experiment with a dog model with an undisclosed laser excitation power level. On the other hand, as discussed in Section 1.4.3, in one of the earliest literatures on *in vivo* Raman glucose sensing with human subjects, only negligible to weak resemblances between the glucose Raman spectrum and the regression vectors were observed, yet strong R^2 values were reported in a single CV scheme [1]. It is therefore not clear whether the glucose signal was the true correlation source and whether the CV and test scheme were used properly in the study. Our results show that despite the large differences observed in skin Raman spectra from different subjects, with detectable source signals even as low as having an overall predictive R^2 of ≈ 0.27 for the case with an amplification factor of 10, a universal calibration model should show *strong* features of the source spectral signal in the regression vector. This should be a routine check for verifying the correlation source for future related studies.

Amplification Factor	0	10	20	30
Correlation Coefficient	-0.08	0.61	0.70	0.72

Table 4.5: Correlation coefficients between the glucose Raman spectrum and the regression vectors in PLSR for amplification factors of 0, 10, 20, and 30.

4.3.2 Individual Session Performance

We further examined the performance of individual sessions under the LOSOCV scheme. Figure 4-15 to Figure 4-18 plot the fingerstick glucose measurements and estimated glucose concentrations from Raman spectra for all the 28 sessions for amplification factors from 0 to 30 in steps of 10. Meanwhile, Table 4.6 to Table 4.9 summarize the corresponding RMSE, MARD, and R^2 for each session under these situations. Session 2, 4, 5, 6, and 28 were for visits where the volunteer drank only water without any glucose. As a result, the glucose measurements stayed at the fasting levels throughout the entire experiment for these sessions.

For an amplification factor of 0, which corresponded to the original dataset, only 5 sessions had a MARD value of less than 15%. These are Session 1, 9, 10, 15, and 22. Other sessions such as 13 and 21 exhibited reasonable glucose trend match with some overall measurement offsets. Examinations into the spectral shape and quality or the glucose tolerance response from these sessions do not reflect any particular

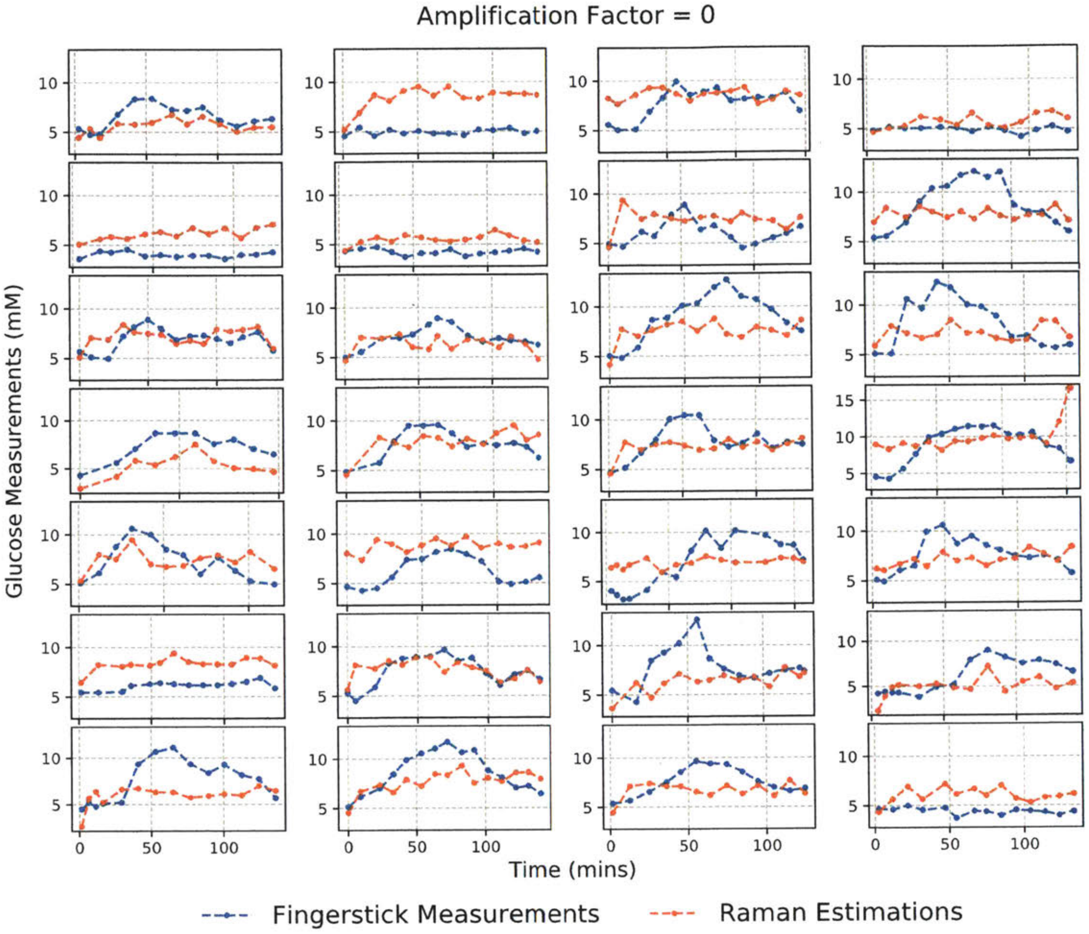


Figure 4-15: Fingerstick glucose measurements and estimated glucose concentrations from Raman spectra for all the 28 sessions with an amplification factor of 0 and the LOSOCV test scheme. The first row is for Session 1 to 4, the second row is for Session 5 to 8, and the rest follows similar orders.

noticeable characteristics. Since these sessions only represent a small fraction of the overall session number, we believe that this is mostly chance correlation and should not be over-interpreted. It is worth mentioning that some have speculated that for many positive correlations observed with non-invasive glucose detection techniques under the OGTT experiments, indirect physiological changes associated with the OGTT experiments are the true sources for the positive correlations [82]. In our case, however, no such correlations were captured with our PLSR processing pipeline under the nested CV scheme and the LOSOCV test method. While it is true that

Session	1	2	3	4	5	6	7
RMSE (mM)	1.22	3.63	1.64	1.06	2.18	1.39	2.05
MARD	14.6%	70.5%	20.2%	17.4%	52.4%	30.5%	30.9%
R^2	-0.10	-171.60	-0.29	-14.31	-63.46	-24.29	-1.96
Session	8	9	10	11	12	13	14
RMSE (mM)	2.58	1.04	1.31	2.61	2.72	2.09	1.44
MARD	23.3%	14.0%	13.6%	24.6%	28.6%	27.5%	16.4%
R^2	-0.24	0.09	-0.59	-0.21	-0.23	-1.31	-0.08
Session	15	16	17	18	19	20	21
RMSE (mM)	1.62	3.33	1.69	2.93	2.32	1.72	2.25
MARD	14.1%	35.2%	21.2%	49.3%	40.0%	18.3%	36.3%
R^2	0.05	-0.94	0.16	-3.06	0.16	-0.08	-30.42
Session	22	23	24	25	26	27	28
RMSE (mM)	1.26	2.51	1.92	2.56	1.96	1.71	1.81
MARD	12.4%	22.6%	24.9%	26.9%	18.2%	18.3%	38.5%
R^2	0.30	-0.67	-0.19	-0.31	0.05	-0.60	-30.67

Table 4.6: RMSE, MARD, and R^2 for all the 28 sessions with a glucose signal amplification factor of 0 and the LOSOCV predictive test scheme. Session 2, 4, 5, 6, and 28 were for visits where the volunteer drank only water without any glucose.

Session	1	2	3	4	5	6	7
RMSE (mM)	0.91	2.97	1.10	0.86	2.48	1.42	1.71
MARD	10.2%	56.5%	11.5%	14.2%	58.2%	30.6%	26.4%
R^2	0.39	-114.48	0.42	-8.94	-83.02	-25.27	-1.05
Session	8	9	10	11	12	13	14
RMSE (mM)	1.81	1.45	1.44	1.61	2.19	2.61	1.89
MARD	16.2%	16.3%	18.2%	17.8%	21.1%	32.8%	20.0%
R^2	0.40	-0.78	-0.91	0.54	0.20	-2.60	-0.85
Session	15	16	17	18	19	20	21
RMSE (mM)	1.39	2.46	1.48	1.96	1.55	1.67	1.33
MARD	13.3%	27.0%	19.4%	32.4%	23.9%	19.0%	19.8%
R^2	0.30	-0.06	0.35	-0.82	0.62	-0.02	-9.92
Session	22	23	24	25	26	27	28
RMSE (mM)	1.22	1.85	2.33	2.53	2.00	0.99	1.47
MARD	14.3%	18.2%	30.4%	28.7%	22.8%	12.4%	31.5%
R^2	0.34	0.10	-0.76	-0.27	0.01	0.46	-19.75

Table 4.7: RMSE, MARD, and R^2 for all the 28 sessions with a glucose signal amplification factor of 10 and the LOSOCV predictive test scheme. Session 2, 4, 5, 6, and 28 were for visits where the volunteer drank only water without any glucose.

spurious correlations can exist especially for within-session measurements, in general only predictive test matters when evaluating the performance of any glucose detection

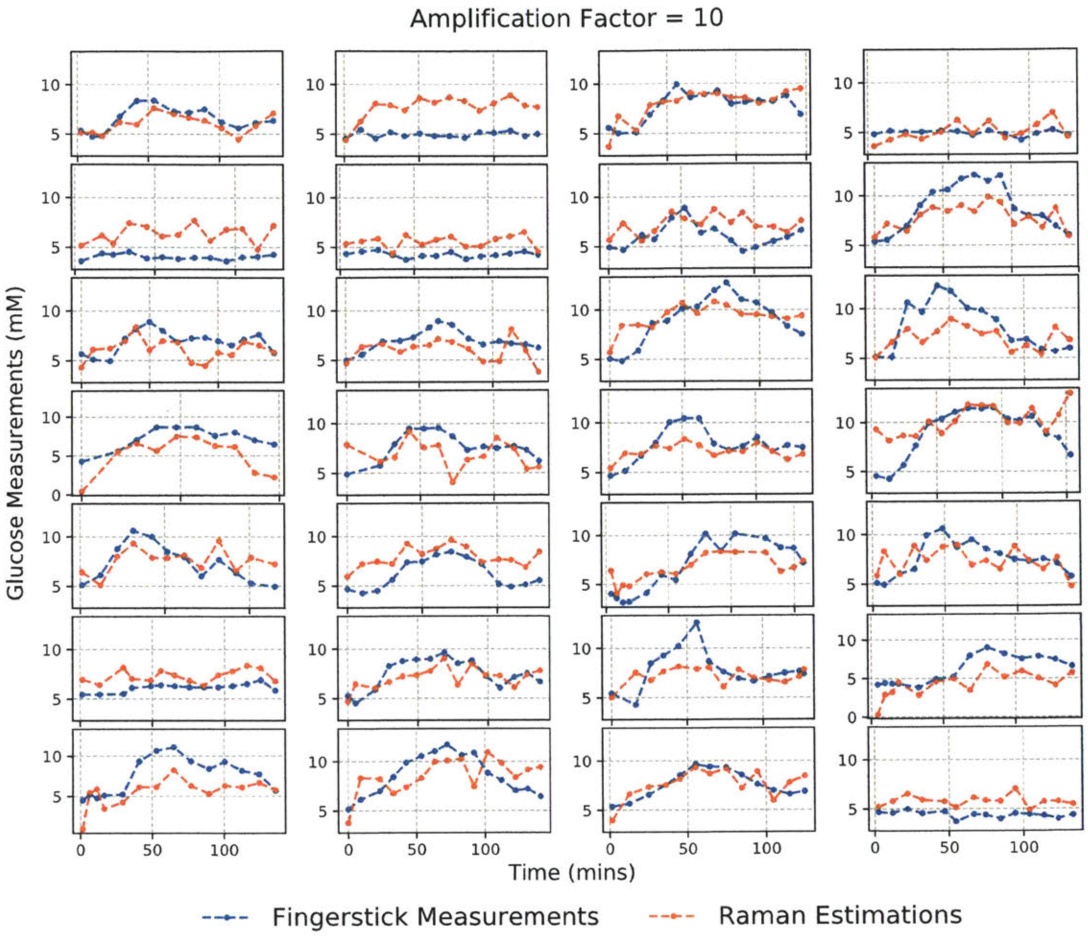


Figure 4-16: Fingerstick glucose measurements and estimated glucose concentrations from Raman spectra for all the 28 sessions with an amplification factor of 10 and the LOSOCV test scheme. The first row is for Session 1 to 4, the second row is for Session 5 to 8, and the rest follows similar orders.

or estimation technology. As a result, one should always be mostly concerned about predictive testing results as opposed to over-interpreting better non-predictive testing results, such as those under the local test scheme and global test scheme discussed in Section 4.2.3, through excessive modeling tuning and parameter fitting. A robust training, validation, and test scheme like ours should be able to differentiate the effect of truly detectable and quantifiable signals from potential confounding factors.

An amplification factor of 10 only improved the overall MARD from 27.3% to 23.6% and R^2 from -0.01 to 0.27 as previously discussed in Table 4.3. For session-wise

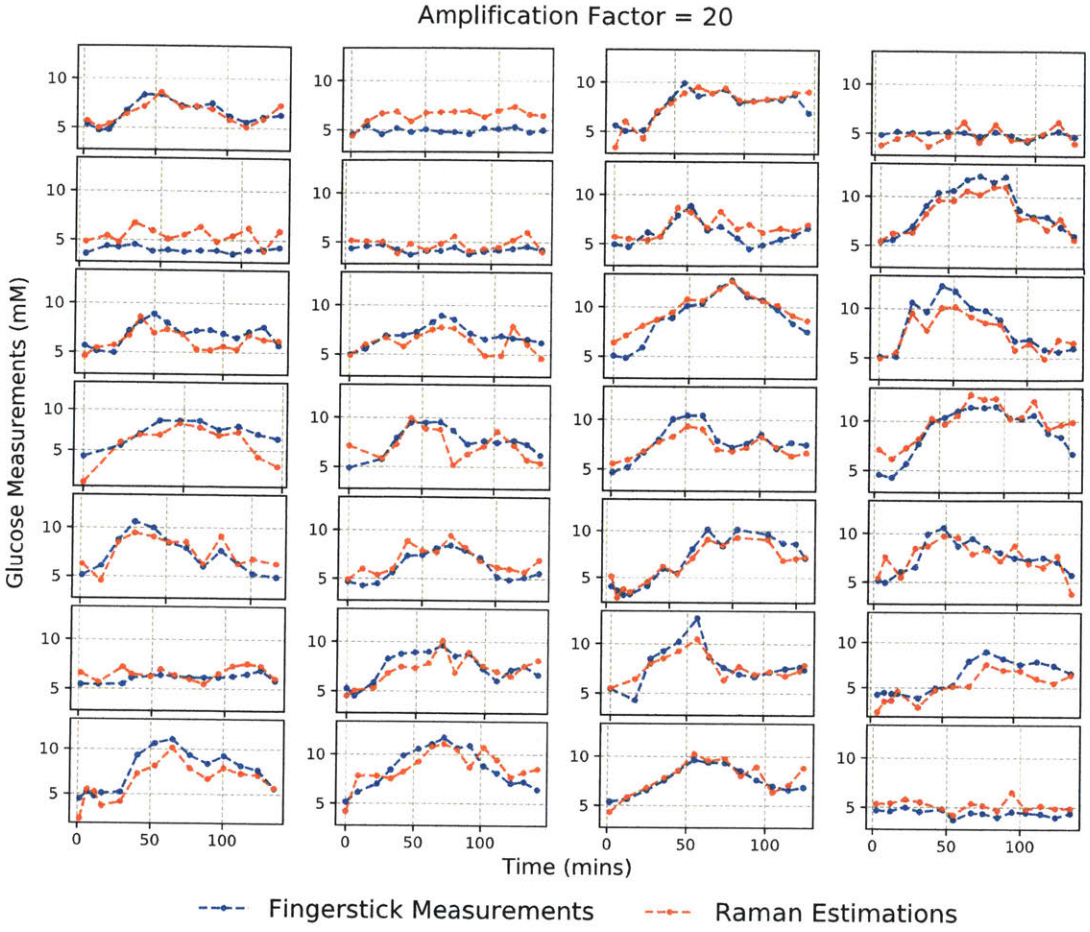


Figure 4-17: Fingerstick glucose measurements and estimated glucose concentrations from Raman spectra for all the 28 sessions with an amplification factor of 20 and the LOSOCV test scheme. The first row is for Session 1 to 4, the second row is for Session 5 to 8, and the rest follows similar orders.

performance, only 6 sessions had a MARD value of less than 15%. While the signal strength in this case was strong enough to be picked up in the regression vector, it was not quantifiable for reliable estimations. As the amplification factor increased to 20, 19 sessions had a MARD value of 15% or less as shown in Table 4.8, which represented $\approx$ 68% of the overall session number. Most sessions exhibited high correlations between the fingerstick measurements and the estimated values at this amplification level. This can be seen in Figure 4-17. At last, an amplification factor of 30, shown in Figure 4-18 and Table 4.9, illustrates accuracies close to the standard strip test

Session	1	2	3	4	5	6	7
RMSE (mM)	0.55	1.68	0.96	0.76	1.61	0.72	1.05
MARD	7.0%	31.9%	9.7%	13.2%	36.6%	14.0%	15.8%
R^2	0.78	-36.22	0.56	-6.85	-34.34	-5.71	0.23
Session	8	9	10	11	12	13	14
RMSE (mM)	0.93	1.14	1.07	1.02	1.16	1.85	1.39
MARD	8.9%	13.7%	13.2%	11.4%	11.8%	22.7%	14.8%
R^2	0.84	-0.09	-0.07	0.81	0.78	-0.81	-0.01
Session	15	16	17	18	19	20	21
RMSE (mM)	0.90	1.39	1.04	0.93	0.87	1.28	0.75
MARD	9.7%	16.2%	13.4%	14.2%	11.0%	15.9%	9.8%
R^2	0.70	0.66	0.68	0.60	0.88	0.40	-2.51
Session	22	23	24	25	26	27	28
RMSE (mM)	1.02	1.00	1.36	1.40	1.31	0.80	0.87
MARD	11.7%	10.3%	18.0%	16.8%	14.6%	8.5%	17.4%
R^2	0.54	0.74	0.41	0.61	0.57	0.65	-6.36

Table 4.8: RMSE, MARD, and R^2 for all the 28 sessions with a glucose signal amplification factor of 20 and the LOSOCV predictive test scheme. Session 2, 4, 5, 6, and 28 were for visits where the volunteer drank only water without any glucose.

techniques based on the overall MARD performance. 25 out of the 28 sessions had a MARD value of 15% or less. 2 out of the 3 sessions that had a higher MARD value were the sessions where the volunteer drank the water-only solution. Consequently, these sessions were more difficult to achieve low MARD values due to the overall low measurement concentrations.

We next investigated the relationship between the estimation error and the spectral characteristics across the 28 sessions under different amplification factor conditions. The estimation RMSE was used to represent the estimation error and the spectral signal mean (SSM) mean, standard deviation, and coefficient of variation were used to represent the spectral characteristics as discussed in Section 4.1.3. The correlation coefficients between the estimation RMSE and the various SSM-derived quantities are shown in Table 4.10. The correlation coefficients start from close to 0 and increase with the amplification factor until it reaches 40. In general, higher SSM-derived quantities correspond to larger overall signal levels and therefore, larger noise backgrounds. This was established in Section 4.1.3. As a result, it is no surprise that session-wise RMSE has a non-negligible correlation with the SSM-derived quantities. Interestingly, the

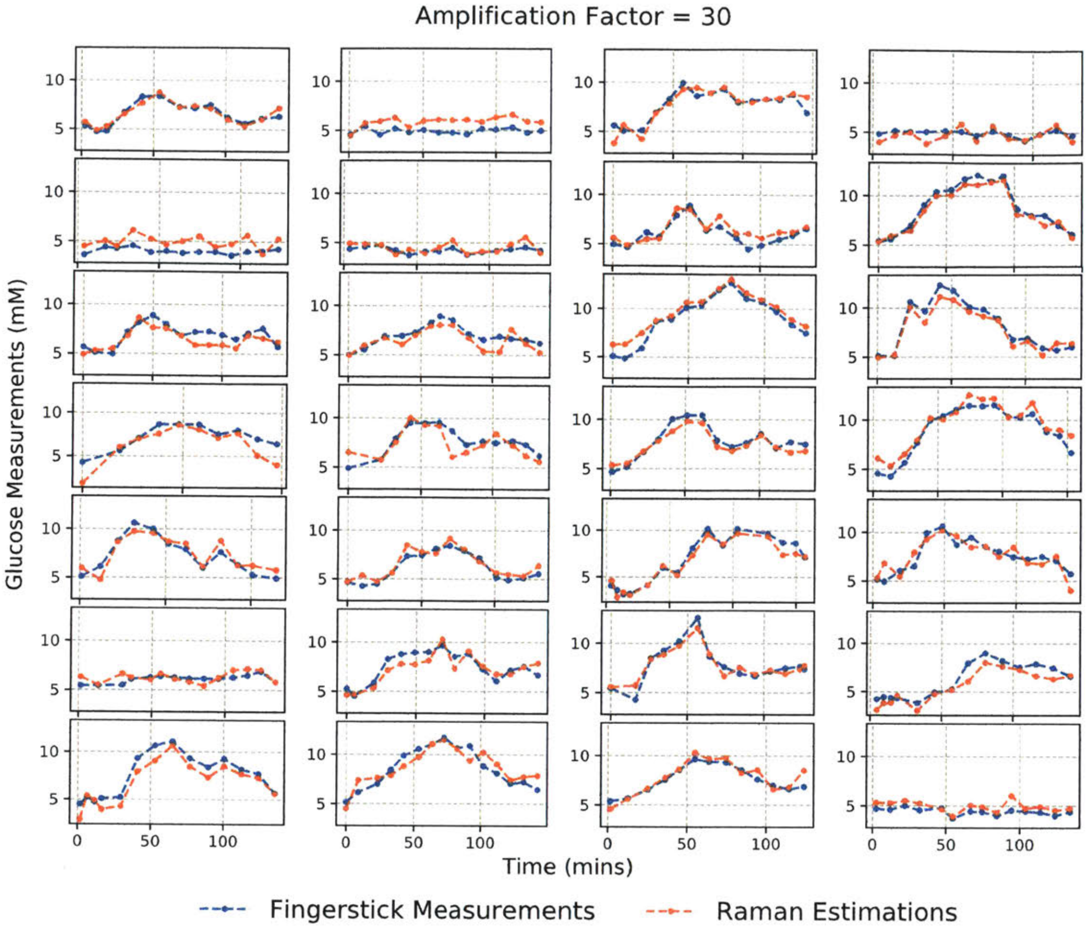


Figure 4-18: Fingerstick glucose measurements and estimated glucose concentrations from Raman spectra for all the 28 sessions with an amplification factor of 30 and the LOSOCV test scheme. The first row is for Session 1 to 4, the second row is for Session 5 to 8, and the rest follows similar orders.

increase of correlation stops when the amplification factor reaches 40. One possible explanation is that with amplification factors smaller than ≈ 40, the limiting factor for estimation accuracy is the background noise in the measurement. As the amplification factor goes pass ≈ 40 that results in high background-limited SNR, other sources become the dominant limiting factor for the estimation accuracy. One possible such source is the interfering spectral signal itself in the skin Raman spectrum. This spectral signal can have contributions from all the skin constitutes and components, as well as the (likely nonlinear) spectral signal distortions that are due to the chemical and

Session	1	2	3	4	5	6	7
RMSE (mM)	0.39	1.05	0.76	0.61	1.08	0.46	0.70
MARD	5.0%	19.8%	7.7%	10.5%	24.5%	8.4%	10.2%
R^2	0.89	-13.55	0.73	-4.08	-14.96	-1.78	0.66
Session	8	9	10	11	12	13	14
RMSE (mM)	0.54	0.81	0.77	0.77	0.68	1.31	1.04
MARD	5.3%	10.0%	9.2%	9.1%	7.1%	15.9%	10.8%
R^2	0.95	0.44	0.46	0.89	0.92	0.09	0.44
Session	15	16	17	18	19	20	21
RMSE (mM)	0.63	0.85	0.76	0.59	0.59	0.96	0.53
MARD	6.7%	9.8%	9.6%	8.0%	7.0%	11.9%	6.6%
R^2	0.86	0.87	0.83	0.84	0.95	0.66	-0.74
Session	22	23	24	25	26	27	28
RMSE (mM)	0.80	0.62	0.86	0.94	0.90	0.64	0.62
MARD	9.1%	7.0%	11.4%	11.3%	9.7%	6.4%	12.0%
R^2	0.72	0.90	0.76	0.82	0.80	0.78	-2.72

Table 4.9: RMSE, MARD, and R^2 for all the 28 sessions with a glucose signal amplification factor of 30 and the LOSOCV predictive test scheme. Session 2, 4, 5, 6, and 28 were for visits where the volunteer drank only water without any glucose.

	SSM Mean	SSM Standard Deviation	SSM Coefficient of Variation
Amplification Factor = 0	-0.05	-0.09	-0.00
Amplification Factor = 10	0.09	0.11	0.16
Amplification Factor = 20	0.33	0.37	0.37
Amplification Factor = 30	0.43	0.48	0.45
Amplification Factor = 40	0.50	0.52	0.45
Amplification Factor = 50	0.50	0.51	0.45

Table 4.10: The correlation coefficients between the estimation RMSE and the spectral signal mean (SSM) mean, standard deviation, and coefficient of variation across the 28 sessions under different amplification factors.

physical inhomogeneities across different volunteers. Unlike the random shot noise that is intrinsic to the spectral signal, it is not dependent on the overall signal level, and therefore its effect may not correlate with the SSM-derived quantities. Consequently, this change of limiting factor for estimation accuracy can be the likely cause for the behavior change of correlation between the estimation accuracy and the SSM-derived quantities under different signal amplification conditions.

4.3.3 Investigations into Processing Variations

In this part, we investigate the estimation performance under different processing variations on test scheme, frame-based smoothing, background estimation and removal, and sample size. Processing recommendations are provided based on our observations and analysis. These can serve as important processing guidelines for future clinical studies with non-invasive skin Raman spectroscopy.

LOSOCV versus Global 10-Fold CV versus Global LOOCV

As the three potential approaches for universal processing with data from all volunteers, we compared the LOSOCV, the global 10-fold CV, and the global LOOCV as the test scheme. More specifically, we quantitatively compared these approaches in the outer CV scheme as the amplification factor changed from 0 to 100 in steps of 10. The inner CV was kept as 10-fold CV for hyperparameter optimization like before in all cases. The RMSE, MARD, and R^2 for these studies are shown in Figure 4-19. The values are also presented in Table 4.11. In general, almost no differences are observed across the 10-fold CV and the LOOCV. While the overall trend of increasing accuracy as the glucose signal strength improves is observed for all test schemes, 10-fold CV and LOOCV consistently provide higher estimation accuracies as compared to the LOSOCV scheme. The differences are more pronounced under lower amplification cases, with the largest performance gap at zero amplification factor. This indicates that the global 10-fold CV and LOOCV scheme can result in more optimistic estimations than the LOSOCV scheme, especially when the glucose signal is close to negligible as in the original spectral data.

As explained previously in Section 4.2.3, skin Raman spectral data from different clinical sessions exhibit strong within-session internal correlation structures. This is likely the cause for the higher correlations observed with the 10-fold CV and LOOCV test scheme, as spectral data and glucose reference measurements from the same session are likely in both the training/validation set and the test set in each test iteration. Only session-wise test schemes can truly isolate such effect from predictive analysis. It

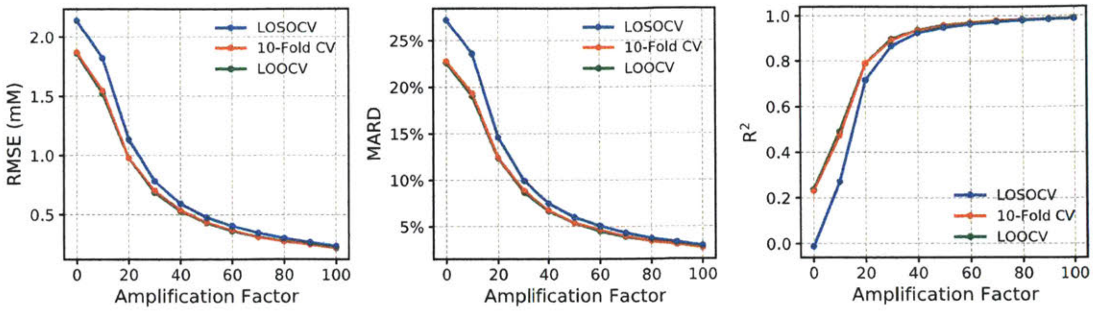


Figure 4-19: RMSE, MARD, and R^2 plots for the overall glucose estimation accuracy with varying glucose signal amplification factors for the LOSOCV, the global 10-fold CV, and the global LOOCV test scheme.

is worth pointing out that with a different Raman instrument that may have a different excitation and collection design, the skin Raman spectra and their photobleaching dynamics may have a drastically different behavior than the ones in this dataset. This may cause the internal correlation structure to change with unpredictable impact on the test performance under the global 10-fold CV and LOOCV scheme. While a non-negligible accuracy performance gap was observed in our dataset, the size of this gap can go either way with a new dataset from a different instrument. As a result, it should be advised to always use session-wise test schemes to exclude such potential spurious correlation sources as much as possible.

Amplification Factor	0	10	20	30	40	50	60	70	80	90	100
LOSOCV											
RMSE (mM)	2.14	1.82	1.14	0.78	0.59	0.48	0.41	0.35	0.30	0.27	0.24
MARD	27.3%	23.6%	14.6%	9.9%	7.5%	6.0%	5.1%	4.4%	3.8%	3.3%	3.0%
R^2	-0.01	0.27	0.72	0.87	0.92	0.95	0.96	0.97	0.98	0.98	0.99
10-Fold CV											
RMSE (mM)	1.87	1.54	0.98	0.70	0.54	0.43	0.37	0.31	0.28	0.25	0.22
MARD	22.8%	19.3%	12.4%	8.8%	6.8%	5.4%	4.6%	3.9%	3.4%	3.1%	2.7%
R^2	0.23	0.47	0.79	0.89	0.94	0.96	0.97	0.98	0.98	0.99	0.99
LOOCV											
RMSE (mM)	1.86	1.52	0.98	0.69	0.53	0.43	0.36	0.31	0.28	0.25	0.22
MARD	22.6%	19.0%	12.4%	8.7%	6.7%	5.4%	4.5%	3.9%	3.4%	3.1%	2.8%
R^2	0.24	0.49	0.79	0.90	0.94	0.96	0.97	0.98	0.98	0.99	0.99

Table 4.11: RMSE, MARD, and R^2 for the overall glucose estimation accuracy with varying glucose signal amplification factors for the LOSOCV, the global 10-fold CV, and the global LOOCV test scheme. Data are also plotted in Figure 4-19.

Frame-based Smoothing

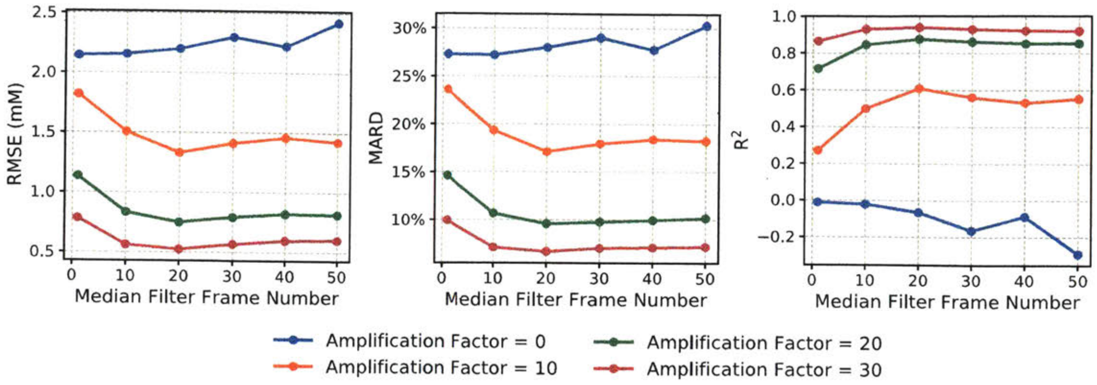


Figure 4-20: RMSE, MARD, and R^2 for the overall glucose estimation accuracy with varying median filter frame number under the LOSOCV test scheme. The median filter was applied symmetrically in the time domain to improve the spectral SNR.

As spectral frames were acquired continuously with 30 s acquisition spacings, spectral smoothing with multiple consecutive frames can be performed to enhance the spectral SNR. For every session data, a median filter with a fixed frame number was applied across the time domain symmetrically for all the spectra corresponding to the fingerstick measurements. These spectra were then selected for future processing. The rest of the preprocessing and nested CV pipeline were the same as the ones presented before. The resulting RMSE, MARD, and R^2 for various amplification factors are shown in Figure 4-20.

From the figure, it can be seen that for the case of zero signal amplification, no increase in prediction accuracy is observed due to the weak glucose signal strength in the original spectra. On the other hand, substantial accuracy improvement can be achieved with frame-based smoothing for cases with amplified glucose signals. The accuracy improvement reaches the highest with 20-frame smoothing. This corresponds to the $\approx$ 10 mins fingerstick measurement spacing. Beyond 20-frame smoothing, the smoothed spectra contain spectral signals closer to adjacent fingerstick measurements than the corresponding fingerstick measurements. Such over-smoothing or over-integration effect is the cause for the reduced prediction accuracy. Overall, frame-based smoothing provides a simple and viable option to increase the prediction accuracy. However, in practical scenarios, one will likely need to balance between the accuracy improvement and the extra time needed for data acquisition. In addition, the accuracy improvement can be limited for situations with sudden glucose changes due to the time averaging aspect of the operation.

Background Estimation and Removal

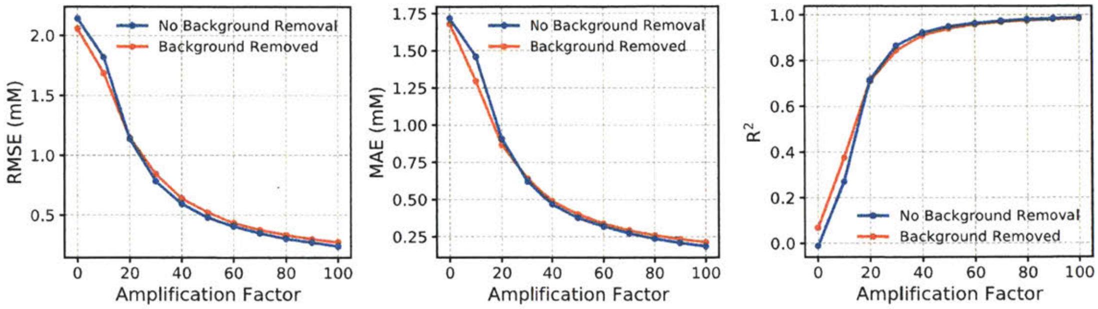


Figure 4-21: RMSE, MARD, and R^2 for the overall glucose estimation accuracy with and without separate background estimation and removal using the Lieber algorithm under the LOSOCV test scheme.

As discussed in details previously from Chapter 3 and Section 4.1.3, handling of the background and baseline signals in Raman spectra has played a significant role in spectral data processing. While many strategies exist to estimate and remove the background and baseline signals, no separate step has been used to explicitly estimate these signals in our processing. This is because in skin Raman spectra, the autofluorescence background signal is mostly a smooth curve (shown in Figure 4-3) that can be represented with a small number of basis functions, such as the polynomial or the B-spline basis functions. With enough training data that have significant variations for the background signal, a training-based algorithm such as PLSR will be able to capture the subspace spanned by the background signal into the model implicitly. This is an example of data-driven handling of the background signal. For explicit background estimation, one has to deal with problems such as the order of the polynomial basis function and the estimation criterion. This can potentially result in either underfitting or overfitting the background signal and introduce error and bias into subsequent training steps. On the contrary, training-based background handling requires no explicit modeling and criterion selection, and gets asymptotically more accurate as the training data size increases. Given the respectable data size collected in our clinical study, we chose this approach for dealing with the autofluorescence background.

To validate our implicit handling of the background signal, Figure 4-21 shows the prediction results with and without a separate background estimation and removal step. The background estimation was performed with the Lieber algorithm with a polynomial basis order of 5 [90]. As can be seen from the figure, only minor accuracy differences are observed from the two methods. This indicates that a separate background estimation step is not necessary for our dataset and processing pipeline. Other than a polynomial basis order of 5, several other polynomial basis orders were also investigated, which did not show much effect on the accuracy outcomes. Our results in this part suggest that for spectroscopy applications with moderate to large amount of spectral data, separate background estimation can be optional with training-based approaches such as PLSR. This can lead to simpler algorithm pipelines and more efficient computational processes.

Sample Size

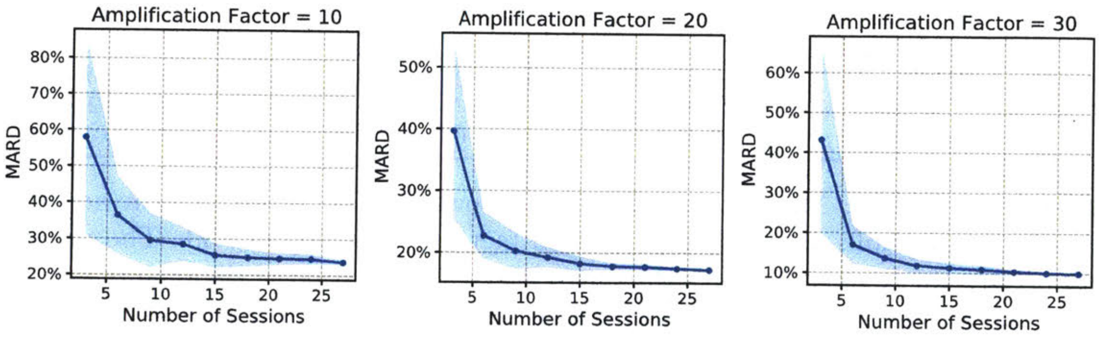


Figure 4-22: LOSOCV test results with varying sample sizes for three different amplification factors. The solid lines and data points are the means of error from 30 independent runs with randomly selected session subsets. The shaded areas around the lines represent the standard deviations of the error from the 30 runs.

Lastly, we investigated the effect of sample size variation on the overall prediction accuracy. For biomedical spectroscopy applications, sample size plays a crucial role in both approach selection as well as model performance evaluation. In general, applied spectroscopy practitioners do not have the luxury of collecting datasets in large volumes due to the often time-consuming and costly nature of the experiments. Therefore, exploring the application and model performance in the sample size domain can be extremely valuable. This can help validate the statistical significance of the conclusion and provide guidelines for future experimental design and budget planning.

Sample Size	6	9	12	15
Amplification Factor = 10				
RMSE (mM)	2.91 ± 0.92	2.40 ± 0.66	2.27 ± 0.41	2.00 ± 0.22
MARD	36.4% ± 10.8%	29.5% ± 7.5%	28.6% ± 4.9%	25.5% ± 3.4%
R^2	−1.15 ± 1.35	−0.42 ± 0.73	−0.21 ± 0.45	0.10 ± 0.22
Amplification Factor = 20				
RMSE (mM)	2.15 ± 0.60	1.65 ± 0.42	1.47 ± 0.23	1.34 ± 0.17
MARD	25.4% ± 7.6%	20.6% ± 6.0%	18.5% ± 3.4%	16.5% ± 2.5%
R^2	−0.17 ± 0.86	0.30 ± 0.43	0.48 ± 0.16	0.58 ± 0.10
Amplification Factor = 30				
RMSE (mM)	1.43 ± 0.44	1.11 ± 0.24	0.95 ± 0.12	0.91 ± 0.10
MARD	17.0% ± 4.9%	13.7% ± 2.8%	11.8% ± 1.8%	11.3% ± 1.2%
R^2	0.44 ± 0.47	0.73 ± 0.10	0.79 ± 0.07	0.81 ± 0.05
Sample Size	18	21	24	27
Amplification Factor = 10				
RMSE (mM)	1.96 ± 0.17	1.93 ± 0.11	1.91 ± 0.08	1.84 ± 0.04
MARD	24.9% ± 2.4%	24.6% ± 1.7%	24.5% ± 1.5%	23.7% ± 0.6%
R^2	0.15 ± 0.15	0.15 ± 0.12	0.20 ± 0.07	0.25 ± 0.03
Amplification Factor = 20				
RMSE (mM)	1.23 ± 0.08	1.22 ± 0.06	1.17 ± 0.05	1.13 ± 0.01
MARD	15.6% ± 1.1%	15.4% ± 1.0%	14.9% ± 0.6%	14.5% ± 0.3%
R^2	0.65 ± 0.05	0.67 ± 0.03	0.70 ± 0.02	0.72 ± 0.01
Amplification Factor = 30				
RMSE (mM)	0.88 ± 0.06	0.82 ± 0.04	0.80 ± 0.03	0.79 ± 0.01
MARD	10.9% ± 1.1%	10.4% ± 0.6%	10.1% ± 0.4%	10.0% ± 0.2%
R^2	0.83 ± 0.03	0.85 ± 0.02	0.86 ± 0.01	0.86 ± 0.01

Table 4.12: RMSE, MARD, and R^2 for the overall glucose estimation accuracy with varying sample sizes for amplification factors of 10, 20, and 30. The number before/after the ± sign indicates the mean/standard deviation of the quantity across the 30 independent runs. The MARD values are also plotted in Figure 4-22.

For this task, we randomly created sub-datasets with predefined data sizes from the original 28-session dataset. For each data size, we created 30 sub-datasets with this size by randomly selecting session data without replacement from the original dataset. We ran our processing pipeline with LOSOCV for each sub-dataset and obtained aggregated performance statistics for each dataset size. The MARD values from our results are plotted in Figure 4-22 for amplification factors of 10, 20, and 30. The overall RMSE, MARD, and R^2 are also shown in Table 4.12. In the figure and the table, we show both the means as well as the standard deviations for the results from the 30 independent runs. It is clear that the overall prediction performance improves with larger dataset sizes. The dependency on the sample size is higher for situations with lower glucose signal strengths. This can be observed by comparing the plots in Figure 4-22. Lower estimation variances are in general observed with larger dataset sizes, though this is partially due to the intrinsic variance property associated with our data sampling approach with a limited overall dataset size of 28. Improvement in prediction accuracy and consistency can be expected if more data are available. However, the increase in prediction accuracy is likely limited for each signal strength based on the overall performance plot trends. Figure 4-22 and Table 4.12 also illustrate the minimum sample size requirements for skin Raman analysis and diagnostics applications under the limitation of the current sample size. For example, with a glucose signal strength equivalent to an amplification factor of 20, at least ≈ 20 sessions are needed to train a model for reliable estimation with MARD lower than $\approx 15\%$ under the LOSOCV test scheme. A weaker signal strength will most likely require a larger sample size to achieve similar accuracy. This may serve as useful reference for future trial design.

4.3.4 Result Implications and Discussions

The investigations performed in this section reveal the many facets related to the signal and calibration requirements for non-invasive glucose estimation with *in vivo* Raman spectroscopy. While the predictive modeling aspect was carried out with manually enhanced glucose signals, useful conclusions and guidelines for future experimental improvement can nonetheless be made. Our predictive test modeling suggests that a signal amplification factor of at least ≈ 20 is needed for reliable estimation of the glucose signal from skin Raman spectra with 30 s acquisition frames. This corresponds to an average glucose peak SNR of ≈ 2.0 to 2.6 in our collected dataset. For the original unamplified glucose signal strength, this means that the light throughput of the instrument has to increase ≈ 400 times in order to match the equivalent peak SNR imposed by the amplified glucose signal, assuming that the measurement is background noise limited. At least ≈ 2.25 times more throughput is needed to future improve the prediction accuracy to match the standard strip test techniques. Such high improvement requirement means that one has to sacrifice the portability aspect of the instrument and use spectrometers that have a much larger slit size with a longer optical path. In addition, excitation and collection optics that are specially designed for diffused light with definitive throughput improvement will be needed [149]. At last, computational throughput enhancement techniques such as the coded aperture design can be used to further increase the spectrometer sensitivity [28]. As any potential glucose signal is extremely dim under the overwhelming skin autofluorescence background, it is likely that one has to exhaust all throughput improvement techniques to meet the stringent SNR requirement target for universal predictive models found in this study.

As discussed in Section 4.1.3, measurement nonidealities such as movement artifacts and ambient light leakage exist in our spectral dataset. Upon examination of the results of our signal generation and processing approach, no noticeable effects were observed from these nonidealities on individual session's predictive accuracy in general. This is not a surprise as the spectral signals from these factors can be represented by a small number of linear spectral bases with limited interfering abilities over the glucose signal of interest. As a result, the contributions from these factors can be implicitly learned with our training process. Having said that, efforts to mitigate these nonidealities are still needed for future experiments. This is because the presence of these spectral components can increase the sample size and quality requirements for model training. Given that any potential signals from glucose or any component of interest are likely to be weak in comparison to the background and interfering spectral signals, all efforts need to be put in place to acquire high quality spectral signals.

It is worth emphasizing that the calculated peak SNR is merely an absolute theoretical lower bound requirement for predictive glucose measurement with Raman spectroscopy. Simple linear models have been assumed for light scattering and the physiological aspects of non-invasive Raman glucose sensing with our treatment. Real *in vivo* glucose signal likely will have characteristics that are different from the glucose solution measurement used in our signal generation process. Nonlinear effects due to light absorption and scattering with the physical and chemical inhomogeneities in skin tissue will likely introduce non-negligible signal distortion and signal dilution. Perhaps more crucially, the physiological aspects of Raman glucose sensing also need careful examinations. It is generally believed that with skin-based Raman glucose sensing techniques, the glucose signal originates from a mix of the interstitial fluid and blood in the skin tissue [2]. However, the extent to which how the interstitial fluid glucose level in different part of the body correlates with the blood glucose level is still under investigation [82]. Furthermore, in order to rigorously evaluate the accuracy of the non-invasive glucose measurement techniques, the reference measurement needs to be performed with venous whole blood samples using standard techniques such as the YSI measurement. This was not implemented in the current work due to personnel and cost constraints. While the overall problem may seem complex with all these complications, modular approaches can be taken to evaluate and investigate the effect from these individual factors independently. Afterwards, a broad picture of the solution can be pieced together with these results.

4.4 Conclusions

In summary, we have performed a comprehensive numerical investigation on the signal requirement for non-invasive glucose sensing with *in vivo* skin Raman spectroscopy. A thorough discussion on the spectral data characteristics and practical issues associated with an experimentally collected skin Raman spectral dataset using a portable clinical Raman instrument was presented. With a numerical signal generation approach, we explored in depth on a universal and predictive processing pipeline for glucose signal estimation from skin Raman spectrum. Processing recommendations as well as quantitative specifications were provided for future references. In the author's opinion, significant technological advances in hardware devices will be required before the next breakthrough in this field. For example, high-speed and large-area photodetector arrays that are able to perform time-gating for fluorescence suppression in Raman spectroscopy can potentially be the building blocks for next-generation Raman instruments [63].

While our analysis has focused on non-invasive glucose estimation in this work, the numerical signal generation framework and the universal and predictive processing pipeline are directly translatable for any potential biomarker and biological analyte of interest in the skin tissue. Numerical signal generation can also be compared with experimental physical signal generation with mixture studies to quantitatively explore the effects of light scattering and photon transport on spectral signal mixing. This can provide useful insights like the case in our study for instrument adjustment and experimental design.

Chapter 5

Future Work

In this thesis, we have presented solutions and investigations into three problems in optical spectroscopy, with an emphasis of merging computational and statistical techniques with physical domain knowledge to break the conventional barriers in spectroscopy technology understanding and development. While complete solutions have been provided for these problems respectively, there are several extension directions which present significant opportunities and values for further advancing the optical spectroscopy applications. These future prospects are discussed in this chapter.

Lensless and Ultra-Compact Talbot Spectroscopy Solutions for Spectrum Sensing and Wavelength Estimation

Due to the near/mid field nature of the Talbot interferogram sampling problem and the grating-sensor positioning requirement as discussed in details in Chapter 2, the optimal solution for compact and high performance Talbot spectroscopy is to directly integrate periodic structures on top of the pixels on the CMOS imager. This sensor-grating integration can be achieved from a bonding process with independently produced image sensor and grating structure pieces. A bonded Talbot sampling device ensures that regions close to the zero path-length difference place are sampled and can enable general spectroscopy applications. Another interesting possibility is through ingenious engineering with the metal and dielectric layers in the CMOS imager fabrication process. This has previously been realized for achieving angle-sensitive image sensors with the Talbot effect [159, 160], where the metal interconnect wire layer was used as the grating structure. A tilted sampling geometry has to be realized in this case, which does not have straightforward solutions though. A possibility is to use multiple metal layers to realize a tilted grating geometry. Regardless of the possible approach, a high level of device integration not only results in an even smaller form factor as compared to the current solution, but also naturally resolves issues related to alignment and positioning. In addition, the resulting mechanical stability from the monolithic structure can potentially offer high robustness towards environmental perturbations – a factor highly valued in high precision wavemeter applications in particular.

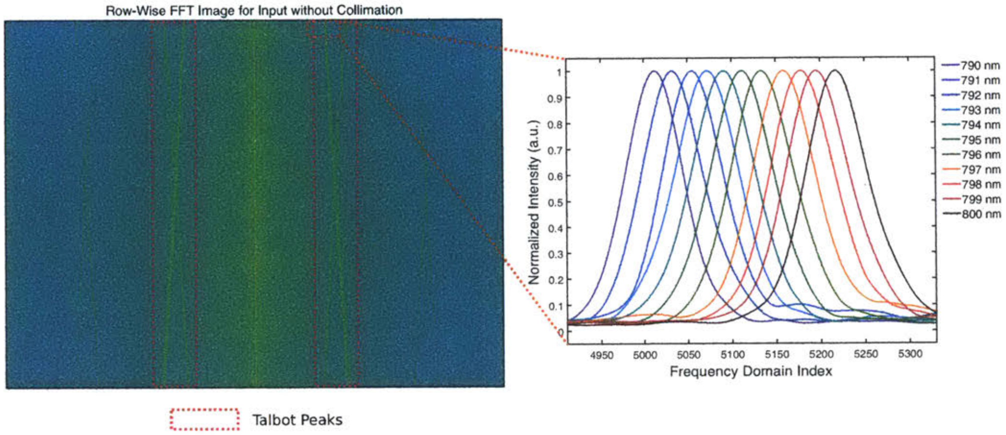


Figure 5-1: (left) A two-dimensional image where each row is the FFT array for an interferogram row with an input laser source directly from a single-mode fiber without any collimation optics. (right) The zoomed-in zero-padded FFT spectra from 30 interferogram rows for a tunable laser source from 790 to 800 nm in steps of 1 nm. The reconstructed spectra show that spectral discrimination is maintained without any collimation optics.

Another area worth exploring is to directly perform spectrum retrieval without using any collimation optics. This is especially attractive and applicable for the Talbot wavemeter work. Figure 5-1 shows a two-dimensional image where each row is the FFT array for an interferogram row with an input laser source directly from a single-mode fiber. Due to the spherical wavefront of the light source, the inverted FFT peaks are not aligned across the column dimension. However, the Talbot peaks still contain the wavelength information as can be seen from the right plot, where a small number of interferogram rows (30) are used for spectra reconstruction across a 10 nm wavelength span in steps of 1 nm. Due to the fact that the input wavefront is well-defined and analytically tractable, there can be algorithmic ways to correct for the systematic frequency and phase transformation encountered in this system, while still maintaining high precision wavelength estimations. The resulting instrument has a lensless geometry and can be extremely attractive for portable applications.

Bayesian Inference and Modeling for Spectral Data Analysis

For our Bayesian modeling work presented in Chapter 3, several potential future directions exist to either improve the computational aspect of the proposed algorithm or to further explore areas of Bayesian modeling for spectral data analysis. While the current algorithm and its execution are able to provide reliable and accurate estimations with both simulated and experimental spectral datasets, they are not able to match the speed performance of regression algorithms such as PLSR. This can be improved by code optimization that better utilizes the multi-thread, parallel processing, and cache locality aspects of the computational process, or through more detailed MCMC computation engineering for speed acceleration [161]. Other than building from the current RJMCMC computation process, another alternative approach is to apply approximation techniques such as variational inference [162] with suitable model selection criteria [127] in the two-stage processing pipeline. Unlike the MCMC sampling approach, variational inference uses optimization to approximate the posterior distribution. As a result, it is more computationally efficient and more applicable for scalable or time-critical applications [163]. However, a rather drastic statistical treatment change is required for the adoption of variational inference as compared to the current approach.

Another possible direction is to demonstrate the usage of our algorithm for application domains that had been challenging for training-based estimation algorithms such as PLSR. For example, in areas such as forensic science, spectral signatures from impurities and backgrounds in the scene can often be problematic for training-based algorithms [164]. This is due to the fact that it is not practical to generate a training dataset that can include all the possibilities for the contents of the impurities and backgrounds in a scene *a priori*. Since these tasks often involve identifying and quantifying spectral signatures from a library of candidate substances of interest in the presence of unknown impurities and backgrounds, a modified version of this algorithm that can deal with a multitude of target substance spectra can be devised for on-the-scene identification and quantification tasks without the need to rely on training-based algorithms.

At last, the extent of the Bayesian modeling techniques on spectral data analysis is by no means restricted to the scope of the current work. As optical spectroscopy datasets are often limited by experimental factors such as sample size and accessibility, mathematical and statistical modeling tools from the Bayesian modeling and inference world can be applied to bypass some of the traditional requirements for regression and classification analysis. For example, Gaussian process has been previously used to facilitate category classification with unseen classes that are not present in the training dataset [165]. As an example of a future exploration, Bayesian model selection can be used for automatic spectral basis identification and selection in CLS-like multi-analyte tracking applications [125], where spectral basis assignment had traditionally been performed through iterative examination of the matrix pseudo-inversion reconstruction residue [33, 4]. While most standard and off-the-shelf conventional tools such as PLSR and SVM have become extremely accessible and effective for applications where high quality and large volume datasets are readily available, a large amount of untapped potential still exists in areas like optical spectroscopy where application-specific and modeling-based algorithms can be more favorable. In some sense, this type of work falls into a broad catalog where co-design of algorithmic processing and experimental protocol can enable reliable target identification and quantification approaches that are also resource-wise optimized.

Raman Spectroscopy for Non-invasive Biomarker Identification and Estimation

For Raman spectroscopy with non-invasive skin analysis and diagnostics applications, the autofluorescence background in the skin Raman spectrum represents a major challenge for signal analysis. This has been discussed extensively with our data analysis in Chapter 4. Skin site pre-bleach with prolonged light exposure has previously been proposed as a way to improve the spectral quality in skin Raman spectroscopy [146]. However, a comprehensive study on developing a robust and universal protocol for a safe and repeatable pre-bleach procedure with quantifiable SNR improvements has not yet been performed. As autofluorescence is one of the central issues surrounding skin Raman spectroscopy, this can be an important topic for future studies.

As discussed in Section 1.4, skin Raman spectroscopy is an extremely attractive option for non-invasive biomarker and physiological analyte detection and estimation. While any glucose signal from skin Raman spectrum is extremely weak for detection, signals from other components or molecules can be much stronger as shown in Figure 4-6. One of the key challenges in the field is to conduct experiments to associate spectral changes to underlying physiological measurements, which are often extremely resource-intensive and costly to obtain, and require professional medical practitioners to conduct the measurements. In addition, for statistical rigor and the possibility of applying powerful non-linear algorithms such as the neural networks, data have to be collected in large volumes. An intensive care unit (ICU) in a hospital is a potential trial environment where many vital and laboratory measurements for the patients are constantly conducted by healthcare professionals and recorded in large-volume on a daily basis. Recently, there has been significant interests from the bio-informatics and computational physiology community for ICU data mining and analysis [166]. A portable clinical Raman instrument like the one developed in this study can be a potential tool to be installed in an ICU alongside of the traditional vital signal monitoring system. This can enable large-volume continuous spectral data collection with reference physiological and laboratory measurements in a critical care environment, where vital signals can have variations much larger than those encountered in healthy individuals. A dataset like this can be monumentally valuable for discovering the utility of skin Raman spectroscopy for non-invasive critical biomarker and biological signal identification and estimation, with the possibility of translatable performance to the more general population outside of the ICU environment. Such experimental trial would require a substantial modification to the existing instrument to allow simpler and more user-friendly light delivery and collection mechanisms with no intervention or adjustment needs.

Appendix A

Far Field Grating Diffraction – Spatially Dispersive Response

Optical field propagation in the far field or Fraunhofer region after passing through a sinusoidal phase grating with a finite aperture is discussed in this part. Our analysis mainly follows Goodman [98]. With the input of the optical system being a uniform plane wave, we assume that the transmission function of the grating $t_G(\xi, \eta)$ introduces a sinusoidal phase change to the input wave in ξ dimension. Assuming a rectangular aperture on the grating, we have the input optical field $U_1(\xi, \eta)$ equal to $t_G(\xi, \eta)$ as

$$U_1(\xi, \eta) = t_G(\xi, \eta) = \exp\left[ja \sin\left(2\pi \frac{\xi}{P}\right)\right] \mathrm{rect}\left(\frac{\xi}{2W}\right) \mathrm{rect}\left(\frac{\eta}{2L}\right),$$

where a is the phase modulation amplitude, P is the grating period, W is the grating half width, L is the grating half length, and rect$(\cdot)$ is the rectangle window function defined with width 1. According to the Jacobi-Anger expansion, the phase modulation term can be expressed as

$$\exp\left[ja \sin\left(2\pi \frac{\xi}{P}\right)\right] = \sum_{q=-\infty}^{\infty} J_q\left(a\right) \exp\left(j2\pi q \frac{\xi}{P}\right), \tag{A.1}$$

where $J_q(\cdot)$ is the Bessel function of the first kind and order q. The Fourier transform of this term can then be expressed as

$$\mathcal{F}\left\{\exp\left[ja\sin\left(2\pi\frac{\xi}{P}\right)\right]\right\} = \sum_{q=-\infty}^{\infty} J_q\left(a\right)\delta\left(f_X - \frac{q}{P}, f_Y\right). \tag{A.2}$$

With this, the Fourier transform of the input optical field follows as

$$\mathcal{F}\left\{U_1(\xi,\eta)\right\} = \sum_{q=-\infty}^{\infty} 4WLJ_q\left(a\right)\operatorname{sinc}\left[2W\left(f_X - \frac{q}{P}\right)\right]\operatorname{sinc}(2Lf_Y),$$

where $\operatorname{sinc}(\cdot)$ is the normalized sinc function. Subsequently, the optical field at x-y plane with location z according to the Fraunhofer diffraction can be written as

$$U_2(x,y) = \frac{4WL}{j\lambda z}e^{jkz}\exp\left[j\frac{k}{2z}(x^2+y^2)\right]\sum_{q=-\infty}^{\infty} J_q\left(a\right)\operatorname{sinc}\left[\frac{2W}{\lambda z}\left(x - q\frac{\lambda z}{P}\right)\right]\operatorname{sinc}\left(\frac{2Ly}{\lambda z}\right).$$

Assuming that $W \gg P$, which means that the width for the sinc function is much smaller than the displacement of different diffraction orders, the intensity $I_2(x,y)$ can then be derived approximately as

$$I_2(x,y) \approx \left(\frac{4WL}{\lambda z}\right)^2 \sum_{q=-\infty}^{\infty} J_q^2\left(a\right)\operatorname{sinc}^2\left[\frac{2W}{\lambda z}\left(x - q\frac{\lambda z}{P}\right)\right]\operatorname{sinc}^2\left(\frac{2Ly}{\lambda z}\right). \tag{A.3}$$

Equation A.3 indicates that if multiple diffraction orders exist, the far field displacement between adjacent orders is $\frac{\lambda z}{P}$. Moreover, within each diffraction order, the spatial shift has a one-to-one mapping with the wavelength. This well-formed and well-conditioned transformation is perhaps one of the main reasons for the popularity of diffraction gratings as the dispersive elements in spectrometers. In practice, only a finite number of diffractive orders exist due to the conditions required for propagating waves, which is discussed in Section 2.1.2.

Appendix B

Derivations of the Talbot Effect Under Tilted Incidence Angles

Tilted Incidence in y-z Plane

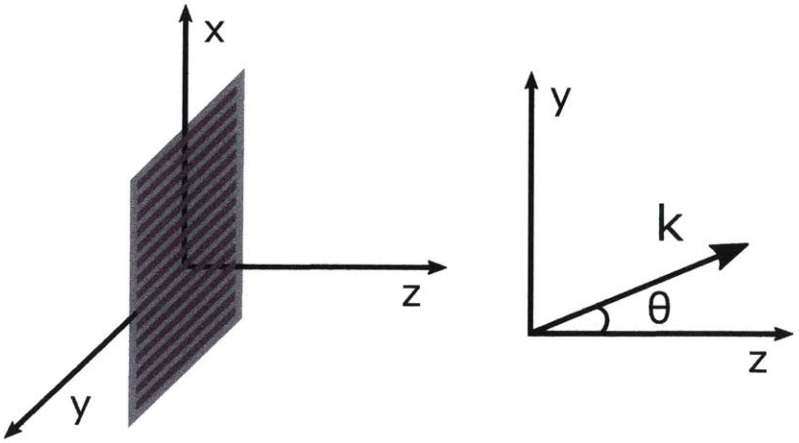


Figure B-1: Illustration for tilted incidence angle θ in the y-z plane.

We first consider tilted plane wave incidence in the y-z plane as shown in Figure B-1. Assume that the tilt angle is θ with respect to the z axis, the transmission function due to the angle tilt is

$$t_T(\xi, \eta) = \exp\left[jk\sin(\theta)\eta\right].$$

Meanwhile, the transmission function with the sinusoidal phase grating is

$$t_G(\xi,\eta) = \exp\left[ja\sin\left(2\pi\frac{\xi}{P}\right)\right].$$

The input optical field $U_1(\xi,\eta)$ can then be expressed as

$$U_1(\xi,\eta) = t_T(\xi,\eta)t_G(\xi,\eta) = \exp\left[jk\sin(\theta)\eta\right]\exp\left[ja\sin\left(2\pi\frac{\xi}{P}\right)\right],$$

and its Fourier transform is

$$\mathcal{F}\left\{U_1(\xi,\eta)\right\} = \mathcal{F}\left\{t_T(\xi,\eta)\right\} \circledast \mathcal{F}\left\{t_G(\xi,\eta)\right\} = \sum_{q=-\infty}^{\infty} J_q\left(a\right)\delta\left[f_X - \frac{q}{P}, f_Y - \frac{\sin(\theta)}{\lambda}\right].$$

With the Rayleigh-Sommerfeld diffraction solution, the Fourier transform of the optical field at the observation plane $U_2(x,y)$ can be expressed as

$$\mathcal{F}\left\{U_2(x,y)\right\} = \mathcal{F}\{U_1(x,y)\}H(f_X,f_Y)$$
$$= \sum_{q=-\infty}^{\infty} J_q\left(a\right)\exp\left[jk\sqrt{\cos^2(\theta) - \left(q\frac{\lambda}{P}\right)^2}z\right]\delta\left[f_X - \frac{q}{P}, f_Y - \frac{\sin(\theta)}{\lambda}\right].$$

With inverse Fourier transform, we have the following equation for the output optical field

$$U_2(x,y) = \sum_{q=-\infty}^{\infty} J_q\left(a\right)\exp\left[jk\sqrt{\cos^2(\theta) - \left(q\frac{\lambda}{P}\right)^2}z\right]\exp\left(j2\pi q\frac{x}{P}\right)\exp\left[j2\pi\frac{\sin(\theta)y}{\lambda}\right].$$

As in previous treatment, by considering only the -1, 0, and $+1$ diffractive orders, we can simplify the optical field as

$$U_2(x,y) = J_0\left(a\right)\exp\left[jk\cos(\theta)z\right]\exp\left[j2\pi\frac{\sin(\theta)y}{\lambda}\right] +$$
$$j2J_1\left(a\right)\sin\left(2\pi\frac{x}{P}\right)\exp\left[jk\sqrt{\cos^2(\theta) - \left(\frac{\lambda}{P}\right)^2}z\right]\exp\left[j2\pi\frac{\sin(\theta)y}{\lambda}\right],$$

and the intensity as

$$I_2(x,y) = J_0^2(a) + 4J_1^2(a)\sin^2\left(2\pi\frac{x}{P}\right) +$$
$$4J_0(a)\,J_1(a)\sin\left(2\pi\frac{x}{P}\right)\sin\left\{k\left[\cos(\theta) - \sqrt{\cos^2(\theta) - \left(\frac{\lambda}{P}\right)^2}\right]z\right\}.$$

Tilted Incidence in x-z Plane

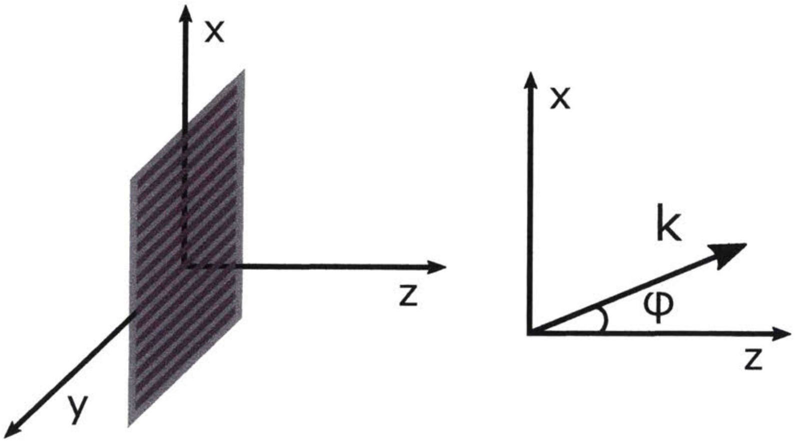


Figure B-2: Illustration for tilted incidence angle ϕ in the x-z plane.

Next we consider the case where the tilted plane wave incidence is in the x-z plane as shown in Figure B-2. In this case, the diffraction solution is more complicated than that from the previous case. The transmission function for tilted plane wave incidence with tilt angle ϕ with respect to the z axis is

$$t_T(\xi,\eta) = \exp\left[jk\sin(\phi)\xi\right].$$

With the same grating transmission function, the input optical field is now

$$U_1(\xi,\eta) = t_T(\xi,\eta)t_G(\xi,\eta) = \exp\left[jk\sin(\phi)\xi\right]\exp\left[ja\sin\left(2\pi\frac{\xi}{P}\right)\right],$$

and its Fourier transform is

$$\mathcal{F}\left\{U_1(\xi,\eta)\right\} = \mathcal{F}\left\{t_T(\xi,\eta)\right\} \circledast \mathcal{F}\left\{t_G(\xi,\eta)\right\} = \sum_{q=-\infty}^{\infty} J_q\left(a\right)\delta\left[f_X - \frac{\sin(\phi)}{\lambda} - \frac{q}{P}, f_Y\right].$$

Similar to previous parts, the Fourier transform of $U_2(x,y)$ and $U_2(x,y)$ are now

$$\begin{gathered}\mathcal{F}\left\{U_2(x,y)\right\} = \mathcal{F}\{U_1(x,y)\}H(f_X,f_Y) \\ = \sum_{q=-\infty}^{\infty} J_q\left(a\right)\exp\left\{jk\sqrt{1-\left[\sin(\phi)+q\frac{\lambda}{P}\right]^2}z\right\}\delta\left[f_X - \frac{\sin(\phi)}{\lambda} - \frac{q}{P}, f_Y\right],\end{gathered}$$

and

$$U_2(x,y) = \sum_{q=-\infty}^{\infty} J_q\left(a\right)\exp\left\{jk\sqrt{1-\left[\sin(\phi)+q\frac{\lambda}{P}\right]^2}z\right\}\exp\left\{j2\pi\left[\frac{\sin(\phi)}{\lambda}+\frac{q}{P}\right]x\right\}.$$

With similar assumptions as before, we reach to the final solution for $U_2(x,y)$ and $I_2(x,y)$ as

$$\begin{gathered}U_2(x,y) = J_0\left(a\right)\exp\left[jk\cos(\phi)z\right]\exp\left[jk\sin(\phi)x\right] + \\ J_1\left(a\right)\exp\left\{jk\sqrt{1-\left[\sin(\phi)+\frac{\lambda}{P}\right]^2}z\right\}\exp\left\{jk\left[\sin(\phi)+\frac{\lambda}{P}\right]x\right\} - \\ J_1\left(a\right)\exp\left\{jk\sqrt{1-\left[\sin(\phi)-\frac{\lambda}{P}\right]^2}z\right\}\exp\left\{jk\left[\sin(\phi)-\frac{\lambda}{P}\right]x\right\},\end{gathered}$$

and

$$I_2(x,y) = J_0^2(a) + 2J_1^2(a) +$$
$$2J_0(a)\,J_1(a)\cos\left\{k\left\{\cos(\phi) - \sqrt{1-\left[\sin(\phi)+\frac{\lambda}{P}\right]^2}\right\}z - 2\pi\frac{x}{P}\right\} -$$
$$2J_0(a)\,J_1(a)\cos\left\{k\left\{\cos(\phi) - \sqrt{1-\left[\sin(\phi)-\frac{\lambda}{P}\right]^2}\right\}z + 2\pi\frac{x}{P}\right\} -$$
$$2J_1^2(a)\cos\left\{k\left\{\sqrt{1-\left[\sin(\phi)-\frac{\lambda}{P}\right]^2} - \sqrt{1-\left[\sin(\phi)+\frac{\lambda}{P}\right]^2}\right\}z - 4\pi\frac{x}{P}\right\}.$$

General Tilt

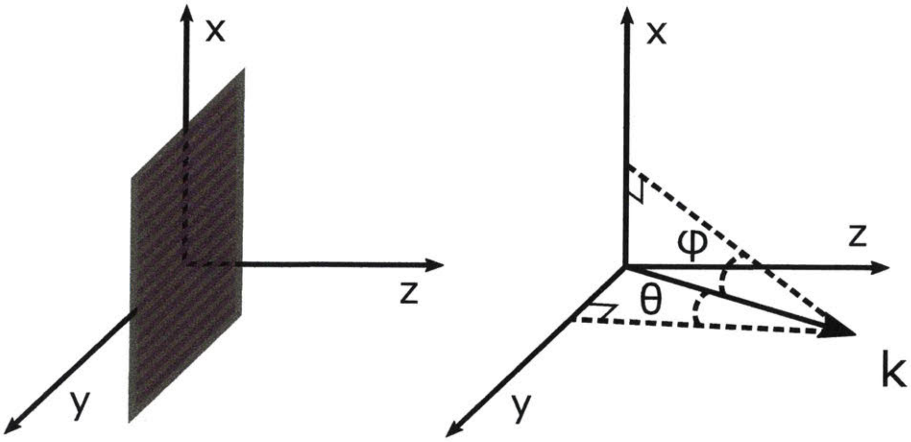


Figure B-3: Illustration for general tilted incidence with both θ and ϕ.

The individual results for tilt in the y-z and x-z plane demonstrate the different pattern formation response towards incidence angle tilt for the Talbot effect. For the case of general tilt with both θ and ϕ, which is illustrated in Figure B-3, we have

$$t_T(\xi,\eta) = \exp\left[jk\sin(\phi)\xi + jk\sin(\theta)\eta\right].$$

The input optical field is now

$$U_1(\xi,\eta) = t_T(\xi,\eta)t_G(\xi,\eta) = \exp\left[jk\sin(\phi)\xi + jk\sin(\theta)\eta\right]\exp\left[ja\sin\left(2\pi\frac{\xi}{P}\right)\right],$$

and its Fourier transform is

$$\mathcal{F}\left\{U_1(\xi,\eta)\right\} = \mathcal{F}\left\{t_T(\xi,\eta)\right\} \circledast \mathcal{F}\left\{t_G(\xi,\eta)\right\} = \sum_{q=-\infty}^{\infty} J_q\left(a\right)\delta\left[f_X - \frac{\sin(\phi)}{\lambda} - \frac{q}{P}, f_Y - \frac{\sin(\theta)}{\lambda}\right].$$

Subsequently, the Fourier transform of $U_2(x,y)$ and $U_2(x,y)$ are

$$\mathcal{F}\left\{U_2(x,y)\right\} = \mathcal{F}\{U_1(x,y)\}H(f_X, f_Y)$$
$$= \sum_{q=-\infty}^{\infty} J_q\left(a\right)\exp\left\{jk\sqrt{\cos^2(\theta) - \left[\sin(\phi) + q\frac{\lambda}{P}\right]^2} z\right\}\delta\left[f_X - \frac{\sin(\phi)}{\lambda} - \frac{q}{P}, f_Y - \frac{\sin(\theta)}{\lambda}\right],$$

and

$$U_2(x,y) = \sum_{q=-\infty}^{\infty} J_q\left(a\right)\exp\left\{jk\sqrt{\cos^2(\theta) - \left[\sin(\phi) + q\frac{\lambda}{P}\right]^2} z\right\}$$
$$\exp\left\{j2\pi\left[\frac{\sin(\phi)}{\lambda} + \frac{q}{P}\right]x\right\}\exp\left[j2\pi\frac{\sin(\theta)y}{\lambda}\right].$$

With the three major diffractive orders as before, the final solution for $U_2(x,y)$ and $I_2(x,y)$ under general incidence tilt is therefore

$$U_2(x,y) = J_0\left(a\right)\exp\left[jk\sqrt{\cos^2(\theta) - \sin^2(\phi)z}\right]\exp\left[j2\pi\frac{\sin(\phi)x}{\lambda}\right]\exp\left[j2\pi\frac{\sin(\theta)y}{\lambda}\right] +$$
$$J_1\left(a\right)\exp\left\{jk\sqrt{\cos^2(\theta) - \left[\sin(\phi) + \frac{\lambda}{P}\right]^2} z\right\}\exp\left\{j2\pi\left[\frac{\sin(\phi)}{\lambda} + \frac{1}{P}\right]x\right\}\exp\left[j2\pi\frac{\sin(\theta)y}{\lambda}\right] -$$
$$J_1\left(a\right)\exp\left\{jk\sqrt{\cos^2(\theta) - \left[\sin(\phi) - \frac{\lambda}{P}\right]^2} z\right\}\exp\left\{j2\pi\left[\frac{\sin(\phi)}{\lambda} - \frac{1}{P}\right]x\right\}\exp\left[j2\pi\frac{\sin(\theta)y}{\lambda}\right],$$

and

$$
\begin{aligned}
I_2(x,y) = \Bigg| & J_0\left(a\right)\exp\left[jk\sqrt{\cos^2(\theta)-\sin^2(\phi)z}\right] + \\
& J_1\left(a\right)\exp\left\{jk\sqrt{\cos^2(\theta)-\left[\sin(\phi)+\frac{\lambda}{P}\right]^2 z}\right\}\exp\left(j2\pi\frac{x}{P}\right) - \\
& J_1\left(a\right)\exp\left\{jk\sqrt{\cos^2(\theta)-\left[\sin(\phi)-\frac{\lambda}{P}\right]^2 z}\right\}\exp\left(j2\pi\frac{x}{P}\right)\Bigg|^2 .
\end{aligned}
$$

Bibliography

[1] Annika MK Enejder, Thomas G Scecina, Jeankun Oh, Martin Hunter, WeiChuan Shih, Slobodan Sasic, Gary L Horowitz, and Michael S Feld. Raman spectroscopy for noninvasive glucose measurements. *Journal of Biomedical Optics*, 10(3): 031114, 2005.

[2] J Lipson, J Bernhardt, U Block, WR Freeman, R Hofmeister, M Hristakeva, T Lenosky, R McNamara, D Petrasek, D Veltkamp, et al. Requirements for calibration in noninvasive glucose monitoring by raman spectroscopy. *Journal of Diabetes Science and Technology*, 3(2):233–241, 2009.

[3] Wei-Chuan Shih, Kate L Bechtel, and Mihailo V Rebec. Noninvasive glucose sensing by transcutaneous raman spectroscopy. *Journal of Biomedical Optics*, 20(5):051036, 2015.

[4] Xu Feng, Austin J Moy, Hieu TM Nguyen, Jason Zhang, Matthew C Fox, Katherine R Sebastian, Jason S Reichenberg, Mia K Markey, and James W Tunnell. Raman active components of skin cancer. *Biomedical Optics Express*, 8 (6):2835–2850, 2017.

[5] Günter Gauglitz and David Steven Moore. *Handbook of Spectroscopy*, volume 2. John Wiley & Sons, 2014.

[6] Michael E Gehm, Scott T McCain, Nikos P Pitsianis, David J Brady, Prasant Potuluri, and Michael E Sullivan. Static two-dimensional aperture coding for multimodal, multiplex spectroscopy. *Applied Optics*, 45(13):2965–2974, 2006.

[7] Etienne Le Coarer, Sylvain Blaize, Pierre Benech, Ilan Stefanon, Alain Morand, Gilles Lérondel, Grégory Leblond, Pierre Kern, Jean Marc Fedeli, and Pascal Royer. Wavelength-scale stationary-wave integrated fourier-transform spectrometry. *Nature Photonics*, 1(8):473, 2007.

[8] Brandon Redding, Seng Fatt Liew, Raktim Sarma, and Hui Cao. Compact spectrometer based on a disordered photonic chip. *Nature Photonics*, 7(9): 746–751, 2013.

[9] Jie Bao and Moungi G Bawendi. A colloidal quantum dot spectrometer. *Nature*, 523(7558):67, 2015.

[10] James D White and Robert E Scholten. Compact diffraction grating laser wavemeter with sub-picometer accuracy and picowatt sensitivity using a webcam imaging sensor. *Review of Scientific Instruments*, 83(11):113104, 2012.

[11] Michael Mazilu, Tom Vettenburg, Andrea Di Falco, and Kishan Dholakia. Random super-prism wavelength meter. *Optics Letters*, 39(1):96–99, 2014.

[12] Nikolaus Klaus Metzger, Roman Spesyvtsev, Graham D Bruce, Bill Miller, Gareth T Maker, Graeme Malcolm, Michael Mazilu, and Kishan Dholakia. Harnessing speckle for a sub-femtometre resolved broadband wavemeter and laser stabilization. *Nature Communications*, 8:ncomms15610, 2017.

[13] Tom Markvart. The thermodynamics of optical étendue. *Journal of Optics A: Pure and Applied Optics*, 10(1):015008, 2007.

[14] Joseph Goodman. *Statistical Optics*. John Wiley & Sons, 2015.

[15] David J Brady. Multiplex sensors and the constant radiance theorem. *Optics Letters*, 27(1):16–18, 2002.

[16] DF Gray, KA Smith, and FB Dunning. Simple compact fizeau wavemeter. *Applied Optics*, 25(8):1339–1343, 1986.

[17] Marek Elbaum. Signal processing for etalon wavemeters. In *Application and Theory of Periodic Structures*, volume 2532, pages 194–209. International Society for Optics and Photonics, 1995.

[18] PJ Fox, RE Scholten, MR Walkiewicz, and Robert E Drullinger. A reliable, compact, and low-cost michelson wavemeter for laser wavelength measurement. *American Journal of Physics*, 67(7):624–630, 1999.

[19] Hira Nasim and Yasir Jamil. Recent advancements in spectroscopy using tunable diode lasers. *Laser Physics Letters*, 10(4):043001, 2013.

[20] N Ismail, L-P Choo-Smith, K Wörhoff, A Driessen, AC Baclig, PJ Caspers, GJ Puppels, RM De Ridder, and Markus Pollnau. Raman spectroscopy with an integrated arrayed-waveguide grating. *Optics Letters*, 36(23):4629–4631, 2011.

[21] Yeonjoon Park, Laura Koch, Kyo D Song, SangJoon Park, Glen King, and Sang Choi. Miniaturization of a fresnel spectrometer. *Journal of Optics A: Pure and Applied Optics*, 10(9):095301, 2008.

[22] Jae Won Hahn, Seung Nam Park, and Chunghi Rhee. Fabry–perot wavemeter for shot-by-shot analysis of pulsed lasers. *Applied Optics*, 32(7):1095–1099, 1993.

[23] ST McCain, ME Gehm, Y Wang, NP Pitsianis, and DJ Brady. Coded aperture raman spectroscopy for quantitative measurements of ethanol in a tissue phantom. *Applied Spectroscopy*, 60(6):663–671, 2006.

[24] Zhaochun Xu, Zhanglei Wang, Michael E Sullivan, David J Brady, Stephen H Foulger, and Ali Adibi. Multimodal multiplex spectroscopy using photonic crystals. *Optics Express*, 11(18):2126–2133, 2003.

[25] Peng Wang and Rajesh Menon. Computational spectrometer based on a broadband diffractive optic. *Optics Express*, 22(12):14575–14587, 2014.

[26] Brandon Redding and Hui Cao. Using a multimode fiber as a high-resolution, low-loss spectrometer. *Optics Letters*, 37(16):3384–3386, 2012.

[27] Noel H Wan, Fan Meng, Tim Schröder, Ren-Jye Shiue, Edward H Chen, and Dirk Englund. High-resolution optical spectroscopy using multimode interference in a compact tapered fibre. *Nature Communications*, 6:7762, 2015.

[28] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. *Applied Optics*, 47(10):B44–B51, 2008.

[29] Per Christian Hansen. *Discrete inverse problems: insight and algorithms*, volume 7. Siam, 2010.

[30] Hui Cao. Perspective on speckle spectrometers. *Journal of Optics*, 19(6):060402, 2017.

[31] Hamamatsu Photonics. Micro-spectrometer c12666ma, Nov. 2017. [Accessed 12-July-2018].

[32] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[33] Harry LT Lee, Paolo Boccazzi, Nathalie Gorret, Rajeev J Ram, and Anthony J Sinskey. In situ bioprocess monitoring of escherichia coli bioreactions using raman spectroscopy. *Vibrational Spectroscopy*, 35(1-2):131–137, 2004.

[34] James N Miller and Jane Charlotte Miller. *Statistics and chemometrics for analytical chemistry*. Pearson Education, 2010.

[35] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[36] Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2): 109–130, 2001.

[37] Tormod Næs and Harald Martens. Principal component regression in nir analysis: viewpoints, background details and selection of components. *Journal of Chemometrics*, 2(2):155–167, 1988.

[38] Federico Marini, R Bucci, AL Magrì, and AD Magrì. Artificial neural networks in chemometrics: History, examples and perspectives. *Microchemical Journal*, 88(2):178–185, 2008.

[39] Richard G Brereton and Gavin R Lloyd. Support vector machines for classification and regression. *Analyst*, 135(2):230–267, 2010.

[40] José M Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2): 354–379, 2012.

[41] Nikos Pasadakis and Andreas A Kardamakis. Identifying constituents in commercial gasoline using fourier transform-infrared spectroscopy and independent component analysis. *Analytica Chimica Acta*, 578(2):250–255, 2006.

[42] José MP Nascimento and José MB Dias. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(4):898–910, 2005.

[43] Lidan Miao and Hairong Qi. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):765–777, 2007.

[44] Satoru Kohno, Ichiro Miyai, Akitoshi Seiyama, Ichiro Oda, Akihiro Ishikawa, Shoichi Tsuneishi, Takahi Amita, and Koji Shimizu. Removal of the skin blood flow artifact in functional near-infrared spectroscopic imaging data through independent component analysis. *Journal of Biomedical Optics*, 12(6):062111, 2007.

[45] Jessica Whelan, Stephen Craven, and Brian Glennon. In situ raman spectroscopy for simultaneous monitoring of multiple process parameters in mammalian cell culture bioreactors. *Biotechnology Progress*, 28(5):1355–1362, 2012.

[46] Jack G Dodd and LK DeNoyer. Curve-fitting: Modeling spectra. *Handbook of Vibrational Spectroscopy*, 2006.

[47] Adrian J Brown. Spectral curve fitting for automatic hyperspectral data analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 44(6):1601–1608, 2006.

[48] Michael S Bradley. Lineshapes in ir and raman spectroscopy: A primer. *Spectroscopy*, 30(11):42–46, 2015.

[49] Michael S Twardowski, Emmanuel Boss, James M Sullivan, and Percy L Donaghay. Modeling the spectral shape of absorption by chromophoric dissolved organic matter. *Marine Chemistry*, 89(1-4):69–88, 2004.

[50] Aaron L Stancik and Eric B Brauns. A simple asymmetric lineshape for fitting infrared absorption spectra. *Vibrational Spectroscopy*, 47(1):66–69, 2008.

[51] DGT Denison, BK Mallick, and AFM Smith. Automatic bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60 (2):333–350, 1998.

[52] R Fischer and V Dose. Analysis of mixtures in physical spectra. In *Monographs of Official Statistics: Bayesian Methods With Applications to Science, Policy, and Official Statistics*, pages 145–154. 2005.

[53] Kenji Nagata, Seiji Sugita, and Masato Okada. Bayesian spectral deconvolution with the exchange monte carlo method. *Neural Networks*, 28:82–89, 2012.

[54] Matthew Moores, Kirsten Gracie, Jake Carson, Karen Faulds, Duncan Graham, and Mark Girolami. Bayesian modelling and quantification of raman spectroscopy. *arXiv preprint arXiv:1604.07299*, 2016.

[55] Satoru Tokuda, Kenji Nagata, and Masato Okada. Simultaneous estimation of noise variance and number of peaks in bayesian spectral deconvolution. *Journal of the Physical Society of Japan*, 86(2):024001, 2016.

[56] S Gulam Razul, WJ Fitzgerald, and C Andrieu. Bayesian model selection and parameter estimation of nuclear emission spectra using rjmcmc. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 497(2):492–510, 2003.

[57] Yuan Wang, Xiaobo Zhou, Honghui Wang, King Li, Lixiu Yao, and Stephen TC Wong. Reversible jump mcmc approach for peak identification for stroke seldi mass spectrometry using mixture model. *Bioinformatics*, 24(13):i407–i413, 2008.

[58] David I Ellis, David P Cowcher, Lorna Ashton, Steve O'Hagan, and Royston Goodacre. Illuminating disease and enlightening biomedicine: Raman spectroscopy as a diagnostic tool. *Analyst*, 138(14):3871–3884, 2013.

[59] Kenny Kong, Catherine Kendall, Nicholas Stone, and Ioan Notingher. Raman spectroscopy for medical diagnosticsfrom in-vitro biofluid assays to in-vivo cancer detection. *Advanced Drug Delivery Reviews*, 89:121–134, 2015.

[60] Nils Kristian Afseth, Vegard Herman Segtnan, and Jens Petter Wold. Raman spectra of biological samples: A study of preprocessing methods. *Applied Spectroscopy*, 60(12):1358–1367, 2006.

[61] Holly J Butler, Lorna Ashton, Benjamin Bird, Gianfelice Cinque, Kelly Curtis, Jennifer Dorney, Karen Esmonde-White, Nigel J Fullwood, Benjamin Gardner, Pierre L Martin-Hirsch, et al. Using raman spectroscopy to characterize biological materials. *Nature Protocols*, 11(4):664–687, 2016.

[62] Kristian Hovde Liland, Trygve Almøy, and Bjørn-Helge Mevik. Optimal choice of baseline correction for multivariate calibration of spectra. *Applied Spectroscopy*, 64(9):1007–1016, 2010.

[63] Juha Kostamovaara, Jussi Tenhunen, Martin Kögler, Ilkka Nissinen, Jan Nissinen, and Pekka Keränen. Fluorescence suppression in raman spectroscopy using a time-gated cmos spad. *Optics Express*, 21(25):31632–31645, 2013.

[64] Meenakshi Gaur, Marek Dobke, and Victoria V Lunyak. Mesenchymal stem cells from adipose tissue in clinical applications for dermatological indications and skin aging. *International Journal of Molecular Sciences*, 18(1):208, 2017.

[65] Steven L Jacques. Optical properties of biological tissues: a review. *Physics in Medicine & Biology*, 58(11):R37, 2013.

[66] Wang Lihong, Steven L Jacques, and Zheng Liqiong. Mcml–monte carlo modeling of light transport in multi-layered tissues. *Computer Methods and Programs in Biomedicine*, 2(47):131–146, 1995.

[67] PJ Caspers, GW Lucassen, R Wolthuis, HA Bruining, and GJ Puppels. In vitro and in vivo raman spectroscopy of human skin. *Biospectroscopy*, 4(S5):S31–S39, 1998.

[68] Peter J Caspers, Hajo A Bruining, Gerwin J Puppels, Gerald W Lucassen, and Elizabeth A Carter. In vivo confocal raman microspectroscopy of the skin: noninvasive determination of molecular concentration profiles. *Journal of Investigative Dermatology*, 116(3):434–442, 2001.

[69] Jianhua Zhao, Harvey Lui, David I McLean, and Haishan Zeng. Integrated real-time raman system for clinical in vivo skin analysis. *Skin Research and Technology*, 14(4):484–492, 2008.

[70] Harvey Lui, Jianhua Zhao, David I McLean, and Haishan Zeng. Real-time raman spectroscopy for in vivo skin cancer diagnosis. *Cancer Research*, pages canres–4061, 2012.

[71] ME Darvin, I Gersonde, M Meinke, W Sterry, and J Lademann. Non-invasive in vivo determination of the carotenoids beta-carotene and lycopene concentrations in the human skin using the raman spectroscopic method. *Journal of Physics D: Applied Physics*, 38(15):2696, 2005.

[72] Monika Gniadecka, Hans C Wulf, Ole F Nielsen, Daniel H Christensen, and Jana Hercogova. Distinctive molecular abnormalities in benign and malignant skin lesions: studies by raman spectroscopy. *Photochemistry and Photobiology*, 66(4):418–423, 1997.

[73] Chad A Lieber, Shovan K Majumder, Darrel L Ellis, D Dean Billheimer, and Anita Mahadevan-Jansen. In vivo nonmelanoma skin cancer diagnosis using

raman microspectroscopy. *Lasers in Surgery and Medicine: The Official Journal of the American Society for Laser Medicine and Surgery*, 40(7):461–467, 2008.

[74] Igor V Ermakov, Maia R Ermakova, Werner Gellermann, and Jürgen Lademann. Noninvasive selective detection of lycopene and β-carotene in human skin using raman spectroscopy. *Journal of Biomedical Optics*, 9(2):332–339, 2004.

[75] Juergen Lademann, Martina C Meinke, Wolfram Sterry, and Maxim E Darvin. Carotenoids in human skin. *Experimental Dermatology*, 20(5):377–382, 2011.

[76] Robbert Meerwaldt, Jasper WL Hartog, Reindert Graaff, Roel J Huisman, Thera P Links, Nynke C den Hollander, Susan R Thorpe, John W Baynes, Gerjan Navis, Rijk OB Gans, et al. Skin autofluorescence, a measure of cumulative metabolic stress and advanced glycation end products, predicts mortality in hemodialysis patients. *Journal of the American Society of Nephrology*, 16(12): 3687–3693, 2005.

[77] Robbert Meerwaldt, Helen L Lutgers, Thera P Links, Reindert Graaff, John W Baynes, Rijk OB Gans, and Andries J Smit. Skin autofluorescence is a strong predictor of cardiac mortality in diabetes. *Diabetes Care*, 30(1):107–112, 2007.

[78] Helen L Lutgers, Reindert Graaff, Thera P Links, Lielith J Ubink-Veltmaat, Henk J Bilo, Rijk O Gans, and Andries J Smit. Skin autofluorescence as a noninvasive marker of vascular damage in patients with type 2 diabetes. *Diabetes Care*, 29(12):2654–2659, 2006.

[79] AZAN Antoine, Peter J Caspers, Tom C Bakker Schut, Séverine Roy, Céline Boutros, Christine Mateus, Emilie Routier, Benjamin Besse, David Planchard, Atmane Seck, et al. A novel spectroscopically determined pharmacodynamic biomarker for skin toxicity in cancer patients treated with targeted agents. *Cancer Research*, pages canres–1733, 2016.

[80] Santhisagar Vaddiraju, Diane J Burgess, Ioannis Tomazos, Faquir C Jain, and Fotios Papadimitrakopoulos. Technologies for continuous glucose monitoring: current problems and future promises. *Journal of Diabetes Science and Technology*, 4(6):1540–1562, 2010.

[81] Sandeep Kumar Vashist. Non-invasive glucose monitoring technology in diabetes management: A review. *Analytica Chimica Acta*, 750:16–27, 2012.

[82] John L. Smith. *The pursuit of noninvasive glucose: "hunting the deceitful turkey"*. 2017.

[83] Martina Sattlecker, Nicholas Stone, and Conrad Bessant. Current trends in machine-learning methods applied to spectroscopic cancer diagnosis. *Trends in Analytical Chemistry*, 59:17–25, 2014.

[84] Henry Fox Talbot. Lxxvi. facts relating to optical science. no. iv. *The London and Edinburgh Philosophical Magazine and Journal of Science*, 9(56):401–407, 1836.

[85] Helen L Kung, Aparna Bhatnagar, and David AB Miller. Transform spectrometer based on measuring the periodicity of talbot self-images. *Optics Letters*, 26(21): 1645–1647, 2001.

[86] Sergio De Nicola, Pietro Ferraro, Giuseppe Coppola, Andrea Finizio, Giovanni Pierattini, and Simonetta Grilli. Talbot self-image effect in digital holography and its application to spectrometry. *Optics Letters*, 29(1):104–106, 2004.

[87] Erika Ye, Amir H Atabaki, Ningren Han, and Rajeev J Ram. Miniature, subnanometer resolution talbot spectrometer. *Optics Letters*, 41(11):2434–2437, 2016.

[88] Ningren Han, Seong-Ho Cho, Amir Atabaki, Erika Ye, William Herrington, and Rajeev Ram. Non-paraxial talbot effect for building compact spectrometers. In *Computational Optical Sensing and Imaging*, pages CM2B–2. Optical Society of America, 2016.

[89] Andreas F Ruckstuhl, Matthew P Jacobson, Robert W Field, and James A Dodd. Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 68(2):179–193, 2001.

[90] Chad A Lieber and Anita Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological raman spectra. *Applied Spectroscopy*, 57(11): 1363–1367, 2003.

[91] Jianhua Zhao, Harvey Lui, David I McLean, and Haishan Zeng. Automated autofluorescence background subtraction algorithm for biomedical raman spectroscopy. *Applied Spectroscopy*, 61(11):1225–1232, 2007.

[92] CM Galloway, EC Le Ru, and PG Etchegoin. An iterative algorithm for background removal in spectroscopy by wavelet transforms. *Applied Spectroscopy*, 63(12):1370–1376, 2009.

[93] Johan J de Rooi and Paul HC Eilers. Mixture models for baseline estimation. *Chemometrics and Intelligent Laboratory Systems*, 117:56–60, 2012.

[94] Shixuan He, Wei Zhang, Lijuan Liu, Yu Huang, Jiming He, Wanyi Xie, Peng Wu, and Chunlei Du. Baseline correction for raman spectra using an improved asymmetric least squares method. *Analytical Methods*, 6(12):4402–4407, 2014.

[95] Mingjun Zhong, Mark Girolami, Karen Faulds, and Duncan Graham. Bayesian methods to detect dye-labelled dna oligonucleotides in multiplexed raman spectra. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(2): 187–206, 2011.

[96] Ningren Han and Rajeev J Ram. Bayesian modeling and computation for analyte quantification in complex mixtures using raman spectroscopy. *arXiv preprint arXiv:1805.07688*, 2018.

[97] David George Voelz. *Computational fourier optics: a MATLAB tutorial.* SPIE press Bellingham, WA, 2011.

[98] Joseph Goodman. *Introduction to Fourier optics.* McGraw-hill, 2008.

[99] Yih-Shyang Cheng and Ray-Chung Chang. Theory of image formation using the talbot effect. *Applied Optics*, 33(10):1863–1874, 1994.

[100] Lord Rayleigh. Xxv. on copying diffraction-gratings, and on some phenomena connected therewith. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 11(67):196–205, 1881.

[101] Grant R Fowles. *Introduction to modern optics.* Courier Corporation, 1975.

[102] ePHOTOzine. Complete guide to image sensor pixel size. https://www.ephotozine.com/article/complete-guide-to-image-sensor-pixel-size-29652, 2016. [Online; accessed 11-April-2018].

[103] Chang-Rok Moon, Duck-Hyung Lee, and Seong-Ho Cho. Cmos image sensors including backside illumination structure, April 24 2012. US Patent 8,164,126.

[104] Derek Tseng, Onur Mudanyali, Cetin Oztoprak, Serhan O Isikman, Ikbal Sencan, Oguzhan Yaglidere, and Aydogan Ozcan. Lensfree microscopy on a cellphone. *Lab on a Chip*, 10(14):1787–1792, 2010.

[105] Seung Ah Lee and Changhuei Yang. A smartphone-based chip-scale microscope using ambient illumination. *Lab on a Chip*, 14(16):3056–3063, 2014.

[106] Hongying Zhu, Sam Mavandadi, Ahmet F Coskun, Oguzhan Yaglidere, and Aydogan Ozcan. Optofluidic fluorescent imaging cytometry on a cell phone. *Analytical Chemistry*, 83(17):6641–6647, 2011.

[107] Stephen A Benton and V Michael Bove Jr. *Holographic imaging.* John Wiley & Sons, 2008.

[108] Thorlabs. X-cite light sources. https://www.thorlabs.com/newgrouppage9.cfm?objectgroup_id=10092, 2018. [Online; accessed 12-April-2018].

[109] Freek van der Meer. The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery. *International Journal of Applied Earth Observation and Geoinformation*, 8(1):3–17, 2006.

[110] Alon Greenbaum, Wei Luo, Ting-Wei Su, Zoltán Göröcs, Liang Xue, Serhan O Isikman, Ahmet F Coskun, Onur Mudanyali, and Aydogan Ozcan. Imaging without lenses: achievements and remaining challenges of wide-field on-chip microscopy. *Nature Methods*, 9(9):889, 2012.

[111] DCBP Rife and Robert Boorstyn. Single tone parameter estimation from discrete-time observations. *IEEE Transactions on Information Theory*, 20(5): 591–598, 1974.

[112] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[113] Richard Roy and Thomas Kailath. Esprit-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):984–995, 1989.

[114] J Smith and Xavier Serra. Parshl an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. In *Proceedings of the 1987 International Computer Music Conference, ICMC*, 1987.

[115] Alain De Cheveigné and Hideki Kȧwahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[116] Phil Gregory. *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica Support.* Cambridge University Press, 2005.

[117] R.O. Schmidt. A signal subspace approach to multiple emitter location and spectral estimation. *Ph.D. Thesis, Stanford University*, 1981.

[118] Boualem Boashash. Estimating and interpreting the instantaneous frequency of a signal. ii. algorithms and applications. *Proceedings of the IEEE*, 80(4):540–568, 1992.

[119] Sailes K Sengijpta. Fundamentals of statistical signal processing: estimation theory, 1995.

[120] AJ Barabell, J Capon, DF DeLong, JR Johnson, and KD Senne. Performance comparison of superresolution array processing algorithms. revised. Technical report, Massachusetts Institute of Technology, Lincoln Laboratory, 1998.

[121] Matteo Frigo and Steven G Johnson. The design and implementation of fftw3. *Proceedings of the IEEE*, 93(2):216–231, 2005.

[122] GK Wertheim, MA Butler, KW West, and DNE Buchanan. Determination of the gaussian and lorentzian content of experimental line shapes. *Review of Scientific Instruments*, 45(11):1369–1371, 1974.

[123] Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

[124] Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, 6:233–243, 1986.

[125] Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

[126] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su, et al. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

[127] Larry Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107, 2000.

[128] Peter J Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[129] Sylvia Richardson and Peter J Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997.

[130] David I Hastie and Peter J Green. Model choice using reversible jump markov chain monte carlo. *Statistica Neerlandica*, 66(3):309–338, 2012.

[131] Ajay Jasra, Chris C Holmes, and David A Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.

[132] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

[133] Peter JM Van Laarhoven and Emile HL Aarts. Simulated annealing. In *Simulated Annealing: Theory and Applications*, pages 7–15. Springer, 1987.

[134] Rolf Wolthuis, Gilbert CH Tjiang, Gerwin J Puppels, and Tom C Bakker Schut. Estimating the influence of experimental parameters on the prediction error of pls calibration models based on raman spectra. *Journal of Raman Spectroscopy*, 37(1-3):447–466, 2006.

[135] J Gerbrand Mesu, Tom Visser, Fouad Soulimani, and Bert M Weckhuysen. Infrared and raman spectroscopic study of ph-induced structural changes of l-histidine in aqueous environment. *Vibrational Spectroscopy*, 39(1):114–125, 2005.

[136] Gajendra P Singh, Shireen Goh, Michelangelo Canzoneri, and Rajeev J Ram. Raman spectroscopy of complex defined media: biopharmaceutical applications. *Journal of Raman Spectroscopy*, 46(6):545–550, 2015.

[137] Laser Institute of America. American national standard for safe use of lasers: Ansi z136. 1–2000, 2007.

[138] Andrew P Shreve, Nerine J Cherepy, and Richard A Mathies. Effective rejection of fluorescence interference in raman spectroscopy using a shifted excitation difference technique. *Applied Spectroscopy*, 46(4):707–711, 1992.

[139] Michael D Morris, Pavel Matousek, Michael Towrie, Anthony W Parker, Allen E Goodship, and Edward RC Draper. Kerr-gated time-resolved raman spectroscopy of equine cortical bone tissue. *Journal of Biomedical Optics*, 10(1):014014, 2005.

[140] MJ Wirth and Shiow Hwa Chou. Comparison of time and frequency domain methods for rejecting fluorescence from raman spectra. *Analytical Chemistry*, 60(18):1882–1886, 1988.

[141] Dong Wei, Shuo Chen, and Quan Liu. Review of fluorescence suppression techniques in raman spectroscopy. *Applied Spectroscopy Reviews*, 50(5):387–406, 2015.

[142] Bin Deng, Colin Wright, Eric Lewis-Clark, G Shaheen, Roman Geier, and J Chaiken. Direct noninvasive observation of near infrared photobleaching of autofluorescence in human volar side fingertips in vivo. In *Biomedical Vibrational Spectroscopy IV: Advances in Research and Industry*, volume 7560, page 75600P. International Society for Optics and Photonics, 2010.

[143] Johannes Schleusener, Jürgen Lademann, and Maxim E Darvin. Depth-dependent autofluorescence photobleaching using 325, 473, 633, and 785 nm of porcine ear skin ex vivo. *Journal of Biomedical Optics*, 22(9):091503, 2017.

[144] Franck Bonnier, Syed Mehmood Ali, Peter Knief, Helen Lambkin, Kathleen Flynn, Vincent McDonagh, Claragh Healy, TC Lee, Fiona M Lyng, and Hugh J Byrne. Analysis of human skin tissue by raman microspectroscopy: dealing with the background. *Vibrational Spectroscopy*, 61:124–132, 2012.

[145] Joannie Desroches, Michael Jermyn, Kelvin Mok, Cédric Lemieux-Leduc, Jeanne Mercier, Karl St-Arnaud, Kirk Urmey, Marie-Christine Guiot, Eric Marple, Kevin Petrecca, et al. Characterization of a raman spectroscopy probe system for intraoperative brain tissue classification. *Biomedical Optics Express*, 6(7): 2380–2397, 2015.

[146] Hequn Wang, Jianhua Zhao, Anthony MD Lee, Harvey Lui, and Haishan Zeng. Improving skin raman spectral quality by fluorescence photobleaching. *Photodiagnosis and Photodynamic Therapy*, 9(4):299–302, 2012.

[147] Michael Jermyn, Joannie Desroches, Jeanne Mercier, Marie-Andrée Tremblay, Karl St-Arnaud, Marie-Christine Guiot, Kevin Petrecca, and Frederic Leblond. Neural networks improve brain cancer detection with raman spectroscopy in the presence of operating room light artifacts. *Journal of Biomedical Optics*, 21(9): 094002, 2016.

[148] M Newberry. Tech note: pixel response effects on ccd camera gain calibration. *Mirametrics Inc., Copy Right*, 1998.

[149] Chae-Ryon Kong, Ishan Barman, Narahara Chari Dingari, Jeon Woong Kang, Luis Galindo, Ramachandra R Dasari, and Michael S Feld. A novel non-imaging optics based raman spectroscopy device for transdermal blood analyte measurement. *AIP Advances*, 1(3):032175, 2011.

[150] Carina Reble, Ingo H Gersonde, Stefan Andree, Hans-Joachim Eichler, and Jürgen Helfmann. Quantitative raman spectroscopy in turbid media. *Journal of Biomedical Optics*, 15(3):037016, 2010.

[151] Wei-Chuan Shih, Kate L Bechtel, and Michael S Feld. Intrinsic raman spectroscopy for quantitative biological spectroscopy part i: theory and simulations. *Optics Express*, 16(17):12726–12736, 2008.

[152] Kate L Bechtel, Wei-Chuan Shih, and Michael S Feld. Intrinsic raman spectroscopy for quantitative biological spectroscopy part ii: experimental applications. *Optics Express*, 16(17):12737–12745, 2008.

[153] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-363-8.

[154] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107, 2010.

[155] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

[156] William L Clarke, Daniel Cox, Linda A Gonder-Frederick, William Carter, and Stephen L Pohl. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care*, 10(5):622–628, 1987.

[157] Joan L Parkes, Stephen L Slatin, Scott Pardo, and Barry H Ginsberg. A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose. *Diabetes Care*, 23(8):1143–1148, 2000.

[158] Andrew J Berger, Tae-Woong Koo, Irving Itzkan, Gary Horowitz, and Michael S Feld. Multicomponent blood analysis by near-infrared raman spectroscopy. *Applied Optics*, 38(13):2916–2926, 1999.

[159] Albert Wang, Patrick Gill, and Alyosha Molnar. Light field image sensors based on the talbot effect. *Applied Optics*, 48(31):5897–5905, 2009.

[160] Albert Wang, Patrick R Gill, and Alyosha Molnar. An angle-sensitive cmos imager for single-sensor 3d photography. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, pages 412–414. IEEE, 2011.

[161] Christian P Robert, Víctor Elvira, Nick Tawn, and Changye Wu. Accelerating mcmc algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1435, 2018.

[162] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pages 1218–1226, 2015.

[163] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112 (518):859–877, 2017.

[164] Emad L Izake. Forensic and homeland security applications of modern portable raman spectroscopy. *Forensic Science International*, 202(1-3):1–8, 2010.

[165] Michael Kemmler, Erik Rodner, Petra Rösch, Jürgen Popp, and Joachim Denzler. Automatic identification of novel bacteria using raman spectroscopy and gaussian processes. *Analytica Chimica Acta*, 794:29–37, 2013.

[166] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.