ACTIVE WAVELENGTH SELECTION FOR CHEMICAL IDENTIFICATION USING TUNABLE SPECTROSCOPY

A Dissertation

by

JIN HUANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ricardo Gutierrez-Osuna
Committee Members,	Yoonsuck Choe
	Dylan Shell
	Suman Chakravorty
Head of Department,	Dilma Da Silva

May 2016

Major Subject: Computer Engineering

ABSTRACT

Spectrometers are the cornerstone of analytical chemistry. Recent advances in micro-optics manufacturing provide lightweight and portable alternatives to traditional spectrometers. In this dissertation, we developed a spectrometer based on Fabry-Perot interferometers (FPIs). A FPI is a tunable (it can only scan one wavelength at a time) optical filter. However, compared to its traditional counterparts such as FTIR (Fourier transform infrared spectroscopy), FPIs provide lower resolution and lower signal-noise-ratio (SNR). Wavelength selection can help alleviate these drawbacks. Eliminating uninformative wavelengths not only speeds up the sensing process but also helps improve accuracy by avoiding nonlinearity and noise. Traditional wavelength selection algorithms follow a training-validation process, and thus they are only optimal for the target analyte. However, for chemical identification, the identities are unknown.

To address the above issue, this dissertation proposes active sensing algorithms that select wavelengths online while sensing. These algorithms are able to generate analyte-dependent wavelengths. We envision this algorithm deployed on a portable chemical gas platform that has low-cost sensors and limited computation resources. We develop three algorithms focusing on three different aspects of the chemical identification problems.

First, we consider the problem of single chemical identification. We formulate the problem as a typical classification problem where each chemical is considered as a distinct class. We use Bayesian risk as the utility function for wavelength selection, which calculates the misclassification cost between classes (chemicals), and we select the wavelength with the maximum reduction in the risk. We evaluate this approach on both synthesized and experimental data. The results suggest that active sensing outperforms the passive method, especially in a noisy environment.

Second, we consider the problem of chemical mixture identification. Since the number of potential chemical mixtures grows exponentially as the number of components increases, it is intractable to formulate all potential mixtures as classes. To circumvent combinatorial explosion, we developed a multi-modal non-negative least squares (MM-NNLS) method that searches multiple near-optimal solutions as an approximation of all the solutions. We project the solutions onto spectral space, calculate the variance of the projected spectra at each wavelength, and select the next wavelength using the variance as the guidance. We validate this approach on synthesized and experimental data. The results suggest that active approaches are superior to their passive counterparts especially when the condition number of the mixture grows larger (the analytes consist of more components, or the constituent spectra are very similar to each other).

Third, we consider improving the computational speed for chemical mixture identification. MM-NNLS scales poorly as the chemical mixture becomes more complex. Therefore, we develop a wavelength selection method based on Gaussian process regression (GPR). GPR aims to reconstruct the spectrum rather than solving the mixture problem, thus, its computational cost is a function of the number of wavelengths. We evaluate the approach on both synthesized and experimental data. The results again demonstrate more accurate and robust performance in contrast to passive algorithms.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Ricardo Gutierrez-Osuna for his support. Academically, his guidance not only inspired this dissertation, but also drove me to become a better researcher. I am especially grateful for his constant guidance and unwavering encouragement for high-quality writing, which helped me to get where I am today.

Without the financial support of the National Science Foundation that provided the scholarship for my graduate studies, this work would not have been possible. I am grateful to Dr. Gutierrez for granting me this rare opportunity.

I am also grateful to my committee members Dr. Yoonsuck Choe, Dr. Dylan Shell, and Dr. Suman Chakravorty for their assistance and feedback on this project. My appreciation also extends to my past and present colleagues; special thanks to Rakesh Gosangi, Sandesh Aryal, Avinash Parnandi, and Joseph Lee for their support and friendship, which helped me survive those stressful times.

Finally, I would like to thank my mother and father for the unconditional love.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

Compact tunable chemical sensors based on Fabry-Perot interferometry have recently become available [1, 2], offering the prospect of low-cost, portable embedded spectroscopy for chemical identification and quantitative analysis. However, compared to traditional spectroscopy such as Fourier transform infrared spectroscopy (FTIR), these compact tunable sensors have lower sensitivity and resolution, which leads to higher sensor noise, greater nonlinearity, and greater collinearity. Wavelength selection can help alleviate these problems; it has been shown both theoretically [3] and experimentally [4-8] that by removing uninformative wavelengths, prediction accuracy can be improved. Additionally, since tunable sensors can only scan one spectral line at a time, wavelength selection can significantly speed up the sensing process by avoiding non-informative wavelengths.

Given the combinatorial complexity of the wavelength selection problem, an efficient searching algorithm is crucial to make the process computationally tractable. Several approaches have been proposed in the literature, including various randomized algorithms such as genetic algorithms [8], simulated annealing [7], colony optimization [9], as well as greedy strategies. A noteworthy greedy search technique is the successive projection method of Araújo et al. [5], which extracts wavelengths that minimize collinearity using the sequential orthogonal projections of the Gram-Schmidt procedure. To further reduce the search space, a common technique is to group wavelengths into individual non-overlapping windows, as in the changeable size moving window scheme proposed by Du et al. [10].

Notwithstanding the effectiveness of these wavelength selection algorithms, their performance is limited by the fact that wavelengths are selected offline using a subset of all possible mixtures to which the device may later be exposed. For mixture analysis problems of even moderate size (e.g., tens of potential chemicals), and in the absence of prior knowledge of the most likely components (and possibly their relative concentrations), the search will generally produce wavelength subsets that are either highly redundant or too specific to the particular mixtures in the training set.

To address the limitations of these methods, this dissertation develops an adaptive algorithm framework that interleaves the wavelength-selection and sensing processes. The new method allows the sensor to adapt its wavelength selection program in response to different chemical stimuli, their concentrations, and to environmental influences, such as sensor noise. We apply the active wavelength selection algorithm to Fabry-Perot interferometry on a chemical gas identification problem to experimentally validate the effectiveness of this framework.

1.1 Contributions

There are three major aims in this dissertation:

- To develop a platform based on the tunable FPI sensor to identify chemical mixtures
- To develop active sensing strategies for chemical mixture identification purposes

- To develop fast and accurate solvers for the multi-component analysis problem

To achieve these aims, this dissertation contains the following contributions:

- An active tunable spectrometer based on FPI sensor including both the hardware and software.
- Three different active wavelength selection algorithms for chemical mixture identification, and we validated them on the FPI platform.
- Two least squares solvers for the multi-component analysis problem: the multi-modal non-negative least squares solver (MM-NNLS), and the BIC shrinkage non-negative least squares solver (BICS-bNNLS).

1.2 Organization of the dissertation

We organize this dissertation as follows: Chapter 2 provides a background of the analytical chemistry, the multicomponent analysis problem, and the wavelength selection problem. Chapter 3 provides a literature review for multicomponent analysis, wavelength selection, and the active wavelength selection problem. Chapters 4 through Chapter 7 explain the three major contributions of this dissertation. First, Chapter 4 describes the active framework for single chemical identification based on Bayesian risk. Second, Chapter 5 explains the active sensing algorithm for chemical mixture identification based on MM-NNLS. Third, Chapters 6 and 7 present the active sensing algorithm that focuses on reducing the computational cost. There are two main aspects of this new development: a faster NNLS solver and a computationally simpler wavelength selection utility. Both aspects are equally important, but also cumbersome to fit into one chapter, so we split it across two: Chapter 6 focuses on the nonlinear BICS-bNNLS for the underlying multi-component analysis problem, and Chapter 7 focuses on the active wavelength selection algorithm. The final chapter, Chapter 8, reviews the contributions and discusses future directions.

2. BACKGROUND

This chapter discusses topics regarding analytical chemistry, such as the physical instruments and techniques used to analyze chemicals. We review various spectroscopic techniques to help readers grasp the scope of the area (Section 2.1). We then discuss absorption spectroscopy and its relation to the underlying numerical problem – multicomponent analysis (Section 2.2). After introducing multicomponent analysis, we explain the benefit of wavelength selection for multicomponent analysis, and we review various optical instruments for wavelength selection (Section 2.3). Finally, we motivate the active wavelength selection and formulate the problem (Section 2.4).

2.1 Analytical chemistry

Analytical chemistry is an interdisciplinary science that has a wide range of topics. As a theoretical science, it studies the molecular structure of different chemicals; as a practical science, it provides qualitative and quantitative information of natural or artificial analytes. Structural analysis studies the actual physical arrangement of the atoms in a molecule; qualitative analysis identifies the species of the atoms, molecules, or biomolecules in the analytes; quantitative analysis provides numerical information for each component present in the analyte. Analytical chemistry finds itself in a wide range of applications, such as forensics, archeology, medicine, food and agriculture, environment, industry, material science, and space science. Given the widespread use of analytical chemistry, many varied laboratory techniques have been developed to analyze or measure analytes. The two most common categories of techniques are physical separation and spectroscopy.

2.1.1 Physical separation

In the domain of analytic chemistry, physical separation is normally referred as Chromatography. Chromatography is a set of techniques that physically separate a particular analyte of interest from potential interferents. Modern chromatography can be traced back to the beginning of 20th century when Russian botanist Mikhail Tsweet developed column chromatography. Column chromatography separates chemicals by exploiting the chemical property that different constituents have distinct affinities to different media (solvent). There are two phases (both are solvents) in the process of chromatography: the *mobile phase* refers to the solvent that serves as a carrier for the analyte; the dissolved compounds are then pushed through a medium called *stationary phase*. If a certain constituent has a greater affinity for the stationary phase than it has for the mobile phase, it moves through the medium slowly. If the constituent has less affinity with the stationary phase, it moves through the media quickly. Due to these various rates of migration for different constituents, the mixture is separated physically. Figure 1 illustrates the process. Chromatography was later re-introduced for biomedical separation during the 1930s and its underlying theory, countercurrent extraction, was later established by Martin and Synge [11] during 1940s.

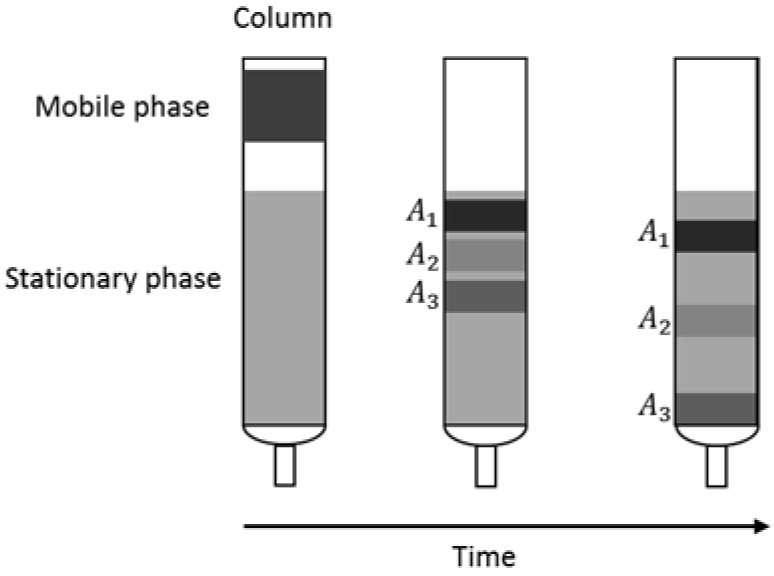
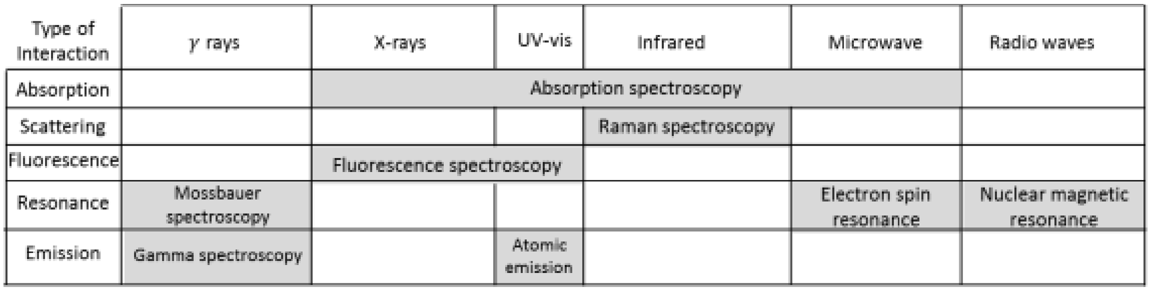


Figure 1: The separation process of column chromatography.

2.1.2 Spectroscopic techniques

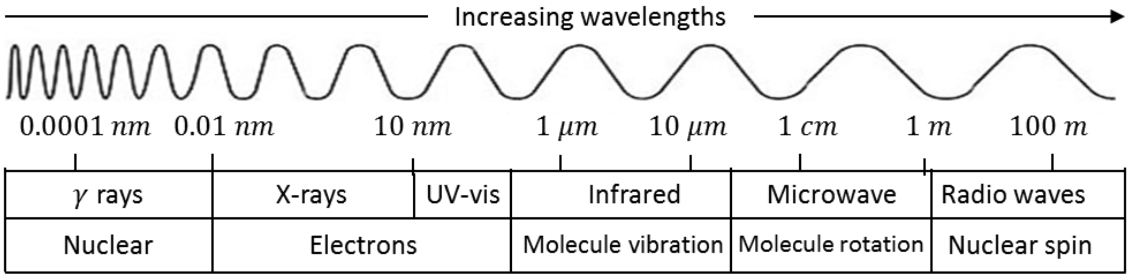
Spectroscopic techniques are the cornerstones of modern analytic chemistry. In general, electromagnetic radiation interacts with molecules in various ways depending on the wavelength of interest. Interactions such as absorbing, emitting, resonating, scattering, and exciting in turn generate or change the radiation intensity at different wavelengths. Measuring the radiation intensity at different wavelengths after the aforementioned interactions is the general principle of spectroscopy. This provides information about the molecule, such as its structure, weight, identity, species of chemical bonds, and quantity of a certain element. Some common spectroscopic methods are summarized in a table, as shown in Figure 2. The first dimension in the table is the wavelengths of interest, which are grouped into radio waves, microwave, infrared, visible, and ultra-violet, x-ray, and γ rays. The second dimension in the table is the interaction principles between the electromagnetic radiation and the analytes. The most popular mechanisms are absorption, scattering, fluorescence, emission, and nuclear magnetic resonance (NMR).



Type of Interaction	γ rays	X-rays	UV-vis	Infrared	Microwave	Radio waves
Absorption		Absorption spectroscopy				
Scattering				Raman spectroscopy		
Fluorescence		Fluorescence spectroscopy				
Resonance	Mossbauer spectroscopy				Electron spin resonance	Nuclear magnetic resonance
Emission	Gamma spectroscopy		Atomic emission			

Figure 2: Summary of different spectroscopic techniques [12]. There are two dimensions in this table. Different columns group techniques into different wavelength ranges following an incremental order from left to right. Different rows represent different electromagnetic radiation interaction principles.

The wavelengths of interest provide different information about the analyte. At the leftmost of the spectrum (highest energy), Mossbauer spectroscopy studies the nuclear structure with the absorption and re-emission of γ rays. At the next level of energy, X-rays and ultraviolet-visible light provide information about electrons. X-rays, at a higher energy level, are more related to the core electrons, whereas ultraviolet and visible light are more related to valence electrons. Core electrons do not participate in bonding, while valence electrons do. Infrared and microwaves provide information about the larger structure, the molecules. Infrared is related to molecule vibration energy and microwaves are linked to molecule rotation energy. Molecule vibration refers to the numerous kinds of vibration of different atomic bonds. Molecule rotation refers to the actual spinning of the whole molecule. At the lowest energy level, radio waves interact with the nuclear spinning and provide information of the atomic bonds in which the target nucleus is involved. Figure 3 summarizes this relationship between spectral regions and the target properties. It is beyond the scope of this dissertation to discuss all spectroscopic techniques in detail. Neverthless, we provide a brief explanation of the most common technologies: absorption spectroscopy, Raman spectroscopy, nuclear magnetic resonance spectroscopy, and mass spectroscopy.



γ rays	X-rays	UV-vis	Infrared	Microwave	Radio waves
Nuclear	Electrons		Molecule vibration	Molecule rotation	Nuclear spin

Figure 3: The wavelengths of interest and the corresponding transition type of the sample

Absorption spectroscopy techniques have the longest history and cover the widest range of wavelengths (X-rays, UV-visible light, infrared, microwave). The instruments and technologies vary significantly depending on the interested spectral region and the state of the sample (such as gas, liquid, solid). However, they share the same general principle: the electromagnetic radiation shines through the analyte; part of the radiation is absorbed by the analyte; the remaining radiation is then measured by a detector placed at the other end. The arrangement is shown in Figure 4.

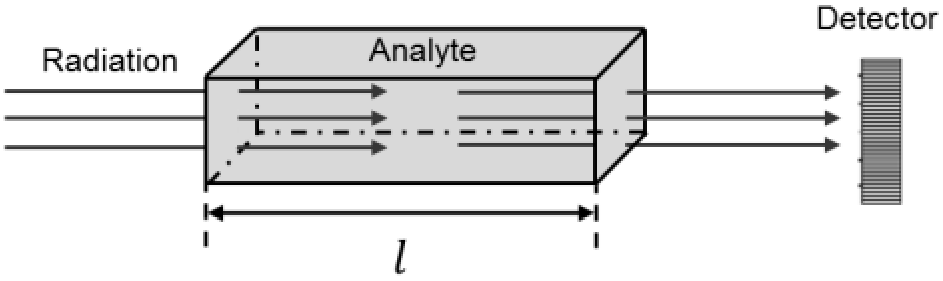


Figure 4: A diagram of the absorption spectrum. l is the length of the effective path of the absorbance.

Raman spectroscopy is another common spectroscopic method and is considered complementary to infrared absorption spectroscopy. While both infrared absorption spectroscopy and Raman spectroscopy interact with the molecules through vibration, certain kinds of vibrations are either Raman active or infrared active, but not both. This mutual exclusion principle makes Raman spectroscopy and infrared absorption spectroscopy complementary to each other. In Raman spectroscopy, a monochromatic light, i.e., laser, shines through the sample, and a very small amount of light is scattered with a slightly shifted frequency, a phenomenon known as Raman scattering. The relationship between the intensity of the Raman scattered light and the shifted frequency reveals some chemical bonds hidden in absorbance spectroscopy. Compared to infrared spectroscopy, Raman spectroscopy is more costly but offers complementary peaks in the spectrum. Raman spectroscopy also has an easier sample preparation process and thus is a suitable solution for portable applications.

The state of the art and the most recent addition to the spectroscopic methods is nuclear magnetic resonance (NMR) spectroscopy, a technology that leads to magnetic resonance imaging (MRI), which widely used in medical applications. The physical principle behind NMR spectroscopy involves the nuclei spin and its precession frequency under a uniform magnetic field. Precession describes the behavior of a gyro spinning under the influence of gravity. Because of gravity, unless the main axis of the gyro is parallel with the direction of gravity, there is a secondary spin, more precisely a wobbling, where the main axis of the gyro is rotating around the direction of gravity. Under a uniform magnetic field, a certain frequency radio wave can excite the gyro from a lower energy level to a higher one. This resonance frequency is a function of the local intensity of the magnetic field that is slightly modified by the surrounding environment of the molecule. Such slight deviations provide clues about the configuration of the molecules, thus, it can be used to deduce the overall structure of a molecule. Since only nuclei with an odd number of protons can interact with this magnetic field, there are two types of commercial NMR - proton NMR (H^1) and C^{13} NMR.

Mass spectrometry is a method that is very commonly coupled with most of the spectroscopic methods. Unlike in the general definition of spectroscopies where electromagnetic radiation plays a crucial part, in mass spectroscopy, the analyte does not interact with the electromagnetic radiation. Rather, the analyte is first ionized, accelerated, and shot through a magnetic field. Ionization separates the analyte molecule into different charged fragments. Different fragments deflect differently in the magnetic field, resulting in fragments distributed continuously in space depending on their mass-to-charge ratios. Manipulating the strength of the magnetic field redirects the different fragments to the fixed-point detector at the other end of the field. As a result, we can plot the abundance of the different ionized fragments with a different mass-to-charge ratio. Mass spectrum may be used to calculate the exact weight of a molecule. In addition, because each molecule has different fragmentation pattern, it also provides some qualitative and quantitative information about the analyte.

2.2 Absorbance and Beer's law

The main domain on which this dissertation focuses is mixture analysis based on absorption spectroscopy. Therefore, this section provides a detailed description of the underlying mathematical problem. It first explains the fundamental theoretical basis, Beer's law, and then extends it for multicomponent analysis.

2.2.1 Transmittance and absorbance

By comparing the partially absorbed radiation intensity I with the reference intensity measured without the analyte I_0 (which can be measured beforehand or after purging the sample cell), we can calculate transmittance in either the transmittance representation or the percent transmittance representation:

$$\begin{gathered} T = \frac{I}{I_0}\ `s \\ \text{or } \%T = \frac{I}{I_0} \times 100\ . \end{gathered} \tag{1}$$

The transmittance can be mapped to a more intuitive measure, absorption, which is how much energy the analyte absorbs:

$$A = log_{10}\frac{1}{T} \tag{2}$$

$$\text{or } A = 2 - log_{10}\,\%T \ \ .$$

Figure 5 illustrates a typical transformation from the raw radiation intensity to the absorbance. The relationship between the absorption intensities and different wavelengths is the absorption spectrum. Since every different chemical often has signature absorption peaks at a certain range of wavelengths, the absorbance provides qualitative information of the analyte.

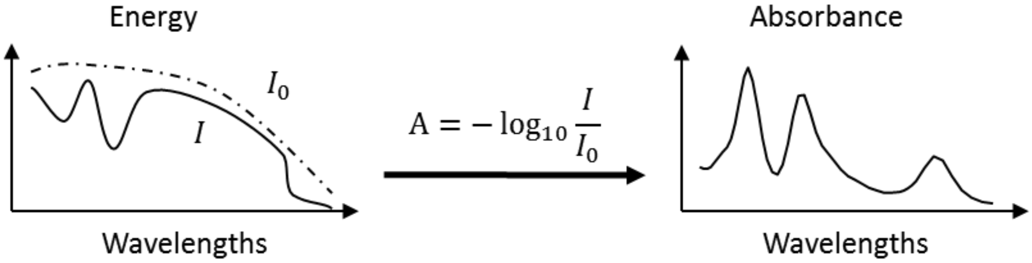


Figure 5: The transformation from transmittance to absorbance.

2.2.2 Beer’s law

Absorbance gives a more intuitive measure, and more importantly, has a linear relationship to the analyte concentration. This linear relationship, known as Beer’s law, was recognized by German physicist, August Beer, during the 1850s, formulated as:

$$\alpha = \epsilon l c \tag{3}$$

where ϵ is the molar absorptivity with a unit of $L/(mol \cdot cm)$; l is the path length as shown in Figure 4; c is the concentration of the analyte. The earliest application of the Beer’s law is the colorimetric analysis developed by Nessler in 1856. This analysis was conducted under visible light, and users visually compared the color of the sample to a reference sample (Nessler tubes) to determine the concentration of the analyte. In the 1930s when photoelectric transducers introduced ultraviolet radiation, absorption spectroscopy was expanded to ultraviolet light. A decade later, during the 1940s, thermocouples introduced infrared radiation; infrared absorption spectroscopy became popular and now serves as one of the most popular techniques in analytic chemistry.

2.2.3 Multicomponent analysis

Beer's law can be extended to multicomponent samples. Assuming component does not chemically react with each other, the absorbances of the different components are additive. Namely, given a N-component mixture with absorbance $\{\alpha_1, \alpha_2, \ldots \alpha_N\}$, the total absorbance of the mixture is simply the summation of the individual absorbances:

$$\beta = \sum_{i=1}^{N} \alpha_i = \sum_{i=1}^{N} \epsilon_i l c_i \ , \tag{4}$$

In mixture analysis, the concentration of the mixture is unknown. Using a reference absorbance α_i for each individual component measured at a known concentration beforehand, we can represent the new total absorbance of the mixture:

$$\beta = \sum_{i=1}^{N} \alpha_i c_i \tag{5}$$

where the relative concentration $c_i = \frac{c_i'}{c_{i0}}$ is the ratio of the true concentration c_i' and the reference concentration c_{i0}. Typically, an absorption spectrum consists of multiple spectral lines at a set of wavelengths $\{\lambda_1, \lambda_2, \dots, \lambda_M\}$. Beer's law stands true for all wavelengths, therefore equation (5) can be written as:

$$\beta_j = \sum_{i=1}^{N} \alpha_{ji} c_i \,, j = \{1, \dots, M\} \tag{6}$$

where β_j is the total absorbance at the wavelength λ_j, and α_{ji} is the absorbance for component i at the wavelength λ_j. It is mathematically more convenient to represent this relationship in a column vector form:

$$\boldsymbol{B} = \sum_{i=1}^{N} \boldsymbol{A}_i c_i \tag{7}$$

where $\boldsymbol{b}$ is the total absorbance with M spectral lines $\mathbf{b} = \{\beta_1, \dots, \beta_M\}$; likewise $\boldsymbol{A}_i$ denotes the absorbances of each individual component $\boldsymbol{A}_i = \{\alpha_{1i}, \dots, \alpha_{Mi}\}$. This relationship can be further reduced to a simpler matrix form if we collect the reference absorbances of different components into a two dimensional matrix $\boldsymbol{A} = \{\boldsymbol{A}_1, \boldsymbol{A}_2, \dots, \boldsymbol{A}_N\}$:

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{c} \tag{8}$$

where $\boldsymbol{C}$ consists of the relative concentration for each individual component $\boldsymbol{c} = \{c_1, \dots, c_N\}$.

In mixture analysis, the problem is to solve this linear system, i.e., given the total absorbance $\boldsymbol{b}$ and the reference potential absorbance $\boldsymbol{A}$, the goal is to estimate the concentration $\boldsymbol{c}$. Since the concentration is the unknown here, let $\boldsymbol{x}$ denote $\boldsymbol{c}$. The concentrations are non-negative, and since there is no energy emitted in the process, the absorbance is also non-negative; therefore, we have non-negative constraints $\mathbf{A}, \mathbf{b}, \mathbf{x} \geq \mathbf{0}$. Finally, the mixture analysis can be formulated as solving such a non-negative constraint linear system:

$$\begin{gathered} \boldsymbol{Ax} = \boldsymbol{b} \\ \text{s.t. } \boldsymbol{x} \geq \mathbf{0} \text{ where } \boldsymbol{A}, \boldsymbol{b} \geq \mathbf{0} \end{gathered} \tag{9}$$

2.3 Wavelength selection

Wavelength selection finds a subset of wavelengths. It reduces the number of dimensions of the problem and improves the model accuracy. There are two aspects of wavelength selection: hardware and software. The hardware must provide optimal wavelength separation; the software must select the most optimal wavelengths to sample and maximize model accuracy. We first provide a brief overview of some common optics for wavelength selection; specifically, we explain the unique advantages of the Fabry-Perot interferometer that is used in this dissertation. We then theoretically justify the necessity of wavelength selection in the context of multicomponent analysis.

2.3.1 Optics for wavelength selection

One advantage of a spectrum with multiple wavelengths is the possibility of quantitative analysis of multicomponent samples. Acquiring multiple wavelengths requires a wavelength selector. However, it is impossible to acquire a single wavelength as its energy converges to zero when the bandwidth is infinitely small. Different wavelength selectors offer different throughput and resolution.

The simplest forms of optical filters are color-coated optics. These optics are **absorptive filters**. They absorb the complementary color of the coated color; for example, a purple filter removes the complementary color – green. Another popular option is the dichroic filter. Dichroic filters are **interference filters**. Rather than absorbing the radiation, dichroic filter work by destructively interfering and reflecting the unwanted wavelengths, only the selected wavelength constructively interfere and pass through. The dichroic filter can achieve a higher throughput and narrower bandwidth in contrast to the absorptive filter, but is more expensive.

Absorptive and dichroic filters only offer very limited choices. Furthermore, to change the wavelengths, the filters need to be physically removed and installed. **Monochromator** offers an alternative that allows continuous adjustment of wavelength selection. The main component of a monochromator is the diffracting grating, which disperses the radiation in space. The direction of the dispersed radiation depends on the wavelength of the radiation. Thus, the radiation at different wavelengths is also separated in space. A second mirror focuses those radiations at a certain angle back to an exit slit. Rotating the grating changes the focusing radiation, hence achieving the effect of wavelength selection.

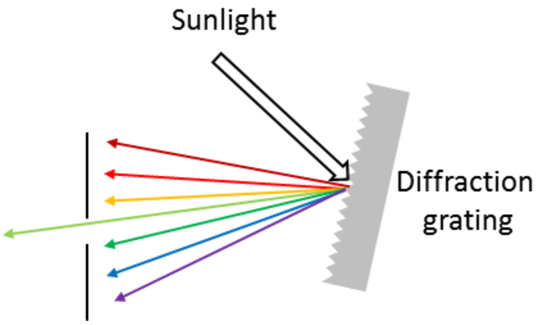


Figure 6: An illustration of the monochromators.

Another wavelength selector that allows continuous wavelength selection is the **interferometer**. The interferometer, like the dichroic filter, is also based on interference. Compared to monochromators, interferometers are superior in one important aspect: higher throughput (*Jacquino's advantage*), which is crucial to achieve higher signal-noise-ratio. Because the interferometer does not scatter radiation as a monochromator does, less energy is lost in the process. The classical interferometer, Michelson interferometer, uses a beam-splitter to split the radiation to two, and then aligns the two beams so that they interfere with each other. By adjusting the positions of the mirrors, some wavelengths are constructively interfered, and others are destructively interfered. A diagram of the Michelson interferometer is shown in Figure 7. By adjusting either l_1 or l_2, one beam is delayed relative to the other by a distance of $2|l_1 - l_2|$. Such delay determines which wavelengths are constructively interfered, and thus achieves wavelength selection.

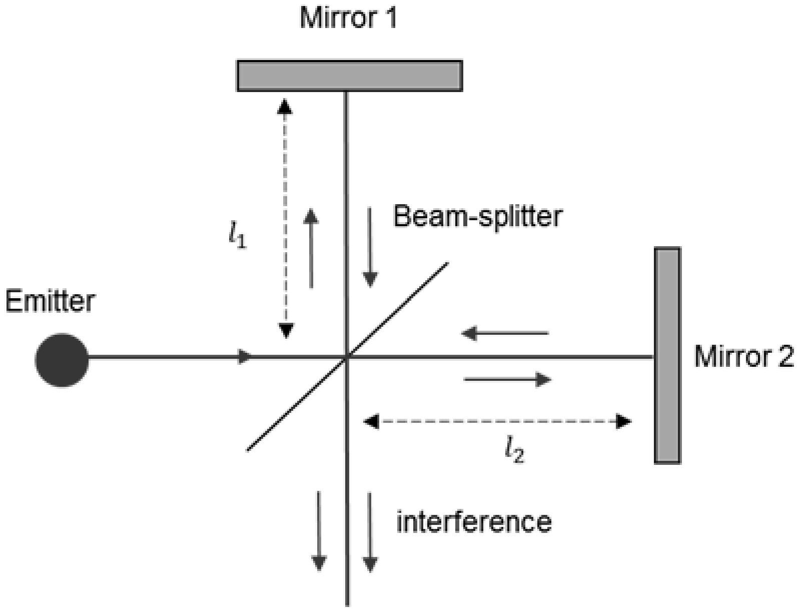


Figure 7: A diagram of the Michelson interferometer.

2.3.1.1. Fabry-Perot interferometer

The Fabry-Perot interferometer, introduced by Charles Fabry and Alfred Perot in 1897, significantly improves the performance of the wavelength selector. The optics are arranged as follows: two halfway mirrors are set up parallel to each other as shown in Figure 8. The mirrors are highly reflective so that a majority of the incoming radiation is reflected and divided many times before it exits the interferometer. The highly reflective mirrors increase the intensity of interference, and when some wavelengths are constructively interfered the signal is much stronger than the two-way interference in Michelson interferometer, resulting in much higher throughput and thus sharper spectral lines. The wavelength being constructively interfered depends on the distance between the two mirrors d as shown in Figure 8. Let λ be the wavelength, if $\frac{2d}{\lambda/2}$ is even ($\frac{2d}{\lambda} = m$, where m is an integer), there is a constructive interference; if $\frac{2d}{\lambda/2}$ is odd ($\frac{4d}{\lambda} = (2m+1)$, where m is an integer), there is a destructive interference. Thus, the wavelength being reinforced by constructive inference is $\lambda = \frac{2d}{m}$ where $m = 1,2,3,\dots$.

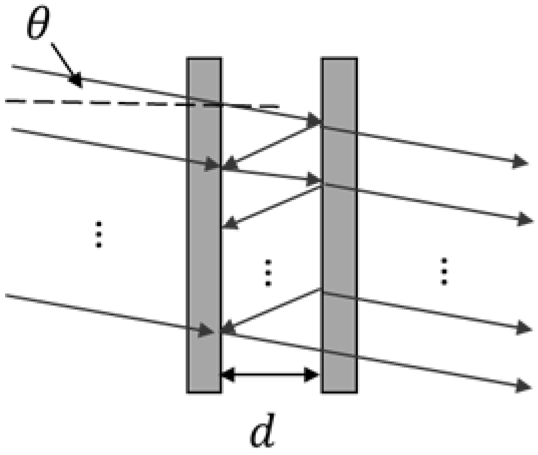


Figure 8: Fabry-Perot interferometer

Therefore, instead of selecting one wavelength, Fabry-Perot interferometer selects multiple wavelengths. This makes the wavelength selector imperfect and its resolving power can be quantified by a metric – finesse F:

$$F = \frac{\Delta\lambda}{\delta\lambda} \tag{10}$$

where $\Delta\lambda$ is the distance between two neighboring constructively interfered wavelengths, and δv is the effective bandwidth (width at half maximum) of the transmission peaks. These two parameters are illustrated in Figure 9.

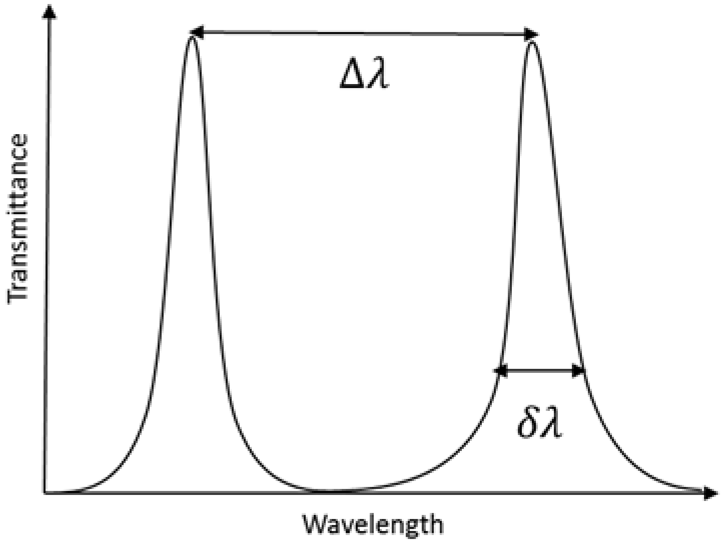


Figure 9: The transmission of the Fabry-Perot interferometer across different wavelengths. $\Delta\lambda$ is the distance between two peaks, $\delta\lambda$ is the effective bandwidth.

The finesse is a function of the reflectance of the mirrors R that can be approximated as:

$$F \approx \frac{\pi\sqrt{R}}{1-R} \; . \tag{11}$$

Therefore, the resolving power is the highest when the reflectance is close to one. For more details about the optics of Fabry-Perot interferometer, please refer to chapter 9 in "Optical Physics" by Lipson [13].

2.3.2 Theoretical justification for wavelength selection

The linearity dictated by Beer's law offers two benefits: computational simplicity and efficiency. However, in the context of multicomponent analysis with spectra that have multiple wavelengths, the calculations often require careful inspection to achieve desirable accuracy. There are two main factors that may affect the accuracy: the first is collinearity (section 2.3.2.1), and the second is nonlinearity (section 2.3.2.2).

2.3.2.1. Collinearity and noise

Infrared spectroscopic data is notoriously collinear because its absorbance peaks are often wide and overlap with each other. Collinearity can significantly reduce the number of components that can be reliably resolved. One metric, effective rank, quantifies such resolvability; however, the original effective rank proposed by Roy et al. [14] is interpreted as the “average significant number of dimensions”, which does not vary with the noise level. We developed an algorithm that also uses noise level as an input parameter: effective rank under noise. More details about the effective rank under noise ($efrank_\sigma$) are described in APPENDIX B:. Effective rank under noise provides insights about the spectra library and the maximum tolerable noise level.

As an example, we calculate the $efrank_\sigma$ of a two-component problem at different noise level and show how noise interleaves with the different wavelengths. Let the two components in the mixture be one flat spectrum and one single-peak spectrum as shown in Figure 10 (a). Three subsets of wavelengths are chosen. All three subset have the same number of wavelengths (35). Their $efrank_\sigma$ are calculated at different noise levels. Figure 10 (b) shows the result. As the noise level increases, the $efrank_\sigma$ of both subset one and subset two decreases, however, the deterioration effect on subset 2 is dramatically larger than the effect on subset 1. Subset 3 is a region with identical values for both components, making the two components indistinguishable. The resulting $efrank_\sigma$ remains at one along different noise levels.

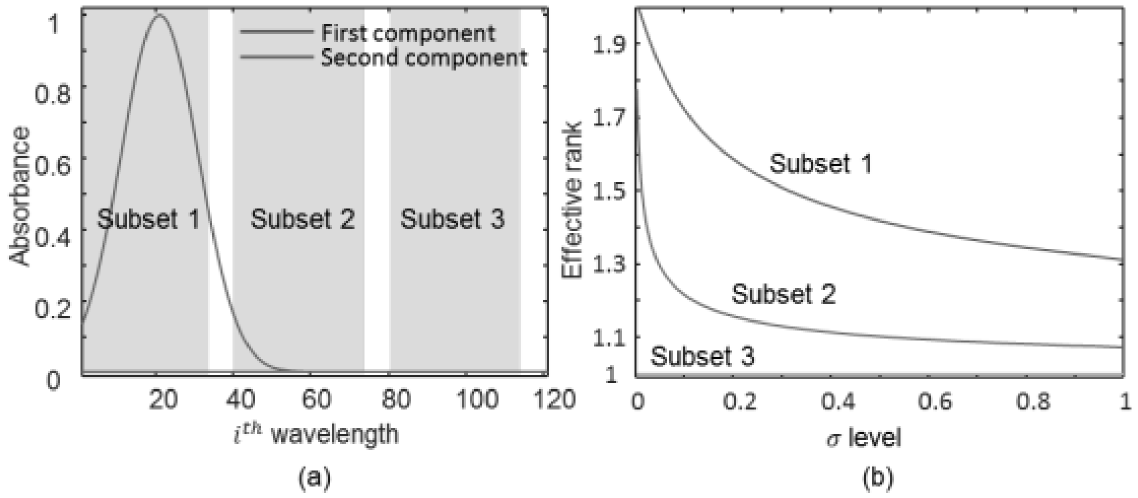


Figure 10: An example of a two-component mixture problem. Three subsets of wavelengths are chosen as shown in the gray areas.

Using the effective rank under noise, we can tell that neither the high noise level nor the collinearity in the spectral library are issues when considered individually; they only become problematic as a combined effect. Consequently, a good conditioned linear system is not necessarily superior when the noise level is low. It only becomes beneficial when noise level grows higher.

2.3.2.2. Nonlinear deviation of Beer's law

In practical applications, nonlinear deviation often occurs due to the limitations of the underlying implementations. Unfortunately, Beer's law is no exception. There can be two main factors causing nonlinearity. The first is the increased interactions between particles of the analyte when the analyte reaches higher concentration. The increased interactions shift the equilibrium state of the analyte such that the relative ratio of the effective individual components in the analyte changes. The second is the fundamental limitation of Beer's law – Beer's law is only valid for pure monochromatic radiation. Polychromatic radiation, on the other hand, always causes a negative deviation from Beer's law as shown in Figure 11. In this dissertation, we assume the concentration is low enough (gas phase close to standard atmospheric pressure) that the shift of the equilibrium state is negligible; however, because the FPI based wavelength selector has a relatively wide bandwidth, the negative deviation emerge when concentration changes.

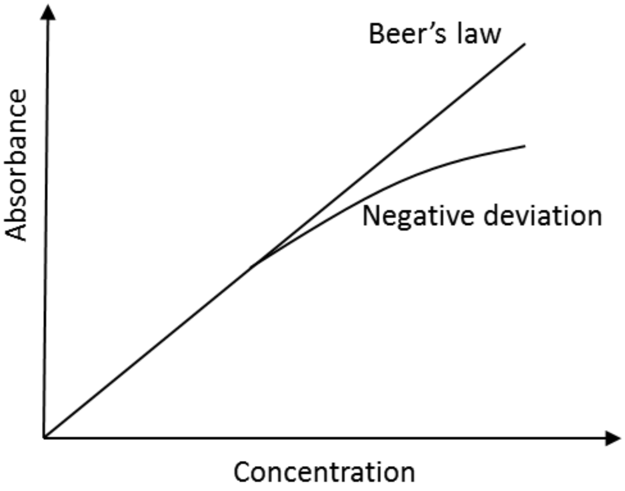


Figure 11: Positive and negative deviations from Beer's law

One may argue that if the deviation scales uniformly across the whole range of wavelengths, then the linear system in equation (9) still remains linear solvable except that the estimated concentration $\boldsymbol{C}$ should be calibrated using the deformed calibration curve as in Figure 11. However, the effect of the negative deviation increases when the local the spectrum is sharper. Please refer APPENDIX A: for the derivation. This causes the spectrum to deform differently at different wavelengths, as shown in Figure 12.

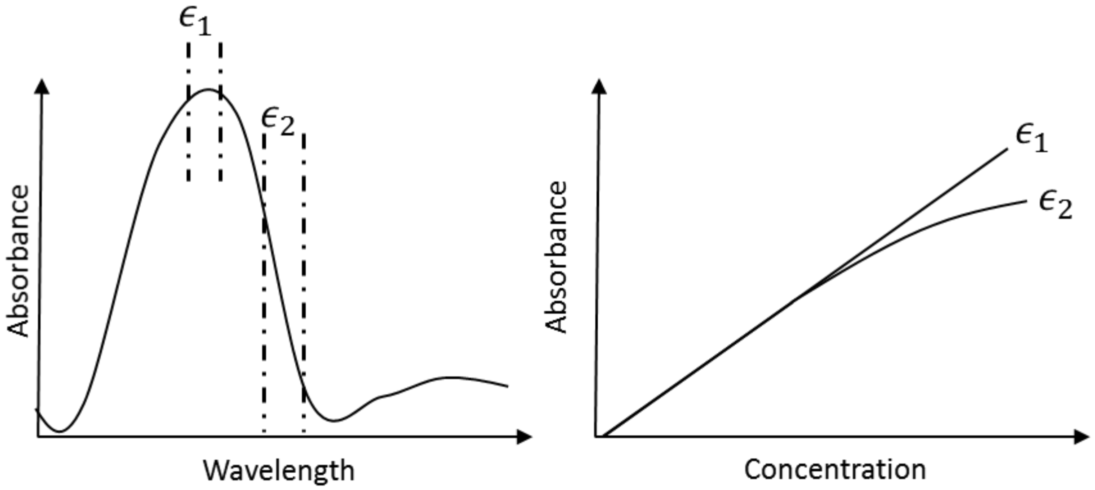


Figure 12: The non-uniform nonlinear deformation of the absorbance due to imperfect wavelength selector. ϵ is the effective absorptivity described in equation (3).

Because of the nonlinear deformation, the spectral matrix $\boldsymbol{A}$ becomes a function of concentration $\boldsymbol{A_C}$. As a result, the error for the estimation is:

$$\boldsymbol{\epsilon_C} = \boldsymbol{C} - \boldsymbol{C}_{true} = \left((\boldsymbol{A}_{\boldsymbol{C}}^{T}\boldsymbol{A}_{\boldsymbol{C}})^{-\boldsymbol{1}}\boldsymbol{A}_{\boldsymbol{C}}^{\boldsymbol{T}} - (\boldsymbol{A}^{T}\boldsymbol{A})^{-1}\boldsymbol{A}^{T}\right)\boldsymbol{B} + (\boldsymbol{A}_{\boldsymbol{C}}^{T}\boldsymbol{A}_{\boldsymbol{C}})^{-\boldsymbol{1}}\boldsymbol{A}_{\boldsymbol{C}}^{\boldsymbol{T}}\boldsymbol{\epsilon}\,, \tag{12}$$

where the first term introduced is a non-zero offset to the estimation; the second term is the zero-mean error induced by observation noise. The non-zero offset is a structural error, which is troublesome because it cannot be diminished by repeated sampling, and as the more wavelengths are observed, the structural error grows.

2.4 Active wavelength selection

Multicomponent analysis numerically solves the linear model described in section 2.2. It transforms the spectroscopic data into useful information such as chemical identities and concentrations. Figure 13(a) illustrates the corresponding mathematical problems. Chemical identification using multicomponent analysis essentially finds the correct columns in matrix $\boldsymbol{A}$, given all the information of the rest in the linear system. On the other hand, wavelength selection finds the minimal number of rows in matrix $\boldsymbol{A}$ that offers the best accuracy as shown in Figure 13(b).

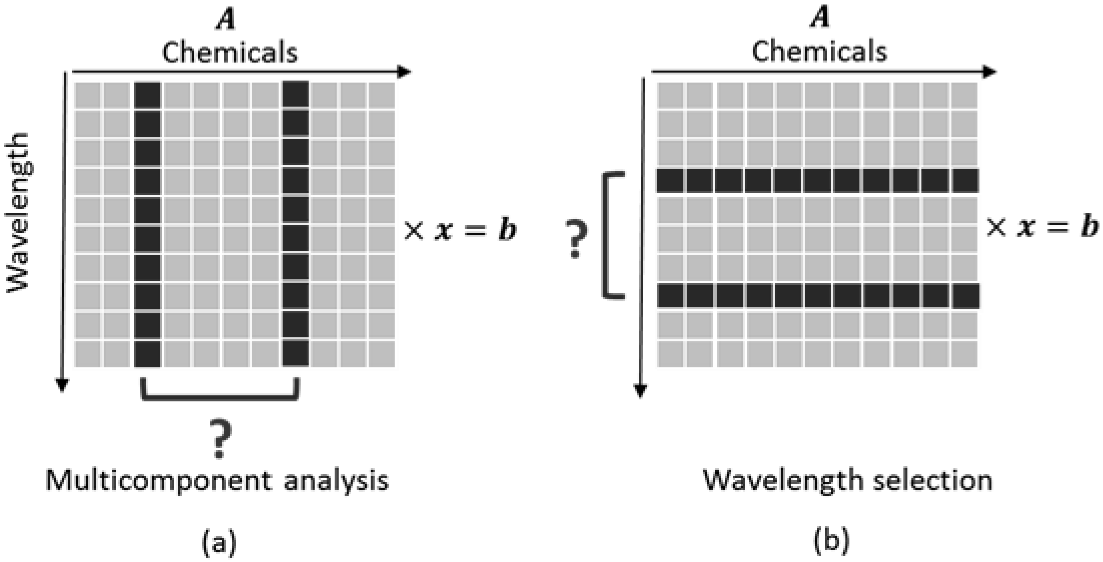


Figure 13: (a) The underlying mathematical problem of multicomponent analysis and (b) the underlying mathematical problem of wavelength selection.

The traditional wavelength selection method assumes that the identity of the chemicals is fixed, i.e., the rows of the matrix are preselected. A supervised method is sufficient for solving the wavelength selection problem so that the selected wavelength gives the best effective rank for those components. However, if the identities of the components are not fixed, the resulting mathematical problem becomes ill-defined because both the rows and the columns of the matrix are unknown as shown in Figure 14.

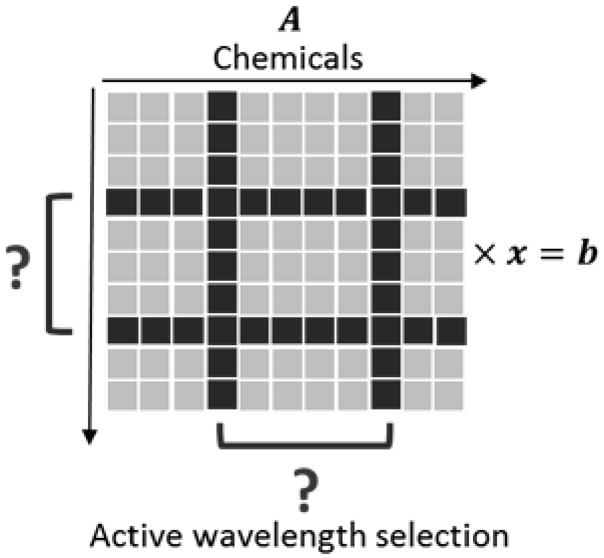


Figure 14: The underlying variable selection for the active wavelength selection problem that is to select the optimal rows (wavelengths) and the optimal columns (components) at the same time.

The active wavelength-selection problem presents a paradox: selecting optimal wavelengths requires knowledge of the component identities; identifying the analyte requires a wavelength set to be measured. We propose an iterative process that solves either problem alternatively, which requires an on-the-fly wavelength selection and sensing process as shown in Figure 15.

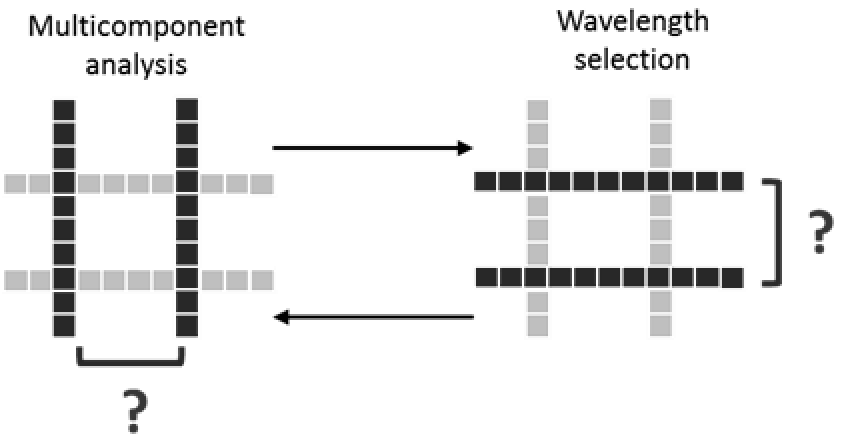


Figure 15: Active wavelength selection by interleaving wavelength selection and multicomponent analysis together.

With this framework, there are two research questions to answer through this entire dissertation: correctness of the multicomponent analysis and efficiency of wavelength selection. As to the former, we ask whether this algorithm correctly identifies the constituents in the analyte. As to the latter, we investigate how efficient this algorithm is compared to a passive wavelength selection strategy, a strategy that conducts wavelength selection offline.

3. LITERATURE REVIEW

Analytical chemistry is an interdisciplinary and application-oriented domain. Much of its research has roots in other domains such as econometrics and pattern analysis. This section reviews multicomponent analysis and wavelength selection at its current state and some prior developments. For a general pattern analysis methods used in chemical sensing, please refer to [15]. Here, we focus on the methods that are relevant to multicomponent analysis and wavelength selection methods. We first review existing numerical methods for multicomponent analysis both inside and outside of analytical chemistry (section 3.1). We then briefly introduce the history of wavelength selection algorithms (section 3.2). We also discuss some similar problems in general machine learning and artificial intelligence (section 3.3). Lastly, we review some works that are directly related to active chemical sensing (section 3.4).

3.1 Numerical methods of least squares and its variations

The most straightforward method for solving linear inversion problem is the ordinary least squares method. That is using the Moore-Penrose pseudo-inverse $\boldsymbol{C} = \boldsymbol{A}^{+}\boldsymbol{B} = (\boldsymbol{A}^{T}\boldsymbol{A})^{\boldsymbol{-1}}\boldsymbol{A}^{T}\boldsymbol{B}$. Pseudo-inverse was first invented by E.H. Moore in 1920 [16], and as computer technologies advanced, it gained popularity in the domain of econometrics during the 1950s [17]. Later on, chemists adopted the method as a tool to calculate the concentrations of an analyte from its spectral data (mostly near infrared spectrum). The method has been called “classical least squares” ever since in chemometrics community.

3.1.1 Non-negative least squares

Classical least squares allows negative values, but negativity and subtraction do not exist in many real problems. In the case of multicomponent analysis, both absorbance and concentration are non-negative. Such problems are formulated as non-negative least squares (NNLS) problem. Lawson first developed the *de facto* standard NNLS algorithm during the 1970s [18]. This algorithm was later improved by Bro and De Jong (FNNLS) in 1997 [19], and Benthem and Keenan in 2004 [20]. Both were developed in the context of chemometrics for calibration purposes where the identities of the chemical components were known. Therefore, the underlying linear problem is over-determined, i.e., the number of observations (wavelengths) is larger than the number of variables (components). There are other approaches to solve this problem more generally. Most recently, Porluru et al. exploited the state-of-art SVM solver and adapted the non-negative least squares problem to an SVM [21].

3.1.2 Sparsity-regularized least squares

Another popular development for linear least squares methods is sparsity regularization. Sparse models are often preferred because they are less prone to overfit. They also provide better interpretability as a parsimonious model often contains the most discriminant latent variables. In the signal processing aspect, sparsity also offers higher compression rate.

The direct metric to measure sparsity is the l_0 norm[1]. With this metric, the sparsity-regularized least squares problem can be formulated as a dual problem:

$$\min_{\boldsymbol{X}}\{\|\boldsymbol{B}-\boldsymbol{AC}\|_2+\lambda\|\boldsymbol{C}\|_0\} \tag{13}$$

where $\|\mathbf{B}-\mathbf{AC}\|_2$ is the l_2 norm error of the model and $\|\mathbf{C}\|_0$ is the l_0 norm – the number of non-zero entries. However, solving this problem is equivalent to variable selection (NP-hard). Interestingly, the aforementioned standard NNLS algorithm by Lawson [18] deployed a greedy variable selection strategy, so, the NNLS algorithm is l_0 norm regularized. At the beginning of the 1990s, sparsity was extensively researched in the signal processing community. Two noteworthy algorithms, matching pursuit [22] and its extension Orthogonal matching pursuit [23], used greedy forward variable selection. They focus on reconstructing a signal using a minimum number of selections from an over-complete wavelets dictionary.

The greedy variable selection strategy often leads to a sub-optimal solution. An alternative solution is to relax the variable selection problem to a convex problem by using l_1 norm or l_2 norm instead of l_0 norm. l_2 norm regularized least squares is also called "ridge regression", which was popularized by Hoerl during the 1960s [24] (It was

[1] l_p norm is defined as: $||\mathbf{C}||_\mathrm{p} = (|\mathrm{C}_1|^\mathrm{p} + |\mathrm{C}_2|^\mathrm{p} + \cdots + |\mathrm{C}_\mathrm{n}|^\mathrm{p})^{\frac{1}{\mathrm{p}}}$. l_0 norm is equivalent to the number of non-zero entries ($||\mathbf{C}||_0 = \sum|C_i|^0$), and l_∞ norm is equivalent to the maximum of all the entries ($||\mathbf{C}||_\infty = \max\{C_i\}$).

first reported by a USSR mathematician Tikhonov during the 1940s). Since l_2 norm is quadratic, the problem has a closed-form solution. However, while ridge regression improves the numerical stability of the solver, it does not necessarily improve the sparsity; i.e., a great number of the entries in the solution can still be non-zero.

A better alternative to l_2 norm is l_1 norm, because l_1 norm offers a heavier penalty for non-sparse solutions. Adding l_1 norm regularization to least squares problems were reported as early as 1973. In the domain of geoscience, Claerbout and Muir proposed a least squares formulation complemented by l_1 norm [25]. In the spectroscopic techniques domain, the same idea was used by Mammone et al. in 1985 [26] for spectrum reconstruction in Fourier transform spectroscopy. The l_1 norm based sparse regularized least squares began to draw attention inside signal processing community during the 1990s, and it was applied in the area of "compressive sensing" (one famous application is the one-pixel camera [27]). Compressive sensing reduces the number of observations beyond the limit of Nyquist frequency. The l_1 norm regularized least squares methods play a crucial role in solving the underlying ill-defined linear system. In 1996, Tibshirani developed *Lasso* [28]. He framed the l_1 norm regularization least squares problem as a quadratic programming problem and was solved using interior-point optimization algorithm (please see [29] for more details about interior-point optimization). Two years later, Chen et al. developed the basis pursuit algorithm [30], which solved the same problem. However, it was reframed as a linear programming problem as the least squares error in equation (9) becomes a constraint:

$$min_{\boldsymbol{C}} \, ||\boldsymbol{C}||_1 \text{ s. t. } \boldsymbol{AC} = \boldsymbol{B}, \boldsymbol{C} \geq \boldsymbol{0} \ . \tag{14}$$

In 2005, Candes et al. theoretically proved that l_1 norm regularized least squares guarantees a good or even exact recovery [31]. In this article, the authors developed the "restricted isometry property constant" ϵ_{2S} that gives clues of how suitable the library matrix $\boldsymbol{A}$ is for an s-sparse[2] reconstruction problem. Their results inspired the following work of randomizing sensing [32], which suggests that a randomly generated reconstruction matrix $\boldsymbol{A}$ from i.i.d Gaussian distributions provide near-optimal signal reconstructions.

To further generalize the sparsity regularization, Fu introduced bridge regression in 1998 [33]. Bridge regression generalizes the sparsity regularization by allowing a continuous selection of l_p norm values where p has to be larger than one and normally smaller than two. Notice that p is not restrained to be an integer, and, as a parameter, it can be auto-tuned using cross-validation. Another noteworthy regularization is to combine l_1 norm and l_2 norm, referred to "elastic net", which was introduced by Zou and Hastie in 2005 [34]. This method linearly combines the two norms as a compromise between smoothness and stability controlled by l_1 norm and l_2 norm.

[2] S-sparse means that there are s non-zero entries in the solution.

3.1.3 Bayesian approach

One major drawback of using sparsity regularized least squares methods is the dependence on the free parameter λ in equation (9), as tuning this parameter can be computationally expensive. A natural solution is a data-driven approach in which the complexity of the model depends on the abundance of the data. Tiao and Zellner proposed the Bayesian multivariate linear regression in 1964, [35]. From a Bayesian point of view, parameters and their distributions are updated iteratively as a sequence of samples. This approach allows the model complexity to grow gracefully as more samples are observed. The Bayesian approach also generalizes the regularization-based methods. Using different l_p norms is equivalent to using special priors to the Bayesian linear regression. For example, l_1 norm is equivalent to Laplace [36, 37]. Table 1 shows their relationships to Bayesian linear regression.

Table 1: Different least squares methods and their equivalents.

Estimator	Ordinary least squares	Variable selection	Lasso	Bridge regression	Elastic net	Ridge regression
Regularization	none	$\boldsymbol{l_0}$	$\boldsymbol{l_1}$	$\boldsymbol{l_{1.x}}$	$\boldsymbol{\alpha l_1 + \beta l_2}$	$\boldsymbol{l_2}$
Bayesian prior	Uniform	Dirac-delta	Laplace	* see[33]	* see [34]	Gaussian

One important development of the Bayesian linear regression is the Gaussian process regression [38]. Instead of using a finite dictionary as in a traditional linear regression problem, Gaussian process uses a dictionary represented by a distribution (commonly a multivariate Gaussian distribution). This method provides a regression method without a concrete dictionary as the dictionary is represented by a distribution. It provides a general tool to reconstruct black-box functions using only a sparse set of samplings. Any knowledge of the latent variables is incorporated in the covariance matrices in the distribution.

3.2 Algorithms for wavelength selection

Given the combinatorial complexity of the wavelength selection problem, exhaustive search methods such as branch-and-bound [39] are computationally prohibitive considering that the modern infrared spectroscopy has thousands of spectral lines. Early work done by Frans and Harris selects only two wavelengths [40]. For more wavelengths, an efficient searching algorithm is crucial to make the process computationally tractable. One approach is to cast the wavelength selection problem to the more general feature selection problem. Many such adaptations have been discussed including genetic algorithms [8], simulated annealing [7], colony optimization [9], back-propagation neural networks [41], and Kohonen neural networks [42]. Nevertheless, there are two major drawbacks of these methods. First, they involve many free parameters, and the performance relies on the parameter tuning and the specific application domain; the second drawback is the computational cost. To address these drawbacks, the analytical chemistry community developed more specialized wavelength selection algorithms. They can be broadly divided into two groups: window/interval selection and greedy feature selection.

3.2.1 Moving window and interval selection

Moving window and interval selection both consider a continuous section of wavelengths as the basic selection unit. This approach has its root from molecule structural analysis since this continuous section normally corresponds to a functional group[3]. However, in modern applications, they are often coupled with PLS. PLS itself is a factor based method, which provides a secondary feature extraction on the result of the wavelength selection. Therefore, discrete wavelength selection is not necessary, and doing so may defeat the purpose of PLS.

Norgaard introduced interval selection partial least squares (iPLS) in 2000 [43]. To reduce the parameter searching size, the spectrum is divided into multiple non-overlapping windows with equal sizes. A PLS model is built for each window, and the most informative window is the one with the minimum error. Norgaard later extended this method and developed backward interval partial least squares [44]. Instead of selecting only one interval, the backward method builds the wavelength set sequentially. It calculates the modeling contribution of an interval by excluding this interval. Thus leaving out the most informative interval leads to the poorest performance. Similarly, Xiao et al. developed a forward iPLS and genetic algorithm guided iPLS in 2007 [45].

Another branch of interval selection is the moving window methods. In 2002, the moving window partial least squares regression was introduced by Jiang et al.[46]. In

[3] A functional group refers to a set of bound atoms such as Benzene ring or alcohol.

this method, a window of fixed size slides through the entire spectrum. To find the best wavelengths, a new PLS model is calculated at each position of the window, and the best position is selected to minimize the residual. Later in 2004, Du et al. [47] made the moving window method more flexible by allowing multiple changeable size windows.

3.2.2 Feature forward/backward selection

Two popular greedy feature selection algorithms are the successive projection algorithm (SPA) [5] and the uninformative variable elimination method (UVE) [48]. The SPA method by Araújo et al. [5] extracts wavelengths that minimize collinearity using the sequential orthogonal projections of the Gram-Schmidt procedure. The original UVA method removes the wavelength with the lowest SNR. A more recent development by Cai in 2008 is the UVE algorithm [49] guided by the PLS algorithm. SPA can be combined with UVE, and Ye proposed the UVE-SPA algorithm [50]. The method is essentially a forward-backward selection method where the number of selected wavelengths is further reduced with similar performance.

3.3 Related problems in other areas

3.3.1 Robotics

Elements of robotics are inspired by the theory of ‘active perception’ [51, 52], which states that organisms actively probe their environments to enhance their ability to extract information. The concept of active sensing was originally proposed during the 1980s in the robotics and vision community [53]. In a classic study, Aloimonos et al. [54] proposed an active vision framework to adjust camera geometric parameters (positions, rotation, and so on) to solve 3D reconstruction problems. The authors showed that, by using an active strategy, an otherwise ill-posed problem became well-posed, which dramatically improved the algorithm's problem-solving efficiency. Other problems in robotics and computer vision were soon adopted through the idea of active sensing, including modeling of facial expressions with temporal information [55], multi-target detection and tracking [56], robot navigation [57], localization [58] and simultaneously map building and localization [59]. These results show that active sensing works exceptionally well compared to passive methods, especially when the observations are noisy, or the problem dynamics must be considered.

3.3.2 Bayesian optimization

Interestingly, a similar concept (adaptive sampling) had already been proposed twenty years earlier in the optimization community. In early work, Kushner and Mockus proposed a stochastic method for function minimization [60, 61], a method later known as Bayesian optimization [62]. The approach samples the objective function sparsely, and then uses a Gaussian process [38] to estimate the objective function and the variance of its estimate at all other locations in sample space. These can then be used to guide the selection process, either by further sampling at the predicted highest/lowest locations to converge to a solution (exploiting) or by sampling areas of high variance to improve the estimation accuracy (exploring).

3.3.3 Multi-armed bandit problem

The multi-armed bandit problem was first introduced by Robbins [63] during the 1950s. The problem consists of a set of K probability distributions, normally Gaussian, with associated expected value and variances. The goal of the player is to extract as much as money as possible by selecting the arm with the highest value. However, at the beginning, none of the distributions are known. Thus, *exploration* is required to discover the most profitable arm. As more and more observations are made, one can *exploit* this information and taking the empirically best actions as often as possible. At each step, the player needs to decide whether next step is exploring or exploiting. This balance is called exploration and exploitation dilemma, and multi-armed bandit problem has been the most investigated problem for exploration and exploitation dilemma. The objective function of multi-armed bandit problem is defined as "regret" – the total loss of using a non-oracle strategy. In 1985, Lai and Robbins proved that no strategy could perform better asymptotically than a logarithmically growing regret [64]. Later in 2002, this result was proved also true over time stripping the asymptotical limitation [63]. mathematically bounded in the example of multi-armed bandit problem in [63]. In the past decade, many algorithms were proposed to solve the problem [65-67]. Most recently, Bubeck and Cesa summarized the performance of different strategies analytically [68], and Kuleshov and Precup conducted a benchmark study some of the most famous methods [69] in which they suggested that some heuristics could outperform many sophisticated and theoretically sound approaches.

3.3.4 Adaptive compressed sensing

More recently, adaptive sampling techniques have been developed for sparse signal recovery, an area known as compressed sensing [70]. Compressed sensing uses ℓ_1 norm regularization to encourage sparsity in the solution of an underdetermined linear system. Compressed sensing algorithms previously use a random selection of the variables to be measured, but recent work by Haupt et al. [71] has shown that the accuracy of the reconstruction can be improved by use of adaptive sampling, particularly in the case of weak signals. Their algorithm assigns each feature an importance measure that is proportional to the value of that feature and diminishes exponentially over the number of times that feature is sampled. In the beginning, all features are assigned uniform importance; as some features are sampled more frequently, their importance diminishes, allowing other features with stronger values to be sampled. The authors evaluated the method on a recovery problem for telescope star images. The proposed adaptive sampling scheme outperformed a non-adaptive scheme, requiring fewer samples to achieve higher star-detection rates.

3.3.5 Active learning

Adaptive sampling concepts have also been developed in the machine learning literature, where they are referred to as active learning. In contrast with active sensing, where the goal is to select an optimal sequence of features for each test case, the goal of active learning is to select training samples to improve the learning of decision boundaries. In a theoretical study, Castro and Nowak [72] compared the minimax bounds[4] of adaptive and passive classification methods and showed that adaptive methods have superior error reduction rates with reasonable complexity. These findings are particularly relevant in chemical sensing applications, where collecting or labeling new samples can be laborious. As such, active learning can alleviate this problem by minimizing the required number of samples without reducing the performance. Motivated by this issue, Lomasky et al. [73] proposed an "active class selection" method for the problem of discriminating vapors with an array of the chemical sensor. Their method generated the next n training instances according to the instability of the class boundaries, the latter being measured by the number of test instances whose classification results change upon inclusion of the previous set of n training samples. In related work, Rodriguez et al. [74] developed an active sampling method for sensor array calibration that selected not only the classes of the vapors to be measured but also their concentrations. The authors modeled the preference for a particular concentration c with the pseudo-distribution $P(c) \propto e^{-kc}$, where parameter k can be used to favor low concentrations $(k > 0)$ or high concentrations $(k < 0)$. Given a sequence of calibration batches $B_1, B_2, \dots, B_n$ and their

[4] The mini-max bound estimates the best possible error reduction rate (mini) under a worst difficulty scenario (max), where difficulty is measured by the dimensionality of the classification problem and the noise level of the measurements.

respective k parameter sequence $k_1, k_2, \dots, k_n$, the algorithm selects the value k_{n+1} that provides the lowest cross-validation error on batch B_{n+1}.

3.4 Active chemical sensing

Active sensing techniques have only recently been used in the context of chemical sensing. To our knowledge, the earliest use is the work of Nakamoto et al. [75, 76] on odor generation. The goal of this work was to reproduce an odor blend by creating a mixture from its individual components. The authors developed an active control algorithm that adjusted the mixture ratio so that the response of a gas sensor array to the mixture matched the response of the array to the target odor blend.

To our knowledge, there are also only a handful of works adopting the idea of adaptiveness into chemical identification. However, with the advancements in microelectronic sensors and microelectromechanical systems (MEMS) technologies, more and more tunable micro-sensors emerged to the market. Please refer to [77] for a thorough review of these adaptive microsystems. Metal-oxide (MOX) gas sensors are the most common and offer an array affordable chemical mixture system. These sensors are especially suitable for specific target analytes using a proper temperature modulation program. Inspired by the biological chemosensory adaptation process, Raman and Gutierrez explored biologically plausible models mimicking the mammalian olfactory system [78, 79]. Gutierrez developed algorithms that adjust the Fisher's linear discriminant functions according to the sensor response from analytes [80, 81]. Gosangi investigated the active sensing framework focused on temperature-modulated metal oxide sensors (MOX). In [82], Gosangi studied the problem of discriminating M

chemicals at a fixed concentration with a single MOX sensor. For this purpose, the author used a partially observable Markov decision process (POMDP) combined with a myopic policy that selected sensing actions based on the expected reduction in Bayesian risk. In [83], the objective function was tuned to minimize the energy consumption for temperature modulation for metal-oxide sensors. Later in [84], a dynamic Bayesian Networks approach was proposed. The Bayesian networks were used to model the dynamic transient response of metal-oxide sensors through temperature modulation. If the concentration of the mixture is discretized, a quantification problem can be easily framed as a classification problem. In [85], Gosangi developed a recursive Bayesian estimation approach that was used to solve the mixture quantification problem. The method was later formalized and experimentally validated for binary mixture problem [86].

MOX sensor are first-order sensors because their measurements are one-dimensional. One interesting development is to expand the number of sensors to sensor arrays, which is higher-order sensor. For a general review of the higher-order sensor, please refer to [87]. In [88], Gosangi developed a Posterior-Weighted Active search method to classify chemical mixture using absorbance spectrum data in simulations and MOX sensor arrays in experiments. Using higher-order sensors based on spectroscopy, in the context of chemical discrimination, Priebe et al. [89] developed a decision tree algorithm for integrated sensing and processing. When evaluated on an optical sensor array exposed to carcinogens, the approach reduced misclassification rates by 50%, while requiring only 20% of total the sensor measurements. An optical implementation of active-sensing principles was proposed by Dinakarababu et al. [90] for rapid identification of chemicals. The authors developed a digital micro-mirror device capable of multiplexing certain spectral bands and directing them onto a photo-detector. The system was able to measure the projection of the incoming spectral density onto a set of basis vectors, rather than measure the spectral density directly. The basis vectors are the eigenvectors of a covariance matrix probabilistically weighted by the likelihood of different classes based on previous measurements.

4. ACTIVE WAVELENGTH SELECTION BASED ON MISCLASSIFICATION COST FOR SINGLE CHEMICAL IDENTIFICATION[5]

In this chapter, we investigate an active wavelength selection strategies based on a Bayesian approach for single chemical identification. The method selects wavelengths sequentially on-the-fly, based on sensor responses obtained thus far. This allows the sensor to adapt its sensing program in response to different chemical stimuli and their concentrations, as well as to environmental influences. Our approach leverages the work by Gosangi et al. on active temperature programming for metal-oxide (MOX) sensors [91], and models active sensing as a probabilistic state estimation process [92]. In this chapter, we apply the active-sensing algorithm to Fabry-Perot interferometry. More importantly, we extend the approach to allow chemical identification at multiple concentrations. The approach consists of generating concentration-independent absorption profiles for each chemical target through non-negative matrix factorization (NNMF) [93], and fitting incoming sensor data to those profiles through linear least squares (LLS) [94]. We evaluate the concentration-independent active sensing algorithm

[5] The descriptionofthe method and the experimental results are reprinted with permission from "Active Concentration-Independent Chemical Identification with a Tunable Infrared Sensor" by Huang and Gutierrez-Osuna, 2012. *IEEE Sensors Journal*, pp. 3135-3142, ©2012 IEEE.

on a database of IR absorption spectra from 27 chemicals, as well as experimentally on an 8-chemical discrimination problem using an FPI prototype.

4.1 Methods

4.1.1 Overview of the approach

During the *active sensing stage*, we first determine the optimal operating wavelength for the sensor at each time step; for this, we use a utility function that measures the difference between the sensing cost at each wavelength and the corresponding expected reduction in Bayes risk. Then we acquire absorption at the chosen wavelength, remove linear concentration effects and update the belief distribution accordingly.

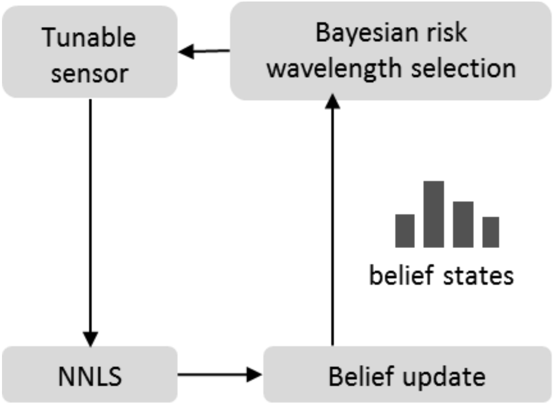


Figure 16: Diagram of the active wavelength selection framework based on Bayesian risk.

Our approach can be broadly divided into two stages: sensor modeling and active sensing. During the *sensor modeling stage* (Section 4.1.3), we create concentration-independent absorption profiles for each chemical; these profiles remove the linear concentration effects while preserving chemical identity information. The resulting concentration-normalized response is then modeled with a Gaussian mixture model.

During the *active sensing stage* (Section 4.1.4), we first determine the optimal operating wavelength for the sensor at each time step; for this, we use a utility function that measures the difference between the sensing cost at each wavelength and the corresponding expected reduction in Bayes risk. Then we acquire absorption at the chosen wavelength, remove linear concentration effects, and update the belief distribution accordingly.

4.1.2 Notations

Given a gas sample of unknown concentration but known to belonging to one of n chemical classes $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ and a tunable IR spectrometer with l spectral lines $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_l\}$, consider the problem of finding a sequence of actions $(a_1, a_2, \dots, a_T)$ that minimizes the cost of discriminating gas samples. For this purpose, each action a_i has an associated cost: tuning the spectrometer to wavelength λ_r incurs a cost $C_s(\lambda_r) = c_r$ (e.g., power consumption), and classifying the gas sample based on available information carries a misclassification cost $C_c(u, v) = c_{uv}$ when a sample from class ω_u is incorrectly assigned to class ω_v.

We model this problem as that of probabilistic state estimation where each of the n classes is represented as a state, and maintain a probabilistic distribution b_T (where T denotes time) that represents our belief that the sample belongs to each class:

$$b_T : \Omega \rightarrow [0,1]; \sum_{\omega_i \in \Omega} b_T(\omega_i) = 1 \tag{15}$$

Given an initial belief distribution $b_0(\omega_i)$, a sequence of actions $a_1, a_2, \dots, a_T$, and the corresponding observations $o_1, o_2, \dots, o_T$, the current belief b_T is defined as:

$$b_T(\omega_i) = p(\omega_i | o_1, o_2, \dots, o_T, a_1, a_2, \dots, a_T) \quad (16)$$

In [91], we used a similar approach to actively modulate the operating temperatures of MOX sensors. However, that formulation did not account for the concentration of the gas sample; the framework assumed that each chemical was presented at a fixed concentration. Here, we extend the framework to identify samples at various concentrations.

4.1.3 Library acquiring

To a first-order approximation, the relation between light absorption and its traveling medium follows the Beer-Lambert law [95] (see more details in Section 2.2). The absorption A of an IR beam transmitted through a spectrometer with a gas chamber of length l filled with a chemical of absorption coefficient ϵ at concentration c can be estimated as $A = \epsilon l c$. Thus, absorption is linearly dependent on analyte concentration. We take advantage of this relation to remove the linear influence of concentration on absorption spectra, thus creating concentration-independent absorption profiles. We start by collecting absorption spectra for each chemical (ω_i) at m different concentrations and organizing the data as a matrix X_i of size $m \times l$, where l is the number of discrete wavelengths in the FPI. We then employ NNMF to factorize X_i into a product of two matrices W_i and H_i such that error function $||W_i H_i - X_i||^2$ is minimized:

$$X_i = W_i H_i + U_i \tag{17}$$

where W_i is a column matrix of size $m \times 1$, H_i is a row matrix of size $1 \times l$, and U_i is a matrix of size $m \times l$. W_i represents the concentration of the m absorption spectra, H_i can be interpreted as the concentration-independent absorption profile of chemical ω_i, and U_i is a residual matrix that captures sensor noise. Given a concentration w_r, we can approximate the absorption spectra of chemical ω_i as $w_r H_i$. We chose to use NNMF for this factorization process because NNMF enforces a constraint that all the elements of W_i and H_i be non-negative; this constraint is necessary because concentrations W_i and absorption spectra H_i are strictly non-negative. This process is repeated n times to obtain absorption profiles for each chemical: $H_1, H_2, \dots, H_n$. Figure 17 illustrates the NNMF process for the absorption spectra of acetone at 50 different concentrations ranging from 0% (air) to 100% (pure chemical) [6].

[6] Concentrations are specified as percentage dilutions of a saturated headspace, e.g., a 10% concentration corresponds to a mixture of 90% air and 10% of the analyte vapors obtained from a saturated headspace.

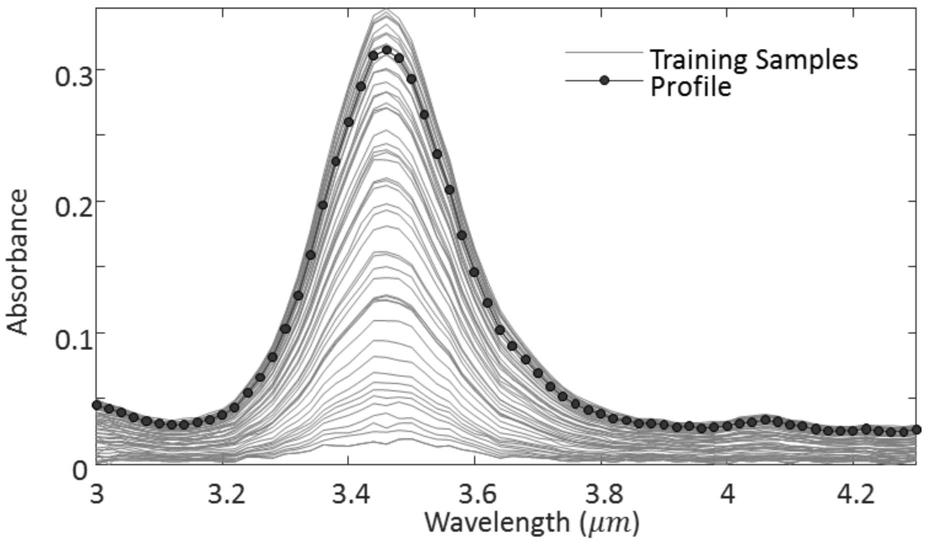


Figure 17: Absorption spectra for acetone at 50 different concentrations (solid lines) and the estimated profile (dotted line).

For each chemical, we then model the concentration-independent spectra (profile plus residual) with a Gaussian mixture model (GMM). These models will be used during the sensing stage to predict sensor responses. For each chemical ω_i, we first create matrix $\hat{X}_i$ as:

$$\hat{X}_i = U_i + I_{m\times 1}H_i \tag{18}$$

where $I_{m\times 1}$ is an identity matrix of size $m \times 1$. Thus $\hat{X}_i$ is the sum of sensor noise and concentration-independent absorption spectra for chemical ω_i. Figure 18 shows an example of this matrix for the acetone dataset in Figure 17.

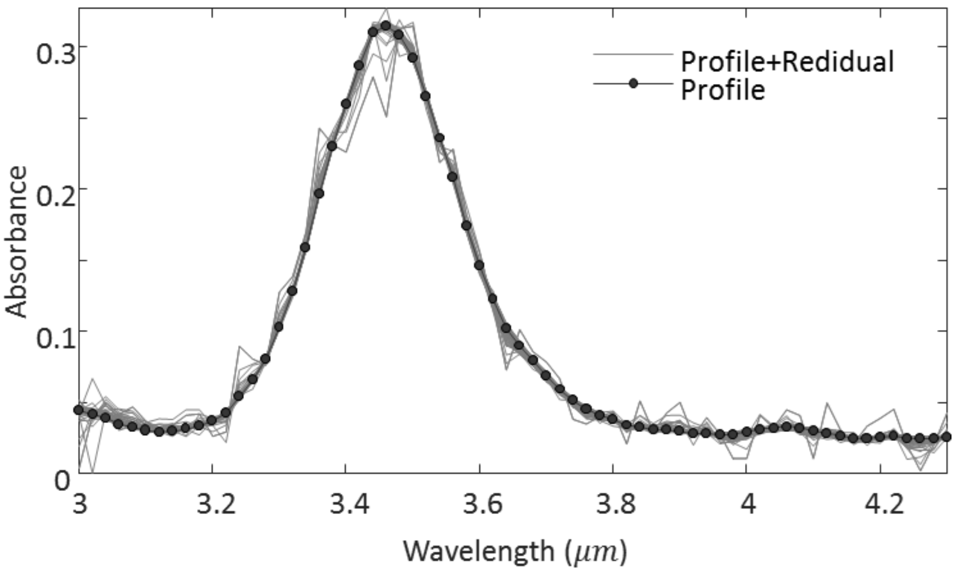


Figure 18: Dotted line is the absorption profile of acetone. The solid green lines are the sum of the residual matrix and the absorption profile.

Using a GMM, the response $\hat{x}$ to chemical ω_i at wavelength λ_j can be expressed as follows:

$$P(\hat{x}|\omega_i, \lambda_j) = \sum_{k=1}^{M_{i,j}} \alpha_{i,j,k} N(o|\mu_{i,j,k}, \sigma_{i,j,k}) \tag{19}$$

where $M_{i,j}$ is the number of Gaussian components. $\alpha_{i,j,k}$, $\mu_{i,j,k}$, and $\sigma_{i,j,k}$ are the mixing coefficient, mean, and standard deviation of each Gaussian component, respectively. These models are trained on $\hat{X}_i$, i.e., the j^{th} column of matrix $\hat{X}_i$ is used to learn the mixture model for chemical ω_i at wavelength λ_j. Model parameters are estimated using Expectation Maximization [96].

4.1.4 Wavelength selection

The first step in the active sensing stage is to select the 'best' wavelength to which the FPI should be tuned. For this, we used a greedy approach that selects a wavelength that maximizes the following utility function:

$$U(b_T,\lambda_i) = \Delta R - c_i = \big(R_C(b_T) - R_E(b_T,\lambda_i)\big) - c_i \qquad (20)$$

where $U(b_T,\lambda_i)$ is the utility of wavelength λ_i, b_T is the current belief distribution, ΔR is expected reduction in Bayes risk, and c_i is the sensing cost at wavelength λ_i. The expected reduction in Bayes risk (ΔR) is defined as the difference between the current Bayes risk $R_c(b_T)$ and the expected risk $R_S(b_T,\lambda_i)$ upon tuning the sensor to λ_i. The current Bayes risk $R_C(b_T)$ is estimated as:

$$R_C(b_T) = \min_{\omega_u} \sum_{\omega_v \in \Omega} c_{uv} b_T(\omega_v) \qquad (21)$$

which reflects the expected risk of classifying the sample. The expected Bayes risk $\mathrm{R_S}(b_T,\lambda_i)$ of tuning the sensor to λ_i at the next time step is computed as:

$$R_S(b_T,\lambda_i) = \sum_{\forall o} \min_{u} \left(\sum_{\omega_v \in \Omega} c_{uv} p(o|\omega_v,\lambda_i) b_T(\omega_v) \right) \qquad (22)$$

$R_s(b_T,\lambda_i)$ averages the minimum Bayes risk over all observations that may result from λ_i. If the utility of all l wavelengths is negative, we halt the sensing process and classify the sample based on equation (21). For a continuous observation space, equation (22) becomes an intractable integral; instead, we discretize the absorption space into a finite set of values (see APPENDIX C:).

The second step in the active-sensing process is to tune the spectrometer to the optimal wavelength λ_T and obtain the corresponding observation o_T. To remove linear concentration effects, we then fitted the observation sequence $\vec{o} = o_1, o_2, \dots, o_T$ (for wavelength sequence $\vec{\lambda} = \lambda_1, \lambda_2, \dots, \lambda_T$) to the profile of each chemical using linear least squares. Namely, given concentration-independent profile H_i for chemical ω_i, we found coefficient k_i that minimizes the sum-squared error $\sum_{j=1}^{T}\left(k_i \cdot H_i(\lambda_j) - o_j\right)^2$; this results in fitted observations $\vec{F}_i = \frac{\overrightarrow{o}}{k_\iota}$. The process is repeated for each chemical, leading to fitted observations $\vec{F}_1, \vec{F}_2, \dots, \vec{F}_n$ that are independent of concentration effects. Figure 19 illustrates the entire process with an example. We first created absorption profiles for acetone and propanol (as described in section 4.1.3) using data collected at 50 different concentrations ranging from 0% to 100%. Then, the sensor was exposed to acetone at a concentration of 20%, and we obtained responses at 66 wavelengths. The responses were then fitted to the concentration-independent profile of each chemical. As shown in the figure, the observations fit better to the acetone profile (mean square error of 3.38×10^{-4}) than to propanol (4.4×10^{-3}).

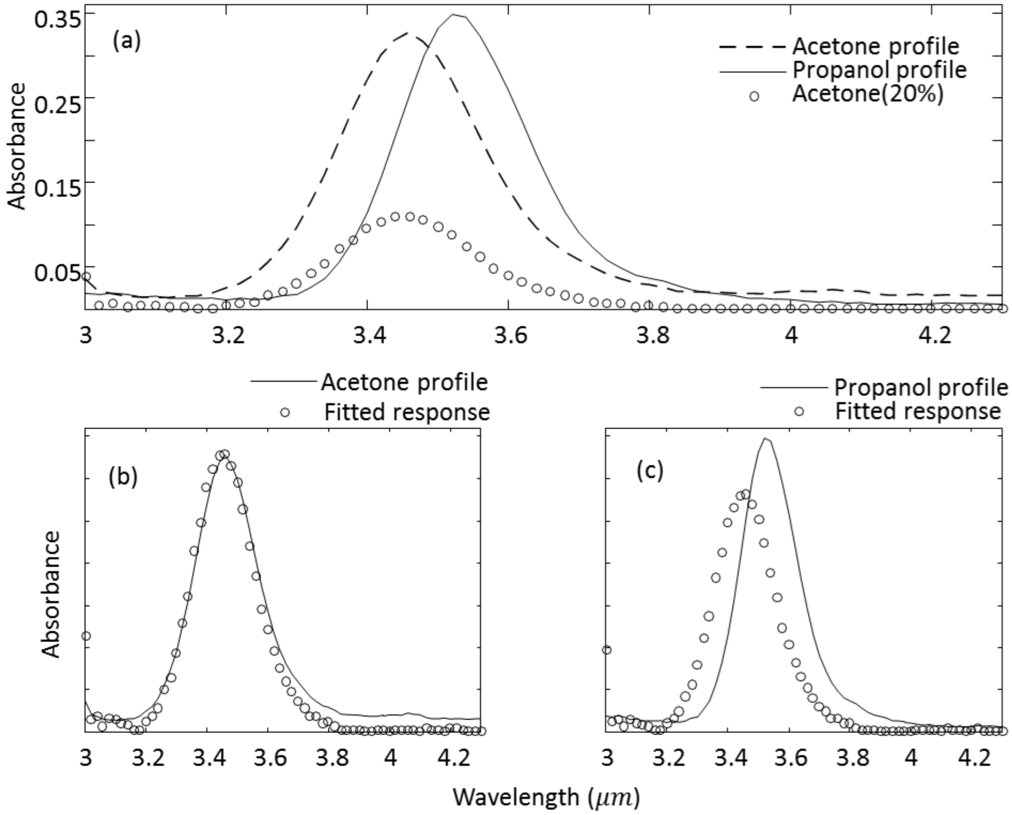


Figure 19: Fitting sensor responses to concentration-independent profiles H_i. (a) Solid lines represent absorption profiles for acetone and propanol, whereas circles correspond to sensor observations in the presence of 20% acetone. Observation fitted to (b) the acetone profile and (c) the propanol profile.

The last step in the sensing process is to update the belief distribution using the fitted observations. Since the normalization step has to be performed on the entire observation sequence $(o_1, o_2, \ldots, o_T)$, we also recalculate the belief from time $t = 0$ using the fitted observation $\vec{F}_i$. The process is summarized in Table 2.

Table 2: Pseudo code for belief update procedure

```
Input: Fitted observations $\vec{F}_1, \vec{F}_2, \dots, \vec{F}_n$ and wavelengths $\vec{\lambda}$
Output: Updated belief $b_t$
Procedure belief_update($\vec{F}_1, \vec{F}_2, \dots, \vec{F}_n, \vec{\lambda}$)
Initialize belief: $b_0(\omega_i) = \frac{1}{n} \forall i$
for $t = 1$ to $T$
        $\eta = 0$
        for $i = 1$ to $n$
                $b_t(\omega_i) = p(\vec{F}_i(t) | \vec{\lambda}(t), \omega_i) b_{t-1}(\omega_i)$
                $\eta = \eta + b_t(\omega_i)$
        end
        $b_t(\omega_i) = \frac{b_t(\omega_i)}{\eta}$
end
```

In the above pseudo-code, η acts a normalization constant that ensures the belief distribution sums to 1. The value $p(\vec{F}_i(t)|\vec{\lambda}(t), \omega_i)$ is obtained using the probabilistic sensor models as described in section 4.1.3.

4.2 Validation on synthetic data

We first tested the active-sensing framework on a large classification problem using simulated data; this allowed us to compare the active-sensing approach against a passive feature selection strategy thoroughly.

4.2.1 Dataset

To simulate the response of the FPI sensor to different chemicals, we used data from the NIST Chemistry WebBook [97], which provides high resolution FTIR spectra in the range 3-21 μm for over 40,000 chemicals in gas phase. We identified 27 chemicals that had at least one absorption peak in the operating range of our FPI (3-4.3 μm). To simulate the spectral resolution of the FPI, we downsampled the FTIR absorption spectra to 66 values, the number of unique wavelength tunings in the FPI. Simulated spectra are shown in Figure 20. Using these spectra, we generated 30 absorption spectra for each chemical by adding Gaussian noise of variance 0.05 at each wavelength. The resulting dataset (30×27 spectra) was then used to train the sensor models, one for each chemical, as described in section 4.1.3. We evenly discretized the observation space at each wavelength into 200 steps (see APPENDIX C:).

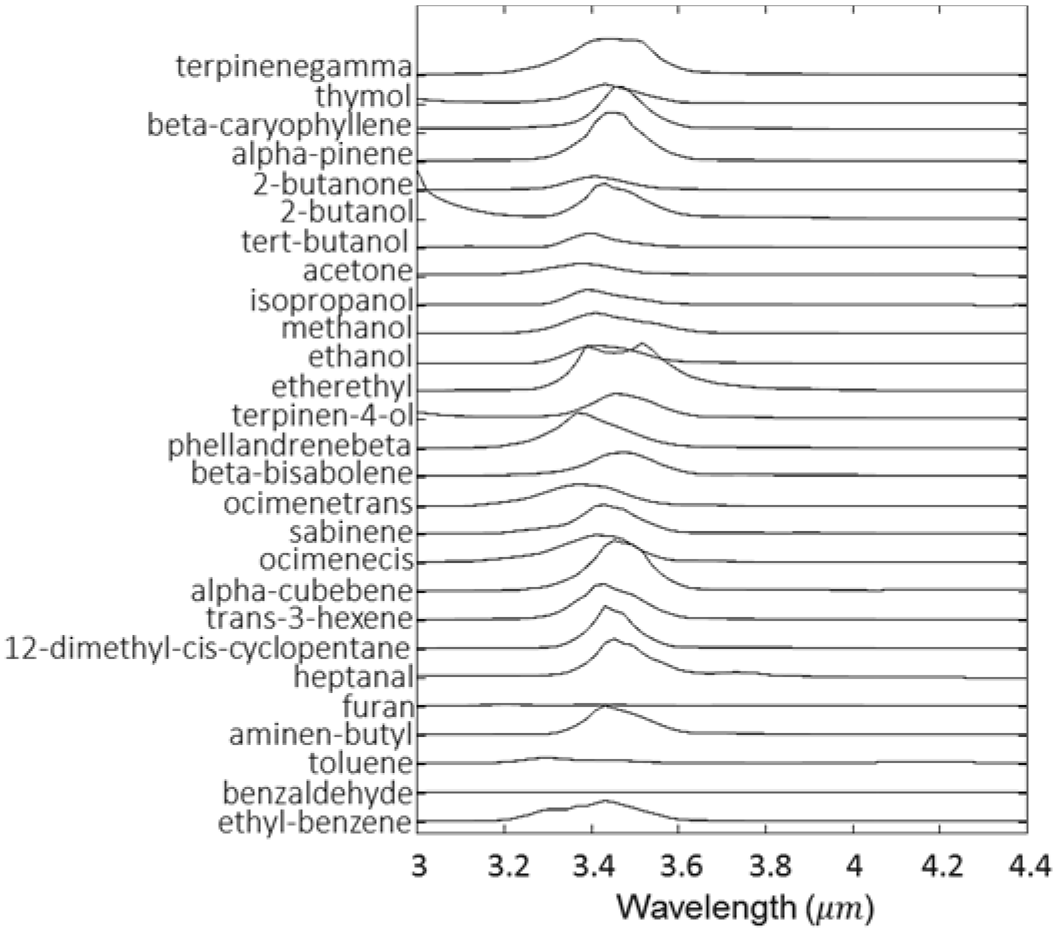


Figure 20: Simulated absorption spectra for 27 chemicals. Spectra are plotted with an offset along the y-axis for visualization purposes.

4.2.2 Test case

First, we present a test case to illustrate the active-sensing process. In this case, the sensor was exposed to trans-3-hexene, and we assumed a 1-0 loss function for the classification costs $c_{uv} = I(u \neq v)$[7], and uniform sensing costs $c_i = 0.02$ for all wavelengths. The algorithm required 15 sensing actions before it classified the sample. Figure 21(a) shows the average absorption spectrum of trans-3-hexene and the 15 operating wavelengths selected by the method, whereas Figure 21(b) shows the belief distribution as a function of time. At time $t = 0$, all chemicals are equally likely. As observations are obtained, the belief for trans-3-hexene, sabinene, and butyl-aminen increase. However, from time $t = 8$ the probability of trans-3-hexene starts dominating. At $t = 15$, the sensing process is halted since sufficient evidence is available where the utility based on equation (20) for any further sensing becomes negative, and the sample is classified as trans-3-hexene.

[7] $I()$ is the indicator variable.

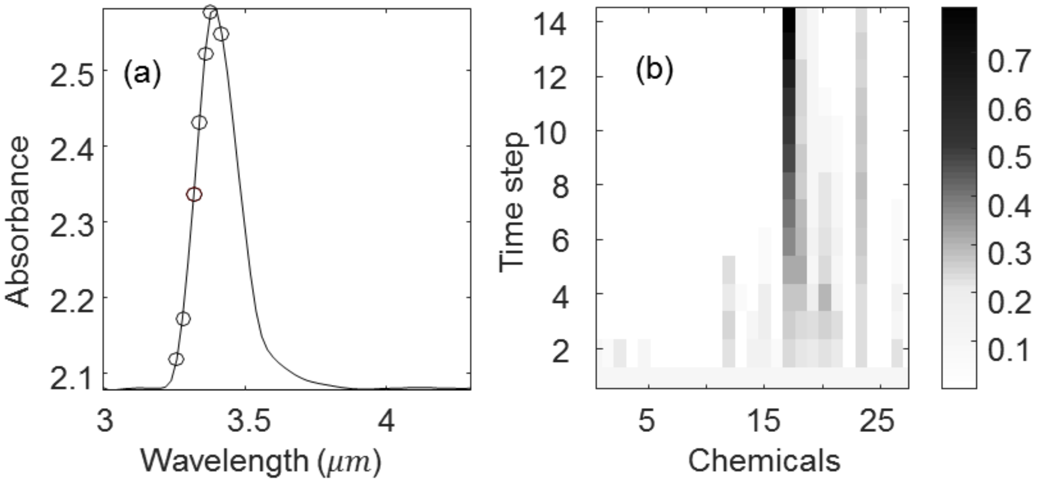


Figure 21: (a) Average absorption spectrum of trans-3-hexene (solid line) and wavelengths chosen by the active sensing algorithm (circles); the following wavelengths (μm) were chosen: 3.4, 3.36, 3.3, 3.34, 3.28, 3.38, 3.38, 3.44, 3.38, 3.44, 3.44, 3.38, 3.38, 3.44, and 3.34. (b) Belief distribution as a function of time.

4.2.3 Performances at sensing budgets

Next, we tested the active sensing algorithm for various settings of the misclassification and sensing costs. Adjusting these costs allows us to balance the total cost of sensing against the potential cost of misclassification. Without loss of generality, we used a 1-0 loss function for the classification costs $c_{uv} = I(u \neq v)$ and varied the sensing costs c_i from 0 to 0.2 in increments of 0.02. At each cost setting, we tested the algorithm 30 times on each of the 27 chemicals resulting in 27x30 = 510 test cases. Results are summarized in Figure 22. As shown in the figure, the classification rate deteriorates with increasing sensing costs. Also, the average number of sensing actions used reduced with increasing sensing costs. Hence, the method balances the classification performance with sensing cost. We also observed that at $c_i \geq 0.16$ the cost of taking a sensing action is higher than the expected reduction in risk. Therefore, the algorithm halts the sensing process immediately, resulting in a classification performance of 3.7% that corresponds to chance level performance for a problem with 27 classes.

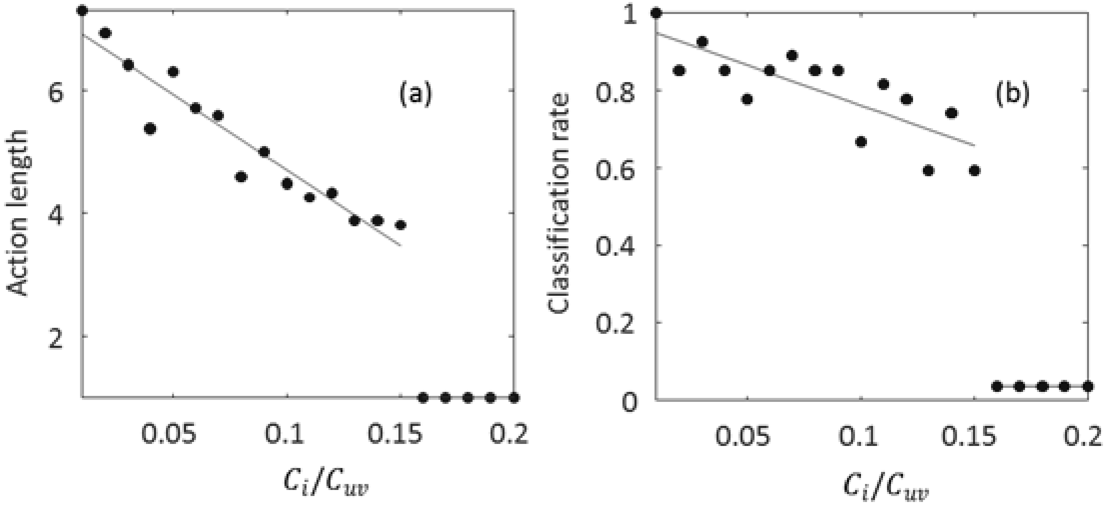


Figure 22: (a) Average number of observations used and (b) classification rate obtained by the active sensing framework as a function of the ratio of sensing cost to misclassification cost (c_i/c_{uv}).

4.2.4 Comparison with passive sensing

We also compared the active sensing method against a feature subset selection strategy. Feature selection is a passive process where the optimal subset is obtained off-line using training data. In contrast, active sensing selects features on-line. We used sequential forward selection (SFS) coupled with a wrapper objective function [15] to obtain 'optimal' feature subsets of different cardinalities (p); the wrapper was based on a naïve Bayes classifier. To ensure a fair comparison between the two methods, we modified the stopping criterion of the active-sensing algorithm such that the algorithm stopped as soon as it acquired p observations.

First, we conducted an experiment to compare the performance of the two methods with increasing levels of measurement noise. For this purpose, we generated training data from FTIR spectra (see section 4.2.1), added Gaussian noise of variance 0.05 and then trained sensor models. To test the active sensing and feature selection methods, we generated 20 test sets, each containing 270 spectra (10 times per chemical), by changing the variance in the noise from 0.05 to 1 in steps of 0.05. Figure 23 compares the classification performance of the two methods for $p = 15$ features. As shown, both strategies obtain nearly perfect classification performance at low noise levels. However, as noise levels increase active sensing consistently outperforms SFS. This is because active sensing selects features at measurement time in a way that adapts to noise levels, whereas SFS uses a pre-specified sequence that was computed off-line under more forgiving noise conditions.

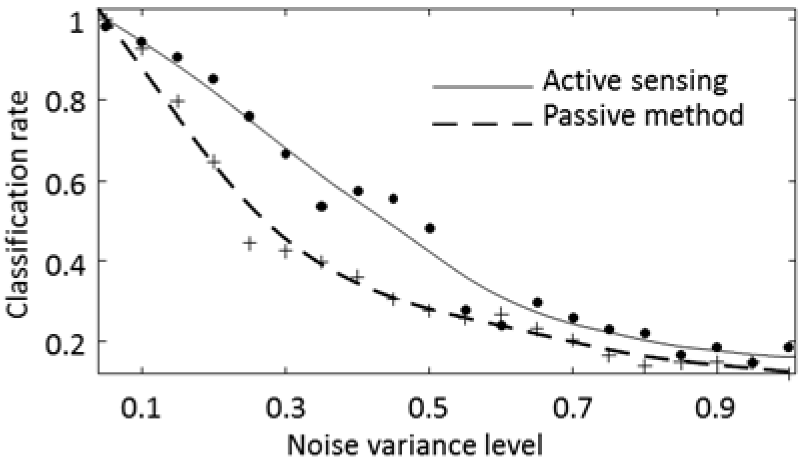


Figure 23: Classification performance of the two methods as a function of the variance of the additive Gaussian noise.

To further emphasize the advantages of active sensing over feature subset selection, we conducted a second experiment that compares the classification performance of the two methods with increasing values of p, the number of features or wavelengths. As before, we generated training data with an additive Gaussian noise of variance 0.05. Then, we evaluated both methods on test data with additive noise of variance 0.4. For each value of p, we ran each method 270 times (10 times per chemical). Results are summarized in Figure 24. As before, active sensing consistently outperforms SFS at all values of p.

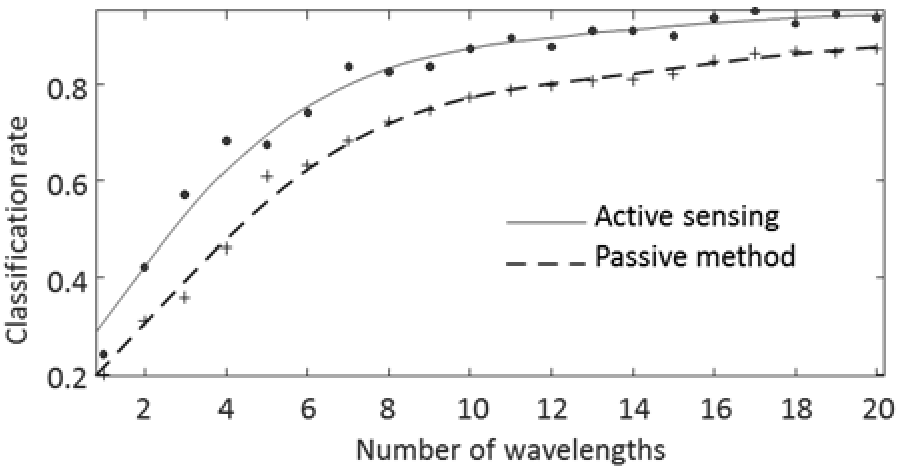


Figure 24 Classification performance of the two methods as a function of the number of observations used.

4.3 Validation on experimental data

4.3.1 Experimental setup

We also evaluated the active sensing framework on experimental data using an FPI device (LFP-3041L-337; Infratec, Inc). This device operates in the range of ($3 - 4.3\mu m$) and has a resolving power $\lambda/\Delta\lambda$ of 60. We used a broadband infrared pulsable source (INTX 20-1000-R; Intex, Inc.) operated at a 10Hz modulation frequency and 60% duty cycle. We mounted a 10cm gas cell (66001-10A; Specac, Inc.) with ZnSe window (602L08; Specac, Inc.) between the sensor and the IR source using an opto-mechanics fixture (Thorlabs, Inc.). This ensured a precise alignment of source, gas cell, and FPI sensor. The sensor was controlled using Matlab™ through a USB based evaluation board provided by the vendor. Chemicals were delivered to the system from 30ml glass vials using negative pressure with a pump connected downstream from the sample cell. Figure 25 shows the configuration of the device.

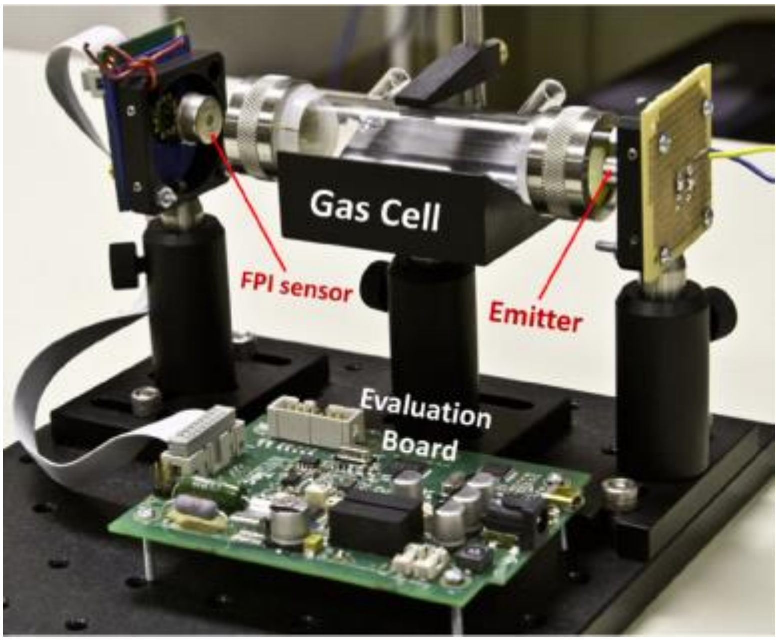


Figure 25: Experimental prototype of the Fabry-Perot spectrometer.

We tested the active-sensing algorithm and FPI prototype on a discrimination problem with eight chemicals; see Table 3. We operated the FPI sensor at 66 different wavelengths ranging from $3\mu m$ to $4.3\mu m$ in steps of $0.02\mu m$. The sensor response was sampled at a rate of 1 KHz. Since the sensor response is modulated by the emitter, we can minimize interferences (e.g., external infrared sources, electronic noise) by extracting the power only at the modulation frequency (10 Hz) using Goertzel's algorithm [98]. We estimated transmittance as the ratio of the sensor response (power at 10Hz) to the sample and to a reference gas (air), and converted transmittance Tr into absorption as $A = \log_{10} \frac{100}{Tr}$. The experiments were conducted in a laboratory environment at a temperature of 22.2 ºC and standard atmospheric pressure of 760 mmHg. Before each experiment, we acquired the sensor's response to air and used these values as reference to estimate the absorption spectra. This helped us minimize the effects caused by changes in temperature, pressure, or humidity between experiments.

Table 3 List of chemicals and their major components

Index	Chemical	Components
1	Air	
2	Brush cleaner	Raffinates, Acetone, Methanol
3	Lacquer thinner	Toluene, Methanol, Hexane, Light aliphatic naphtha
4	Denatured alcohol	Ethyl alcohol, Methanol
5	Acetone	Acetone
6	Xylene	Xylene (mixed isomers), Ethylbenzene
7	Isopropyl alcohol	Isopropyl alcohol
8	Propanol	Propanol

4.3.2 Experimental data

We collected 50 absorption spectra for each chemical by varying the concentration from 0% to 100% in steps of 2%. Figure 26 (a) shows the average absorption spectra for the eight chemicals, obtained by averaging the 50 absorption spectra. We then applied NNMF to obtain the concentration-independent absorption profiles, shown in Figure 26 (b). We generated training data for each chemical using the NNMF profiles and residual matrices, as described in section 4.1.3, from which GMMs were trained. We experimented with various numbers of Gaussian components per GMM, but GMMs with a single component proved sufficient since the NNMF residual noise was approximately Gaussian-distributed.

We tested the framework 20 times for each chemical, resulting in 160 test cases. For each chemical, the concentration of the test sample was randomly varied in the range of

20% to 80%. Chemicals were introduced in a randomized order to avoid any systematic errors or drift, and the gas cell was flushed with air for 2 minutes between exposures to remove residuals from the previous sample. When introducing a sample into the gas cell, we monitored the sensor response continuously until it stabilized; this ensured that the sample concentration had reached equilibrium. We used the sensor's average response (over 20 repetitions) to air as the reference.

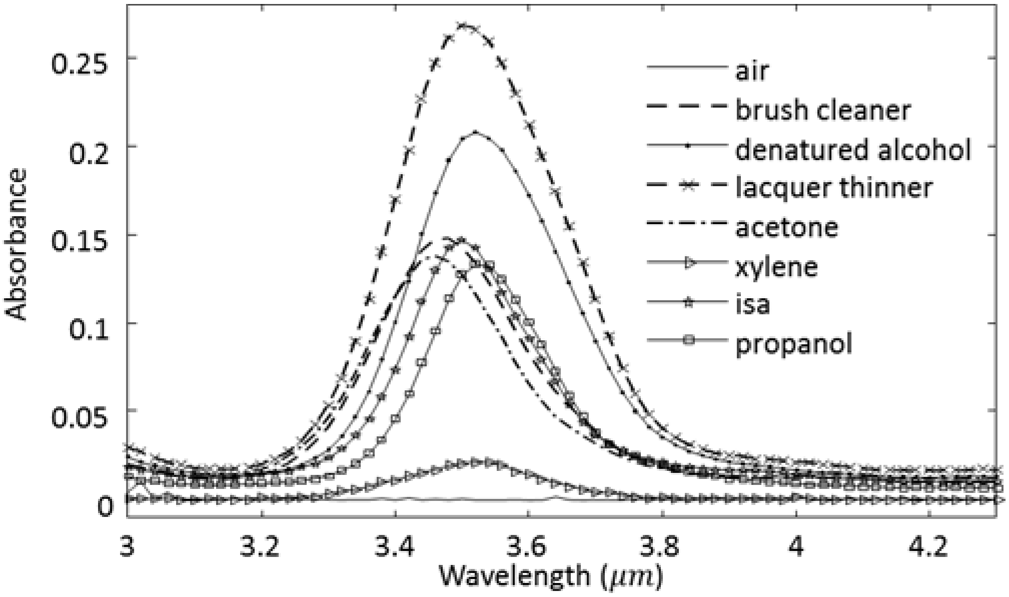


Figure 26 (a) Average absorption spectra of all the chemicals including air. (b) Concentration-independent absorption profiles of all the chemicals.

4.3.3 Results

Based on the simulation results obtained in section 4.2.3, we chose to set the ratio of sensing cost and misclassification cost as 0.02 to promote high classification performance. Classification results and the corresponding confusion matrix are shown in Figure 27(a) and (b), respectively. Denatured alcohol and lacquer thinner are correctly classified 100% of the times, followed by air, brush cleaner, acetone, and isopropyl alcohol, which are classified accurately on more than 75% of the test cases. In contrast, propanol is classified 50% of the times as lacquer thinner; the two chemicals have highest absorption strength (peaks) at the same wavelength. Xylene is most often misclassified as air because of its low absorption strength in the sensor's spectral range, which is comparable to that of sensor noise variance. Also, the absorption profile of xylene obtained using NNMF is significantly noisy compared to other chemicals (see Figure 26 (b)), especially in the range $3-3.3\mu m$ and $3.7-4.3\mu m$. We also observed that brush cleaner and acetone are often misclassified as air at low concentrations.

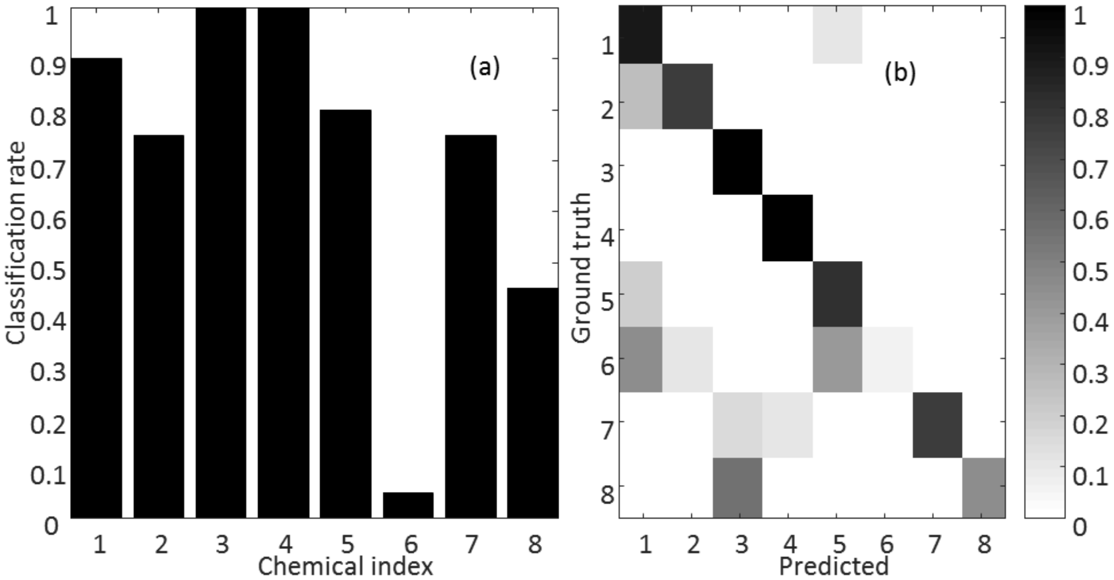


Figure 27 (a) Classification performance (true positives) for different chemicals and (b) the corresponding confusion matrix.

To test the robustness of the concentration-normalization method, we tested the framework on ten samples of acetone at each of 10 concentrations, ranging from 100%

(pure sample) down to 10% in steps of 10%, for 100 test cases. Acetone concentration was controlled using a gas diluter [4]. Results are summarized in Figure 28. The active sensing method accurately identified all samples in the concentration range 20-100% and only failed at a 10% concentration, in which case all samples were classified as air. Figure 28 (b) shows the average number of actions used at different concentrations. The average number of observations used increased as concentration decreased, from 5 observations at a 100% concentration up to 9.3 observations at a 20% concentration. This result is consistent with the fact that the SNR decreases with concentration, and shows how the active sensing method can adapt the number of measurements required in order to obtain sufficient evidence for classification.

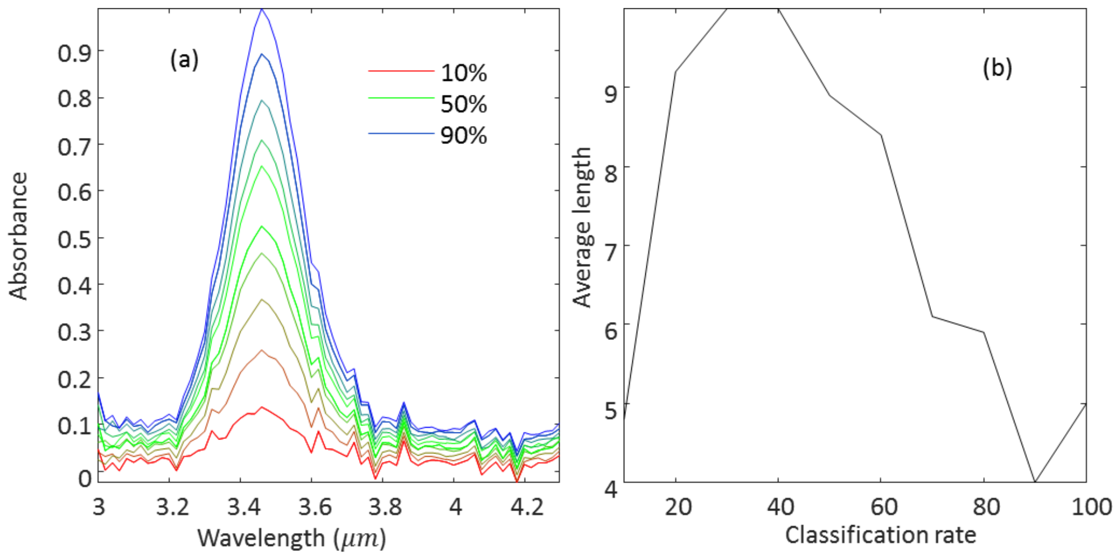


Figure 28 (a) Average absorption spectra of acetone and (b) average number of observations at the ten concentrations.

4.4 Conclusion and discussion

We have presented an approach to actively select absorption wavelengths for a tunable IR interferometer in the context of concentration-independent discrimination of chemical samples. Our approach first creates concentration-independent absorption profiles for each target chemical using non-negative matrix factorization (NNMF). The resulting normalized responses are then modeled using Gaussian mixture models. We formulated sensing as a decision-theoretic process, where we sequentially select wavelengths that are expected to provide the best reduction in Bayes risk. We validated the proposed method on both simulated and experimental data. Results on simulated data show that the passive sensing can outperform active sensing regarding classification rates for various sensing budgets and at various levels of sensor noise. Using a Fabry-Perot interferometer, we further validated experimentally that the active sensing method can identify chemical samples independently of their concentration.

Our experimental results also show an interesting anomaly: samples at 10% concentration only triggered 4.8 observations on average and were all misclassified as air. We believe this result reflects the limitations of sensor's sensitivity. At low concentrations, the light beam interacts with fewer analyte molecules, which results in very weak absorption spectra; as shown in Figure 17, the peak absorption values become comparable to sensor noise. As a result, the first few observations obtained drives up the belief associated with air, which tricks the algorithm into bringing the sensing process to an early termination. This limitation can be addressed, and the overall sensitivity of the system improved, by either: a) increasing the length of the optical path, which would increase the number of analyte molecules encountered by the light beam and thus increase the signal strength; or b) incorporating a pre-concentrator (PCT) into the gas delivery system, which would increase the concentration of the gas samples by 1-2 orders of magnitude.

To create concentration-independent absorption profiles, our approach uses NNMF, a technique that it is only applicable to non-negative matrices. However, at low concentrations (or if an analyte has low absorption strength at certain wavelengths), the sensor response can become negative. Whenever this occurs, our current implementation replaces negative values with zeros before applying NNMF. This affects the factorization process, since the negative noisy responses are replaced by zeros but their positive counterparts are not. This effect can be seen in xylene's absorption profile of

Figure 26 (b), which is jagged as compared to the other chemicals. This is more evident in the wavelength ranges of $3 - 3.2\mu m$, and $4 - 4.3\mu m$, where many negative values have been replaced. An alternative solution to this problem would be to apply a denoising technique to the spectra prior to performing NNMF.

5. ACTIVE WAVELENGTH SELECTION BASED ON MULTI-MODAL LEAST SQUARES METHOD FOR CHEMICAL MIXTURE IDENTIFICATION[8]

This chapter presents an adaptive algorithm that interleaves the wavelength-selection and sensing processes based on misclassification risk estimated using a multi-modal least squares solver. Given a set of previous measurements (absorption at specific wavelengths), the algorithm generates a pool of candidate solutions, each solution representing a vector of concentrations across all the chemicals in the library, then selects the next wavelength that maximizes discrimination among the candidate solutions. In this fashion, the algorithm can be viewed as adaptively generating a training set of chemical mixtures for the wavelength-selection process. As the sensing process continues, the training set becomes closer to the test sample, and the selected wavelengths grow more relevant. A weighting function over candidate solutions according to their fitness (consistency with the measured wavelengths) can then be used to bias the algorithm towards exploration (e.g., during the initial stages) or exploitation (e.g., to promote convergence in the final steps).

[8] The description of the method and the experimental results are reprinted with permission from "Active Wavelength Selection for Mixture Analysis with Tunable Infrared Detectors" by Huang and Gutierrez-Osuna, 2015. Sensors and Actuators B: Chemical, pp. 245-257, ©2015 Elsevier.

We validated the approach through a series of experiments with real and synthetic data. First, we established proof-of-concept on real data from an FPI prototype exposed to binary mixtures (background/foreground) randomly selected from a small library of volatile organic compounds. To make the problem challenging, the foreground component was set to have a concentration 1-2 orders of magnitude below that of the background component; at these levels, repeated sampling at the informative wavelengths for the weaker foreground is necessary before it can be detected. Then, we characterized the approach through simulations of binary mixtures with different degrees of numerical ill-conditioning (increasingly similar mixture components). Finally, we extended the simulation to complex mixture problems containing up to 15 chemical components from a library of 500 analytes. In all three sets of experiments, a passive algorithm (sequential forward selection) was used as a baseline for comparison purposes. Our results indicate that the active strategy outperforms the passive strategy systematically, particularly in the presence of noise or numerical ill-conditioning.

The remaining sections of this chapter are organized as follows. Section 5.1 describes the proposed active sensing framework, including a strategy to balance the exploration and the exploitation. Section 0 presents an experimental evaluation of our framework on two-chemical mixture problems with a Fabry-Perot interferometer. Section 5.3 provides a thorough evaluation of the framework on high-order mixtures using synthetic spectra from a library of 500 chemicals. This chapter concludes with a discussion of results and directions for future work.

5.1 Methods

5.1.1 Overview of the approach

To solve the constrained linear system in equation (9), the proposed algorithm operates in two broad strokes: at each sensing step, it generates a pool of sparse solutions based on previous measurements, then selects the next wavelength to maximize separability among the solutions in the pool. The approach is illustrated in Figure 29. With the arrival of the i-th measurement at wavelength λ_i, the algorithm generates a number of candidate solutions $\{\boldsymbol{x}_1, \boldsymbol{x}_2 \dots \boldsymbol{x}_K\}$ through a non-negative least squares solver (see section 5.1.2); the use of multiple solutions is needed given the instability of the problem. Each solution (a vector containing the concentration of the C chemicals in the library) is then transformed into the estimated full spectrum $\boldsymbol{b}_i = \boldsymbol{A}\boldsymbol{x}_i$, and each spectrum is weighted according to its fitness based on the Akaike information criterion (AIC) to prevent overfitting (see section 5.1.3). The wavelength $\lambda'$with largest weighted variance $\sigma_{\boldsymbol{W}}$ across the solution pool (each solution weighted by its fitness) is then chosen as the next measurement, and the process is repeated. Though other selection criteria may be used (e.g., maximize correlation, goodness of fit, mutual information [99]), weighted variance $\sigma_{\boldsymbol{W}}$ is fast to compute and is an approximation of the misclassification risk of the candidates (see APPENDIX C:). The individual steps are described in more detail in the following subsections.

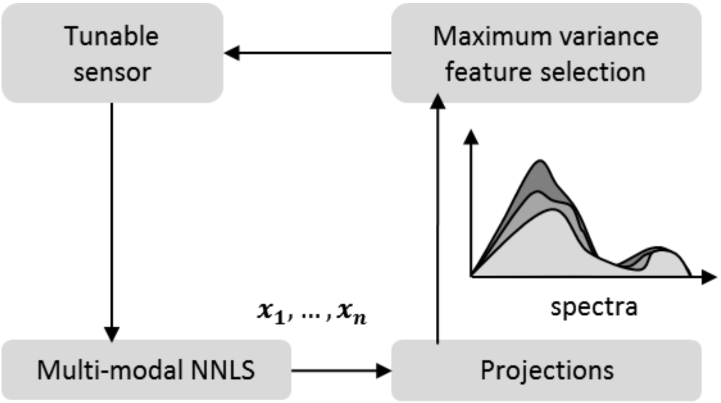


Figure 29: Overview of the multimodal wavelength selection approach.

5.1.2 Generating candidate solutions: multi-modal non-negative least squares

The most critical component of the active sensing framework is the multivariate solver for the underlying linear system in equation (9). Our implementation uses a non-negative least squares (NNLS) algorithm [18] based on matching pursuit, a numerical technique that finds the best matching projections onto an over-complete dictionary [100]. To promote sparsity, the NNLS algorithm starts with an all-zero solution vector $\boldsymbol{x} = [0\ 0 \ldots 0]^T$**,** and adds one nonzero entry to $\boldsymbol{x}$ at a time, each entry representing a chemical component; namely, the algorithm computes the gradient term $\boldsymbol{\nabla}_{\boldsymbol{x}} = \boldsymbol{A}^T(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})$ for each chemical in the library, and adds the one whose gradient is largest. Once a new entry has been selected, the pseudo-inverse solution[9] is used to find the minimum – minus the constraints. At this point, if any of the non-zero elements is negative, the

[9] Using Matlab notation, the pseudo-inverse solution can be computed as $\left[\boldsymbol{x}'^T \delta\right]^T = [\mathbf{A}(:,\mathbf{P}')\ \mathbf{1}]\backslash$ **b**

solution $\boldsymbol{x}'$ is shifted back to the closest point in the feasible area (along any one axis) using the iteration:

```
while $\exists i \mid \boldsymbol{x}'(i) < 0$
   $\boldsymbol{x}' \leftarrow \boldsymbol{x}' - \alpha(\boldsymbol{x}' - \boldsymbol{x})$
   where $\alpha = \min\left\{\frac{x'(j)}{x'(j) - x(j)} : j \in \boldsymbol{P}'\right\}$        (23)
end while
```

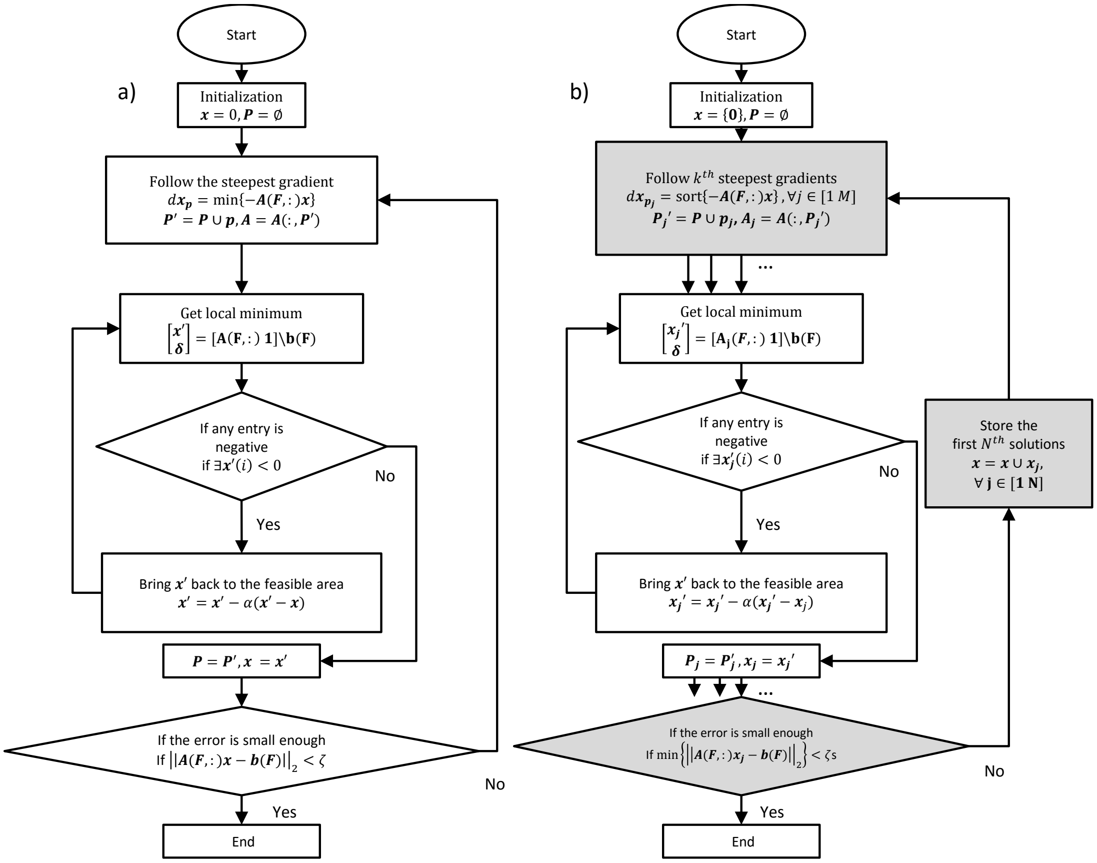
where $\boldsymbol{P}$ and $\boldsymbol{P}'$ are the set of indices of nonzero entries in $\boldsymbol{x}$ and $\boldsymbol{x}'$, respectively, $\boldsymbol{x}$ is the previous feasible solution, i.e., before the non-zero entry was added, and $\boldsymbol{x}'(i)$ represents the estimated concentration for the i-th chemical in the library. The iteration in equation (23) is repeated until all entries in $\boldsymbol{x}'$ are non-negative.

The process of estimating the local minimum (via the pseudo-inverse) and adjusting to the non-negative constraint[10] is repeated until no more negative entries exist. If the error term over measured wavelengths $\hat{\epsilon} = \|\boldsymbol{b}(\boldsymbol{F}) - \boldsymbol{A}(\boldsymbol{F},:)\boldsymbol{x}'\|_2$ is large, a new entry is added to the solution vector, and the NNLS process is repeated; otherwise, the process is terminated. A flowchart of the overall process is illustrated in Figure 30 (a).

[10] Note that the offset δ is not included in the adjusted term $\boldsymbol{x}'$ since δ has no constraints. Instead, δ is re-estimated together with the adjusted term $\boldsymbol{x}'$ (through the pseudo-inverse) after the constraint adjustment. Using the pseudo-inverse ensures that the offset is local-optimal –this is in contrast to setting δ to the minimum observed value in $\boldsymbol{x}'$, a common heuristic in NNLS.

5.1.2.1. Multi-modality: tracking multiple solutions

The NNLS algorithm generates a single solution, which is problematic for two reasons. First, the underlying linear problem is often ill-defined since only a few wavelengths $\boldsymbol{F}$ are measured; this is particularly severe at the beginning of the sensing process. Second, the process of wavelength selection requires a training set of mixtures from which to rate individual wavelengths. To address these issues, our approach wraps the NNLS algorithm around a multi-modal loop to generate multiple solutions. In contrast with heuristic multi-modal optimization techniques, which use random seeds for the search [101], our algorithm takes advantage of the gradient information $\left(\boldsymbol{\nabla_x} = \boldsymbol{A^T}(\boldsymbol{b} - \boldsymbol{Ax})\right)$ and the closed-form solution of ordinary linear least squares (the pseudo-inverse), which significantly improve computational efficiency. Starting with an all-zero solution vector $\boldsymbol{x}$, the algorithm selects the top M entries (according to their gradient $\boldsymbol{\nabla_x}$), and transforms each into a feasible solution, as described earlier –see equation (23). The result is a set of M solutions with one non-zero entry (a single chemical). The best N of these M solutions according to their sum squared error ($\epsilon = \|\boldsymbol{b}(\boldsymbol{F}) - \boldsymbol{A}(\boldsymbol{F},:)\boldsymbol{x}'\|_2$) are saved to the candidate pool, and used as seeds for the next step to yield $N \times M$ solutions with 2 non-zero entries (two chemicals). The best N of those according to their error ϵ are added to the pool, and used as seeds for the next step (solutions with 3 non-zero entries). The process continues until the candidate pool contains a solution whose error ϵ is below a pre-specified threshold; duplicate solutions, which may occur due to the parallel searches, are removed in a final step. Parameters M and N are set by the user depending on the computing resources available. A flowchart of the resulting MM-NNLS algorithm is shown in Figure 30 (b).



a)
Start
Initialization
$\boldsymbol{x} = 0, \boldsymbol{P} = \emptyset$
Follow the steepest gradient
$d\boldsymbol{x_p} = \min\{-\boldsymbol{A}(\boldsymbol{F},:)\boldsymbol{x}\}$
$\boldsymbol{P}' = \boldsymbol{P} \cup \boldsymbol{p}, \boldsymbol{A} = \boldsymbol{A}(:,\boldsymbol{P}')$
Get local minimum
$\begin{bmatrix}\boldsymbol{x}'\\ \boldsymbol{\delta}\end{bmatrix} = [\mathbf{A}(\mathbf{F},:)\ \mathbf{1}]\backslash\mathbf{b}(\mathbf{F})$
If any entry is negative
if $\exists \boldsymbol{x}'(i) < 0$
No
Yes
Bring $\boldsymbol{x}'$ back to the feasible area
$\boldsymbol{x}' = \boldsymbol{x}' - \alpha(\boldsymbol{x}' - \boldsymbol{x})$
$\boldsymbol{P} = \boldsymbol{P}', \boldsymbol{x} = \boldsymbol{x}'$
If the error is small enough
If $||\boldsymbol{A}(\boldsymbol{F},:)\boldsymbol{x} - \boldsymbol{b}(\boldsymbol{F})||_2 < \zeta$
No
Yes
End
b)
Start
Initialization
$\boldsymbol{x} = \{\mathbf{0}\}, \boldsymbol{P} = \emptyset$
Follow k^{th} steepest gradients
$d\boldsymbol{x}_{\boldsymbol{p}_j} = \mathrm{sort}\{-\boldsymbol{A}(\boldsymbol{F},:)\boldsymbol{x}\}, \forall j \in [1\ M]$
$\boldsymbol{P_j}' = \boldsymbol{P} \cup \boldsymbol{p_j}, \boldsymbol{A_j} = \boldsymbol{A}(:,\boldsymbol{P_j}')$
...
Get local minimum
$\begin{bmatrix}\boldsymbol{x_j}'\\ \boldsymbol{\delta}\end{bmatrix} = [\mathbf{A_j}(\boldsymbol{F},:)\ \mathbf{1}]\backslash\mathbf{b}(\mathbf{F})$
If any entry is negative
if $\exists \boldsymbol{x}_j'(i) < 0$
No
Yes
Bring $\boldsymbol{x}'$ back to the feasible area
$\boldsymbol{x_j}' = \boldsymbol{x_j}' - \alpha(\boldsymbol{x_j}' - \boldsymbol{x_j})$
$\boldsymbol{P_j} = \boldsymbol{P}_j', \boldsymbol{x_j} = \boldsymbol{x_j}'$
...
If the error is small enough
If $\min\left\{||\boldsymbol{A}(\boldsymbol{F},:)\boldsymbol{x_j} - \boldsymbol{b}(\boldsymbol{F})||_2\right\} < \zeta$s
No
Yes
End
Store the first N^{th} solutions
$\boldsymbol{x} = \boldsymbol{x} \cup \boldsymbol{x_j}$,
$\forall\ \mathbf{j} \in [\mathbf{1}\ \mathbf{N}]$

Figure 30: Flowchart of the non-negative least-squares (NNLS) algorithm (a) and multi-modal NNLS algorithm (b). The shaded blocks in (b) highlight differences between both algorithms.

5.1.3 Wavelength selection

The MM-NNLS algorithm returns a pool of candidate solutions, where each solution represents a mixture by its concentration vector across all chemicals in the library. Thus, each concentration vector $\boldsymbol{x_i}$ can be used to reconstruct the full spectrum of the corresponding mixture:

$$\boldsymbol{b}_i = \boldsymbol{A}\boldsymbol{x}_i \tag{24}$$

It is this collection of spectra that can be used as a "training set" for wavelength selection. Caution must be exercised, however, because these spectra are synthesized from solutions $\boldsymbol{x}_i$ that were obtained by fitting a small number of absorption wavelengths $\boldsymbol{b}(\boldsymbol{F})$ – see equation (9). As a result, there is the distinct possibility that some of the solutions overfit the noisy measurements $\boldsymbol{b}(\boldsymbol{F})$. This is particularly problematic at low concentrations, where measurement noise can dominate the sensor response, which may lead the MM-NNLS algorithm to include additional entries (chemicals) in the solutions.

We illustrate this problem with an example. For this purpose, we randomly selected a binary mixture from a spectral library containing 500 chemicals, then randomly sampled absorption at 20 wavelengths, and added 1% white noise to each of the 20 measurements –refer to section 5.3 for details on the spectral library. Then, we allowed NNLS to generate a number of solutions. Results are illustrated in Figure 31(b), with solutions ranked according to the <u>sampled error</u> ($\hat{\epsilon} = ||\boldsymbol{b}(\boldsymbol{F}) - \boldsymbol{A}(\boldsymbol{F},:)\boldsymbol{x}\,||_2$), as well as the <u>true error</u> ($\epsilon = ||\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}||_2$) with respect to the noise-free ground-truth full spectrum (assumed known in this case). Figure 31(a) shows the solution $\boldsymbol{x}_1$ with the lowest sampled error $\hat{\epsilon}_1$, and the solution $\boldsymbol{x}_2$with the lowest true error ϵ_2. Even though solution $\boldsymbol{x}_1$ has lower sampled error ($\hat{\epsilon}_1 < \hat{\epsilon}_2$), it overfits the noisy measurements $\boldsymbol{b}(\boldsymbol{F})$ by using 31 non-zero entries, as shown in Figure 31(c); notice also how solution $\boldsymbol{x}_1$ deviates quite significantly from ground truth at $\lambda = 8.87, 9.07\ \mu m$. In contrast, solution $\boldsymbol{x}_2$ has only 4 non-zero entries and has smaller true error ($\epsilon_2 < \epsilon_1$) but unfortunately ranks in position #89 according to the sampled error (the only error that can be measured in practice).

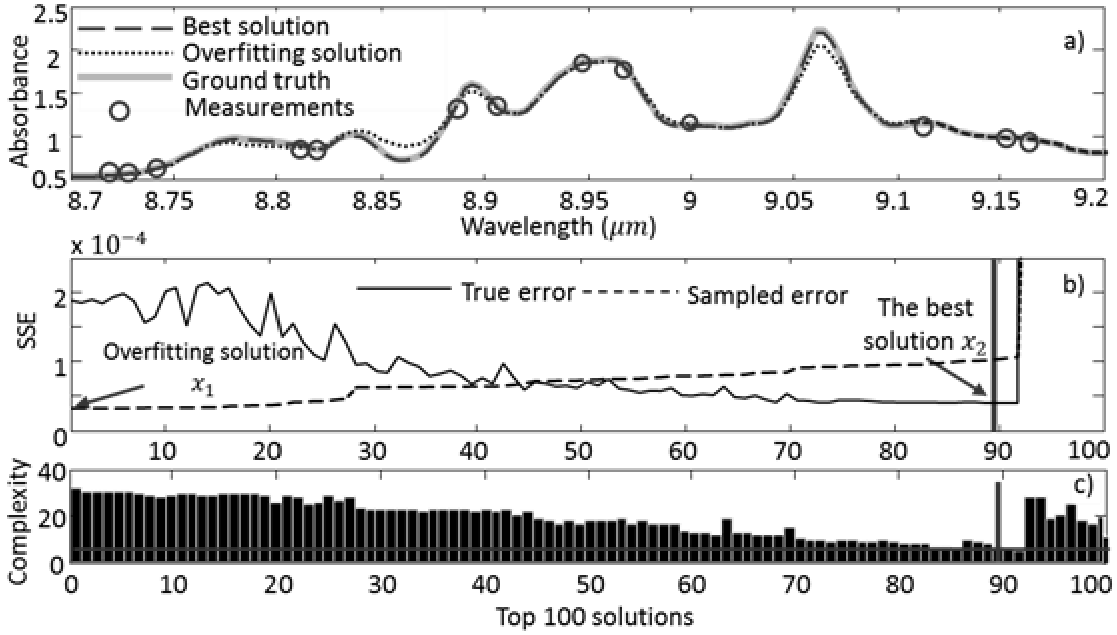


Figure 31: (a) Projected spectra of the solutions ranked by the sampled error $(\|\boldsymbol{b}(\boldsymbol{F}) - \boldsymbol{A}(\boldsymbol{F},:)\boldsymbol{x}'\|_{\boldsymbol{2}})$ and the true error over the full spectrum $(\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}'\|_{\boldsymbol{2}})$; only the range $\boldsymbol{\lambda} = 8.7 - 9.2\ \mu m$ is shown for illutration purposes. (b) The top 100 solutions according to the sampled error, and the corresponding true error. (c) Complexity of each model (number of components in the mixture) for the top 100 solutions.

The above example illustrates how, in the presence of noise, the solution with the smallest sampled error can overfit the measurements by using extra complexity (the number of chemicals in the mixture). To address this issue, we rank solutions based on the Akaike information criterion (AIC), an information-theoretic measure that takes into account both error and parsimony [102]. The AIC score can be computed as:

$$AIC_i = -2logL_i + 2V_i \tag{25}$$

where L_i is the likelihood for candidate model i, and V_i is its number of free parameters (number of components in the mixture). For linear regression, assuming the sensor noise at each measurement is uncorrelated with the sensor noise of previous measurements, the formula can be further simplified as [103]:

$$AIC_i = n \times log\left(\frac{\hat{\epsilon}_i}{n}\right) + 2V_i \tag{26}$$

where n is the number of measured wavelengths, and $\hat{\epsilon}_i$ is the sum-squared error for model i, i.e., $\hat{\epsilon}_i = \|\boldsymbol{b}(\boldsymbol{F_i}) - \boldsymbol{A}(\boldsymbol{F_i},:)\boldsymbol{x_i}\|_2$. It is these AIC scores that we use to rank candidate solutions, following conversion into likelihoods [102]:

$$w_i = e^{-\frac{1}{2}\Delta(AIC_i)} \tag{27}$$

where $\Delta(AIC_i)$ is the difference in AIC scores between model We and the best candidate: $\Delta(AIC_i) = AIC_i - \min_{\forall i} AIC_i$. We then normalize the weights to ensure they add up to one:

$$w_\mathrm{i} \leftarrow \frac{w_\mathrm{i}}{\sum_{\forall i} w_\mathrm{i}} \tag{28}$$

5.1.3.1. Selection criterion

A number of traditional selection criteria (maximize correlation, the goodness of fit, mutual information [99]) can be used at this point to determine the next wavelength to be sampled. In prior work [104] we used variance as a measure of uncertainty, choosing as the next measurement the wavelength having the highest variance across spectra in the solution pool $\{\boldsymbol{b_i}|\forall i\}$:

$$\lambda_{opt} = \max_{\lambda_j \mid \forall j \in [1\, N]} \left\{ \sigma^2 \{ \boldsymbol{b_i}(\lambda_j) | \forall i \} \right\} \tag{29}$$

where λ_{opt} is the next wavelength to be sampled, $\boldsymbol{b_i}(\lambda_j)$ is the (estimated) absorption at wavelength λ_j for mixture solution i, and σ^2 is the variance operator. As shown in APPENDIX C:, the variance at each wavelength approximates the misclassification risk. Thus, by selecting the wavelength with largest variance the algorithm can be viewed as minimizing the risk of choosing the wrong candidate.

However, this approach treats all candidate solutions equally, regardless of their fitness. To address this issue, we then weigh each candidate solution as:

$$\lambda_{opt} = \max_{\lambda_j \mid \forall j \in [1\, N]} \left\{ \sigma^2_{\boldsymbol{W}} \{ \boldsymbol{b_i}(\lambda_j) | \forall i \} \right\} \tag{30}$$

where $\sigma^2_{\boldsymbol{W}}$ is the weighted variance, which can be calculated as:

$$\sigma^2_{\boldsymbol{W}}(\boldsymbol{x}) = \frac{1}{1 - \sum_i w_i^2} \sum_{i=1}^{N} w_i (x_i - \mu_{\boldsymbol{W}})^2 \tag{31}$$

and $\mu_{\boldsymbol{W}}$ is the weighted mean, calculated as:

$$\mu_{\boldsymbol{W}} = \sum_{i=1}^{N} w_i x_i \tag{32}$$

5.1.3.2. Balancing exploitation and exploration

During the initial stages, when only a few measurements are available, the sampling process can be dominated by a few solutions. Whenever this happens, the algorithm invests a large number of measurements to investigate a narrow region of the spectrum (to discriminate among the few early solutions) rather than explore the global structure of the spectrum in search of new candidates. This often leads to premature convergence. To guard against this problem, our implementation includes a parameter ς^2 that can be used to adjust the spread of the AIC weights:

$$w_i = e^{-\frac{\Delta(AIC_i)}{2\varsigma^2}} \tag{33}$$

Selecting ς^2 is non-trivial, as its value reflects on the credibility of the estimated solutions. If ς^2 is small, the solutions will be weighted aggressively, and active sensing runs the risk of premature convergence to a suboptimal solution. In contrast, if ς^2 is large, the weights become uniform regardless of how different they were before normalization, which may lead to irrelevant features being introduced. Thus, the weighting scheme controls how many solutions to consider for the next measurement, a trade-off commonly referred to as the exploration-exploitation dilemma [105].

To balance this exploration-exploitation dilemma, we propose an entropy-guided method that adjusts the offset parameter ς^2 such that the entropy of the weight landscape $H(\boldsymbol{W}_O)$ remains constant. Assuming l candidate solutions, the highest entropy ($\log(l)$) is achieved when the weights are uniformly distributed, whereas the lowest entropy[11] (0) is obtained when only one of the l solutions has a non-zero weight. Finally, we select a value between these two extremes:

[11] $\sup\{H(\boldsymbol{W})\} = -\sum_{i=1}^{l} \frac{1}{l} \log\left(\frac{1}{l}\right) = \log(l);\ \inf\{H(\boldsymbol{W})\} = -1\log(1)\ = 0.$

$$\dot{H} = (\sup\{H(\boldsymbol{W})\} - \inf\{H(\boldsymbol{W})\})\,\alpha + \inf\{H(\boldsymbol{W})\} = (\log(l) - 0)\alpha + 0 = \alpha\log(l) \tag{34}$$

where the multiplier α $(0 \leq \alpha \leq 1)$ controls the balance between exploitation and exploration: i.e., $\alpha = 0$ leads to extreme exploitation whereas $\alpha = 1$ leads to extreme exploration. In the studies reported here, we use $\alpha = 0.5$ to balance exploration and exploitation. Once the desired entropy $\dot{H}$ has been fixed, the last step is to find the corresponding parameter ς^2. Since entropy grows monotonically as ς^2 increases, this can be easily done with a continuous linear binary search; see Table 4.

Table 4: bSearch

```
Input: $f(*)$, $\varsigma^2_{min}$, $\varsigma^2_{max}$, $\dot{H}$
Output: $\varsigma^2$
if $\varsigma^2_{max} \geq \varsigma^2_{min}$
    if $\dot{H} < f(\varsigma^2_{min})$
        return $\varsigma^2_{min}$
    $\varsigma^2 = \left(\dot{H} - f\left(\varsigma^2_{min}\right)\right)\frac{(\varsigma^2_{max} - \varsigma^2_{min})}{f(\varsigma^2_{max}) - f(\varsigma^2_{min})} + \varsigma^2_{min}$
    if $\varsigma^2_{max} - \varsigma^2_{min} > 10^{-2}$ // Resolution of the search
        if $\left|f(\varsigma^2) - \dot{H}\right| < 10^{-2}$
            return $\varsigma^2$;
        elseif $f(\varsigma^2) < \dot{H}$
            return $bSearch(f(*), \varsigma^2, \varsigma^2_{max}, \dot{H})$
        elseif $f(\varsigma^2) > \dot{H}$
            return $bSearch(f(*), \varsigma^2_{min}, \varsigma^2, \dot{H})$
return $\emptyset$
```

5.2 Validation on experimental data

5.2.1 Experimental setup

For the experiments described here, we used a long-wave FPI sensor (LFP-80105, Infratec, Inc) with 107 tunings (absorption lines) in the range $8 - 10.5\mu m$, coupled with a collimated broadband IR source (INTX 20-1000-R; Intex, Inc.) modulated at 10Hz and 50% duty cycle. We mounted a 10cm gas cell (66001-10A; Specac, Inc.) with ZnSe window (602L08; Specac, Inc.) and a ZnSe focusing lens (LA7542-F, Thorlabs, Inc.). The FPI, IR source and sample cell were mounted onto an opto-mechanics fixture (Thorlabs, Inc.) to ensure precise alignment. The FPI device was controlled using Matlab™ through a USB evaluation board provided by the vendor.

The sample delivery system is illustrated in Figure 32. Vapors from the headspace of 30mm glass vials are delivered using negative pressure with a pump connected downstream from the sample cell. The pump is modulated at 0.125 Hz with 20% duty cycle to avoid exhausting the headspace and keep the sample concentration relatively stable. Two diluters (1010 precision gas diluter, Custom Sensor Solutions, Inc.) independently mix the foreground and background sample vapors with dry air. Since water and carbon dioxide have major peaks outside of the sensor's range, air has a negligible contribution to the spectrum.

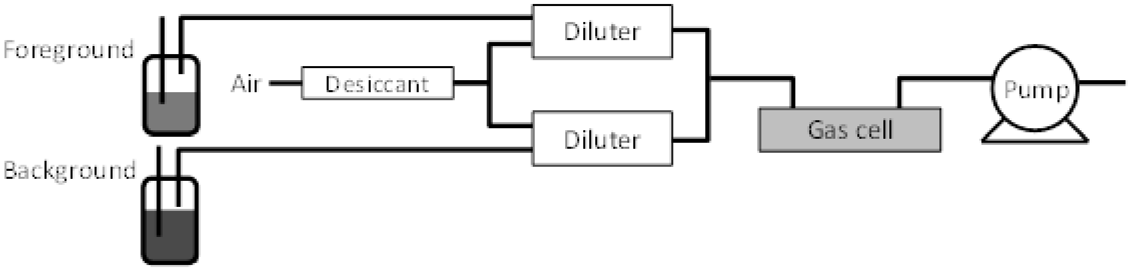


Figure 32: Schematic diagram of the headspace vapor sampling system.

Eight different volatile commercial chemicals that show absorption peaks in the range of our sensor ($8 - 10.5\ \mu m$) were used for the experiments; see Table 5. Of those, acetone was chosen as the strong background because it has the strongest absorption peak of all. The remaining seven chemicals were randomly chosen as the weak foreground. Experiments were conducted in a laboratory environment at a temperature of 22.2 ºC and standard atmospheric pressure of 760 mmHg.

Table 5: List of chemicals used in the experiments, and their major components

Chemical	**Components**
Propanol	Propanol
Acetone (background)	Acetone
Ethyl alcohol	Ethyl alcohol
Isopropyl alcohol	Isopropyl alcohol
Tert-Butyl alcohol	Tert-butyl alcohol
Air	Air and sensor drift
Denatured alcohol	Ethyl alcohol, methanol
Brush cleaner	Raffinates, acetone, methanol
Lacquer thinner	Toluene, methanol, hexane, light aliphatic naphtha

5.2.2 Experiment 1: test case

In a first experiment, we illustrate the performance of the active wavelength selection algorithm on a two-chemical mixture problem containing acetone at 2.5% dilution as background, and isopropyl alcohol at 5% as foreground. Figure 33 (a) shows the full spectra of the background, foreground, and the final mixture; circles represent the actual measurements that took place during the sensing process. The background chemical shows a major peak at $8.3\mu m$, while the foreground has a minor peak at $8.8\mu m$. Figure 33 (d) shows the rank of the correct solution as iterations progress; the correct solution is added to the pool at the 5-th measurement, and is confirmed (ranked #1) at the 20-th measurement.

Figure 33 (c) shows the total number of solutions considered by the algorithm as the iterations progress, whereas Figure 33 (b) shows the distribution[12] of selected wavelengths before and after the 20-th measurement. We observe a typical two-stage pattern emerging from the active sensing process: at first, the algorithm performs a broad sampling of absorption peaks for both chemicals (the exploration stage), then performs a focused search on spectral details and smaller peaks to confirm the identity of the weaker chemical (the exploitation stage). The most selected wavelength is around $8.8\mu m$, which is consistent with an absorption peak for the weak foreground chemical. It is important to note that this shift from exploration to exploitation is not programmed but rather an emerging behavior of the algorithm, driven by the lower SNR from the weak foreground contributing to most of the uncertainty, which the algorithm seeks to minimize.

[12] This distribution was obtained by applying a Gaussian kernel with 0.1 μm standard deviation to smooth the discrete distribution of measurements.

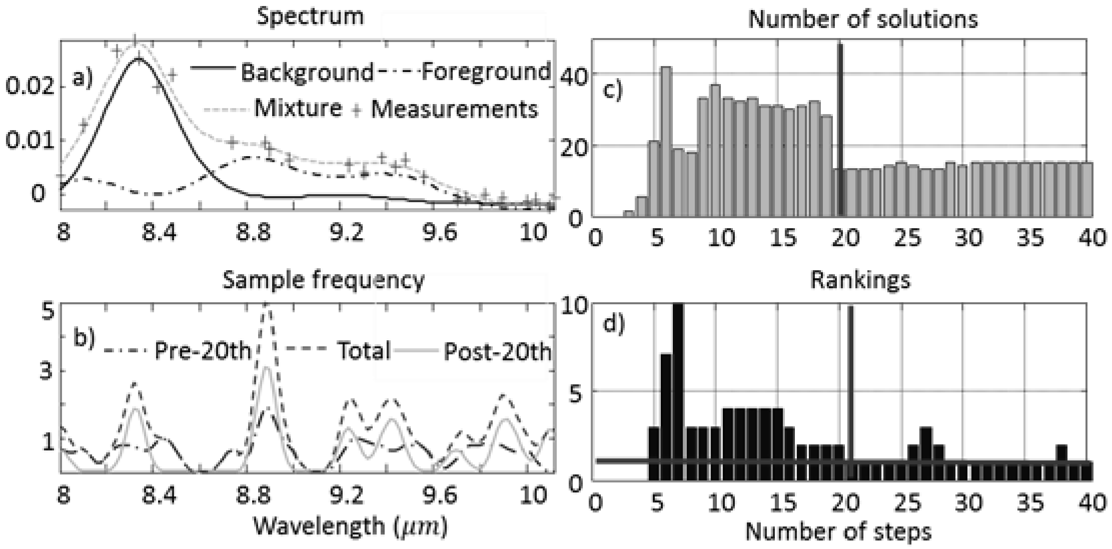


Figure 33: A test case with acetone as background and isopropyl alcohol as foreground. (a) Background, foreground and the mixture; (b) Sampling frequency distribution before and after confirming the ground truth (20-th iteration); (c) Total number of solutions generated, 20-th iteration (vertical line); (d) Ranking of the correct solution, ranking #1 (horizontal line).

5.2.3 Experiment 2: active vs. passive

In a second experiment, we compared the active wavelength selection algorithm against a "passive" baseline algorithm based on sequential forward selection [106]. The passive algorithm selects a fixed subset of wavelengths that best represents the average absorption spectrum across all chemicals in the library: $\bar{\boldsymbol{b}} = \frac{\mathbf{1}}{\boldsymbol{C}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{C}} \boldsymbol{A}(:,\boldsymbol{i})$**.** The passive algorithm works as follows:

- The first wavelength λ_1 is selected (deterministically) as the one with the highest variance in absorption across all chemicals in the library: $\lambda_1 = max_{\lambda_{j \mid \forall j \in [1\ N]}} \{\sigma^2\{\boldsymbol{A}(:,\boldsymbol{i})|\forall i\}\}$.
- To select the second wavelength λ_2, the passive algorithm estimates the

concentration of each individual spectrum in the library to fit that first measurement $\overline{\boldsymbol{b}}(\lambda_1)$; doing so exposes variance at other wavelengths but the first one –which can be fitted with zero error. The second wavelength is then selected as the one with the highest weighted variance across all newly fitted one-chemical spectra.

- To select the third wavelength λ_3, the passive algorithm randomly generates 10,000 two-chemical mixtures to fit the two measurements $\overline{\boldsymbol{b}}([\lambda_1\ \lambda_2])^{\boldsymbol{T}}$, and selects the wavelength with the highest weighted variance across the 10,000 fitted mixture spectra.
- The process is repeated until the desired number of wavelengths has been selected: to select the (n+1)-th wavelength, the passive algorithm randomly generates 10,000 n-chemical mixtures to form a full-rank linear system to fit all previous measurements. This ensures that neighboring wavelengths, which are correlated to those already selected, are not selected before the whole range of the spectrum has been sampled at least once. This idea of decorrelating observations is common in passive wavelength selection methods such as successive orthogonal projection [5].

In contrast with the passive algorithm outlined above, our active algorithm requires no training. To ensure a fair comparison, both methods used the same evaluation function and solver and were stopped after 20 sensing steps. To measure performance, we computed the rank of the correct solution among those returned by the MM-NNLS solver, averaged over all tests cases. The lower the average rank, the better the method, i.e., a rank of one indicates that the solver has placed the correct solution as the first one in its pool. As a supplementary measure, we also used the classification rate, measured as the percentage of trials where the correct solution was ranked as #1.

To perform the comparison, we randomly selected five chemicals as foregrounds and diluted them at multiple levels while keeping the background acetone at 100% concentration. Each sample was tested five times at dilutions ratios of 1/50, 1/33, 1/20, 1/10 and 1/5, for a total of 125 tests samples (5 chemicals × 5 dilutions × 5 replicates). Figure 34 shows the average classification rate (1 if the correct solution is ranked as #1; 0 otherwise) and the average ranking of the correct solution achieved by both methods. There is no significant difference at dilution ratios above 1/20; at such concentrations the problem becomes trivial, and both approaches can find the correct solution with only 4-5 measures. At low concentrations, however, active sensing outperforms its passive counterpart regarding classification rate, and more significantly when considering the average ranking of the correct solution. At the lowest dilution ratio (1/50), active sensing ranks the correct solution as #3 on average, whereas the passive algorithm ranks it at #16. This is largely because active sensing samples the most informative wavelengths repeatedly, avoiding the introduction of new irrelevant wavelengths. This results in a much more compact feature set and, as a consequence, fewer distortions due to noise are introduced to the solver.

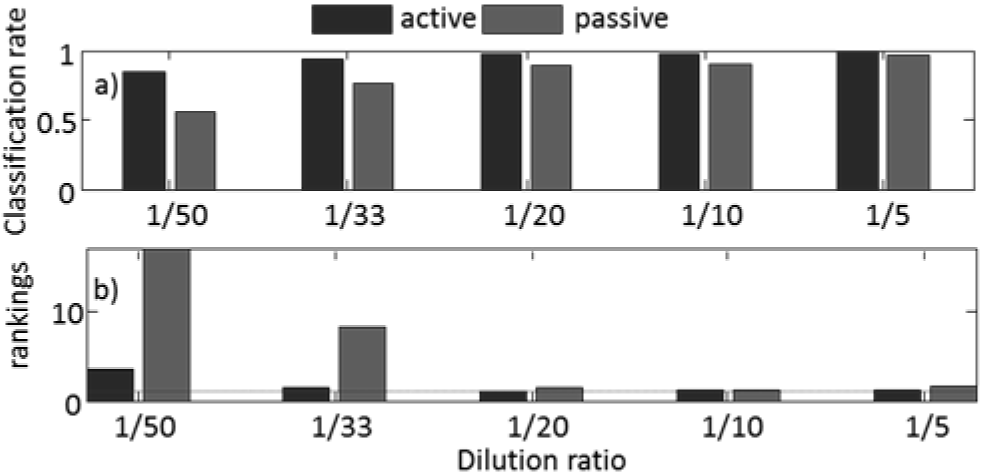


Figure 34 (a) Average classification rate for the active and passive wavelength selection algorithms as a function of the foreground dilution ratio. (b) The average ranking of the correct solution as a function of the foreground dilution ratio; the dashed line represents a ranking of one, indicating that the correct solution was found.

5.2.4 Experiment 3: analyzing the exploration-exploitation tradeoff

In a third experiment, we evaluated the effect of the entropy setting for the AIC weightings described in section 5.1.3.2. For this purpose, we randomly picked one chemical five times as the foreground (out of seven chemicals) while keeping the background fixed (acetone). To make the problem more challenging, the foreground/background ratio was set to 1/20 (background twenty times stronger than foreground). For each of the five foreground cases, we ran experiments with five different AIC weighting entropies of 0.1, 0.3, 0.5, 0.7 and 0.9; settings of 0 (converge to the first solution found) and 1 (never converge) were not considered since they lead to trivial strategies. Each setting was tested ten times for each foreground, for a total of 250 experiments (5 chemicals × 5 dilutions × 10 replicates), or 50 experiments for each entropy setting. From these experiments, we then counted the number of tests for which the correct solution appeared in the candidate pool (N_{FIND}) and the number of tests for which the correct solution ranked as #1 in the pool(N_{CONF}). Denoting the total number of tests by N_{TOT}, we then calculated three measures:

(1) The discovery rate, measured as the proportion of times that the correct solution is included in the pool (N_{FIND}/N_{TOT}),

(2) The resolution rate, measured as the number of times the correct solution is confirmed given that it was included in the pool (N_{CONF}/N_{FIND}), and

(3) The confirmation rate, measured as the proportion of times the correct solution is selected (N_{CONF}/N_{TOT})

Results are shown in Figure 35. When the algorithm uses a higher explorative setting, the discovery rate in Figure 35 (a) increases, but at the cost of reducing the confirmation rate in Figure 35 (b). The final classification rate shows an asymmetric inverted U curve, suggesting that a tradeoff between exploitation and exploration may be found at an entropy setting around 0.5. Interestingly, as shown in Figure 35 (c), too much exploitation appears to be more dangerous than too much exploration. In our case, extreme exploitation leads the algorithm to stop gathering information prematurely, eliminating any chance of discovering the correct solution; in contrast, extreme exploration will tend to evaluate the whole spectrum, with repeated sampling to compensate for noise, allowing the algorithm to converge slowly to the correct solution.

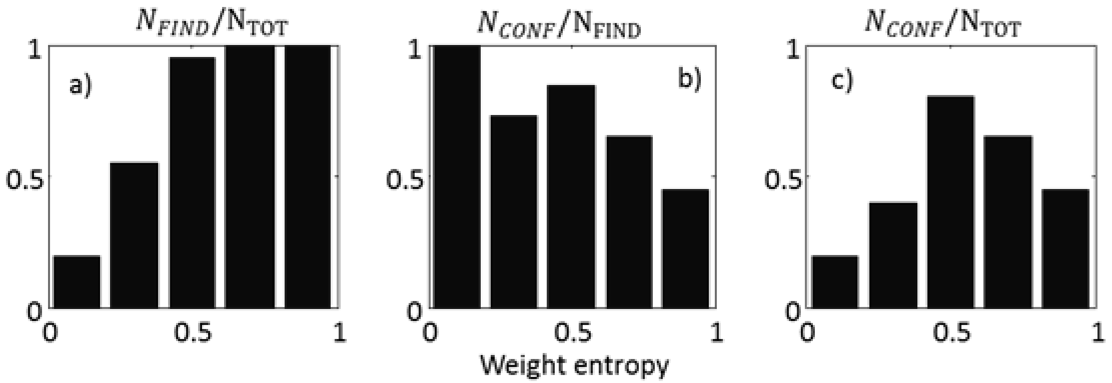


Figure 35: (a) Discovery rate, (b) resolution rate and (c) confirmation rate. The entropy controls the balance between exploitation (entropy being zero) and exploration (entropy being one).

5.3 Validation on synthetic data

To provide a more thorough evaluation than what can be afforded experimentally, we also analyzed the active wavelength selection algorithm on a large dataset of synthetic IR spectra. The dataset consisted of FTIR spectra (660 spectral lines) from 500 chemicals in the NIST Webbook infrared absorption spectrum database [107]. To simulate the spectral resolution of FPIs, we convolved the FTIR spectra with a Gaussian filter of $0.1\mu m$ spread, and added white noise (details included in section 5.3.1) to each individual wavelength. Each spectrum was normalized to sum up to one. For the subsequent experiments, we compared the proposed active wavelength selection algorithm against the passive algorithm described in section 5.2.3. In all cases, we allowed the algorithms to sample each wavelength multiple times.

5.3.1 Binary mixtures

In a first experiment, we tested the algorithm on a similar two-chemical mixture problem as in section 0. However, instead of using a fixed background, both foreground and background were randomly chosen from the library. We then evaluated the algorithm at increasing levels of difficulty by adding Gaussian noise with standard deviation from 0% to 35% of the median value of the complete absorption spectrum library. We also evaluated the algorithm as a function of the degree of collinearity between the foreground and background analytes, measured as the condition number of the column matrix containing the spectra of the two chemical components in the target mixture; the higher the condition number, the more collinear the two chemicals are. Figure 36 (e-h) illustrates pairs of spectra at different condition numbers: for lowest condition number[13] the two spectra are nearly orthogonal, whereas for the highest condition number all major peaks from both chemicals overlap.

To measure performance, we considered the number of iterations (wavelength measurements) required for the algorithm to converge to the correct solution, with convergence strictly defined as the correct solution being ranked as #1 among all solutions <u>and</u> being ten times more likely than the second most likely solution. Results are shown in Figure 36 (a-d). At low noise levels, there are no significant differences between both algorithms. As noise levels increase, performance degrades for both

[13] The lowest condition number for any two pairs of chemicals in our library is 1.2.

algorithms. The effects of noise are considerably amplified by collinearity: when the two spectra are very dissimilar (condition number close to 1) noise has minimal impact, whereas, for similar spectra (condition number of 10), the number of required steps increases significantly with noise. The active algorithm consistently outperforms its passive counterpart in all cases.

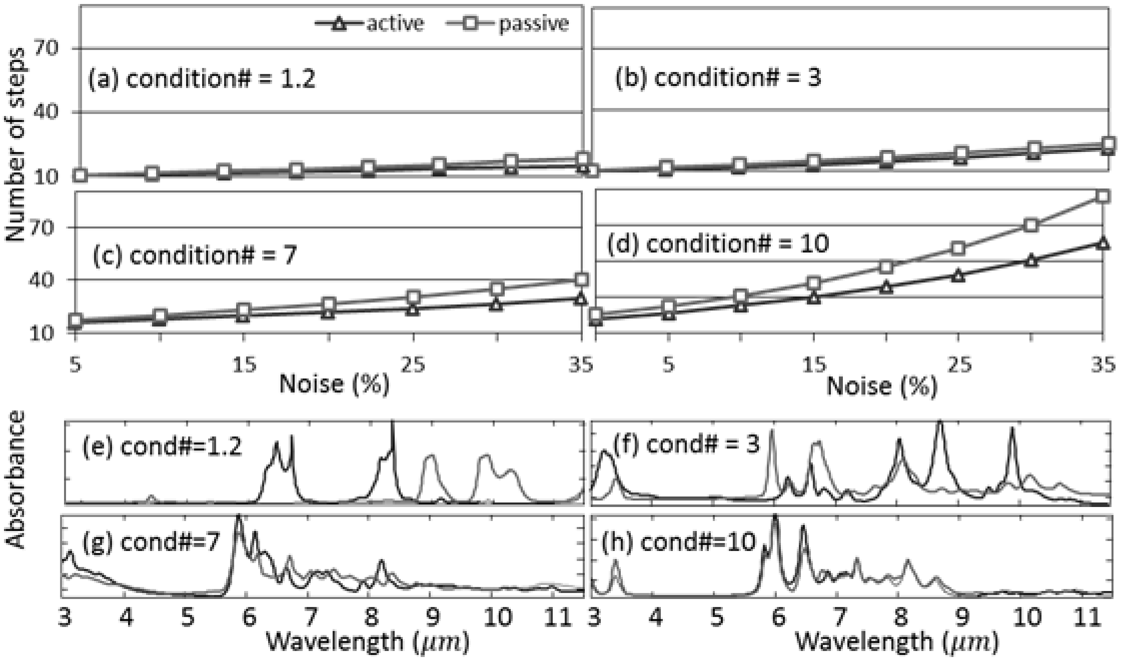


Figure 36: (a-d) Number of steps needed to converge to the correct solution. (e-h) The corresponding foreground and background for each condition number; spectra were normalized to sum up to one.

5.3.2 Higher-order mixtures

In a second and final experiment, we tested the algorithms on higher-order mixture problems containing up to 15 chemical components. In this case, the noise level was fixed at 1% of the median value of all absorption spectra in the library. As the difficulty of a mixture problem can vary dramatically (a badly conditioned two chemical mixture can become unsolvable), we designed a mixture construction policy so that the chosen problems would not be arbitrarily easy or hard. For this purpose, instead of condition numbers we used a random wavelength-selection algorithm to rate the difficulty of one hundred randomly-selected 15-chemical mixtures, and selected five mixtures that could be correctly classified $1\% - 10\%$ of the times using a maximum of 200 randomly-selected measurements; this ensured that the highest-order mixture problems were solvable but non-trivial. For each of these five 15-chemical mixtures, we sequentially removed one component at a time to form chemical mixtures of a lower order; this process ensured a graded transition in problem complexity from hard to easy. For each of the resulting 45 mixtures (15 × 5), we evaluated the active and passive algorithms 40 times, each time with different added noise, for 3,000 cases.

Results are shown in Figure 37 regarding the number of measurements needed for the correct solution to be ranked as #1, up to a maximum of 200 measurements. Since the noise level is low (1%), there is no significant difference between active and passive algorithm for problems with up to four chemicals. With five or more chemicals, the active algorithm gradually outperforms the passive algorithm. As expected, the number of measurements needed grows exponentially for both algorithms with the number of chemicals, but the active algorithm can solve a significantly more complex problem than its passive counterpart for a fixed sensing budget can. As an example, given 100 measurements the active algorithm can solve an 11-chemical mixture problem whereas the passive algorithm can only solve an 8-chemical problem at best. Likewise, to solve a

9-chemical problem the active algorithm requires 70 measurements on average, whereas the passive algorithm requires 130 measurements.

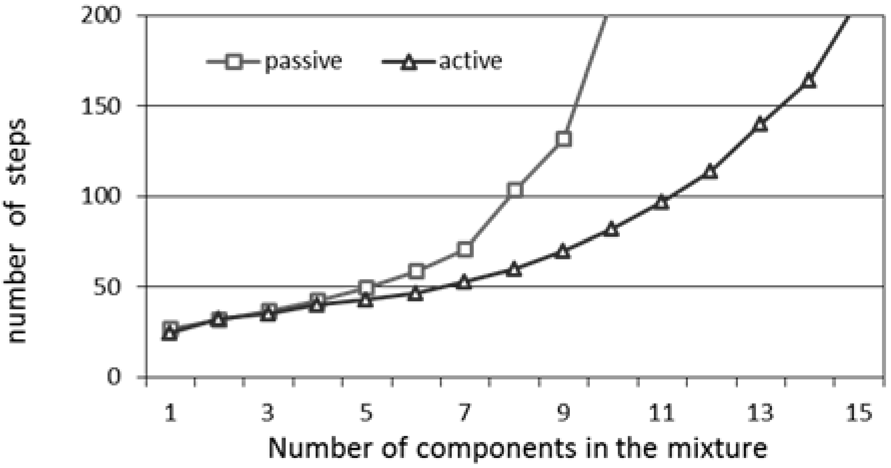


Figure 37: The number of steps used to converge to the correct solution with mixture problems up to 15 chemicals.

5.4 Conclusions and discussion

We have proposed an active wavelength selection algorithm for mixture analysis with tunable chemical sensors. The algorithm uses a multi-modal solver to maintain a pool of likely candidate solutions based on previous measurements, then selects its next wavelength as the one which maximizes discrimination among all the candidates in the pool. To address the ill-conditioned nature of the problem, the algorithm promotes sparse solutions with two complementary strategies. First, the algorithm adds mixture components to the candidate solutions in an incremental fashion, from single analytes, to binary mixtures, to ternary mixtures, and so on. Second, the algorithm promotes sparse candidates using a weighting function based on the Akaike information criterion. To prevent the search from converging prematurely, the algorithm also uses an entropy-guided normalization method that rebalances the AIC weights such that the strongest candidate solutions do not dominate the wavelength-selection process during the early stages.

The algorithm is first validated experimentally on binary mixture problems with a Fabry-Perot Interferometer. Our results show that active wavelength selection outperforms its passive counterpart, particularly at low concentrations and low foreground-background ratios. We also characterized the algorithm on synthetic data at increasing levels of ill-conditioning and higher-order mixtures and compared it with a passive algorithm. Active wavelength selection provides higher and more stable performance than passive selection, and more importantly, shows higher tolerance to noise and collinearity. Compared against passive wavelength-selection techniques, which require retraining if additional chemicals are added to (or removed from) the library, active wavelength selection can also be trivially adapted to problems of varying library sizes.

Correlation between neighboring wavelengths can make the library matrix $A(F,:)$ close to singular. In practice, however, the system rarely selects neighboring wavelengths before the underlying linear system reaches full rank: once observations have been made at certain wavelengths, the NNLS solver will fit the candidate models at those measurements with zero error because the system is under-determined. As a result, variance at those wavelengths will be minimized, and so will be the variance at neighboring wavelengths, significantly reducing the chances that they will get selected at the next iteration. It is not until the linear system becomes full rank that the algorithm begins to sample neighboring frequencies to average out noise.

6. BIC SHRINKAGE NON-NEGATIVE LEAST SQUARES

In the previous chapter, we described an active wavelength selection framework based on a multi-modal solver that generates multiple solutions. We leveraged such multi-modality to calculate the uncertainty of the sampling space (wavelengths), and then to select those wavelengths with the highest uncertainty.

However, the multi-modal solver is computationally costly, especially when the chemical library grows large or the mixture is complex. When the complexity grows, the computational cost soon becomes prohibitive. In addition to its lack of computational efficiency, the solver is also incapable of adapting for nonlinearity and emitter drift. Such nonlinearity and drift introduce structural errors that break the assumed underlying linear model.

To address this issue, we present a single-modal solver that also accommodates for nonlinearity and emitter drift. Note that this solver is also built for the faster wavelength selection based on GPR, which is described in the next chapter (Chapter 7). It consists of a sparse linear solver, a search algorithm to accommodate for nonlinearity, and a first order Taylor approximation to compensate for emitter drift. We refer to the first component, the linear solver, as “BIC shrinkage batch NNLS” (BICS-bNNLS, see Section 6.1) where BIC stands for Bayesian information criterion, and bNNLS stands for batch non-negative least squares, a modification of the classical NNLS algorithm developed by Lawson [18]. The second component, on top of BICS-bNNLS, is an iterative procedure to search a spectral library from a data cube that captures nonlinear distortions across all concentrations (see Section 6.2). Lastly, we compensated for emitter drift using a non-uniform offset vector in the spectral library (see Section 6.3).

We conducted experiments to validate the nonlinear BICS-bNNLS. First, we compared the batch NNLS for its computational efficiency against other solvers in Section 6.4.2. Second, we tested its effectiveness in searching sparse solutions in Section 6.4.3. Last, in Section 6.4.4, we validated the effectiveness to accommodate nonlinearity and emitter drift using experimental spectral data.

6.1 BICS-bNNLS

BICS-bNNLS uses a forward-backward variable (constituent) selection strategy: batch NNLS, and then BIC shrinkage. Namely, batch NNLS first adds constituents in batch to fit the observations, and then BIC shrinkage eliminates any insignificant constituents guided by Bayesian information criterion. This forward-backward variable selection strategy avoids the common pitfalls of convergence to a local optimum. Figure 38 illustrates an example of this process. The algorithm starts with empty solution vector $\boldsymbol{x}_0$ and forward-selects constituents until the error is zero (with corresponding solution x_{bNNLS}). Then, the shrinkage process begins to eliminate the insignificant constituents guided by BIC. This process continues until the BIC score stops improving.

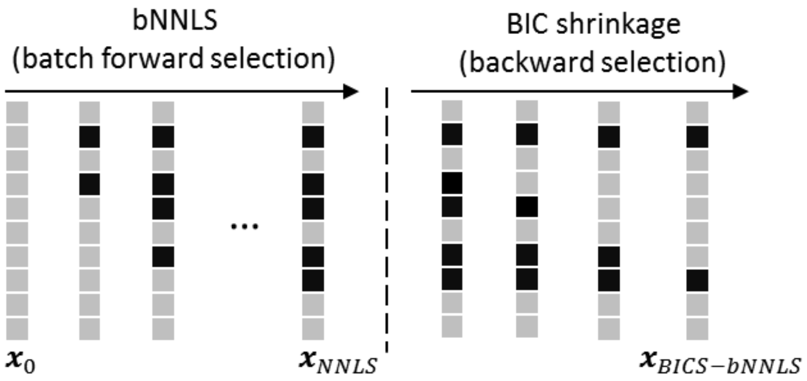


Figure 38: The variable (constituent) forward-backward selection process.

6.1.1 Batch NNLS

Batch NNLS is a modification of the classical NNLS algorithm written by Lawson. For a detailed description of the original NNLS algorithm, please refer to [18]. Lawson's algorithm is a variable forward selection algorithm that adds one variable at a time. Every time a variable is added, the algorithm checks feasibility of the solution and adjusts the solution to maintain feasibility. bNNLS uses the same variable selection strategy, but in batch. Table 6 presents the pseudo-code for bNNLS. The bNNLS algorithm consists of an outer-loop and an inner-loop: the outer-loop selects and adds variables to the solution, and the inner-loop calculates a feasible solution given the selected variables.

Table 6: The pseudo-code of batch NNLS

```
Input: spectral library $\boldsymbol{A}$, observation $\boldsymbol{b}$, maximum #variables to update each time $\boldsymbol{n_{max}}$
Output: solution $\boldsymbol{x}$
procedure bNNLS($\boldsymbol{x}$, $\boldsymbol{b}$, $\boldsymbol{n_{max}}$)
    $\boldsymbol{N} \leftarrow \#\, \boldsymbol{columns\ of\ A}$ //#constituents
    $\wp \leftarrow \emptyset$, $\mathbb{Z} \leftarrow \{1,2,\dots,N\}$
    $\boldsymbol{x} \leftarrow \boldsymbol{0}$
    while True
      $\boldsymbol{w} \leftarrow \boldsymbol{A^T}(\boldsymbol{b} - \boldsymbol{Ax})$ //calculate gradient
      if $\mathbb{Z} \neq \emptyset$ and $\boldsymbol{w_j} > \boldsymbol{0}$, $\exists \boldsymbol{j} \in \mathbb{Z}$ //gradient signals improvements
        return $\boldsymbol{x}$ //return if no more improvement
      $\boldsymbol{n} \leftarrow \begin{cases} \frac{\boldsymbol{n_b}}{\boldsymbol{2}}\ \boldsymbol{if}\ \wp' = \wp \\ \boldsymbol{n_{max}}\ \boldsymbol{otherwise} \end{cases}$ //reduce step-size if infeasible
      $\boldsymbol{J} \leftarrow \{\boldsymbol{the\ n\ largest\ w_j, j} \in \mathbb{Z}\}$ //tentative variables
      $\wp_{\boldsymbol{lst}} \leftarrow \wp$ //record the non-zeros for later comparison
      $\boldsymbol{w_t} \leftarrow \boldsymbol{w_J}$
      $\mathbb{Z} = \mathbb{Z} - \boldsymbol{J}$, $\wp = \wp + \boldsymbol{J}$ //move indices from Z to P
      while iter < $\boldsymbol{MAX\_ITER}$ //find a local feasible solution
        Let $\boldsymbol{A}_\wp$ defined by:
          column j of $\boldsymbol{A}_\wp \leftarrow \begin{cases} \boldsymbol{column\ j\ of\ A\ if j} \in \wp \\ \boldsymbol{0\ if\ j} \in \wp \end{cases}$
        $\boldsymbol{z} \leftarrow \left(\boldsymbol{A}_\wp^{\boldsymbol{T}} \boldsymbol{A}_\wp\right)^{-\boldsymbol{1}} \boldsymbol{A^T b}$ //pseudo-inverse
        if $\boldsymbol{z_j} > \boldsymbol{0}\ \forall \boldsymbol{j} \in \wp$ //check feasibility
          $\boldsymbol{x} \leftarrow \boldsymbol{z}$
        else //set infeasible variables to zero
          $\boldsymbol{J} \leftarrow \{\boldsymbol{j} \in \wp, \boldsymbol{z_j} \leq \boldsymbol{0}\}$
          $\boldsymbol{x_J} \leftarrow \boldsymbol{0}$
          $\wp \leftarrow \wp - \boldsymbol{J}$, $\mathbb{Z} \leftarrow \mathbb{Z} + \boldsymbol{J}$
        endif
      endwhile
    endwhile
    return $\boldsymbol{x}$
end
```

To improve its computational efficiency, we modified both the outer-loop (variable selection) and the inner-loop (feasibility). In the outer-loop, bNNLS adds multiple variables at each step rather than just one variable. The number of variables n added at each iteration is initialized to be n_{max} [14], but can be adjusted when no feasible solution exists (see Line 9). In the inner-loop, bNNLS removes multiple variables with infeasible values altogether (see lines from 21 to 23), compared to just one variable in the original NNLS.

6.1.2 BIC shrinkage

bNNLS generates a feasible solution that normally fits the observations within machine epsilon[15]. However, for mixture identification with noisy observations, overfitting often leads to false-positives in the solutions. We address this issue by sparsifying the solutions. The following describes the sparsifying process.

Two common model selection methods to measure overfitting are the Akaike information criterion (AIC) [108] and the Bayesian information criterion (BIC) [109]. Both criteria encourage parsimony by penalizing model complexity, with the penalty of BIC growing stronger as the number of measurements increases. BIC was developed assuming that only one true model exists[16]. Considering our goal in this work is to

[14] By default, $n_{max} = \min\{N, M\}$, where N is the number of constituents, and M is the number of wavelengths.

[15] Machine epsilon is the upper bound of the relative error due to rounding in floating point arithmetic.

[16] Asymptotically, the BIC score reaches the lowest point when a true model is found in [110].

recover the mixture constituents (thus a true model must exist), we chose BIC for the shrinkage criterion. Table 7 shows the pseudo-code of the BIC guided shrinkage method. Once the NNLS algorithm generates a solution, the shrinkage algorithm greedily tests and eliminates the least significant component (the one with the lowest concentration) until the BIC score stops improving.

Table 7: Pseudo-code for the BIC guided shrinkage procedure

```
Input: solution x, observation b
Output: new sparsified solution x′

procedure BIC_shrinkage(x, b)
   BIC ← getBIC(x,b)
   do
      x′ ← eliminate(x)
      BIC′← BIC
      BIC ← getBIC(x′,b) // equation(35)(36)
   while  BIC ≤ BIC'
   return x
end

/** Find the minimal non-zero element and set it to zero **/
sub-procedure eliminate(x)
   minX ← ∞
   minI ← 0
   for i←0 to length(x)-1
      if minX > x(i) and x(i)>0
         minxX ← x(i)
         minI ← i
      endif
   endfor
   x′ ← x
   x′(minI) ← 0
   return x′
end
```

The BIC score can be calculated as:

$$BIC = -2\log(\mathcal{L}) + nlog(m) \tag{35}$$

where n, as a measure of model complexity, is the number of non-zero components in the solution $\boldsymbol{x}$; m is the number of measurements; and $\mathcal{L}$ is the likelihood of the model, which can be calculated as:

$$\mathcal{L} = (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{b}-\boldsymbol{Ax})^T(\boldsymbol{b}-\boldsymbol{Ax})\right\}. \tag{36}$$

where σ is the spread of the Gaussian noise, and $(\boldsymbol{b}-\boldsymbol{Ax})^{\boldsymbol{T}}(\boldsymbol{b}-\boldsymbol{Ax})$ is the sum squared error.

6.2 Nonlinear BICS-bNNLS

The imperfection of optical filter can cause nonlinear deviations from Beer's law (see Section 2.3.2.2). As a result, the absorbances at different wavelengths scale differently as the concentrations changes. Such nonlinearity can be compensated by building a spectral library that captures the nonlinear distortions at the corresponding concentration. However, the concentration itself is unknown beforehand. We developed an iterative process where the spectral library construction and concentration estimation run alternatively to search both variables to improve the regression. The end result of this algorithm is a concentration vector $\boldsymbol{x}$ and a nonlinearly distorted spectral library $\boldsymbol{A}$.

6.2.1 Spectral library

The first step to building a spectral library that captures the nonlinear distortions is to acquire such a library throughout the range of concentrations for each constituent.

However, acquiring clean spectra is especially challenging at low concentrations, at which absorbance is dominated by sensor noise. Hence, to build a usable spectra library at lower concentrations, we smoothed the sampled spectra between different concentrations using a two-dimensional Gaussian process. Please refer to [110] for the computation of Gaussian process regression. Here, we briefly describe this procedure. Gaussian process regression is an interpolation technique that exploits smoothness in the data. Such smoothness assumption is met in absorption spectra since we can safely assume that the spectra at neighboring concentrations are close to each other. The smoothness constraints can also be added to a second dimension (wavelengths) because absorbances at adjacent wavelengths are also close. Thus, given a sparse set of noisy samplings of spectra at different concentrations, Gaussian process regression calculates smooth spectra at any concentration.

However, because of the interpolation nature of Gaussian process regression, the lowest concentration at which the spectrum can be calculated is limited by the acquired spectral data. Acquiring spectra at extremely low concentrations is challenging because of the overwhelmingly low signal to noise ratio. Fortunately, the spectra at zero concentrations are known to be zeros unless the emitter drifts (which will be discussed in the next section 6.3). The extrapolation problem becomes an interpolation once the spectrum at the lowest concentration, zero percent, is known. As a result, using Gaussian process regression, we can acquire a clean spectral library across the whole range of concentrations for each constituent. Such procedure is repeated for each constituent, and the final spectral library is a data cube with three dimensions of constituents, wavelengths, and concentrations.

Figure 39 shows an example of Tert-Butyl Alcohol samples interpolated by Gaussian process regression. We collected sample spectra at concentrations of 10%, 20%, 50%, and 100%.

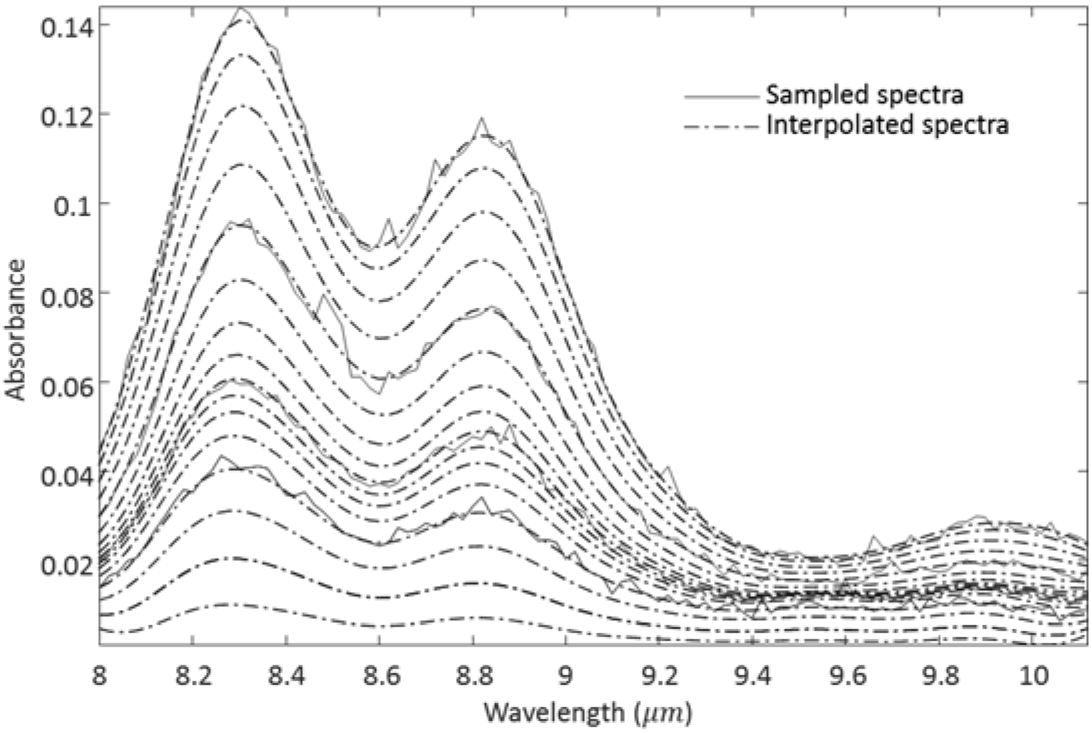


Figure 39: Two-dimensional Gaussian process regression reconstructs clean spectra at different concentrations for lacquer thinner.

6.2.2 Nonlinear solver

Once we have the data cube, a search algorithm constructs an ad-hoc two-dimensional spectral library from the data cube based on the intermediate solutions generated from BICS-bNNLS. BICS-bNNLS then recalculate the estimation using the new library. The algorithm repeats this process until convergence, i.e., when the reconstructed library stops changing. Figure 40 illustrates the process.

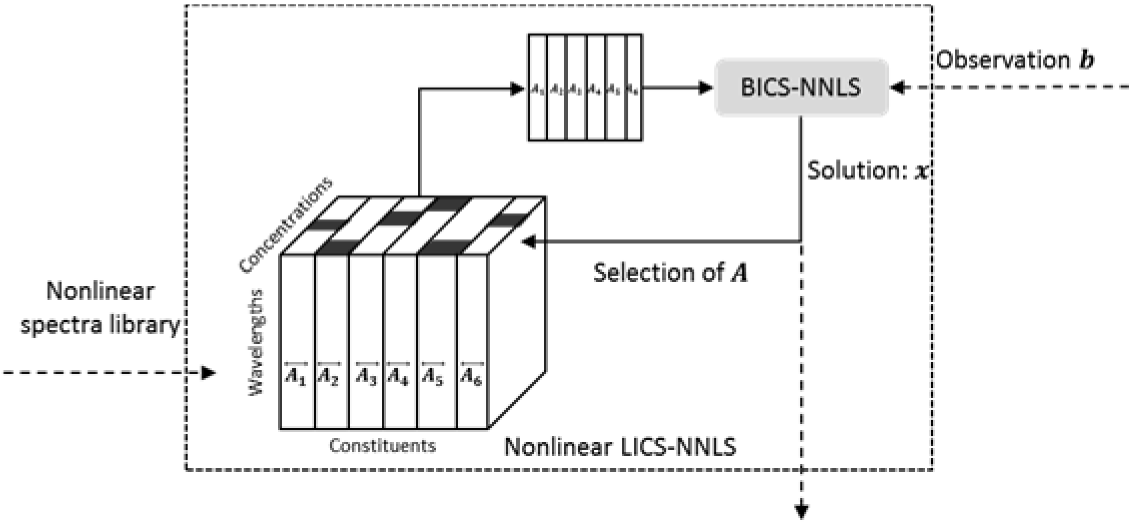

A_1 A_2 A_3 A_4 A_5 A_6
BICS-NNLS
Observation $\boldsymbol{b}$
Solution: $\boldsymbol{x}$
Concentrations
Selection of $\boldsymbol{A}$
Nonlinear spectra library
Wavelengths
$\bar{A}_1$ $\bar{A}_2$ $\bar{A}_3$ $\bar{A}_4$ $\bar{A}_5$ $\bar{A}_6$
Constituents
Nonlinear LICS-NNLS

Figure 40: Diagram of the nonlinear BICS-bNNLS.

We illustrate the pseudo-code for this nonlinear algorithm in Table 8. The algorithm begins with 100% concentration for all constituents in the library (Line 2: $\boldsymbol{x} \leftarrow \mathbf{1}$). Currently, the spectral library $\boldsymbol{A}$ consists of the spectra at the highest concentrations in the data cube (Line 4). Given $\boldsymbol{A}$, BICS-bNNLS then calculates a solution $\boldsymbol{x}'$ (Line 5). Since $\boldsymbol{x}'$ is calculated using the spectral library $\boldsymbol{A}$ extracted at the latest estimation $\boldsymbol{x}$, $\boldsymbol{x}'$ is a relative concentration. The absolute concentration is updated correspondingly $\boldsymbol{x} \leftarrow \boldsymbol{x}\boldsymbol{x}'$ (Line 8). A new spectral library $\boldsymbol{A}$ is then extracted from the data cube at this updated concentration, and the iteration continues until convergence, which is when the relative concentration is close to one with an user-defined margin ϵ_x (Line 6) so that there is no more need to reconstruct a new spectral library.

Table 8: Pseudo-code for nonlinear spectral library search algorithm

```
Input: data cube $\{\overleftrightarrow{\mathbf{A}}_{\mathbf{1}},\dots,\overleftrightarrow{\mathbf{A}}_{N}\}$, observation $\boldsymbol{b}$, solver BICS-bNNLS
Output: solution $\boldsymbol{x}$

procedure nonlinear_NNLS ($\{\overleftrightarrow{\mathbf{A}}_{\mathbf{1}},\dots,\overleftrightarrow{\mathbf{A}}_{N}\}$, $\boldsymbol{b}$, BICS-bNNLS)
   $\boldsymbol{x} \leftarrow \mathbf{1}$
   do
     $\boldsymbol{A} \leftarrow$ extract($\{\overleftrightarrow{\mathbf{A}}_{\mathbf{1}},\dots,\overleftrightarrow{\mathbf{A}}_{N}\}$, $\boldsymbol{x}$)
     $\mathbf{x}' \leftarrow$ BICS-bNNLS($\boldsymbol{A}$, $\boldsymbol{b}$)
     if abs($\boldsymbol{x}' - \mathbf{1}$) < $\epsilon_{\boldsymbol{x}}$
        return $\boldsymbol{x}\boldsymbol{x}'$
     $\boldsymbol{x} \leftarrow \boldsymbol{x}\boldsymbol{x}'$
   while True
end

// Extract the spectra at the concentration $\boldsymbol{x}$
sub-procedure extract($\{\overleftrightarrow{\mathbf{A}}_{\mathbf{1}},\dots,\overleftrightarrow{\mathbf{A}}_{N}\}$, $\boldsymbol{x}$)
   for We = 0 to $\boldsymbol{N}$
     $\boldsymbol{A_i} \leftarrow \overleftrightarrow{\mathbf{A}}_{\boldsymbol{i}}(\lfloor \boldsymbol{x} \rceil)$
   return $\{\boldsymbol{A}_{\mathbf{1}},\dots,\boldsymbol{A}_{\boldsymbol{N}}\}$
end
```

6.3 Drift compensation

Another problem we were facing during experimentations is emitter drift. As shown in equation (1), the absorption value of a chemical is computed using both energy readings $\boldsymbol{I}$ and $\boldsymbol{I}_0$. During experimentation, the power of the emitter can drift slightly by $\Delta\boldsymbol{I}$, which is caused by changes in the emitter surface temperature. Ideally, if $\Delta\boldsymbol{I}$ is known, the absorption should be corrected as $-log(\frac{\boldsymbol{I}+\Delta\boldsymbol{I}}{\boldsymbol{I}_0+\Delta\boldsymbol{I}})$. However, the sensor can only measure $\boldsymbol{I}+\Delta\boldsymbol{I}$ as whole. Without knowing $\Delta\boldsymbol{I}$, the absorption value is incorrectly calculated as $-\log\left(\frac{\boldsymbol{I}+\Delta\boldsymbol{I}}{\boldsymbol{I}_0}\right)$. Although the drift ΔI is relatively small ($\frac{\Delta I}{I_0} < 1\%$), it can still be troublesome for analyte with low concentrations or low sensitivity. In the previous work (Chapter 5), a uniform offset (a vector of ones) was added to the library matrix $\boldsymbol{A}$ to compensate for this drift. This offset was beneficial, especially when the emitter drift shifted some absorbances below zeros. However, since the drift $\Delta\boldsymbol{I}$ is different at various wavelengths, the offset should be non-uniform. In addition, logarithm is a nonlinear operator, so the emitter drift is a non-uniform nonlinear transformation of the original spectrum. We propose to compensate for the transformation using a first order Taylor series approximation by the first derivative of the absorption $\boldsymbol{\Delta} = -\frac{\boldsymbol{d}\log\left(\frac{(I+\Delta I)}{I_0}\right)}{\boldsymbol{d}I_0} \propto \frac{1}{\boldsymbol{I}_0}$. Thus, we add the column vector $\frac{\mathbf{1}}{\mathbf{I_0}}$ to the linear system:

$$\boldsymbol{A} = \left[\boldsymbol{A}_1, \boldsymbol{A}_2, \dots, \boldsymbol{A}_N, \frac{\mathbf{1}}{\mathbf{I_0}}\right]. \tag{37}$$

The amplitude of the drift x_0 is then solved together with concentration $\boldsymbol{x} = [x_1, \dots, x_N, x_0]^T$. The projection $\frac{\mathbf{1}}{\mathbf{I_0}} \cdot x_0$ shifts the original zero absorbances when $x_0 \neq 0$. Note that unlike concentration, this offset coefficient does not conform to the non-negative constraint. Figure 41 illustrates an example of a spectrum with such emitter drift for denatured alcohol with a 2% concentration. Note that the absorption spectra are partially negative because of the drift.

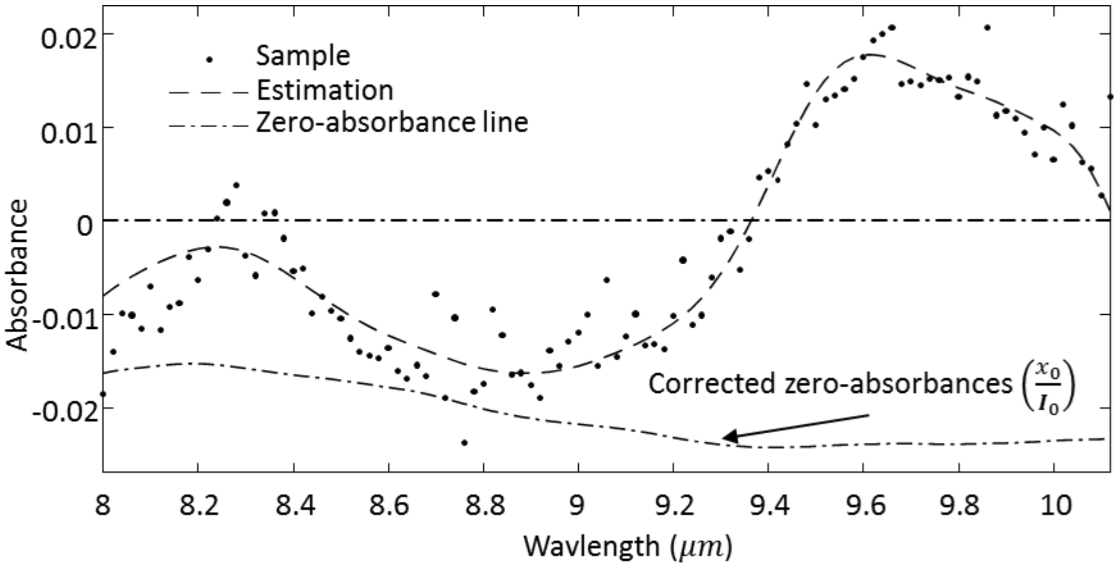


Figure 41: Corrected zero-absorbance line with emitter drift compensation.

6.4 Results

6.4.1 Experimental setup

The calculations were performed on a 2.8GHz i7 860 desktop computer with 32GB of RAM implemented in MATLAB® 8.5a. We used two types of data for experimentation: a randomly generated spectra from a uniform distribution, and instrumental data from the FPI sensor. For detailed experimental apparatus, please refer to Section 5.2.1.

6.4.2 bNNLS speed comparison

To test the computational speed of the algorithm, we compared bNNLS against three alternatives: the fast NNLS implementation of Bro [19], the *lasso* implementation of Kim et. al. [111], and the classical NNLS implementation 600of Lawson [18]. We conducted a simulation for large-scale non-negative least squares with various library sizes from 600 to 1.2M components. With 600 features (wavelengths), the biggest library occupied 5.76 GB ($1.2 \times 600 \times 8\ MB$) of memory. The library consisted of entries drawn from zero to one with a uniform distribution. We synthesized mixture problems with 600 components (full rank) out of randomly picked, and we added a white noise with a spread of 0.01 to the observation. We ran all algorithms with the same stopping criterion: the mean squared error had to be smaller than 10^{-6} or reached the time limit of one hour. Table 9 summarizes the average computation time over 20 runs for each library sizes.

Table 9: Time consumption and relative speed-up of different algorithms with different library sizes averaged over 20 runs.

Library size	bNNLS	NNLS	fNNLS	*Lasso*
$\mathbf{6 \times 10^2}$	$\mathbf{0.8s}$ (1X)	$\mathbf{10.2s}$ (12X)	$\mathbf{1.95s}$ (2X)	$\mathbf{9.98s}$ (12X)
$\mathbf{6 \times 10^3}$	$\mathbf{2.2s}$(1X)	$\mathbf{15.46s}$(7X)	$\mathbf{3932s}$(1787X)	$\mathbf{86.8s}$(39X)
$\mathbf{6 \times 10^4}$	$\mathbf{4.6s}$(1X)	$\mathbf{31.2s}$(7X)	∞	$\mathbf{2113s\,(459X)}$
$\mathbf{6 \times 10^5}$	$\mathbf{29.2s}$(1X)	$\mathbf{232.4s}$(8X)	∞	∞
$\mathbf{1.2 \times 10^6}$	$\mathbf{54.2s}$(1X)	$\mathbf{430.3s}$(8X)	∞	∞

As can be seen, bNNLS outperforms all other algorithms in all cases. Upon closer inspection, for the smallest problem (600 components), fNNLS is the second best performer. However, as the library size increases, fNNLS scales terribly (it ran out of time at a problem with 60,000 components). *Lasso* scales slightly better but is still a lot worse than bNNLS and NNLS. NNLS scales as well as bNNLS, but is about eight times slower than bNNLS.

6.4.3 BICS-bNNLS sparsity comparison

To test the effectiveness of the BIC shrinkage method, we compared it against four alternatives: the pseudo-inverse solution (without sparse regularization, as a control condition), the classical NNLS, the *lasso* (with l_1 norm sparse regularization), and ridge regression (with l_2 norm guided sparse regularization). We generated a library $\boldsymbol{A}$ with each entry drawn from a uniform distribution $U(0,1)$. We then synthesized mixture problems consisting of one to ten components randomly selected from this library. We also added Gaussian noise with a standard deviation of 0.01 to the observations. We solved the mixture problems using these algorithms and compared the sparsity of their solutions. To illustrate what the generated solutions, Figure 42 shows an example of the solution in both linear scale and logarithmic scale.

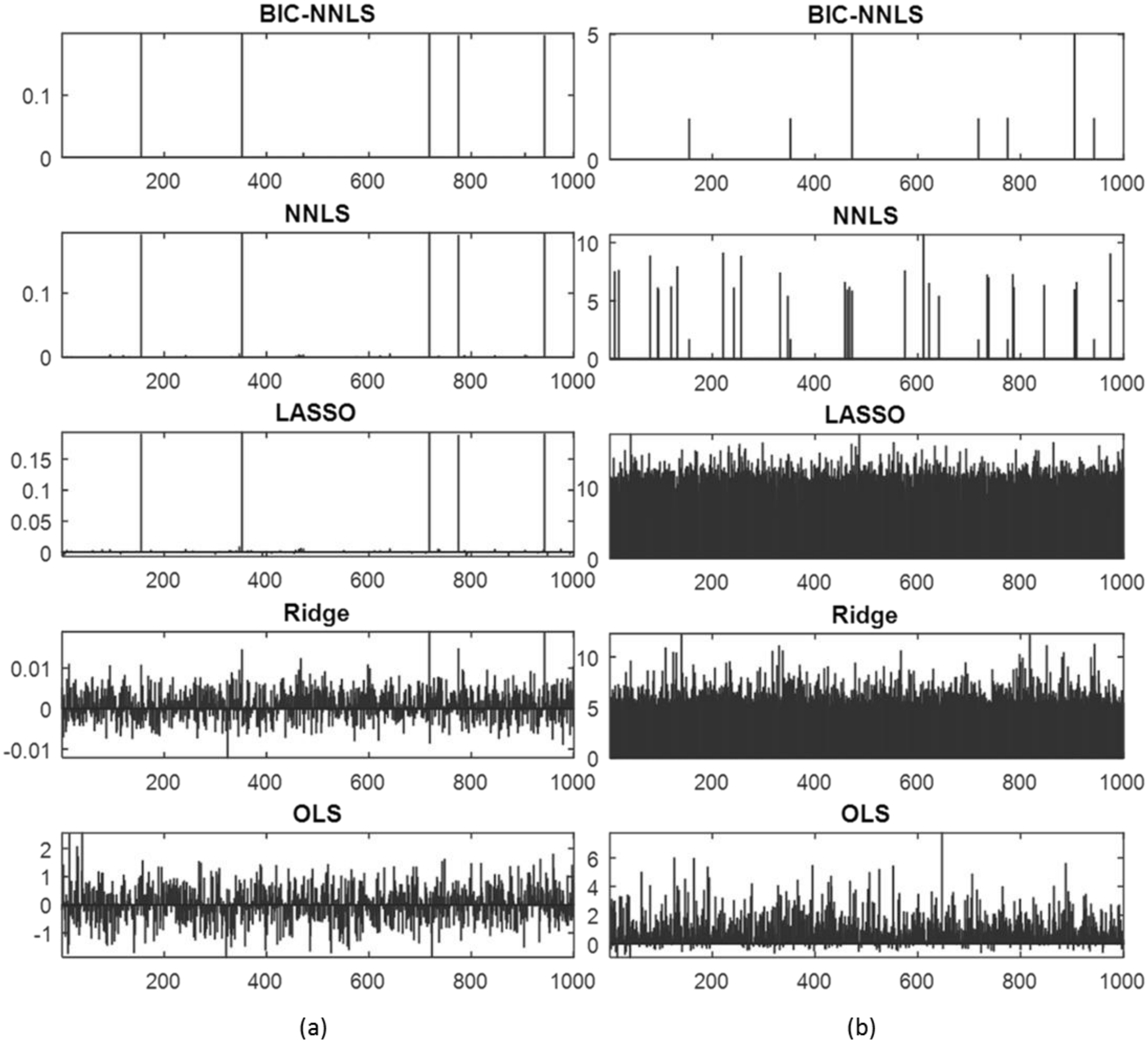


Figure 42: (a) Sample solutions from different solvers; (b) The same solutions in logarithmic scale.

A superficial inspection of results in Figure 42(a) suggests that BIC-NNLS, NNLS, and *lasso* were able to generate sparse solutions. However, as shown in see Figure 42(b), in logarithmic scale the *lasso* solution does not appear sparse while BICS-bNNLS and NNLS maintain their sparsity.

We also repeated this experiments over 100 times, and computed the average relative sparsity – the ratio between the number of non-zero entries and that of the ground truth ($\|\widehat{\boldsymbol{x}}\|_{\mathbf{0}}/\|\boldsymbol{x}\|_{\mathbf{0}}$). The result is consistent with the previous test case. As shown in Figure 43. The *lasso* algorithm generated solutions are just as dense as the solutions generated by ridge regression and OLS. For signal reconstruction purposes, the sparsity of the solution is not critical, but in our work, the true sparsity is critical because it represents the number of constituents being identified as present in the mixture. Excessive non-zero entries are most likely false-positives. Therefore, the l_1 norm guided *lasso* algorithm is insufficient for the purpose of mixture identification. On the other hand, the NNLS solver we used developed by Lawson [18] is a forward variable selection algorithm. Such forward variable selection implicitly implemented the l_0 norm regularization[17]. Finally, it is worthwhile to mention that BICS-bNNLS generated solutions with a sparsity ratio of 1, indicating perfect mixture identification without introducing any false positive.

[17] l_0 norm provides the strongest sparse regularization of all norms (such as l_1 norm and l_2 nom).

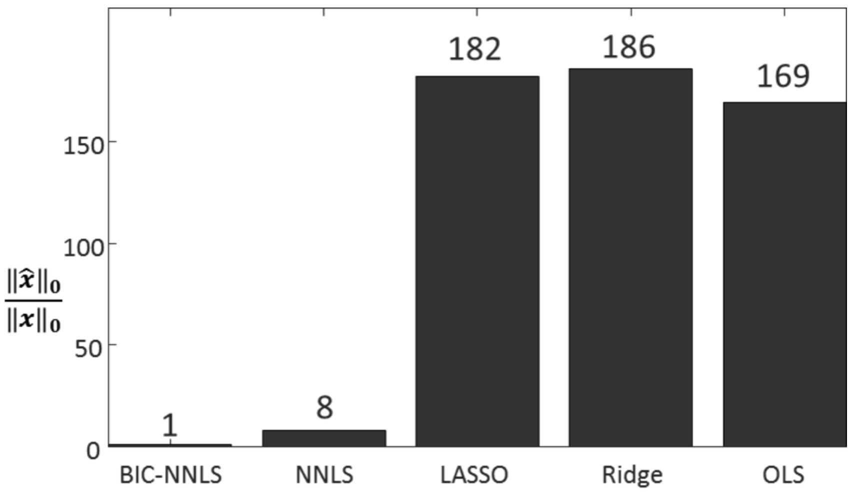


Figure 43: Sparsities of the solutions in l_0 norm.

6.4.4 Nonlinear BICS-bNNLS accuracy comparison

To test the accuracy of nonlinear BICS-bNNLS, we conducted an experiment with acetone at a concentration of 10%. We tested the algorithm with 10-100 randomly sampled wavelengths. We repeated the experiment 20 times and used the average classification rate as the performance metric. We conducted a 2×2 comparison with conditions: linear solver only (L) vs. nonlinear solver (NL) and uniform drift (D) vs. non-uniform drift (ND). Results are shown in Figure 44.

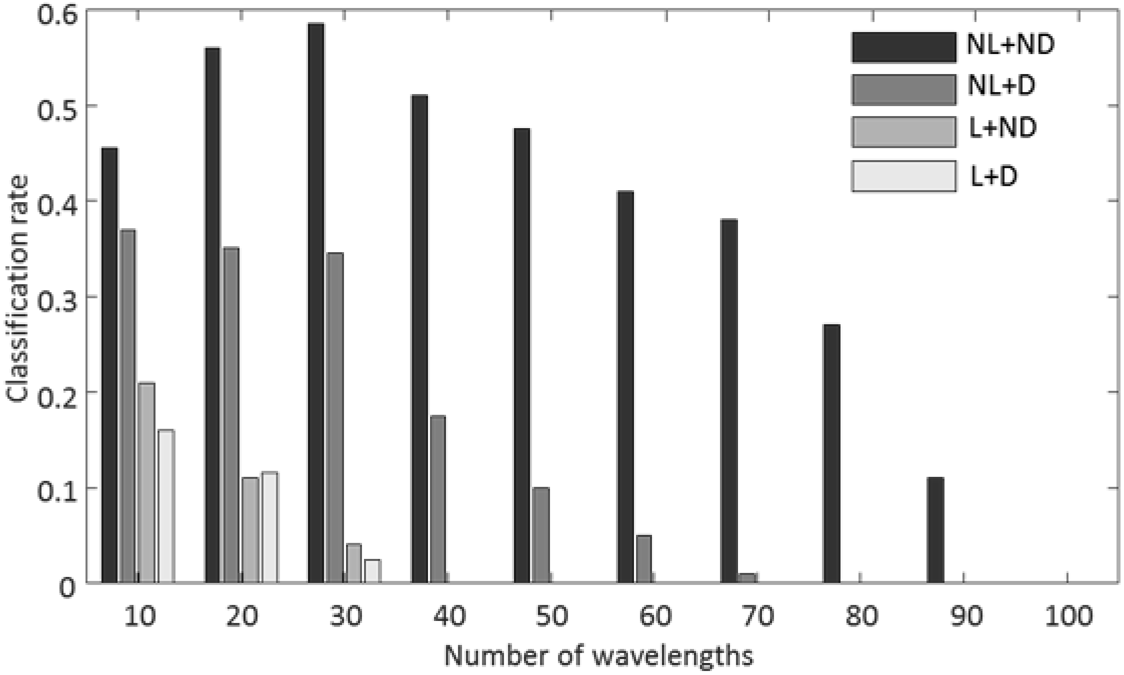


Figure 44: Number of misclassified components with and without the nonlinear search algorithm.

As shown in the result, with the exception of the NL+ND method (nonlinear solver with non-uniform drift), the performances of all methods deteriorated as more wavelengths are introduced. This trend is consistent with the fundamental limitation Beer's law[18] as described in Section 2.3.2.2. As more wavelengths are introduced, the structural error caused by nonlinearity begins to emerge. Importantly, the NL+ND method outperformed all other methods significantly across all wavelengths. Improved performance grants a selection of a larger number of wavelengths, indicating less structural error is

[18] Beer's law is only linear for infinitely narrow wavelengths, but in practice all wavelengths selectors are imperfect; consequently, more wavelengths introduce more nonlinearity.

introduced. In a final analysis, we also examined the effect of drift compensation. While the non-uniform drift compensation benefits both linear and nonlinear solver, it provides a higher performance improvement for nonlinear solver than to the linear solver. This suggests that both dynamic factors (e.g., emitter drift) and static factors (e.g., nonlinear distortions) have to be considered to model the mixture problem accurately.

6.5 Conclusion and discussion

In this chapter, we presented a nonlinear BICS-bNNLS algorithm that is tailored to solve mixture identification problem with a large number of constituents in the spectral library. As a linear solver, bNNLS is faster than the classical NNLS other counterparts mentioned in the experiments. Interestingly, in contrast to the *lasso* algorithm and fNNLS, both NNLS and bNNLS are faster. This can be explained by the smaller regression problem variable due to the forward selection strategy.

BICS-bNNLS also provides a more sparse solution in contrast to NNLS due to the BIC shrinkage. The *lasso* algorithm provides a sparse solution in linear scale; however, upon closer inspection, the solution is densely filled by small values, rendering ineffective for our purpose of mixture identification.

Lasly, we also improved the accuracy of our solver by accommodating nonlinear distortions and compensating for emitter drift. To collect the spectral data cube that captures the nonlinear distortions, we used the two-dimensional Gaussian process regression to interpolate spectra at low concentrations at which the spectra would be otherwise too noisy. Combining nonlinearity and emitter drift, the solver provides the best performance and resolving power (solutions correctly identifying the analytes).

Admittedly, there are limitations of this method. First, it only considers the nonlinearity caused by the change of concentration while ignoring the nonlinearity caused by interactions between constituents. However, since our target analyte is relatively simple with only a few constituents, we can safely ignore such interactions. Second, the algorithm can only compensate for minor emitter drift due to the limitation of the first-order Taylor expansion. Introducing higher-order expansions can help alleviate the problem, but they also introduce additional model complexities that demand higher spectral resolutions to keep the underlying linear system well defined.

7. FAST ACTIVE WAVELENGTH SELECTION GUIDED BY GAUSSIAN PROCESS REGRESSION

In our previous approach of active wavelength selection (MM-NNLS) described in Chapter 5, the wavelength selection strategy relies on feedback from the multi-modal solver. However, the multi-modal solver is computationally expensive, which slows down the sensing process especially when the spectral library grows large.

To address this issue, we developed a faster wavelength selection algorithm. With the faster sparse NNLS solver (nonlinear BICs-bNNLS) described in the last chapter (Chapter 6), in his chapter, we present a wavelength selection strategy guided by Gaussian process regression (GPR) and linear discriminant analysis (LDA). Both methods are advantageous because their computational complexity is a function of the number of wavelengths in the spectrum, whereas the complexity of MM-NNLS based approach is a function of the library size. Like in MM-NNLS, the wavelength selection is divided into two stages: exploration and exploitation; GPR guides the explorative stage while LDA guides the exploitative stage. To further speed up the computation, we calculate concentrations using the more efficient BICS-bNNLS solver in Chapter 6 to replace the MM-NNLS solver.

The rest of the chapter is organized as follows: Section 7.1.1 first provides an overview of the active wavelength selection process. Section 7.1.2.1 explains GPR in details, and Section 7.1.2.3 explains the how GPR guides the wavelength selection. Next, Section 7.1.3 describes the exploitative selection strategy guided by LDA. Section 7.2 shows the experimental results on both experimental and synthetic data. This chapter concludes at Section 7.4.

7.1 Active wavelength selection

7.1.1 Overview

Figure 45 illustrates a diagram of the active wavelength selection process. It consists of an inner-loop and an outer-loop. The inner-loop conducts the active wavelength selection and sensing. The wavelength selection has two stages: it first aims at reconstructing the entire spectrum (exploration), then it targets at more subtle but distinctive regions of the spectrum (exploitation). The inner-loop is computationally cheap compared to the outer-loop. The outer-loop is computationally expensive but recovers the concentration of the analyte. The new estimated concentration not only helps identify the analyte but also improves the model accuracy of the utility functions (GPR and LDA) of the wavelength selection.

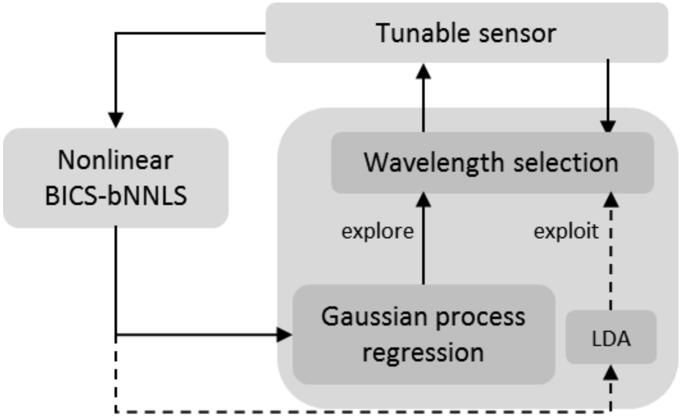


Figure 45: Diagram of the active wavelength selection framework for mixture identification problems.

7.1.2 Explorative wavelength selection

The explorative wavelength selection (inner-loop) is guided by the GPR. GPR allows us to interpolate a smooth arbitrary function using a set of sparse samplings. This procedure leverages the underlying smoothness of the target function. Figure 46 illustrates the GPR recovering an arbitrary one-dimensional function. In this example, only ten features are observed; however, because of the inherent smoothness of the function, GPR recovers it accurately especially at sampled regions. Conveniently, GPR also estimates the variance of the reconstruction, which is shown as the shaded area in Figure 46. The variance indicates how uncertain the estimation is across all features. We used this variance as the utility function of our explorative wavelength selection. The following explains GPR in details.

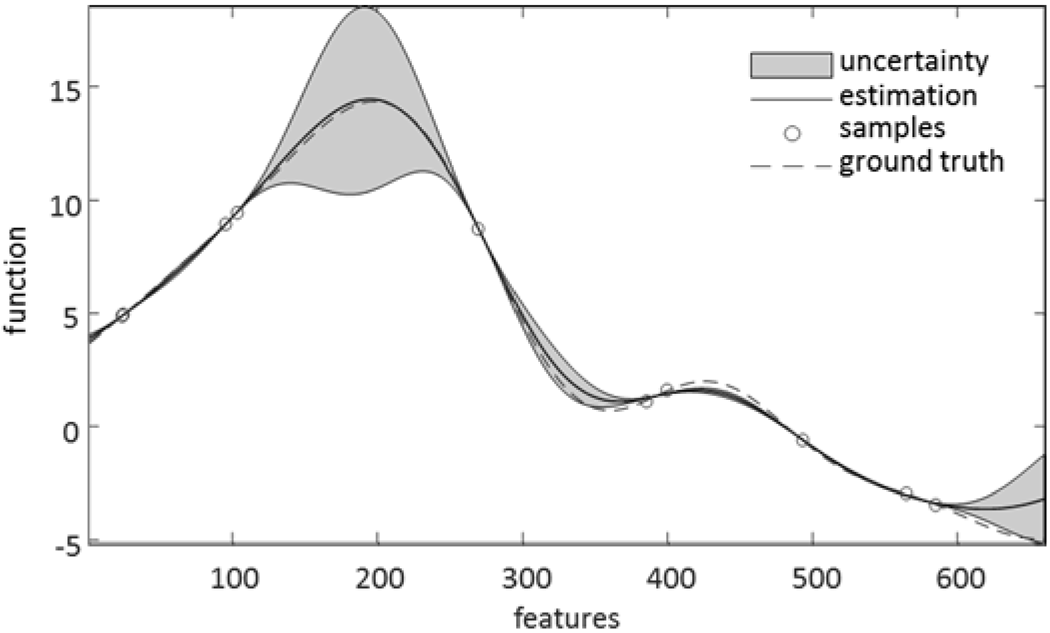


Figure 46: An example of Gaussian process regression.

7.1.2.1. Gaussian process regression

Consider the case where we have wavelengths $\boldsymbol{\lambda}_m = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$ and obtained corresponding observations $\boldsymbol{b}_{\lambda_m} = \left\{b_{\lambda_1}, b_{\lambda_2}, \dots, b_{\lambda_m}\right\}$. The goal of GPR is to reconstruct the full spectrum $\boldsymbol{b}_{GP} = \{b_{\lambda_1}, b_{\lambda_2}, \dots, b_{\lambda_M}\}$ with $M \gg m$, and calculate the variance of the estimation $\boldsymbol{S}_{GP}^2 = \left\{S_{\lambda_1}^2, S_{\lambda_2}^2, \dots, S_{\lambda_M}^2\right\}$. Gaussian processes model an arbitrary function as a multivariate random vector that follows a multivariate normal distribution $\boldsymbol{b}_{GP} \sim \mu + \mathcal{N}(\mathbf{0}, \boldsymbol{R_0})$ where μ is a scalar, and $\boldsymbol{R_0}$ is a covariance matrix $cov(\boldsymbol{\lambda}_M, \boldsymbol{\lambda}_M)$. The output of GPR is a multivariate distribution $\mathcal{N}(\boldsymbol{b}_{GP}, \boldsymbol{\Sigma}_{GP})$**.** Given the input measurements $\boldsymbol{b}_{\lambda_m}$**,** the best linear unbiased predictor to reconstruct the spectrum can be calculated as:

$$\boldsymbol{b}_{GP} = \mu + \boldsymbol{r}^T \boldsymbol{R}^{-1}(\boldsymbol{b}_{\lambda_m} - \mathbf{1}\mu) \tag{38}$$

where $\boldsymbol{R}$ denotes the $m \times m$ (auto)covariance matrix of the sampled wavelengths $\boldsymbol{\lambda}_m$ ($cov(\boldsymbol{\lambda}_m, \boldsymbol{\lambda}_m)$); $\boldsymbol{r}$ denotes the $m \times M$ covariance matrix between the sampled wavelengths $\boldsymbol{\lambda}_m$ and the output wavelengths $\boldsymbol{\lambda}_M$ ($cov(\boldsymbol{\lambda}_m, \boldsymbol{\lambda}_M)$). Although possible, calculating the full covariance matrix $\boldsymbol{\Sigma}_{GP}$ is not necessary in our case. According to [112, 113], we can directly calculate the variance of the estimation, the diagonal elements of $\boldsymbol{\Sigma}_{GP}$:

$$\boldsymbol{S}_{GP}^2 = \sigma^2 \left(1 - \boldsymbol{r}'\boldsymbol{R}^{-1}\boldsymbol{r} + \frac{(1 - \boldsymbol{1}'\boldsymbol{R}^{-1}\boldsymbol{r})^2}{\boldsymbol{1}'\boldsymbol{R}^{-1}\boldsymbol{r}} \right). \tag{39}$$

The inputs of this function are the covariance matrices $\boldsymbol{R}, \boldsymbol{r}$.

7.1.2.2. Covariance function

Constructing the covariance matrices are non-trivial because they need to be positive semi-definite. To construct the covariance matrices, a covariance function $cov(\lambda_i, \lambda_j)$ calculates the covariance value of a pair of wavelengths (λ_i, λ_j). Note that all three covariance matrices $\boldsymbol{R}_0$, $\boldsymbol{R}$, and $\boldsymbol{r}$ are constructed using the same covariance function $cov(\lambda_i, \lambda_j)$. In this work, we design the covariance function with three components:

$$cov(\lambda_i, \lambda_j) = a_{SE} \exp\left(-(\lambda_i - \lambda_j)^2/\rho\right) + a_{prod}\boldsymbol{b}_{NNLS}(\lambda_i) \cdot \boldsymbol{b}_{NNLS}(\lambda_j) + \delta_{ij}\sigma^2 \tag{40}$$

The first component, $\exp\left(-(\lambda_i - \lambda_j)^2/\rho\right)$, corresponds to the global smoothness of the spectrum. It is the squared exponential covariance function explained in [114], weighted by a scalar α_{SE}; The second component, $\boldsymbol{b}_{NNLS}(\lambda_i) \cdot \boldsymbol{b}_{NNLS}(\lambda_j)$, corresponds the product covariance function explained in [115]. This component incorporates the intermediate estimate (a spectrum) projected from the solution ($\boldsymbol{b}_{NNLS} = \boldsymbol{Ax}$). It is weighted by a scalar α_{prod}. The third component, $\delta_{ij}\sigma^2$, corresponds to the sensor's noise level, which indicates how reliable each measurement is. It is added to the diagonal of the covariance matrix with $\delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$. We provide some remarks for these three components:

- The first component, smoothness, is determined by the effective resolution of the spectrum regarding Gaussian process. It is fixed because although different chemicals have different spectral signatures, but they share the same effective resolution that is determined by the sensor.
- The second component is the estimation of the analyte spectrum given the information collected so far. The projected spectrum is incorporated into the before improving the model accuracy of the Gaussian process.
- The third component is the sensor noise. The higher the sensor noise is, the less credible each observation is, and the less drastic the model responds to each new observation. We measure the sensor's noise level beforehand assuming that it is analyte independent. This assumption holds in the case of our absorption spectrometer as the analyte is physically separated from the sensing mechanics.

7.1.2.3. Wavelength selection

As mentioned in section 7.1, the Gaussian process calculates a variance of the estimation $\boldsymbol{S}_{GP}^2 = \{\sigma_{\lambda_1}^2, \sigma_{\lambda_2}^2, \dots, \sigma_{\lambda_M}^2\}$. Thus, we can use this information to guide the feature selection. Namely, we use a myopic strategy that selects the wavelength that maximally reduces this variance. Since the only variable in equation (39) is the sampled feature set $\boldsymbol{\lambda}_m$, if we write the covariance matrix as a function of $\boldsymbol{\lambda}_m$: $\boldsymbol{S}_{GP}^{\mathbf{2}}(\boldsymbol{\lambda}_m)$, we can compute the total reduction in variance when one more wavelength is introduced:

$$\Delta_{\sigma_i^2} = \mathbf{1}' \times \left(\mathbf{S}^2(\boldsymbol{\lambda_m}) - \mathbf{S}^2(\boldsymbol{\lambda_m} \cup \lambda_i)\right), \tag{41}$$

where $\mathbf{1}$ is a column vector of ones. Using this total reduction of variance as the utility of each wavelength, we select the wavelength randomly following a probability of the utility function:

$$f_{expr} \sim \frac{\Delta_{\sigma^2}}{\sum_{i=1}^{M} \Delta_{\sigma_i^2}}. \tag{42}$$

Based on this strategy, the sampling process adapts to the previous observations and keeps sampling unexplored areas. The covariance function is also periodically updated once a new solution is solved by the BICS-bNNLS.

7.1.3 Exploitative wavelength selection

The goal of explorative wavelength selection is to reconstruct the spectrum as closely as possible. As such, it overfits the observations, causing false-positive constituents in the solution. This problem occurs when the ground truth is sparse (only a few components constitute the analyte) and becomes worse when the size of the reference library grows. To address this issue, we designed an exploitative selection strategy to sparsify the solution. Although it is impossible to know the false-positives without knowing the ground truth, we used the shrinkage result of BICS-bNNLS as an approximation.

Our approach works as follows. Recall that BICS-bNNLS generates an overfitting solution $\boldsymbol{x}'$, and then sparsifies it to $\boldsymbol{x}$ via shrinkage. The chemical constituents eliminated during shrinkage are potentially false-positives. During exploitation, we select wavelengths according to their ability to discriminate between these false-positives and remaining constituents in the solution, which we treat as an approximation of the ground truth. Figure 47 illustrates this process. Let $\boldsymbol{x}'$ (p' non-zero entries) be the BICS-bNNLS solution, $\boldsymbol{x}$ ($p < p'$ non-zero entries) be the sparsified BICS-bNNLS solution. Let $\boldsymbol{x}_{\mathbb{Z}}$ be the concentration of the eliminated entries $\boldsymbol{x}_{\mathbb{Z}} = \{x_{\mathbb{Z}_1}, x_{\mathbb{Z}_2}, \dots, x_{\mathbb{Z}_z}\}$ with indices $\mathbb{Z} = \{\mathbb{Z}_1, \dots, \mathbb{Z}_z\}$. To identify the next wavelength, we project each eliminated component back to absorbance $\boldsymbol{b}_{\mathbb{Z}_i} = \boldsymbol{A}_{\mathbb{Z}_i}\boldsymbol{x}_{\mathbb{Z}_i}$where $\boldsymbol{A}_{\mathbb{Z}_i}$ is the corresponding $\mathbb{Z}_i th$ column vector in the library matrix $\boldsymbol{A}$. Then, we calculate the LDA solution for the binary discrimination problem ($class_1 = \{\boldsymbol{b}\}$, $class_2 = \{\boldsymbol{b}_{\mathbb{Z}_1}, \dots, \boldsymbol{b}_{\mathbb{Z}_z}\}$), where $\boldsymbol{b} = \boldsymbol{b}' - \sum_{i=1}^{z} \boldsymbol{b}_{\mathbb{Z}_i}$. The LDA solution (a rotation vector $\boldsymbol{w}$**)** provides the direction maximum discrimination between the final mixture $\boldsymbol{b}$ and the eliminated components $\{\boldsymbol{b}_{\mathbb{Z}_1}, \dots, \boldsymbol{b}_{\mathbb{Z}_z}\}$ – see Figure 47. Accordingly, the exploitative wavelength selection follows a random sampling scheme with sampling probability proportional to the absolute value of the linear discriminant $|\boldsymbol{w}|$. This is an approximation of feature extraction process:

$$f_{expt} \sim \frac{|w_i|}{\sum |w_i|}. \tag{43}$$

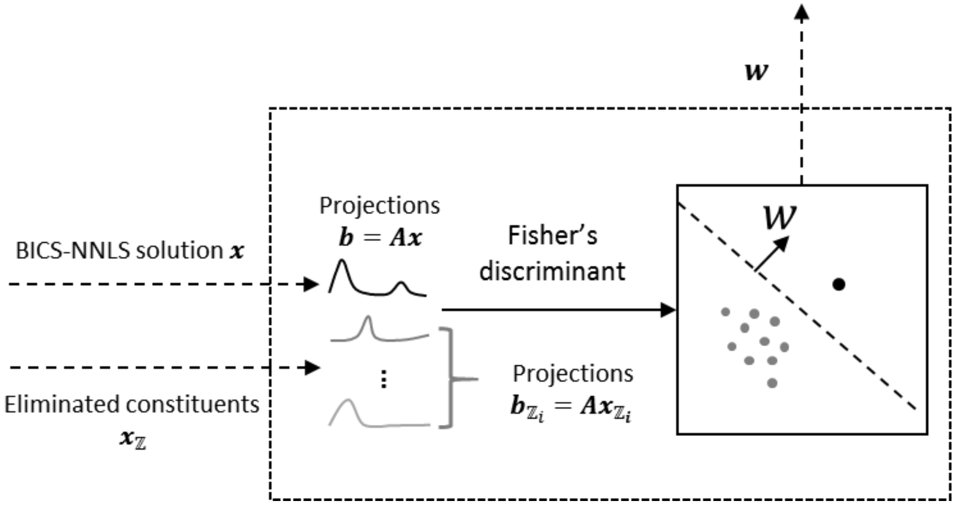


Figure 47: Exploitative wavelength selection diagram.

7.1.3.1. Switching between exploration and exploitation

The transition from exploration to exploitation (and vice versa) is signaled by the complexity of the solution. Namely, exploration stage continues for as long as the complexity of the solution continues to increase when more wavelengths are added. Denoting by $p^{(t)}$ the number of non-zero elements at step t, exploration continues for as long as $p^{(t)} > p^{(t-1)}$, and exploitation starts whenever $p^{(t)} \leq p^{(t-1)}$. The algorithm can return at any time from exploitation to exploration (if $p^{(t)} > p^{(t-1)}$), though in practice happens.

7.2 Validation on experimental data

The experimental apparatus is same as those in Chapter 5. Please refer to the Section 5.2.1 for more details.

7.2.1 Smoothness parameter tuning

In a first experiment, we investigated the global smoothness parameter ρ in equation (40). Once learned, this smoothness parameter is fixed throughout the experiment because spectra collected from the same sensor share similar smoothness. To run the cross-validation, we first collected five spectra for each chemical at 100% concentrations to achieve the highest signal-to-noise ratio. Using one setting of the parameter ρ, we generated one smoothed spectrum for each sampled spectrum. If we randomly leave one of the five spectra as the test sample (leave-one-out cross-validation), we then calculated the average of the other four smoothed spectra. The mean squared error (MSE) between the test sample and averaged spectrum served as the parameter metric. We repeated this process with parameters from $\rho = 0.2\mu m$ to $\rho = 2\mu m$ for each chemical and calculated the average MSE as the performance metric. Figure 48 illustrates the result.

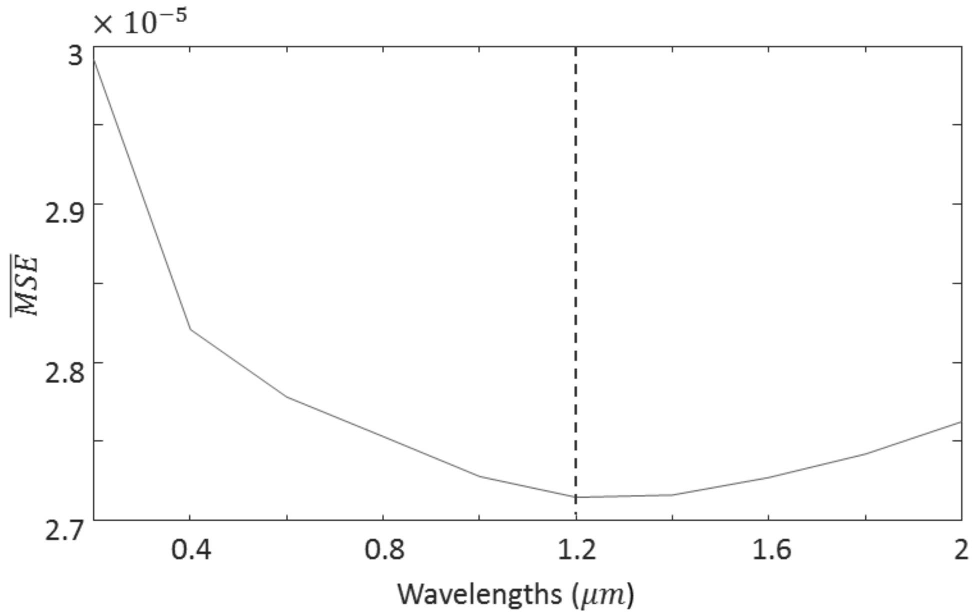


Figure 48: The fitness measured by MSE at different smoothness settings across all chemicals.

As we can see, $\rho = 1.2\mu m$ gives the optimal performance suggesting that the spectrum data collected using our sensor gives an effective resolution of $1.2\mu m$ in the context of GPR.

7.2.2 Comparison with passive algorithms

In a second experiment, we compared the active wavelength selection algorithm against a passive algorithm. As our baseline method, we used successive projection algorithm (SPA), a well-known passive algorithm [116]. SPA selects features that minimize collinearity using the sequential orthogonal projections of the Gram-Schmidt procedure. SPA is greedy: it iteratively adds one wavelength at a time, the one that is minimally correlated to the previously selected wavelengths. By minimizing correlation, the selected wavelength set is minimally redundant regarding collinearity. To avoid assumptions about which chemical is present, we trained SPA on the reference spectra of all eight chemicals. As a result, the features selected by SPA capture the signatures of all the eight chemicals.

Both algorithms were stopped when they converged to the ground truth, which we defined as the algorithm identifying the analyte correctly for ten steps in a row. In this experiment, we set 200 steps as the maximum allowable steps before converging. We tested three aspects of the algorithm: efficiency, stability, and reliability.

- Efficiency: we measured efficiency as the total number of steps before converging (excluding the ten steps required for confirmation). A smaller number implies a more efficient wavelength selection strategy.
- Stability: we measured stability by the standard deviation of the number of steps until convergence between tests. The lower the standard deviation is, the more stable the algorithm is.
- Reliability: we measured reliability by the classification rate that the algorithm successfully converges before the maximum 200 steps. The higher the classification rate is, the more reliable the algorithm is.

7.2.2.1. Selecting testing mixtures

Since there are many combinations $\left(\sum_{i=1}^{8}\binom{i}{8}=255\right)$ of the constituents, due to the lack of resolution of FPI sensor only a small portion of these testing mixtures are solvable. An ideal group of analyte should satisfy following constraints: a) The analyte is not trivial, i.e., the analytes can be identified within reasonable number of measurements; b) the selected analytes should represent a wide spread of different difficulties. Thus, we propose a two-way metric to measure the difficulty of a mixture based on the condition number. A condition number is a function of a set of spectrum $\boldsymbol{A}$; it calculates how stable a solution of a linear system $\boldsymbol{x}$ is with respect to the disturbance of the observations $\boldsymbol{b}$. A l_2 norm condition number can be calculated as:

$$cond(\boldsymbol{A}) = \frac{\sigma_{max}(\boldsymbol{A})}{\sigma_{min}(\boldsymbol{A})} \tag{44}$$

where $\sigma_{max}(\boldsymbol{A})$ and $\sigma_{min}(\boldsymbol{A})$ are the maximum and minimum eigenvalues of the matrix $\boldsymbol{A}$. We used the condition number to calculate two aspects of the analyte. One is to measure how resolvable the analyte is:

$$R_{\boldsymbol{A}_s} = cond(\boldsymbol{A}_s) \tag{45}$$

where $\boldsymbol{A}_s$ is a column matrix with the spectra of the constituents of the analyte. The other metric of the difficulty is to measure distinguishable the analyte is:

$$D_{\boldsymbol{A}_s} = cond(\boldsymbol{A}) - cond(\boldsymbol{A}_{-S}) \tag{46}$$

where $\boldsymbol{A}$ has all the reference spectra and $\boldsymbol{A}_{-S}$ has all reference spectra except the ones of the constituents in the analyte. This directly calculates the condition number contribution of the analyte constituents. Intuitively, the smaller the number is, the more correlated the analyte is to the other constituents in the library, thus harder to be distinguished from other constituents using linear solvers. With this measure, we picked following analytes as shown in Table 10.

Table 10: The analytes components and their abbreviations

Analyte	**Abbreviation**	**# components**	$R_{\boldsymbol{A}_s}$	$D_{\boldsymbol{A}_s}$	$R_{\boldsymbol{A}_s} + D_{\boldsymbol{A}_s}$
Tert-Butyl alcohol	TBA	1	1	48	49
Ethyl alcohol	EA	1	1	55	56
Isopropyl alcohol	IA	1	1	57	58
Denatured alcohol	DA	1	1	75	76
Tert-Butyl alcohol & Brush cleaner	TBA+BC	2	5	65	70
Denatured alcohol & tert-Butyl alcohol	DA+TBA	2	3	96	99
Lacquer thinner	LT	1	1	98	99
Lacquer thinner & isopropyl alcohol	LT+ISA	2	4	105	109
Brush cleaner & acetone	BC+ACT	2	11	120	131

As an example, we illustrate the most difficult mixture, brush cleaner & acetone (BC+ACT), in Figure 49. The largest difference between these two chemicals is the minor absorption from 9.4 to 10 microns. During our experiments, complex mixtures that have a higher difficulty (measured by $R_{\boldsymbol{A}_s} + D_{\boldsymbol{A}_s}$) become unsolvable using the spectral data collected from our sensor.

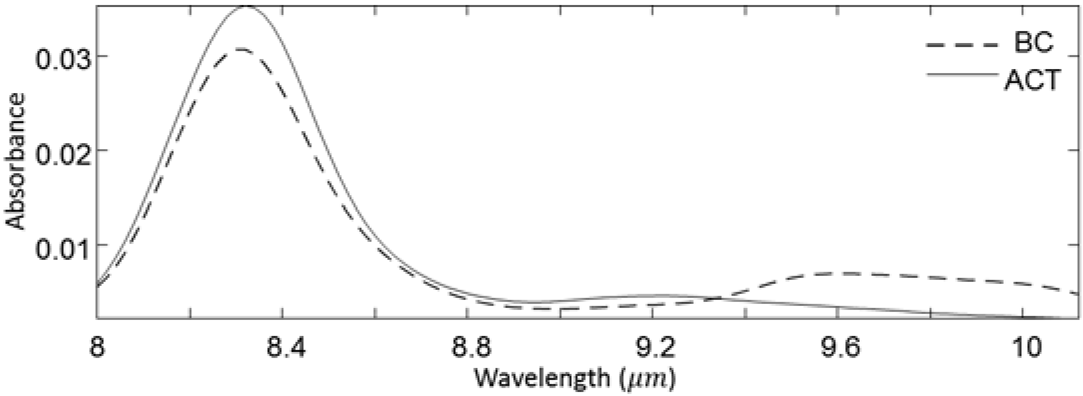


Figure 49: The spectra of acetone and brush cleaner. They are both very similar to each other (hard to resolve).

7.2.2.2. Performance comparison

Once the analytes were chosen, we tested both algorithms (active and passive) on each analyte 25 times for a total of $9 \times 25 = 225$ tests. The sequence of the tests was selected randomly to eliminate ordering effects, and the gas cell was purged with air before each test to avoid any residual. Figure 50 shows the result of efficiency comparison. As we can see, the active framework outperforms the passive algorithm across all analytes.

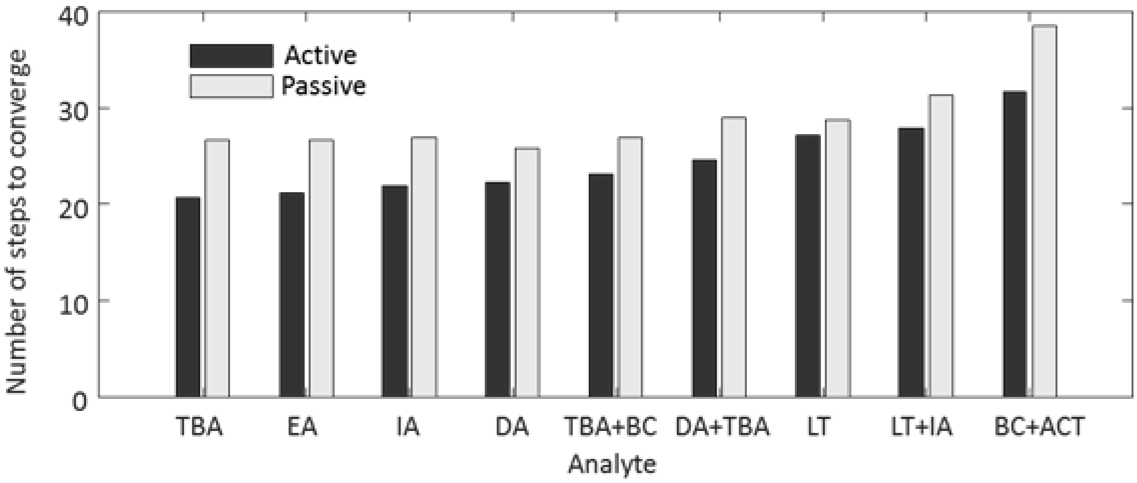


Figure 50: Efficiency test – the number of steps measurements used to converge to the correct solution.

Figure 51 shows the result for stability. As we can see, active sensing framework is also more stable compared to SPA for all analytes. Notice that the performance gap diminishes when the complexity of the analyte becomes higher. This is as expected because a more complicated chemical requires coverage of more spectral signatures, which is equivalent to a passive algorithm.

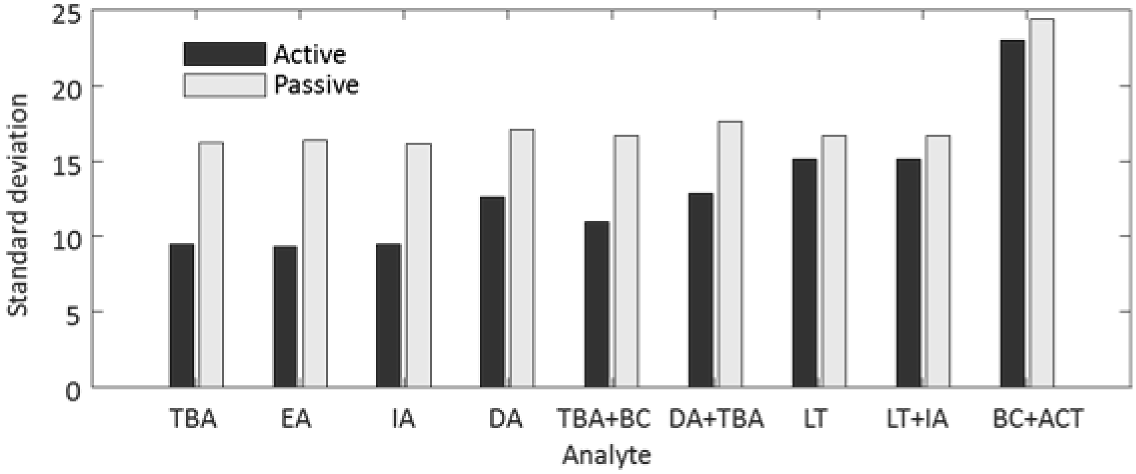


Figure 51: Stability test – the standard deviation of the number of steps required before convergence.

Efficiency and stability provide a performance metric for the algorithms when they converge. In some tests, the algorithm never converges before a finite number of measurements. As a result, results of those tests have to be excluded. Hence, we use another metric – the classification rate – to measure how reliable the algorithm is. Figure 52 shows results.

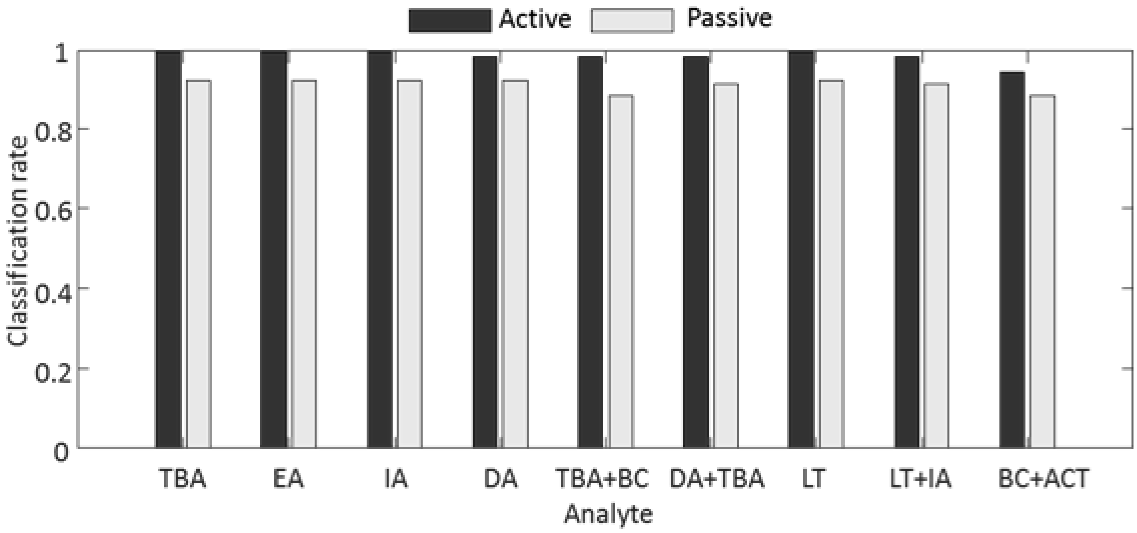


Figure 52: Reliability test – classification rate for the two algorithms. The algorithm needs to converge before 200 steps; otherwise, its result is considered misclassified.

As we can see, the active strategy has a better chance to identify the analyte correctly. It is noteworthy to mention that even for simple analytes (such as single chemical TBA, EA, IA), the passive algorithm did not reach 100% classification rate while the active strategy successfully identifies the chemical every time.

7.3 Validation on synthetic data

To provide a more thorough evaluation than what can be afforded experimentally, we also analyzed the active wavelength selection algorithm on a large dataset of synthetic IR spectra. The dataset consisted of FTIR spectra (660 spectral lines) from 500 chemicals in the NIST Webbook infrared absorption spectrum database [107]. To simulate the spectral resolution of FPIs, we convolved the FTIR spectra with a Gaussian filter of 0.1μm spread. Each spectrum was normalized to sum up to one. For the subsequent experiments, we compared the proposed active wavelength selection algorithm against the passive algorithm described in section 7.2.1. In all cases, we allowed the algorithms to sample each wavelength multiple times.

7.3.1 Performance comparison

In a third experiment, we tested the algorithm on mixtures with more than 50 constituents. We added white noise with standard deviation at 1% of the median value of all absorption spectra in the library. However, the classification rate collapsed to nearly zero as the number of constituents went beyond 50. That suggests that the solver reached the maximum effective resolvability of the spectral library. As the difficulty of a mixture problem can vary dramatically (e.g., a badly conditioned two-component mixture can be unsolvable while a well-conditioned 20-component mixture can be easily identified), we designed a mixture construction policy so that the chosen problems would be neither too trivial nor unsolvable. For this purpose, we randomly selected many 50-component mixtures and calculated their classification rate with the set noise level. We then selected the five mixtures that can be correctly classified $1\% - 10\%$ of the time. For each of these five 50-component mixtures, we sequentially removed one component at a time to form chemical mixtures of a lower order; this process ensures a gradual transition in problem complexity from hard to easy. For each of the resulting 250 mixtures (50 × 5), we evaluated the active and passive algorithms 40 times, each time with randomly added noise, for 10,000 cases. The maximum number of allowable measurements is 5000.

Similar to the procedures and the metrics used in experimental validation in section 7.2.2.2, we used the average steps to converge, the variance, and the classification rate as the measures for efficiency, stability, and reliability respectively. First, Figure 53 and

Figure 54 show the average and the standard deviation of the number of measurements taken before convergence. The result is the statistics of 200 (40×5) tests for each order of mixture.

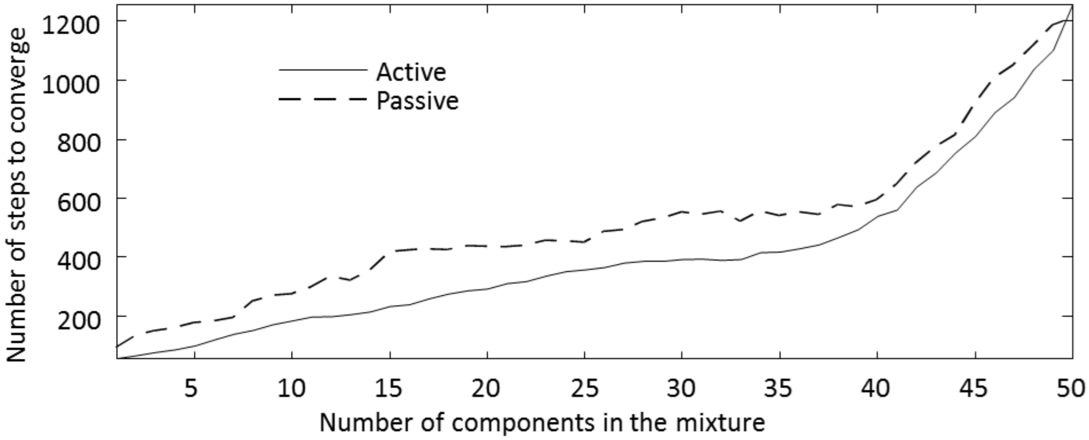


Figure 53: Efficiency test – total number measurements required before convergence from 1-component chemical to the 50-component chemical mixture.

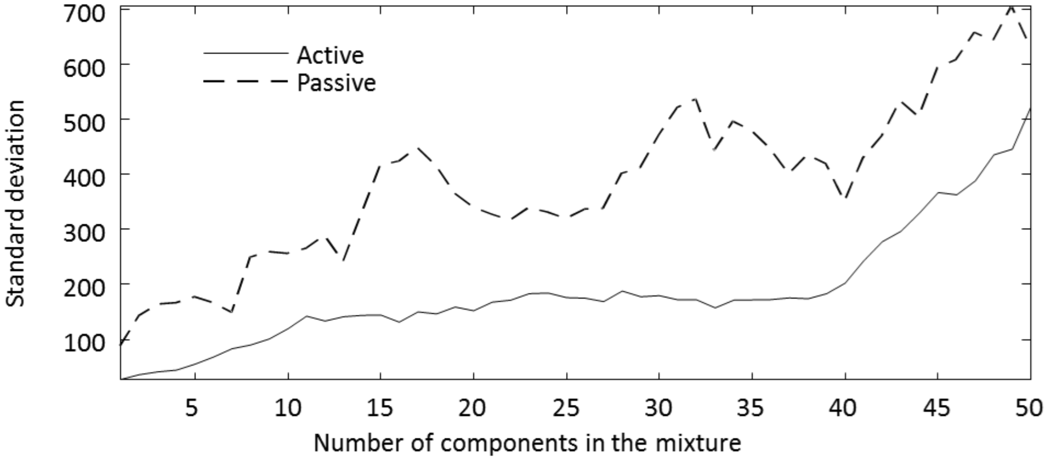


Figure 54: Stability test – the standard deviation of the number of measurements before convergence.

As we can see, the result is consistent with the ones in the experimental section. Active approach outperforms the passive algorithm for all analytes. The active algorithm is also much more stable than the passive approach. This is an expected outcome of an analyte adaptive approach. To illustrate the improvements better, we also calculated the improvement ratio: $\frac{P-A}{P}$. This ratio tells how much percentage the improvement is on the basis of the passive algorithm. Figure 55 shows the result.

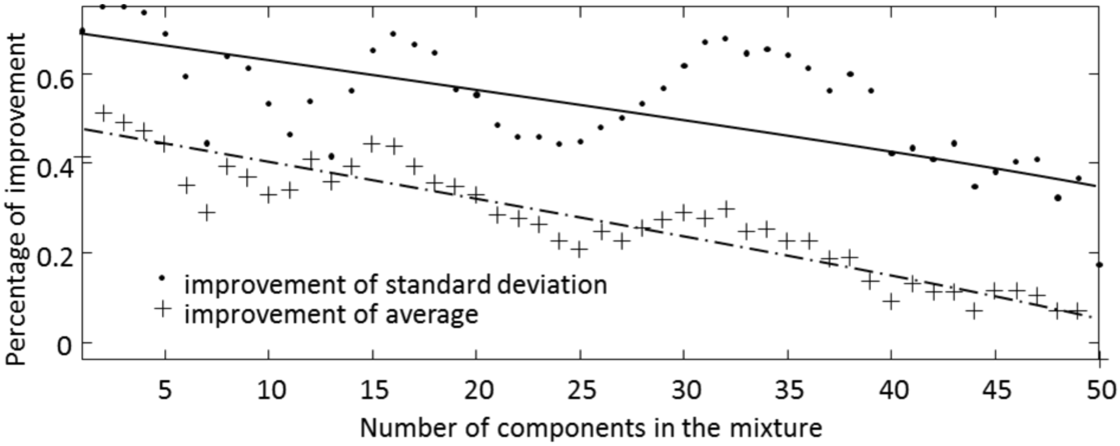


Figure 55: Improvement of the active approach over the passive approach regarding both the efficiency and stability.

As we can see here, the result is also consistent with the ones in experimental validation. Active approach outperforms passive approach especially in the aspect of stability. It is also consistent with the experimental result that the higher complexity of the analyte eventually diminishes the advantage of active approach for efficiency. The stability advantage of the active approach is more prominent than the efficiency advantage.

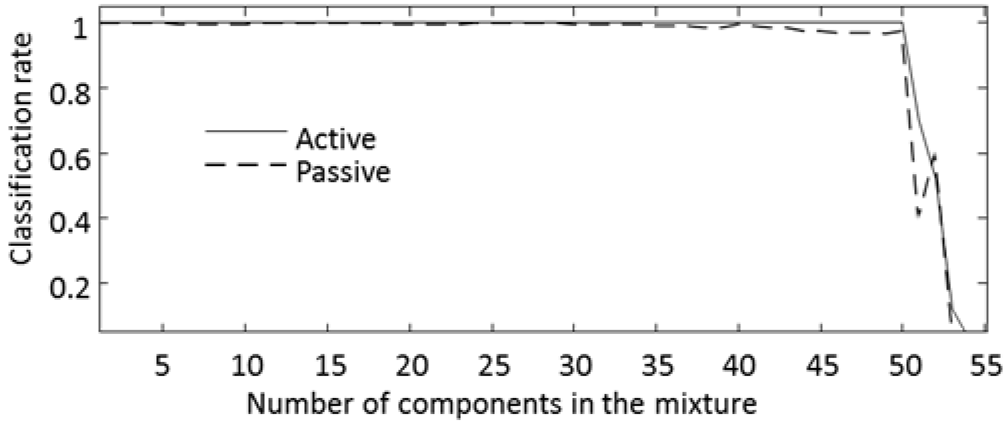


Figure 56: Reliability test – the classification rates of active approach and passive approach.

Lastly, we calculate the classification rate to measure the reliability when the complexity of the analyte becomes higher. Figure 56 shows the result. As we can see, active approach maintains a 100% classification rate until passing 50-component mixture while passive approach usually fluctuates but rarely reach a 100% classification rate (see Figure 57 as a zoomed-in version). Another interesting observation is that the classification rates of both algorithms collapsed drastically after reaching 50-component mixture. This might suggest that the measurements have reached the intrinsic dimensionality of the underlying linear system – effective rank under noise explained in section 2.3.2.1. In the experiments, we found that three factors could influence the classification rate: the measurement noise level, the collinearity of the linear system, and the maximum number of measurements allowed. The maximum number of measurements plays a less important role asymptotically because averaging samplings reduces the noise level quadratic-hyperbolically ($\sigma\left(\frac{\sum_{i=1}^{N} x_i}{N}\right) \sim \frac{1}{\sqrt{N}}\sigma_0$ where $x_i \sim$

$\mathbb{N}(0,\sigma_0)$). Equivalently, to compensate the noise level, the number of samples needs to grow quadratically, which soon becomes infeasible with real instruments.

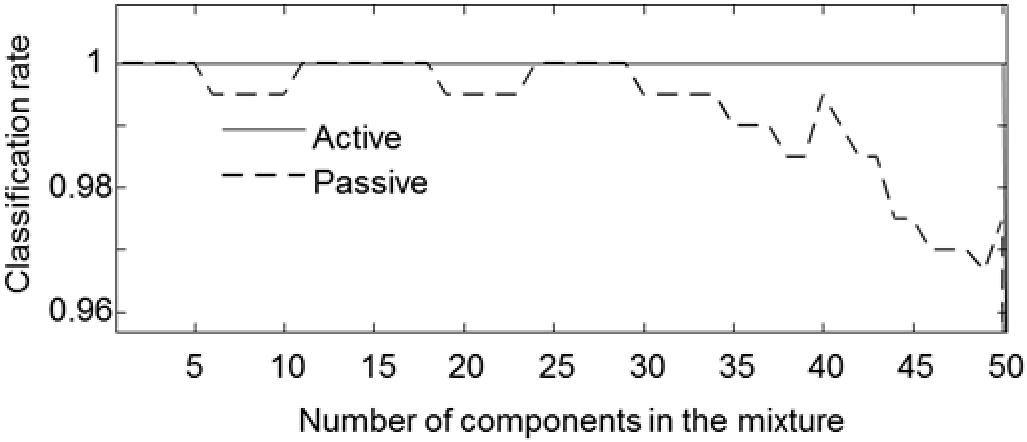


Figure 57: A zoomed-in version of Figure 56.

7.3.2 Convergence rates comparison

In a fourth experiment, we analyzed the performance of exploration stage and exploitation stage independently. In the concentration space, the exploration stage adds more constituents to recover the spectrum, whereas exploitative stage removes constituents to accelerate the convergence. Thus, the sparsity of the solution is a good indicator of how fast and well each stage performed. Let $\|\cdot\|_0$ be the l_0 norm of a vector that is equivalent to the number of non-zero entries in the vector. Let the relative sparsity be $p = \frac{\|x\|_0}{\|x_{true}\|_0}$ where $\boldsymbol{x_{true}}$ denotes the ground truth and $\boldsymbol{x}$ denotes the estimated concentration. Figure 58 shows the average relative sparsity for 1-component analyte. At the exploration stage, the solution complexity kept increasing until approximately eight measurements. This result suggests that the eight measurements of 660 wavelengths can capture the overall structure of the entire spectrum using GPR. Compared with passive approach, active method grows the solution slower at with the same amount of measurements during exploration. Consequently, the active approach overshoots less and helps accelerate convergence during exploitation, On the other hand, at the exploitation stage, the active approach is able to select the wavelengths that reduce the solution complexity much faster than the passive method. As a result, the active method converges to the solution much earlier than the passive method.

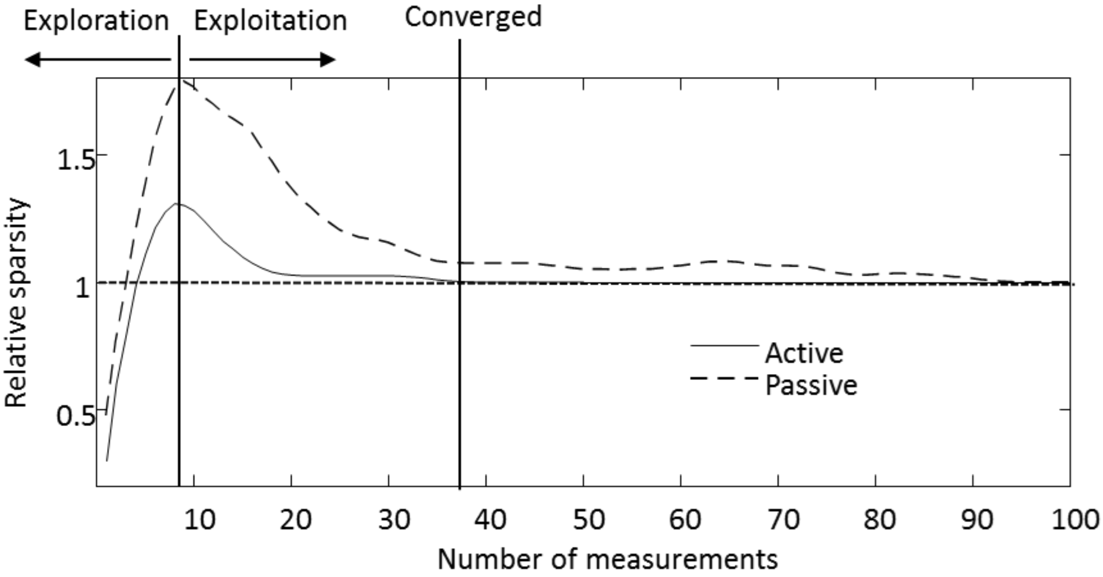


Figure 58: An example of the relative sparsity trajectory through the sensing process.

Next, we also investigated the relative sparsity across different orders of mixtures. Figure 59 shows the average trajectory of the 1-component mixtures to 51-component mixtures with an increment of 10. The advantage above of active wavelength selection is most prominent at lower-order mixtures. As the complexity of the mixture grows, the convergence rate begins to decrease at exploitation stage. At the extreme case of a 51-component mixture, the exploitative wavelength selection becomes ineffective. This result suggests that the system has reached the maximum resolvability of the spectrum given the spectral library and this noise level. Interestingly, the GPR guided explorative wavelength selection manages to maintain a more controlled complexity growth rate than the SPA passive algorithm.

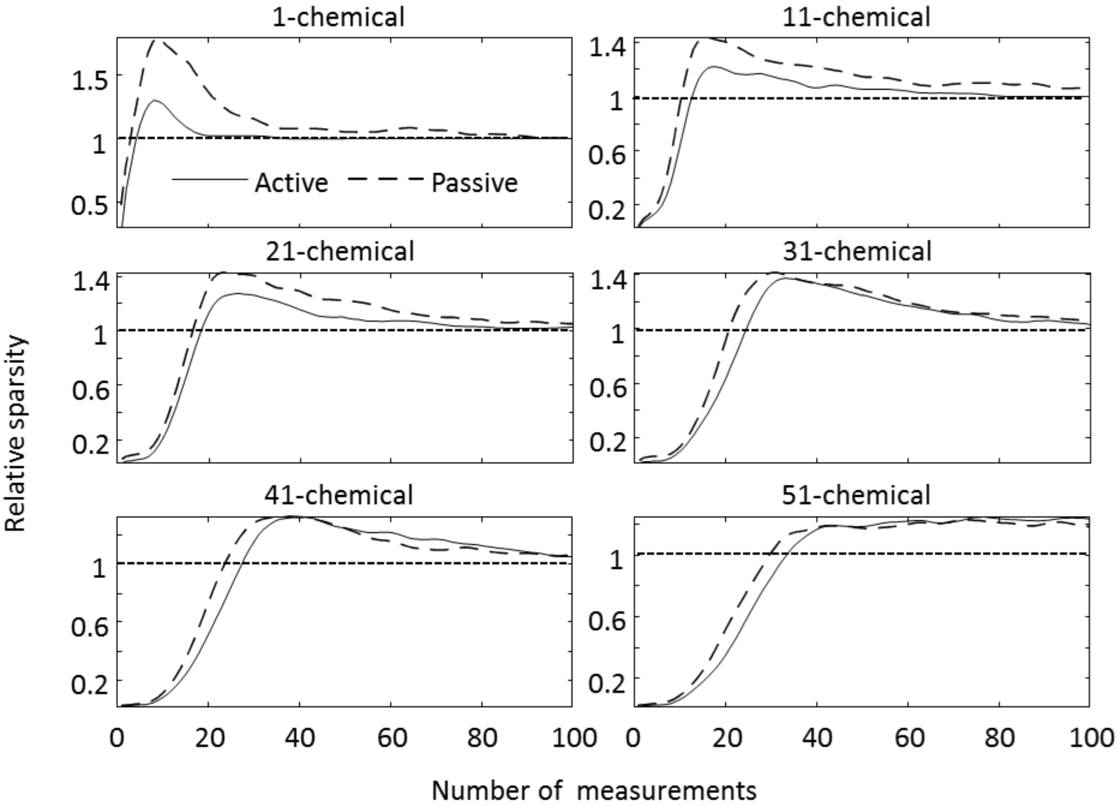


Figure 59: The relative sparsity during the first 100 measurements across different orders of mixtures.

7.4 Conclusion and discussion

In this chapter, we have presented an active wavelength selection based on GPR and LDA method. The wavelength selection method introduced a feedback loop from the mixture estimation from the nonlinear BICS-bNNLS solver, and the feedback adjusts the prior of the wavelength selection algorithm, thus the selected wavelength set is analyte dependent. The wavelength selection algorithm consists of two stages: exploration and exploitation. The exploration stage is based on Gaussian processes aiming to select wavelengths for spectrum recovery; the exploitation stage is based on a variation method to sparsify the least squares solution calculated using the measurements explored in the exploration stage. Both stages are unsupervised, and they do not require a typical training-validating process. The method is also computationally efficient, suitable for portable platforms with limited computation resources.

We evaluated our approach on both Fabry-Perot interferometer sensor and synthetic data. Experimentally, we tested our approach on up to two-chemical mixture problem out of a library of eight chemicals. Using the cross-validation method, we quantified the resolution of our Fabry-Perot sensor with a resolution of 1.2 μm. For the more comprehensive studies, we also tested the approach on synthetic data with up to fifty-chemical mixture out of a library of five hundred chemicals. We compared our method with a passive method, successive projection algorithm. Both experimental and simulation results suggest that the active approach outperforms passive approach. The active method has a faster convergence rate, and, more importantly, performed much more stable with the presence of different analytes.

However, both experimental and simulation results showed that the performance gain of the active approach became smaller as the number of the constituents in the mixture became larger. This is as expected because the passive method was trained to cover the signatures of the spectra in the entire reference library. Interestingly, the active method still consistently performed better regarding stability. This highlights the advantage of selecting an analyte-dependent feature. With the identities and concentrations of the analyte changing, there is no global optimal feature set, since the absolute optimal feature set can only be found by oracle methods that already have the knowledge of the identities and concentrations of the analyte. Therefore, an iterative active approach is the solution for more general wavelength selection problems.

8. CONCLUSION

We developed a set of techniques to identify chemical or chemical mixture with online wavelength selection and sensing. There are two main aspects of the contribution: the adaptive wavelength selection for mixture identification and the non-negative least squares solvers for mixture analysis.

As to the adaptive wavelength selection, we developed three approaches that are based on Bayesian risk, multi-modal solver, and the Gaussian process regression respectively. The Bayesian risk based approach can only identify single chemicals; the multi-modal solver based approach can identify chemical mixtures but is computationally costly. However the proposed Gaussian process regression approach improves the computational speed by leveraging the smoothness of the spectrum.

Additionally, we developed two sparse NNLS (non-negative least squares) algorithms. The first solver is the multi-modal NNLS that generates multiple solutions. It generates more sparse solutions than what classical NNLS does, and its multi-modality enables our first adaptive wavelength selection to identify chemical mixture. The second solver is the nonlinear BICS-bNNLS. It speeds up the classical NNLS using batch updating, and BIC shrinkage method provides additional sparsity. Furthermore, it provides a greater accuracy by accommodating nonlinearity and emitter drift.

We evaluated the three active wavelength selection algorithms together with the corresponding solvers on both synthesized and experimental FPI data. The results lead to several insights:

- The optimal wavelengths for identification are analyte-dependent. Thus, in addition to improving the sensing speed, active wavelength selection improves the accuracy.
- Similar to many other applications, the linear model on which IR spectral mixture analysis relies on can be broken due to the limitation of experimental setup. Factors such as nonlinearity and emitter drift introduce structural error that becomes the bottleneck of the platform.
- The active sensing strategy becomes superior under two circumstances: when the model is nonlinear and when the observations are noisy. By sampling a smaller number of wavelengths, the algorithm introduces less nonlinearity and it is able to leverage repeated sampling to compensate for noise.
- In contrast to passive wavelength selection algorithms, the analyte-dependent advantages of active sensing diminish as the analyte consists of more components, or numerically speaking, the concentration vector grows dense. This is as expected because higher-order mixture covers a large number of spectral signatures that are shared with many constituents.

8.1 Future work

8.1.1 Studies of nonlinearity in chemical interactions

Because the nonlinearity becomes the bottleneck of further improvement of accuracy, accommodation for nonlinearity can boost the performance. In this dissertation, we accommodate nonlinearity caused by changes in concentrations; however, we did not consider the nonlinearity caused by chemical interaction. In real applications, mixing two chemicals together often leads to nonlinearity through two mechanisms: reversible reactions and irreversible reactions. Irreversible reactions between chemicals essentially generate new chemicals. Reversible reaction shifts the equilibrium state of the mixture, changing the effective concentration of the constituents. Much effort needs to be made to investigate such interactions with spectral data in order to conduct analysis and identification for higher-order mixtures. Since the mixture is combinatorial, studying them is labor-intensive.

8.1.2 Active sensing based on Bayesian multivariate linear regression

The Bayesian approach of linear regression represent both the solutions and the spectral library in terms of distribution. The distribution representation is a natural development the multi-modal NNLS approach in Chapter 5. Instead of generating multiple solutions that are sampled from a distribution, the Bayesian multivariate linear regression offers the whole distribution. Active wavelength selection can leverage the information of such distributions as the utility function of the wavelength selection process.

8.1.3 Generalized effective rank

The effective rank provides a theoretical bound of the number of resolvable constituents. As an extension of the original effective rank developed by Roy et al. [14], we developed an effective rank with observation noise level in consideration (see 2.3.2.2). However, this method does not consider correlation between features; neither does this method consider possible noisy readings in the spectral library. To generalize the effective rank, both the spectral library and the observations need to be represented by multivariate distribution. The generalized effective rank should shine light on the maximum resolvable constituents under the context of generalized least squares.

8.1.4 Effective rank under nonlinearity

Since nonlinearity in the spectrum also contributes to the deterioration of the effective rank, further generalization of the effective rank requires incorporation of nonlinearities. Studies of different nonlinear transformation or nonlinear operators and their impact on effective rank can shine light on future experimental design so that the experiment can avoid the detrimental effects caused by nonlinearity.

8.1.5 Active chemical verification

So far, we have investigated the problem of chemical identification. It is a general framework without any prior knowledge of the analytes. In many real world applications, we are typically interested in only certain constituents, such as some specific pollutants in the atmosphere. In this case, the target analyte is known. Since we can exploit the characteristics of the target analyte, it is a simpler problem to verify its existence than identifying every constituent in the mixture.

Admittedly, when all the constituents of both the analyte and the backgrounds are known, the optimal wavelength selection strategy regresses to the traditional passive wavelength selection. However, when the identities of backgrounds are unknown, active wavelength selection is still beneficial. In this case, the optimal wavelength set is background-dependent. Depending on the similarities and differences between the background constituents and the target constituent, the most distinctive wavelengths for the analyte of interest vary.

REFERENCES

[1] H. Ding, J. Liang, J. Cui *et al.*, “A novel fiber Fabry-Perot filter based mixed-gas sensing system,” *Sensors and Actuators B: chemical*, pp. 154-159, 2009.

[2] N. Neumann, M. Ebermann, S. Kurth *et al.*, "Novel MWIR microspectrometer based on a tunable detector."

[3] C. H. Spiegelman, M. J. McShane, M. J. Goetz *et al.*, “Theoretical Justification of Wavelength Selection in PLS Calibration: Development of a New Algorithm,” *Anal. Chem.*, vol. 70, no. 1, pp. 35-44, 1998.

[4] B. Hemmateenejad, R. Ghavamia, R. Mirib *et al.*, “Net analyte signal-based simultaneous determination of antazoline and naphazoline using wavelength region selection by experimental design-neural networks,” *Talanta,* vol. 67, no. 4, pp. 1222–1229, 2006.

[5] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão1 *et al.*, “The successive projections algorithm for variable selection in spectroscopic multicomponent analysis,” *Chemometrics and Intelligent Laboratory Systems,* vol. 57, no. 2, pp. 65–73, 2001.

[6] B. Hemmateenejad, M. Akhond, and F. Samari, “A comparative study between PCR and PLS in simultaneous spectrophotometric determination of diphenylamine, aniline, and phenol: Effect of wavelength selection,” *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,* vol. 67, no. 3-4, pp. 958–965, 2007.

[7] H. Swierengaa, F. Wülfertb, O. E. de Noordc *et al.*, “Development of robust calibration models in near infra-red spectrometric applications,” *Analytica Chimica Acta,* vol. 311, no. 1-2, pp. 121–135, 2000.

[8] R. Leardi, “Genetic algorithms in chemometrics and chemistry: a review,” *Journal of Chemometrics,* vol. 15, no. 7, pp. 559-569, 2001.

[9] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad *et al.*, “Ant colony optimisation: a powerful tool for wavelength selection,” *Journal of Chemometrics,* vol. 20, no. 3-4, pp. 146-157, 2006.

[10] Y. P. Du, Y. Z. Liang, J. H. Jiang *et al.*, “Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares,” *Analytica Chimica Acta,* vol. 501, no. 2, pp. 183-191, 1/16/, 2004.

[11] A. Martin, and R. M. Synge, “A new form of chromatogram employing two liquid phases: A theory of chromatography. 2. Application to the micro-determination of the higher monoamino-acids in proteins,” *Biochemical Journal,* vol. 35, no. 12, pp. 1358, 1941.

[12] D. Harvey, "Spectroscopic methods of analysis," *Modern analytical chemistry*, pp. 368 - 460: McGraw-Hill New York, 2000.

[13] A. Lipson, S. G. Lipson, and H. Lipson, *Optical physics*: Cambridge University Press, 2010.

[14] O. Roy, and M. Vetterli, "The effective rank: A measure of effective dimensionality." pp. 606-610.

[15] R. Gutierrez-Osuna, “Pattern analysis for machine olfaction: a review,” *Sensors Journal, IEEE,* vol. 2, no. 3, pp. 189-202, 2002.

[16] E. H. Moore, “On the reciprocal of the general algebraic matrix,” *Bulletin of the American Mathematical Society,* vol. 26, no. 9, pp. 394-395, 1920.

[17] R. L. Basmann, “A generalized classical method of linear estimation of coefficients in a structural equation,” *Econometrica: Journal of the Econometric Society*, pp. 77-83, 1957.

[18] C. L. Lawson, and R. J. Hanson, "Problem NNLS," *Solving least squares problems*, pp. 160 - 165: Prentice-hall, Inc., 1974.

[19] R. Bro, and S. De Jong, “A fast non-negativity-constrained least squares algorithm,” *Journal of chemometrics,* vol. 11, no. 5, pp. 393-401, 1997.

[20] M. H. Van Benthem, and M. R. Keenan, “Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems,” *Journal of chemometrics,* vol. 18, no. 10, pp. 441-450, 2004.

[21] V. K. Potluru, S. M. Plis, S. Luan *et al.*, "Sparseness and a reduction from totally nonnegative least squares to svm." pp. 1922-1929.

[22] S. G. Mallat, and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *Signal Processing, IEEE Transactions on,* vol. 41, no. 12, pp. 3397-3415, 1993.

[23] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition." pp. 40-44.

[24] A. E. HOERL, “Application of ridge analysis to regression problems,” *Chemical Engineering Progress,* vol. 58, pp. 54-59, 1962.

[25] J. F. Claerbout, and F. Muir, “Robust modeling with erratic data,” *Geophysics,* vol. 38, no. 5, pp. 826-844, 1973.

[26] R. Mammone, O. McKee, and D. Schilling, "Frequency resolution enhancement of a compressive receiver by spectral estimation." pp. 713-719.

[27] M. F. Duarte, M. A. Davenport, D. Takhar *et al.*, “Single-pixel imaging via compressive sampling,” *IEEE Signal Processing Magazine,* vol. 25, no. 2, pp. 83, 2008.

[28] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 267-288, 1996.

[29] S. Boyd, and L. Vandenberghe, "Interior-point methods," *Convex optimization,* pp. 561-323: Cambridge university press, 2004.

[30] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM journal on scientific computing,* vol. 20, no. 1, pp. 33-61, 1998.

[31] E. J. Candes, J. K. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Communications on pure and applied mathematics,* vol. 59, no. 8, pp. 1207-1223, 2006.

[32] E. J. Candes, and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?,” *Information Theory, IEEE Transactions on,* vol. 52, no. 12, pp. 5406-5425, 2006.

[33] W. J. Fu, “Penalized regressions: the bridge versus the lasso,” *Journal of computational and graphical statistics,* vol. 7, no. 3, pp. 397-416, 1998.

[34] H. Zou, and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 67, no. 2, pp. 301-320, 2005.

[35] G. C. Tiao, and A. Zellner, “On the Bayesian estimation of multivariate regression,” *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 277-285, 1964.

[36] P. M. Williams, “Bayesian regularization and pruning using a Laplace prior,” *Neural computation,* vol. 7, no. 1, pp. 117-143, 1995.

[37] T. Park, and G. Casella, “The bayesian lasso,” *Journal of the American Statistical Association,* vol. 103, no. 482, pp. 681-686, 2008.

[38] C. E. Rasmussen, and C. K. I. Williams, *Gaussian processes for machine learning*: The MIT Press, 2006.

[39] P. M. Narendra, and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *Computers, IEEE Transactions on,* vol. 100, no. 9, pp. 917-922, 1977.

[40] S. D. Frans, and J. M. Harris, "Selection of analytical wavelengths for multicomponent spectrophotometric determinations," *Analytical Chemistry,* vol. 57, no. 13, pp. 2680-2684, 1985.

[41] J. Smits, W. Melssen, L. Buydens *et al.*, "Using artificial neural networks for solving chemical problems: Part I. Multi-layer feed-forward networks," *Chemometrics and Intelligent Laboratory Systems,* vol. 22, no. 2, pp. 165-189, 1994.

[42] R. Todeschini, D. Galvagni, J. Vılchez *et al.*, "Kohonen artificial neural networks as a tool for wavelength selection in multicomponent spectrofluorimetric PLS modelling: application to phenol, o-cresol, m-cresol and p-cresol mixtures," *TrAC Trends in Analytical Chemistry,* vol. 18, no. 2, pp. 93-98, 1999.

[43] L. Norgaard, A. Saudland, J. Wagner *et al.*, "Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy," *Applied Spectroscopy,* vol. 54, no. 3, pp. 413-419, 2000.

[44] R. Leardi, and L. Nørgaard, "Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions," *Journal of chemometrics,* vol. 18, no. 11, pp. 486-497, 2004.

[45] Z. Xiaobo, Z. Jiewen, H. Xingyi *et al.*, "Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of 'Fuji'apple based on different PLS models," *Chemometrics and Intelligent Laboratory Systems,* vol. 87, no. 1, pp. 43-51, 2007.

[46] J.-H. Jiang, R. J. Berry, H. W. Siesler *et al.*, "Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data," *Analytical chemistry,* vol. 74, no. 14, pp. 3555-3565, 2002.

[47] Y. Du, Y. Liang, J. Jiang *et al.*, "Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares," *Analytica chimica acta,* vol. 501, no. 2, pp. 183-191, 2004.

[48] V. Centner, D.-L. Massart, O. E. de Noord *et al.*, “Elimination of uninformative variables for multivariate calibration,” *Analytical chemistry,* vol. 68, no. 21, pp. 3851-3858, 1996.

[49] W. Cai, Y. Li, and X. Shao, “A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra,” *Chemometrics and intelligent laboratory systems,* vol. 90, no. 2, pp. 188-194, 2008.

[50] S. Ye, D. Wang, and S. Min, “Successive projections algorithm combined with uninformative variable elimination for spectral variable selection,” *Chemometrics and Intelligent Laboratory Systems,* vol. 91, no. 2, pp. 194-199, 2008.

[51] J. J. Gibson, *The ecological approach to visual perception*, 1st ed., New Jersey: Lawrence Erlbaum Associates, 1986.

[52] J. J. Gibson, “Observations on active touch,” *Psychological Review,* vol. 69, no. 6, pp. 477-491, 1962.

[53] R. Bajcsy, “Active perception,” *Proceedings of the IEEE,* vol. 76, no. 8, pp. 966-1005, 1988.

[54] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, “Active vision,” *International Journal of Computer Vision,* vol. 1, no. 6, pp. 333-356, 1988.

[55] Z. Yongmian, and J. Qiang, “Active and dynamic information fusion for facial expression understanding from image sequences,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 27, no. 5, pp. 699-714, 2005.

[56] T. H. Chung, V. Gupta, J. W. Burdick *et al.*, "On a decentralized active sensing strategy using mobile sensor platforms in a network." pp. 1914-1919 Vol.2.

[57] N. Roy, W. Burgard, and S. T. D. Fox, "Coastal navigation-mobile robot navigation with uncertainty in dynamic environments." pp. 35-40.

[58] D. Fox, W. Burgard, and S. Thrun, “Active Markov localization for mobile robots,” *Robotics and Autonomous Systems,* vol. 25 no. 3-4, pp. 195 - 207, 1998.

[59] J. J. Leonard, and H. F. Durrant-Whyte, "Simultaneous map building and localization for an autonomous mobile robot." pp. 1442-1447.

[60] H. J. Kushner, “A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise,” *Journal of Basic Engineering,* vol. 86, no. 1, pp. 97-106, 1964.

[61] J. Močkus, "On Bayesian methods for seeking the extremum." pp. 400-404.

[62] J. Mockus, “Application of Bayesian approach to numerical methods of global and stochastic optimization,” *Journal of Global Optimization,* vol. 4, no. 4, pp. 347-365, 1994/06/01, 1994.

[63] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine learning,* vol. 47, no. 2-3, pp. 235-256, 2002.

[64] T. L. Lai, and H. Robbins, “Asymptotically efficient adaptive allocation rules,” *Advances in applied mathematics,* vol. 6, no. 1, pp. 4-22, 1985.

[65] J. L. T. Zhang, “The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits.”

[66] J.-Y. Audibert, R. Munos, and C. Szepesvári, “Exploration–exploitation tradeoff using variance estimates in multi-armed bandits,” *Theoretical Computer Science,* vol. 410, no. 19, pp. 1876-1902, 2009.

[67] R. Kleinberg, A. Slivkins, and E. Upfal, "Multi-armed bandits in metric spaces." pp. 681-690.

[68] S. Bubeck, and N. Cesa-Bianchi, “Regret analysis of stochastic and nonstochastic multi-armed bandit problems,” *arXiv preprint arXiv:1204.5721*, 2012.

[69] V. Kuleshov, and D. Precup, “Algorithms for multi-armed bandit problems,” *arXiv preprint arXiv:1402.6028*, 2014.

[70] E. J. Candes, and T. Tao, “Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies,” *IEEE transaction on Information Theory,* vol. 52, no. 12, pp. 5406-5425, 2006.

[71] J. Haupt, R. Castro, and R. Nowak, "Distilled Sensing: Selective Sampling for Sparse Signal Recovery."

[72] R. M. Castro, and R. D. Nowak, “Minimax bounds for active learning,” *Information Theory, IEEE Transactions on,* vol. 54, no. 5, pp. 2339-2353, 2008.

[73] R. Lomasky, C. E. Brodley, M. Aernecke *et al.,* "Active Class Selection," *Machine Learning: ECML 2007,* Lecture Notes in Computer Science J. Kok, J. Koronacki, R. Mantaras *et al.,* eds., pp. 640-647: Springer Berlin Heidelberg, 2007.

[74] I. Rodriguez-Lujan, J. Fonollosa, A. Vergara *et al.*, “On the calibration of sensor arrays for pattern recognition using the minimal number of experiments,” *Chemometrics and Intelligent Laboratory Systems,* vol. 130, pp. 123-134, 2014.

[75] T. Nakamoto, S. Ustumi, N. Yamashita *et al.*, "Active gas/odor sensing system using automatically controlled gas blender and numerical optimization technique," *Sensors and Actuators B: Chemical,* vol. 20, no. 2–3, pp. 131-137, 1994.

[76] T. Nakamoto, N. Okazaki, and H. Matsushita, "Improvement of optimization algorithm in active gas/odor sensing system," *Sensors and Actuators A: Physical,* vol. 50, no. 3, pp. 191-196, 1995.

[77] R. Gutierrez-Osuna, and A. Hierlemann, "Adaptive microsensor systems," *Annual Review of Analytical Chemistry,* vol. 3, pp. 255-276, 2010.

[78] B. Raman, A. Gutierrez-Galvez, A. Perera-Lluna *et al.*, "Sensor-based machine olfaction with a neurodynamics model of the olfactory bulb." pp. 319-324.

[79] R. Gutierrez-Osuna, and P. Sun, "A biologically-plausible computational architecture for sensor-based machine olfaction." pp. 57-59.

[80] R. Gutierrez-Osuna, N. Powar, and P. Sun, "Chemosensory adaptation in an electronic nose." pp. 223-229.

[81] R. Gutierrez-Osuna, and N. U. Powar, "Odor mixtures and chemosensory adaptation in gas sensor arrays," *International Journal on Artificial Intelligence Tools,* vol. 12, no. 01, pp. 1-16, 2003.

[82] R. Gosangi, and R. Gutierrez-Osuna, "Active Temperature Programming for Metal-Oxide Chemoresistors," *IEEE Sensors Journal,* vol. 10, no. 6, pp. 1075 - 1082, 2010.

[83] R. Gosangi, and R. Gutierrez-Osuna, "Energy-aware active chemical sensing." pp. 1094-1099.

[84] R. Gosangi, and R. Gutierrez-Osuna, "Data-driven Modeling of Metal-oxide Sensors with Dynamic Bayesian Networks." pp. 135-136.

[85] R. Gosangi, and R. Gutierrez-Osuna, "Quantification of gas mixtures with active recursive estimation." pp. 23-24.

[86] R. Gosangi, and R. Gutierrez-Osuna, "Active temperature modulation of metal-oxide sensors for quantitative analysis of gas mixtures," *Sensors and Actuators B: Chemical,* vol. 185, pp. 201-210, 2013.

[87] A. Hierlemann, and R. Gutierrez-Osuna, "Higher-order chemical sensing," *Chemical reviews,* vol. 108, no. 2, pp. 563-613, 2008.

[88] R. Gosangi, and R. Gutierrez-Osuna, "Active classification with arrays of tunable chemical sensors," *Chemometrics and Intelligent Laboratory Systems*, 2014.

[89] C. E. Priebe, D. J. Marchette, and D. M. Healy, "Integrated sensing and processing decision trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 26, no. 6, pp. 699-708, 2004.

[90] D. V. Dinakarababu, D. R. Golish, and M. E. Gehm, "Adaptive feature specific spectroscopy for rapid chemical identification," *Optics Express,* vol. 19, no. 5, pp. 4595-4610, 2011.

[91] R. Gosangi, and R. Gutierrez-Osuna, "Active Temperature Programming for Metal-Oxide Chemoresistors," *Sensors Journal, IEEE,* vol. 10, no. 6, pp. 1075 - 1082, 2010.

[92] S. Ji, and L. Carin, "Cost-sensitive feature acquisition and classification," *Pattern Recognition,* vol. 40, no. 5, pp. 1474-1485, 2007.

[93] D. D. Lee, and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature,* vol. 401, no. 6755, pp. 788-791, 1999.

[94] C. L. Lawson, and R. J. Hanson, *Solving least squares problems*: SIAM, 1995.

[95] Beer, "Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten," *Annalen der Physik und Chemie,* vol. 86, pp. 78-88, 1852.

[96] A. Dempster, N. Laird, and D. Rdin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B,* vol. 39, no. 1, pp. 1-38, 1977.

[97] H. A. Seipel, and J. H. Kalivas, "Effective rank for multivariate calibration methods," *Journal of chemometrics,* vol. 18, no. 6, pp. 306-311, 2004.

[98] G. Goertzel, "An Algorithm for the Evaluation of Finite Trigonomentric Series," *The American Mathematical Monthly,* 1958.

[99] I. Guyon, and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[100] G. S. Mallat, and Z. Zhang, "Matching Pursuits With Time-Frequency Dictionaries," *IEEE Transactions on signal processing,* vol. 41, no. 12, pp. 3997-3415, 1993.

[101] S. Das, S. Maity, B. Qu *et al.*, "Real-parameter evolutionary multimodal optimization — A survey of the state-of-the-art," *Swarm and Evolutionary Computation,* vol. 1, no. 2, pp. 71-88, 2011.

[102] E. J. Wagenmakers, and S. Farrell, “AIC model selection using Akaike weights,” *Psychonomic Bulletin & Review,* vol. 11, no. 1, pp. 192-196, 2004/02/01, 2004.

[103] K. P. Burnham, and D. R. Anderson, "Model selection and multimodel inference: a practical information-theoretic approach," p. 63: Springer, 2002.

[104] J. Huang, and R. Gutierrez-Osuna, “Active analysis of chemical mixtures with multi-modal sparse non-negative least squares,” in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013, pp. 8756-8760.

[105] S. W. Wilson, "Explore/exploit strategies in autonomy." pp. 325-332.

[106] A. W. Whitney, “A direct method of nonparametric measurement selection,” *Computers, IEEE Transactions on,* vol. 100, no. 9, pp. 1100-1103, 1971.

[107] P. M. Chu, F. R. Guenther, G. C. Rhoderick *et al.*, "Qunatitative Infrared Database," *NIST Chemistry WebBook, NIST Standard Reference Database Number 69,* N. Eds. P.J. Linstrom and W.G. Mallard, ed.

[108] H. Akaike, “A new look at the statistical model identification,” *Automatic Control, IEEE Transactions on,* vol. 19, no. 6, pp. 716-723, 1974.

[109] G. Schwarz, “Estimating the dimension of a model,” *The annals of statistics,* vol. 6, no. 2, pp. 461-464, 1978.

[110] C. K. Williams, "Prediction with Gaussian processes: From linear regression to linear prediction and beyond," *Learning in graphical models*, pp. 599-621: Springer, 1998.

[111] S. Kim, K. Koh, M. Lustig *et al.*, “A method for large-scale l1-regularized least squares problems with applications in signal processing and statistics,” *IEEE J. Select. Topics Signal Process,* vol. 1, no. 4, pp. 606-617, 2007.

[112] D. R. Jones, M. Schonlau, and W. J. Welch, “Efficient global optimization of expensive black-box functions,” *Journal of Global optimization,* vol. 13, no. 4, pp. 455-492, 1998.

[113] J. Sacks, W. J. Welch, T. J. Mitchell *et al.*, “Design and analysis of computer experiments,” *Statistical science,* pp. 409-423, 1989.

[114] C. E. Rasmussen, and C. K. I. Williams, "Squared exponential covariance function," *Gaussian processes for machine learning,* pp. 83-84: The MIT Press, 2006.

[115] C. E. Rasmussen, and C. K. I. Williams, "Dot product covariance funcitons," *Gaussian processes for machine learning,* pp. 89-90: The MIT Press, 2006.

[116] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão *et al.*, “The successive projections algorithm for variable selection in spectroscopic multicomponent analysis,” *Chemometrics and Intelligent Laboratory Systems,* vol. 57, no. 2, pp. 65-73, 2001.

[117] F. Hayashi, "Econometrics," Princeton University Press Princeton, NJ, 2000, p. 27.

[118] H. V. Poor, "An introduction to signal detection and estimation," pp. 5-9: Springer, 1994.

APPENDIX A: NONLINEAR DEVIATION OF BEER'S LAW

The linear relationship in Beer's law dictates that for a certain analyte, the slope ϵ mentioned in equation (3) is a constant. However, in practice, deviation can occur. Furthermore, the effect of the negative deviation increases the spectrum is sharper. To give an intuitive proof, let us assume that one neighboring spectral line is leaked to the detector because of the imperfect wavelength selector. Using equation (1), (2), and (3), the effective absorbance is:

$$\alpha = log_{10}\frac{I_0+I_0'}{I_0 10^{-\epsilon lc}+I_0' 10^{-\epsilon\prime lc}}\ , \tag{47}$$

where I_0 is the targets spectral lines, and the I_0' is a neighboring "leaked" radiation energy; ϵ and ϵ' are the corresponding molar absorptivity at these two neighboring wavelengths. We can calculate the second derivative of the absorbance α as to the concentration c:

$$\frac{d^2\alpha}{dc^2} = -\frac{I_0 I_0' l^2 10^{cl(\epsilon+\epsilon')}\, log(10)(\epsilon-\epsilon')^2}{\left(10^{\epsilon cl}I_0 + 10^{\epsilon' cl}I_0'\right)^2} \leq 0. \tag{48}$$

As we can see, firstly, Beer's law holds ($\frac{d^2\alpha}{dc^2} = 0$) if and only if $\epsilon = \epsilon'$; secondly, the negative deviation worsens when the slope is steeper, $\frac{d^2\alpha}{dc^2} \propto -(\epsilon - \epsilon')^2$. Figure (6) gives an illustration of such non-uniform nonlinear deformation.

APPENDIX B: THE EFFECTIVE RANK UNDER NOISE

The intrinsic dimensionality of a linear system refers to the maximum number of dimensions resolvable for a linear inversion problem. The rank, which measures the number of linearly independent components, is a traditional metric for such intrinsic dimensionality. An alternative, effective rank, was proposed by Roy et al. [14]. Like the rank, the effective rank provides an indication of the intrinsic dimensionality. Unlike the rank, the effective rank offers a continuous measure (so it is possible to have 3.4 dimensions) by computing the entropy of the eigenvalues of the matrix $\boldsymbol{A}$ in the linear system in equation (8). For example, when all eigenvalues are the same and non-zero, the effective rank is the highest. However, the method does not put sensor noise into consideration. Here, we propose a method that considers both the traditional rank of the linear system and the observational noise level: the effective rank with the consideration of noise.

Given a linear system $\boldsymbol{Ax} = \boldsymbol{B}$, assuming the sensor noise follow Gaussian distribution $\epsilon \sim \mathcal{N}(0,\ \sigma^2)$, the estimation of $\boldsymbol{x}$ also follows a multivariate normal distribution $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}_{true},\ \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1})$ [117]. This joint distribution of the estimation $\boldsymbol{x}$ provides us the insight of how reliable the estimation is going to be. The effective rank of this linear system can be then represented as the expected number of components being "reliably" estimated. We quantify Reliability as the probability of the variables $\boldsymbol{x}$

estimated within a range of error, i.e., probability of the estimation being in a hyper-cube $\left(\boldsymbol{T}: x_i \in \left[x_{i_{true}} - \frac{\Delta_i}{2}, x_{i_{true}} + \frac{\Delta_i}{2}\right],\ \mathrm{i} = 1 \dots \mathrm{N}\right)$.

$$p(\boldsymbol{x} \in \boldsymbol{T}) = \int_{\boldsymbol{T}} p(\boldsymbol{x}) d^n\boldsymbol{x} \ . \tag{49}$$

For example, assuming the total number of components $N = 2$, the two dimensional normal distribution can be illustrated in Figure 60. Then the integrating region is the square area marked as T. We define the length of the cube being $\frac{1}{N}$, inspired by the fact that the total concentration is one $\boldsymbol{I}^T \times \boldsymbol{x} = 1$.

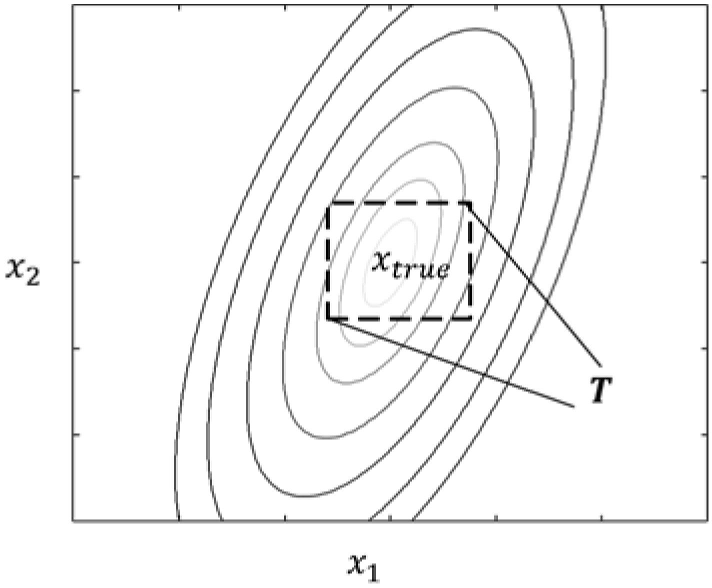


Figure 60 The hypercube regional integration over a two-component linear system.

Given N individual chemical components, there are at most $N_{mix} = 2^N - 1$ possible analytes from single chemicals to N-component mixtures. If we can calculate the expected number of chemical mixtures $\bar{N}_{mix} = E(N_{mix} | \epsilon_{\boldsymbol{x}} \in \left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right])$ being reliably estimated as mentioned before, we can recover the intrinsic dimensionality:

$$E\left[N | \epsilon_{\boldsymbol{x}} \in \left[-\frac{\boldsymbol{\Delta}}{2}, \frac{\boldsymbol{\Delta}}{2}\right]\right] = log_2(\bar{N}_{mix} + 1) \tag{50}$$

where $\boldsymbol{\Delta}$ denotes the margin in the hyper-cube that defines reliable estimation.

Calculating $\boldsymbol{\bar{N}_{mix}}$ is computationally expensive because the number of potential mixtures are combinatorial ($\boldsymbol{2^N - 1}$). Furthermore, calculating the integrals of the multivariate distributions is computationally prohibitive. To alleviate this problem, we rotate the mixture components using SVD (singular-value-decomposition), since each rotated component is orthogonal to each other, we can simplify the multivariate integration to a product of a sequence of univariate integration:

$$p\left(\boldsymbol{x}_j \in \boldsymbol{T} | \boldsymbol{A}, \sigma\right) = \prod_{i=1}^{N} \left(-1 + 2\Phi_{\frac{\sigma^2}{\lambda_i}}\left(\frac{\Delta}{2}\right)\right) \tag{51}$$

where $\Phi_{\frac{\sigma^2}{\lambda_i}}(\cdot)$ is the cumulative distribution function of a zero-mean normal distribution with variation $\frac{\sigma^2}{\lambda_i}$. The final equation for effective rank under noise is:

$$erankN(\boldsymbol{A},\sigma) = log_2\left(\sum_{k=1}^{N} \sum_{j \in \binom{N}{k}} p\left(\boldsymbol{x}_j \in \boldsymbol{T} | \boldsymbol{A}, \sigma\right) + 1\right) \tag{52}$$

where $\boldsymbol{x}_j$ is the estimated solution of one of the $\binom{N}{k}$ possible k-component problem; $\boldsymbol{T}$ is a hyper-cube centering at the ground truth of the corresponding mixture problem. $\boldsymbol{A}$ and σ are the key parameters in the linear system described in equation (9).

To illustrate new effective rank under different noise levels, we compute the new effective rank on the same positive semi-definite circulant matrix $\boldsymbol{A}$ used in the original effective rank [14]. The matrix is defined as:

$$\boldsymbol{A} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho & \rho^2 & \rho & 1 \end{bmatrix}, \tag{53}$$

where the parameter $\rho \in [-1,1]$. When $\rho = 0$, the matrix has the highest rank of four, as ρ approaches either 1 or -1, the matrix becomes more ill-conditioned and eventually reaches rank 1. The result of the new effective rank is illustrated in Figure 61. As the noise level decreases, the effective rank converges to the rank that assumes noiseless measurement. As the noise level increases, the condition of the matrix becomes less relevant as the effective rank function flattens. It is worthwhile to mention that, when the noise is overwhelmingly high ($\sigma^2 \gg 1$), i.e., the variance of the noise dominates the information of the linear system in matrix $\boldsymbol{A}$, the effective rank eventually converge to zero.

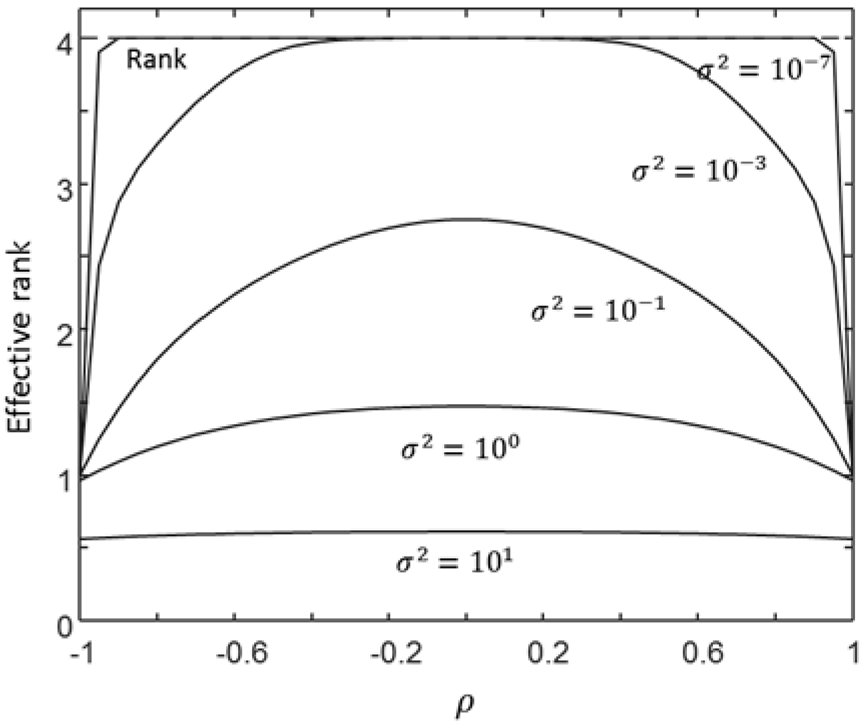


Figure 61 Effective rank vs ρ under different noise level. The red dashed line is the rank. The solid lines illustrate different effective ranks at different noise level. σ^2 denotes the variance in the normal distribution for the noise.

APPENDIX C: OBSERVATION DISCRETIZATION

equation (22) is only applicable in discrete observation domains; in a continuous space, it becomes an intractable integration. To address this problem, we discretize the continuous observation space into a finite number of discrete values. For each wavelength λ_i, we uniformly discretize the corresponding observation space into a sorted set of d discrete values $\{\bar{o}_{i,1}, \bar{o}_{i,2} \dots, \bar{o}_{i,d}\}$. The posterior probability of the k^{th} discrete observation for chemical ω_j is calculated as:

$$p\left(\bar{o}_{i,k} \middle| \lambda_i, \omega_j\right) = \int_{\frac{\bar{o}_{i,k}+\bar{o}_{i,k-1}}{2}}^{\frac{\bar{o}_{i,k}+\bar{o}_{i,k+1}}{2}} \sum_{g=1}^{M} N\left(\bar{o}_{i,k} \middle| \mu_{i,j,g}, \sigma_{i,j,g}\right) \tag{54}$$

The number of discrete observations d influences the accuracy and computational complexity of equation (22). Therefore, after some experiments, we choose the d to be 200 to approximate the continuous observation space and still allow real-time operation.

APPENDIX D: APPROXIMATION OF MISCLASSIFICATION COST

To reduce computational costs, the variance of each wavelength across all candidate observations in equation (29) is used as an approximation of the misclassification risk. Assume the binary classification problem illustrated in Figure 62. Using Bayesian decision theory [118], its misclassification risk $\mathcal{R}$ can be calculated as:

$$\mathcal{R} = C_{12}P_{12} + C_{21}P_{21} \tag{55}$$

where C_{ij} is the cost of wrongly assigning a sample to class j when i is the correct class, P_{ij} is the probability of such misclassification: $P_{ij} = \int_{\Gamma_{ij}} p(x)dx$, and Γ_{ij} represents the region where such misclassifications may occur (see Figure 62). Assume that the noise in observation space is independent and normally distributed, and that both costs are equal$(C_{12} = C_{21} = C)$. Since the two distributions are symmetric relative to the classification boundary, the misclassification risk is monotonically related to the distance between the two means (μ_1, μ_2):

$$\mathcal{R} = C_{12}P_{12} + C_{21}P_{21} = \frac{2C}{\sigma\sqrt{2\pi}} \int_{\Gamma_{12}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx \ \propto \ -\Phi\left(\frac{|\mu_1 - \mu_2|}{2}\right) \tag{56}$$

where $\Phi(*)$ is a monotonically increasing function. Intuitively, this means that the further the two Gaussian means are, the easier the binary classification problem is.

In our case, we have M candidates, so the problem becomes one of $M - ary$ classification with total misclassification risk given by:

$$\mathcal{R} = \sum_{\forall i,j,i\neq j} C \times P_{ij} \quad \propto \quad - \sum_{\forall i,j,i\neq j} \Phi\left(\frac{|\mu_i - \mu_j|}{2}\right). \tag{57}$$

This computation is expensive when M is large, as is our case. However, since $\Phi(*)$ is monotonic, there also exists a monotonic function $\mathrm{G}(*) = \Phi\left(\frac{\sqrt{*}}{2}\right)$ such that

$$\Phi\left(\frac{|\mu_i - \mu_j|}{2}\right) = \Phi\left(\frac{\sqrt{|\mu_i - \mu_j|^2}}{2}\right) = \mathrm{G}\left(|\mu_i - \mu_j|^2\right) \tag{58}$$

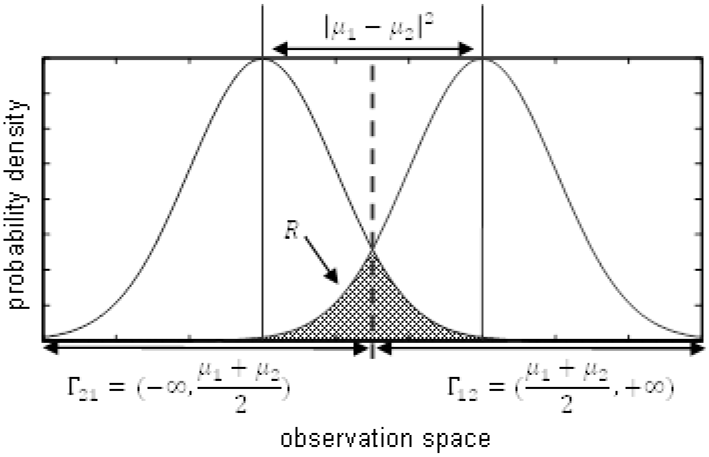

$|\mu_1 - \mu_2|^2$
probability density
R
$\Gamma_{21} = (-\infty, \frac{\mu_1 + \mu_2}{2})$
$\Gamma_{12} = (\frac{\mu_1 + \mu_2}{2}, +\infty)$
observation space

Figure 62: Misclassification risk of a binary classification problem.

Using the 1st order Taylor approximation: $\mathrm{G}(\mathrm{z}) \approx G(0) + G'(0) \times z$, we have:

$$R \propto - \sum_{\forall i,j,i\neq j} \Phi\left(\frac{|\mu_i - \mu_j|}{2}\right) \approx - \left(G(0) + G'(0) \times \sum_{\forall i,j,i\neq j} |\mu_i - \mu_j|^2 \right) \propto -\sigma^2(\boldsymbol{\mu}) \tag{59}$$

where $\sigma^2(\boldsymbol{\mu})$ is the variance defined in equation (29). Thus, by selecting the wavelength with maximum variance we minimize the misclassification risk.

APPENDIX E: SIGNAL PROCESSING FOR THE FPI SENSOR

The FPI sensor platform consists of emitter, gas cell, FPI detector, evaluation board, and a computer as illustrated in Figure 63. The evaluation board is responsible for driving the emitter with modulated signal, sending tuning signal to FPI sensor and receiving/processing output signal from the FPI sensor. The tuning wavelength and the final processed data transmit through USB to a computer.

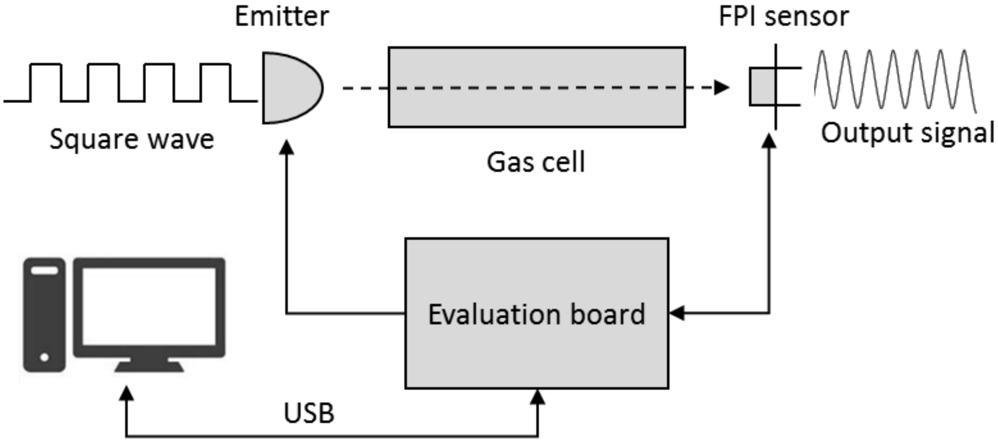


Figure 63: Diagram of the FPI platform for chemical identification.

The FPI detector is based on pyroelectric effect. Pyroelectric effect is a property of certain materials to generate a voltage as a response to the change of its temperature. A common method to utilize this property for measuring signal is through modulation. As shown in Figure 63, we send a modulated signal (a square wave) to the emitter. Then, the whole chain of the sensing platform (emitter, lens, gas, lens, and FPI sensor) serves as low pass filters that smooth the modulated signal. As a result, the FPI sensor acquires a sinusoid-like signal. Figure 64(a) shows a typical example of the raw modulated signal collected from the FPI sensor. In this example, a new tuning is set at around the 700 milliseconds, the signal shifts to a signal with larger amplitude suggesting a stronger transmittance at this new wavelength. To measure the detected energy is to calculate the average amplitude of these periodical signals at the modulation frequency. Discrete-time Fourier transform (DTFT) with continuous frequency solves the problem:

$$P_f = T \sum_{n=-\infty}^{\infty} s(nT) \cdot e^{-i2\pi f nT} \,. \tag{60}$$

where f corresponds to the frequency of interest, $s(\cdot)$ represents the time-series signal, T corresponds to the sampling interval, n denotes the sample index.

A common problem of Fourier transform is the spectral leakage. Spectral leakage is the blurring effect in frequency domain where a portion of the energy at other frequencies "leaked" into the frequency of interest. As to the raw FPI signal as shown in Figure 64(a), the signal is accompanied by a low-frequency component, a drift, when the sensor is settling for new tunings.

Fortunately, since the signal is periodical and its interval is fixed. We eliminate the low-frequency component by subtracting the moving average calculated with a sliding window at a size of the interval. The extracted moving average is shown in Figure 64(a) and the compensated signal is shown in Figure 64(b).

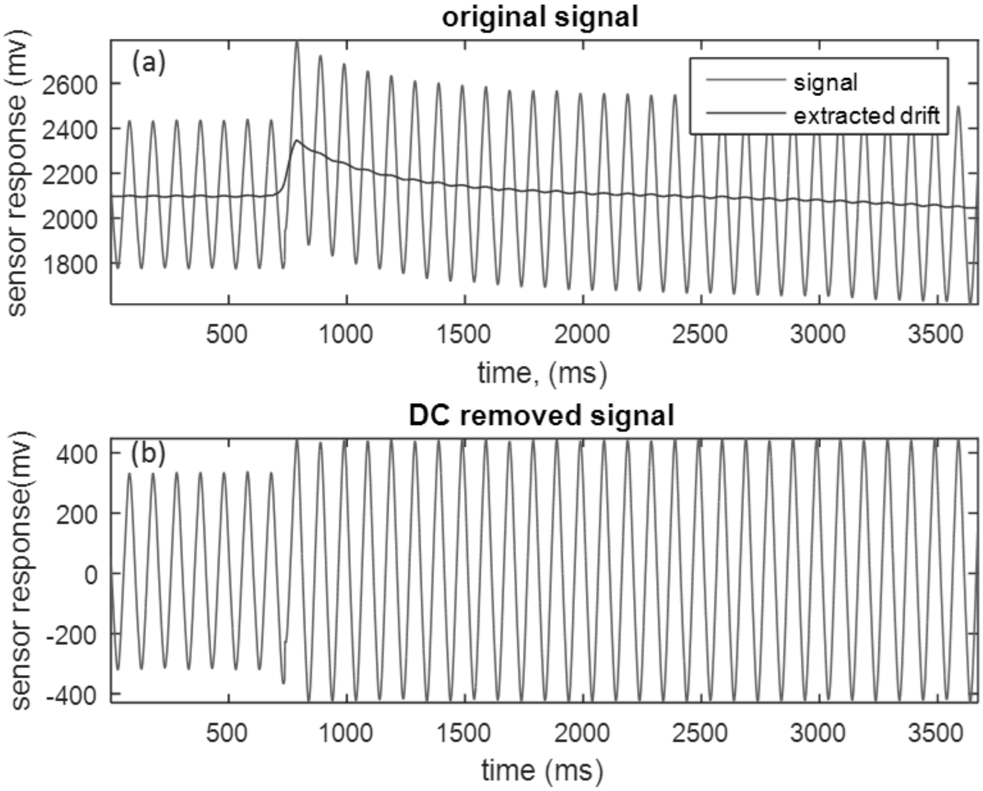


Figure 64: The raw signals before and after drift compensations.