ABSTRACT OF DISSERTATION

Bin Dai

The Graduate School

University of Kentucky

2007

SIMULATIONS-GUIDED DESIGN OF PROCESS ANALYTICAL SENSOR USING MOLECULAR FACTOR COMPUTING

ABSTRACT OF DISSERTATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry at the University of Kentucky

By
Bin Dai

Lexington, Kentucky

Director: Dr. Robert A. Lodder, Professor of Chemistry

Lexington, Kentucky

2007

ABSTRACT OF DISSERTATION

SIMULATIONS-GUIDED DESIGN OF PROCESS ANALYTICAL SENSOR USING MOLECULAR FACTOR COMPUTING

Many areas of science now generate huge volumes of data that present visualization, modeling, and interpretation challenges. Methods for effectively representing the original data in a reduced coordinate space are therefore receiving much attention. The purpose of this research is to test the hypothesis that molecular computing of vectors for transformation matrices enables spectra to be represented in any arbitrary coordinate system. New coordinate systems are selected to reduce the dimensionality of the spectral hyperspace and simplify the mechanical/electrical/computational construction of a spectrometer.

A novel integrated sensing and processing system, termed “Molecular Factor Computing (MFC)” based near infrared (NIR) spectrometer, is proposed in this dissertation. In an MFC -based NIR spectrometer, spectral features are encoded by the transmission spectrum of MFC filters which effectively compute the calibration function or the discriminant functions by weighing the signals received from a broad wavelength band. Compared with the conventional spectrometers, the novel NIR analyzer proposed in this work is orders of magnitude faster and more rugged than traditional spectroscopy instruments without sacrificing the accuracy that makes it an ideal analytical tool for process analysis.

Two different MFC filter-generating algorithms are developed and tested for searching a near-infrared spectral library to select molecular filters for MFC-based spectroscopy. One using genetic algorithms coupled with predictive modeling methods to select MFC filters from a spectral library for quantitative prediction is firstly described. The second filter-generating algorithm designed to select MFC filters for qualitative classification purpose is then presented.

The concept of molecular factor computing (MFC)-based predictive spectroscopy is demonstrated with quantitative analysis of ethanol-in-water mixtures in a MFC-based prototype instrument.

KEYWORDS: chemometrics, process analytical technology (PAT), near-infrared (NIR), genetic algorithm (GA), integrated sensing and processing (ISP).

Bin Dai

March 5, 2006

SIMULATIONS-GUIDED DESIGN OF PROCESS ANALYTICAL SENSOR USING MOLECULAR FACTOR COMPUTING

By

Bin Dai

Robert A. Lodder

Director of Dissertation

Robert B. Grossman

Director of Graduate Studies

March 5, 2007

RULES FOR THE USE OF DISSERTATIONS

Unpublished dissertations submitted for the Doctor's degree and deposited in the University of Kentucky Library are as a rule open for inspection, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but quotations or summaries of parts may be published only with the permission of the author, and with the usual scholarly acknowledgments.

Extensive copying or publication of the dissertation in whole or in part also requires the consent of the Dean of the Graduate School of the University of Kentucky.

A library that borrows this dissertation for use by its patrons is expected to secure the signature of each user.

Name Date

DISSERTATION

Bin Dai

The Graduate School

University of Kentucky

2007

SIMULATIONS-GUIDED DESIGN OF PROCESS ANALYTICAL SENSOR USING MOLECULAR FACTOR COMPUTING

DISSERTATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry at the University of Kentucky

By
Bin Dai

Lexington, Kentucky

Director: Dr. Robert A. Lodder, Professor of Chemistry

Lexington, Kentucky

2007

For Xuebei (Kathy) Shi

Acknowledgements

I must first thank my advisor, Rob, for all his support and patient guidance over the course of this research.

I also thank other members of my committee, Sylvia Daunert, Leonidas G. Bachas and Craig C. Douglas, for their help and time.

I would like to thank all fellow members of the Analytical Spectroscopy Research Group (ASRG) at the University of Kentucky, whose support and cooperation were essential for this research work. Special thanks to my best friends and colleagues, Aaron Urbas and Joseph Medendorp, for their invaluable support and friendship during these years.

To my family; I owe everything to them for making me what I am today.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

ANN	Artificial neural network
AOTF	Acousto-optical tunable filter
BEST	Bootstrap error-adjusted single sample technique
CI	Chemical imaging
CV	Canonical variables
DARPA	Defense Advance Research Projects Agency
DMA	Digital micromirror array
DoE	Design of experiment
FDA	Food and Drug Administration
FFT	Fast Fourier transform
FTIR	Fourier transform infrared
FT-NIR	Fourier transform near infrared
GA	Genetic algorithm
HICI	Hyperspectral integrated computational imaging
ICI	Integrated computational imaging
IR	Infrared
ISP	Integrated sensing and processing
KNN	K nearest neighbor
LDA	Linear discriminant analysis
LOOCV	Leave-one-out cross validation
MFC	Molecular factor computing
MOE	Multivariate optical element
MLR	Multiple linear regression
MSC	Multiplicative signal correction
MIR	Mid infrared
MS	Mass spectrometry
NIR	Near infrared
NMR	Nuclear magnetic resonance
PAT	Process analytical technology
PCA	Principal component analysis
PCR	Principal component regression
PLS	Projection to latent structure
PLSDA	Partial least squares discriminate analysis
RMSEC	Root mean square error of calibration
RMSECV	Root mean square error of cross validation
RMSEP	Root mean square error of prediction
RSD	Relative standard deviation
SD	Standard deviation

SEC	Standard error of calibration
SEE	Standard error of estimation
SEP	Standard error of prediction
S/N	Signal-to-noise ratio

List of Files

File name	size
BinDai_dissertation.pdf	1.48Mb

Preface

Motivation

Process analytical technology (PAT) has existed for several decades in different areas. Recently promoted by the US Food and Drug Administration (US FDA) through the new regulations, PAT is now gaining its popularity in the pharmaceutical industry. Modern process analyzers and multivariate data processing tools are critical parts for the successful implementation of PAT in industry.

Instrumentation in analytical chemistry has dramatically changed during the last few decades. Development of new sensors and parallelizing of sensors allow the acquisition of a large amount of information with extraordinary speed. Such an advance in analytical instrumentation enables deeper understanding of complex systems. However, the flood of data generated by the new instruments threatens to cause a bottleneck in process analysis, drug discovery and development, and other research areas. Data flow rate therefore brings a new challenge to the field of data analysis.

The need for developing a stable process analyzer that enables frequent or continuous monitoring of the manufacturing process with integrated data processing capability is clear. Spectroscopy-based analytical techniques are the most important and commonly used techniques in PAT. Examples include absorption spectroscopy, polarimetry, diffuse reflectance spectroscopy and Raman spectroscopy. Among them, near infrared (NIR) spectroscopy and imaging is the most widely deployed technology. Therefore the design of NIR-based process sensors is the core of this dissertation work. The background of NIR spectroscopy and imaging is given in Chapter 1.

Conventional NIR imaging spectrometers used to acquire full-spectrum information suffer two major shortages in PAT applications. First, too much information is collected by full-spectrum imaging spectrometers, creating a huge computing demand, a need for large data storage, and high maintenance cost. Second, most full-spectrum spectrometers are not rugged enough for PAT applications due to a more 'hostile' environment in the manufacturing process. Multiple interference filters-based NIR analyzers are still the predominant analyzers installed in the field for process monitoring because of their simplicity and stability. However, although frequently used in PAT, interference filter-based analyzers have an optical throughput limit. In the interference filter-based analyzer, only a narrow band of light (typically 10 nm FWHM) passes through the interference filter. Under real-life circumstances that might include fiber-optic interfaces, highly absorbing or highly scattering samples, and occasionally needed for large pathlength, PAT applications can become detector noise-limited due to limited optical throughput.

Dissertation Objectives

Clear understanding of the conceptual and technical challenge is crucial for success in developing an innovative MFC PAT analyzer. The most important challenge is identified as creating a reliable method to generate the molecular filters needed for different applications. To fulfill this need, two novel MFC-filter generating algorithms for searching a near-infrared spectral library to select molecular filters for MFC-based spectroscopy are proposed. One algorithm is designed for quantitative applications while the other algorithm is designed for qualitative applications. In these algorithms, transmission spectra of a library of reference compounds are multiplied by transmission spectra of training samples to obtain simulated library scores. The algorithm treats the simulated library scores as variables and searches the library-scores space to select a few variables that represent the sample spectra in a new coordinate system. The MFC filter-generating algorithms are tested on both simulated and actual experimental data. Methods that allow objective assessments of the algorithm are introduced. Therefore, upon reading this dissertation, the reader should have a deeper understanding of the proposed MFC concept and the methods to solve the challenge they present.

Subsequently, a novel MFC-based prototype instrument is developed and tested via quantitative analysis of ethanol-in-water mixtures to demonstrate the concept of molecular factor computing (MFC)-based predictive spectroscopy.

Finally, this dissertation intends to convince the reader that the new MFC approach has many advantages over conventional NIR spectrometers in PAT applications. These advantages include significantly reducing the computational demand (the integrated sensing and processing, or ISP, advantage), shorter data collection and analysis time with higher signal-to-noise ratio (S/N) (especially for imaging spectrometry, through the Fellgett advantage), higher optical throughput (the Jacquinot advantage), and more rugged instrumentation with a considerably lower cost. In real PAT applications, an MFC-based process analytical sensor should be orders of magnitude faster and more rugged than traditional spectroscopic instruments without sacrificing the analytical accuracy.

Content Summary

The outline of this dissertation work starts by introducing PAT with a focus on NIR spectroscopy and chemometrics in chapter 1. The new FDA regulations for promoting the implementation of PAT in pharmaceutical industry are introduced, followed by NIRS technology and its applications in PAT. The chapter ends with an overview of chemometrics in PAT.

In the following chapter, the concept of molecular factor computing-based spectroscopy and imaging is introduced and explained. The initial prototype instrument and potential applications are proposed.

In chapter 3, a novel genetic algorithm-based MFC filter generation algorithm is presented. This algorithm is mainly designed for quantitative applications. It uses genetic algorithms coupled with predictive modeling methods to select MFC filters from a spectral library for quantitative prediction. For the qualitative applications, where pattern recognition-based sample classification is sought, a new algorithm for selecting desired MFC filters by searching a reference spectral library is proposed and tested in chapter 4.

Finally, in chapter 5 the construction and design of the prototype analyzer is described. A test of the instrument for concept demonstration purposes is introduced, and the results are discussed. The dissertation ends with the conclusions from this work and some suggestions for future research.

Chapter One – Near Infrared Spectroscopy and Chemometrics in Process Analytical Technology

1.1 Background

Conventional pharmaceutical manufacture and quality control are normally completed by the batch processing and off-line laboratory testing that is carried out on samples collected from the manufacturing line. This approach has been successfully employed in pharmaceutical industry over decades in pharmaceutical industry and has been supplying quality pharmaceutical products for the public. However, recent developments in process analytical technology and multivariate modeling methods provide a significant opportunity to improve pharmaceutical manufacture and product quality assurance. Therefore, the pharmaceutical industry has arrived at a crossroad; [1] one path goes toward a desired state by implementing the new process analytical technologies and the other path maintains the current state. Unfortunately, the path that leads to the desired state by introducing the innovative system into manufacturing was unfavorable due to the rigidly interpreted cGMP(current Good Manufacture Practice) regulations. As stated in an article[2] published on *Wall Street Journal* in Sep. 2003, "the pharmaceutical industry has a little secret: even as it invents futuristic new drugs, its manufacturing techniques lag far behind those of potato-chip and laundry soap makers". Such pharmaceutical industrial reluctance to innovation is undesirable from a public heath perspective. Innovation and cutting-edge scientific and engineering knowledge are needed in pharmaceutical manufacturing to provide more innovative medicine more safely and cheaply. Regulatory policies are therefore needed to permit industry response to the opportunity. [3]

In September of 2004, the US FDA released the guidance for industry, PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance.[4] This guidance is designed to facilitate innovation in process development and quality assurance. PAT will help in better design, monitor and control of pharmaceutical manufacturing processes by integrating multivariate modeling, sensor design and process optimization with the goal of ensuring final product quality. [5]

The FDA PAT initiative and other regulatory initiatives[3] are the key drivers for the current research and manufacturer interest. These initiatives also promote the implementation of NIR and chemometrics in pharmaceutical industry.[6, 7]

1.2 Process Analytical Technology

The field of PAT has undergone huge growth over the last decade. Its growth in the chemical and pharmaceutical industry is driven by the need for productivity improvement and better quality assurance as well as new industry regulations as mentioned previously.

Most significant differences between PAT and laboratory analysis refer to the approach of analysis in the process line versus remote off-line analyses in a controlled laboratory. The manufacturing advantages of PAT over laboratory analysis are obvious, and they can be summarized as following four characters:

- The PAT enables one to gain a deeper understanding of the real-time manufacture process by direct measurement of important process parameters.
- PAT reduces the analytical time, thus making real-time process control possible.
- The PAT eliminates the sample handing and transportation, and thereby gains operational safety and reduces the operator error.
- The PAT offers the low –cost quality assurance by fully automating the process analyzers.

The PAT is defined in the US FDA guidance[4] as:

"*A system for designing, analyzing and controlling manufacturing through timely measurements (i.e. during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality*".

According to the definition above, the term analytical in PAT is not limited to chemical analysis; instead, the term is defined broadly to include physical, biological and mathematical analysis. Another important phrase in the PAT definition is the "timely measurements". FDA encourages the pharmaceutical manufacturing community to implement in-line or on-line process analyzers and use the information acquired from these analyzers, which has many significant advantages over the traditional off-line quality test performed in the pharmaceutical laboratory today. The FDA has also defined PAT tools that include many current technologies and methods that could provide effective scientific ways for improving pharmaceutical development, manufacture and quality assurance. In the PAT frame work,[4] these tools can be categorized as:

- Multivariate data acquisition and analysis tools;
- Modern process analyzers or process analytical chemistry tools;
- Process and endpoint monitoring and control tools;
- Continuous improvement and knowledge management tools;

This dissertation work contributes to the first three of the categories listed above, with a focus on NIR process analyzer design and chemometrics.

1.3 Near Infrared Spectroscopy and Imaging in PAT

1.3.1 Theory of near infrared spectroscopy

The light absorbance in NIR region (1000nm-2500nm) is mostly due to the overtones and combinations of the fundamental vibration bands from mid-infrared (MIR) region.[8] Vibrational spectroscopy is based on the concept that molecular bonds vibrate at certain frequencies. The vibration of the atoms in a molecule is confined with by a potential energy well. When light at energy is absorbed, the molecular vibrators are excited to a higher energy level. The fundamental vibrational frequency of a molecule with two atoms connected by a bond can be calculated by assuming that the vibration obeys the diatomic harmonic oscillator model (Hooke's law).

$$\upsilon = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \qquad \textbf{1.1}$$

where υ is the vibrational frequency in Hz , k is the bonding force constant, and μ is the reduced mass (m_1m_2/m_1+m_2).

Using the harmonic oscillator as an example, the energy levels of a diatomic molecule can be expressed as:

$$E_n = (n+\frac{1}{2})h\frac{1}{2\pi}\sqrt{\frac{k}{\mu}} = (n+\frac{1}{2})\mathrm{H}\omega \qquad (n=0,1,2,3\ldots.) \qquad \mathbf{1.2}$$

where $H=h/2\pi$, h is planck's constant, and $\omega=\sqrt{\frac{k}{\mu}}$.

Based on the harmonic oscillator model, the energy gap between adjacent energy levels is exactly the same:

$$\Delta E = E_{n+1} - E_n = H\omega \qquad \mathbf{1.3}$$

and ω is called the fundamental frequency.

As mentioned earlier, the harmonic oscillator model is only an illustration. In reality, the anharmonic oscillation and higher terms need to be considered. In the realistic anharmonic model, the rigorous selection rule △n= ± 1 is relaxed, therefore, the weaker absorptions can occur with △n= ± 2 (first overtone) and even with △n= ± 3 (second overtone), etc.

In polyatomic molecules, although weak, the transition to excited states involving the two vibrational modes simultaneously does occur. Such a transition is called a combination band.

Most NIR absorbance bands are overtone and combination bands of a bond involving hydrogen

(C-H, N-H, O-H, S-H). These functional groups have their fundamental frequencies in the mid-IR. Because the absorption of overtone and combination bands are much weaker compared with the fundamental IR bands, an NIR light beam can penetrate deeper into a sample, and thus make the measurement of physically thick condensed phase samples possible. The weak absorption of NIR also virtually eliminates the need for sample preparation, and fortunately so, because most PAT applications required direct measurement of the process stream in situ.

The mainstream NIR spectrometers in current application include scanning grating monochromators, grating polychromator photodiode array spectrometers, acousto-optical tunable filter (AOTF) based spectrometers, Fourier transform spectrometers and discrete interference filter based spectrometers.

1.3.2 Application of near infrared in process analytical technology

There are many excellent spectroscopic techniques other than near-infrared available to the process analyst. These techniques include mid-range infrared (MIR),[9] ultraviolet and visible (UV/Vis),[10] fluorescence,[11] nuclear magnetic resonance (NMR),[12, 13] mass spectrometry (MS),[14] and, increasingly, Raman spectroscopy and imaging.[15] Although these techniques are applicable for some particular tasks, it almost seems that NIR is now *de rigueur* for pharmaceutical PAT analyses. What make the NIRS the number one technique in PAT are the unique features of the NIR spectrum (physics) and the development of chemometrics (mathematics & computers). The physics of NIRS has been introduced in this section. The chemometrics will be covered in next section.

Near-infrared spectroscopy is now a well established analytical technique and has been routinely used in the chemical and pharmaceutical industry.[16-22] The applications of NIRS range in complexity from simple moisture content determination to ambitious PAT applications that include determining the potency of the active ingredient or concentration of the major compound in the process stream. The advantages of speed of analysis and nondestructiveness, in addition to the flexibility of the sampling interface (transmission, diffusion reflectance, transreflectance), make NIR spectroscopy an ideal technology for at-line, on-line or in-line process analysis.

NIR chemical imaging (CI) also receives much attention these days.[23, 24] The additional spatial information provided by imaging offers a better understanding of the manufacturing process especially for that of the complex mixtures. The imaging technique not only shows the chemical composition within a mixture, but the spatial relationships of the individual materials. Typical applications of NIR chemical imaging include characterization of solid dosage forms and mixing endpoint determination.

The successful implementation of NIRS for chemical or pharmaceutical process analysis depends upon the NIRS instrumentation, chemometric modeling and sampling interfaces. Although there are a number of different NIR analyzer technologies available in the current market, the majority of NIR analyzers implemented in the manufacture process stream for process monitoring purpose are based on discrete bandpass interference filter.[25] Due to the more ‘hostile’ environment in manufacturing process applications, rugged instruments are required for analyzer automation, stability and low-cost. Analyzer automation, stability and low-cost are exactly the motivation for this dissertation work and are drivers for the molecular factor computing filter-based NIR spectrometer.

1.4 Chemometrics in PAT

1.4.1 Overview of the chemometrics

There are a number of different stories about how the field of chemometrics came into being. Most people cite that the term chemometrics originated with Swedish chemist Svante Wold in the early 1970s.[26] The definition of chemometrics by the International Chemometrics Society (founded in 1974) is: “Chemometrics is the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical and statistical methods”.

Despise its successes in other fields, chemometrics was not introduced to the field of analytical chemistry until the late 1970s. [27] The first appearance of modern chemometrics in analytical chemistry actually began in the agriculture and food industry, as pioneered by Norris at the US Department of Agriculture (USDA).[28, 29] The work in USDA demonstrated that multivariate calibration could be successfully implemented for rapid nondestructive measurements of agricultural materials based on multi-wavelength NIR spectra. Historically, chemometrics has also been closely linked to near-infrared process analyzers. The pioneering work [30, 31]was carried out by the Center for Process Analytical Chemistry, led by Prof. Kowalski in Seattle, Washington, in the 1980s.

As pointed out by Charles Miller,[8, 32, 33] there are three key elements that are consistently used in nearly every application of chemometrics: empirical modeling, multivariate modeling and chemical data. Empirical modeling indicates the data-driven nature of chemometric modeling. Multivariate modeling that describes relationships between two groups of multivariate data is the essential part of the scientific field of chemometrics. The two most important functions of the multivariate modeling are: (1) Instrument calibration. Calibration includes multivariate calibration model construction for the multivariate data generated by an analytical instrument and using a multivariate model to guide better instrument design (the second part is the motivation of this dissertation work). (2) Information extraction. Multivariate analytical tools are used to explore complex data and gain better understanding of the chemistry and process.

1.4.2 Design of Experimental (DoE)

Depending on the experimental objectives, there are a wide variety of experimental designs.[34, 35] Calibration samples that sufficiently cover the type and concentration range of the analytes of interest and of other interferences are always required. Careful experimental design is very useful for effectively optimizing the measurement space spanned by the independent variables with the most representative training samples.

After the experimental objective is specified, the next step in experimental design is determining how many appropriate variables need to be included and how many levels of each variable need to be set. When full factorial designs are used, where the concentration of each analyte is varied independently, the total number of samples needed in the experiment increase rapidly when the number of levels used in each variable increases.

Once the number of the level for each design variable is determined, the next step is to choose an experimental design type. There exist several different design types, which include full-factorial design, central composite designs, [36] D-optimal design[37] and Box_Behnken design. In this work, full factorial design[38] with equidistant levels is used for synthesizing the calibration and validation data set. The details of the dataset generation will be covered in Chapter 3.

1.4.3 Data preprocessing

The purpose of data preprocessing (pretreatment) is to improve the qualitative or quantitative model by systematically modifying the raw data. The typical benefits for employing the data preprocessing include alleviating the baseline variation, noise removal and spectra smoothing. In this work, several data preprocessing methods are used to improve the calibration and classification model based on NIR data. These methods are briefly introduced in the following paragraphs.

1.4.3.1 Autoscaling

Autoscaling is a z-scoring preprocessing procedure that combines the mean centering and variance scaling. Autoscaled data have the characteristic that each of the variables has zero mean and a standard deviation of one. The following equation shows how autoscaling is done:

$$\hat{x}_{i,j} = \frac{x_{i,j} - \bar{x}_j}{\sigma_{x_j}} \qquad \textbf{1.5}$$

where $\hat{x}_{i,j}$ is the autoscaled data, $x_{i,j}$ is the unprocessed data, $\bar{x}_j$ is the mean of j^{th} column data and σ_{x_j} is the standard deviation of j^{th} column data.

1.4.3.2 Derivative

The derivative is useful for removing the baseline and slope variations. The second derivative is commonly used in NIR diffuse reflectance spectroscopy.[39] The mathematical expression of the second derivative is given below:

$$f''(x) = \frac{f'(x+h) - f'(x)}{h} \qquad \textbf{1.6}$$

1.4.3.3 Multiplicative scatter correction (MSC)

In NIR diffuse reflectance and transmission spectroscopy, multiplicative scattering variations between samples exist due to the effective pathlength difference. Such variation cannot be removed by other preprocessing methods (such as derivative or autoscaling) without distorting the appearance of the spectra. Multiplicative scatter correction (MSC) was developed to reduce the effect of multiplicative scattering.[40, 41]

The MSC model is given by the following equation:

$$x_j = a + b\bar{x}_j + \varepsilon \qquad \textbf{1.7}$$

where a and b is the constant for all j wavelengths in the sample. $\bar{x}_j$ is the mean of j wavelengths. The factors *a* and *b* are estimated by regressing each spectrum onto the mean spectrum. Once the factors a and b are determined, the corrected spectrum can be expressed as:

$$X_{MSC} = \frac{(X - b \cdot 1_N)}{a} \qquad \textbf{1.8}$$

1.4.4 Quantitative modeling

In the field of chemometrics, the quantitative modeling refers to building a function that relates multivariate instrument response with the properties or concentrations of interest. There are many possible methods to build quantitative model. The most frequently used methods include multiple linear regression (MLR), principal component regression (PCR), projection to latent structure regression (PLSR) and artificial neural networks (ANN). The chosen method is usually the one that is easiest to perform, offers the best calibration performance, and is simplest to explain.

1.4.4.1 Multiple linear regression

A multiple linear regression (MLR) model can be setup when the assumption of a linear relationship between the one or more properties of interest (Y variables) and the instrument responses (X independent variables) hold. When MLR is used to provide such a predictive model, it is usually referred to as an inverse least squares regression. In an MLR model, the Y variable (concentration, etc) is expressed as a function of x variables (absorbance, etc.):

$$Y = XB + f \qquad \textbf{1.9}$$

The inverse MLR regression coefficients $\hat{B}$ are estimated using the least squares method:

$$\hat{B} = (X^t X)^{-1} X^t Y \qquad \textbf{1.10}$$

Once the regression coefficients are determined, the MLR model is set. The properties of interest of the unknown new samples can be predicted based on the measured spectra.

$$\hat{Y}_{New} = X_{New}\hat{B} \qquad \textbf{1.11}$$

As can be seen from Equation 1.10, to estimate the correlation coefficients $\hat{B}$, the variance-covariance matrix needs to be inverted. Matrix inversion requires that the number of sample is not less than the number of independent variables.

For applications with a small number of independent variables (e.g., spectral data from an interference filter-base spectrometer or a molecular factor computing based spectrometer), MLR is the most applicable calibration method[42]. MLR also shows great modeling power when coupled with variable selection.

The biggest disadvantage of MLR is its instability with collinearity in the data. Collinearity in a spectral data matrix X makes the inverse of variance-covariance matrix an ill-conditioned problem. As a result, MLR often leads to an unstable estimation of correlation coefficients, and the unstable correlation coefficients often cause large prediction error for new samples.

To overcome the shortcomings of MLR, several alternative calibration methods have been developed. Among these methods, PCR and PLS are the two most frequently used alternatives to MLR.

1.4.4.2 Principal component regression

The principal component regression (PCR)[43] is an extension of principal component analysis (PCA).[44] In PCR, instead of being regressed on all measured original variables, the y variable is regressed on the selected principal component scores. So the first step of PCR is PCA, where the spectral data X are factored by using singular value decomposition (SVD).

$$X = TP^t + E \qquad \textbf{1.12}$$

where *T* is the score matrix, and *P* is the loading matrix and *E* is the residual.

After PCA, the number of significant PCs to retain in the model is determined by cross -validation, a procedure where the data are repeatedly split into training data and testing data. Once the number of significant PCs is determined, the y variable is regressed on the selected PC scores.

$$y = \overline{T}b + f \qquad \textbf{1.13}$$

where y denotes the known values of the property of interest, $\overline{T}$ is the selected principal component scores matrix, b is the regression coefficients, and f is the residual. The estimation of b can be obtained by the following equation:

$$\hat{b} = (\overline{T}^t\overline{T})^{-1}\overline{T}^t y \qquad \textbf{1.14}$$

Once the regression coefficients are determined, the property of interest of new unknown samples can be predicted based on the measured spectral data by:

$$\hat{y}_{New} = X_{New}\overline{P}\hat{b} \qquad \textbf{1.15}$$

The major advantage of PCR over MLR is that it overcomes the collinearity problem and thus avoids the potential model computing complications. The PCR also eliminates the condition that requires the number of sample can not be less than the number of independent variables (although the number of sample can not be less than the number of model PCs). A disadvantage of PCR is that it aims to maximize the description of variation of X-data but not to optimize the prediction. This limitation of PCR leads to PLS.

1.4.4.3 Projection to latent structure regression

Projection to latent structure regression (PLS),[45, 46] which was originally developed in the field of econometrics, gained huge success in process analytical chemistry due to its speed, robustness and clear model interpretation. Like PCR, PLS performs the regression on the Y variables (i.e., concentrations) with selected latent variables rather than the original independent variables X (i.e., absorbance). Like PCR, in PLS, the selected latent variables define a new coordination system with a smaller dimensionality. However, unlike PCR, which emphasizes maximization of the variance of the independent variables X, the latent variables in PLS are constructed to explain the most variance in both X and Y. In another word, PLS determines each latent variable to simultaneously optimize the variance in X and the correlation with Y variables. The iterative algorithms such as NIPALS and SIMPLS can be found in most chemometric textbooks[26, 47, 48] and the primary literatures.[49, 50]

There are some advantages of PLS over PCR. First, PLS generally achieves the comparable model performance (i.e., prediction accuracy) with fewer latent variables, and therefore results in a more parsimonious calibration model. Second, the PLS model is sometimes more stable over time due to its simplicity compared with PCR, and fewer latent variables ease qualitative interpretation.

1.4.5 Qualitative modeling

It can be very difficult to obtain a satisfactory quantitative model when the inherent quantitative relationship between concentration and spectra is nonlinear or very complicated. A qualitative model[51] can be very useful in PAT in circumstances where the quantitative models are not available or only qualitative information is needed.

Because this dissertation focuses on chemometrics-guided instrument design, the qualitative modeling introduced here focuses on classification methods that are in the category of *supervised pattern recognition*. The counterpart, *unsupervised pattern recognition* (e.g., hierarchical cluster analysis, or HCA), which is widely used for explorative data analysis, is not included in this work.

The commonly encountered supervised classification methods[52] include K-nearest neighbor (KNN), PLS discriminant analysis (PLSDA), linear discriminant analysis (LDA)[53] and soft independent modeling of class analogies (SIMCA).[54] LDA is used in this work for designing a

MFC filter selection algorithm (chapter 4) and is therefore is introduced here.

In LDA, the classification is done in a new coordination system with lower dimensionality. The new coordination system is defined by a unique set of vectors called canonical variables (CVs) or linear discriminants (LDs), which are the linear combination of original variables. The canonical variables are determined such that the ratio of between-class dispersion to within-class dispersion in the training dataset is maximized.

Once the relevant CVs are selected and the new coordinate system is set, the LDA modeling for the training set is constructed. To predict the class a new unknown sample belongs to, the new sample is projected into the new coordination system and distances of the unknown sample from each of the class centers are calculated. Then the sample is assigned to a certain class based on the distances and within-class standard deviation.

Because the CVs are constructed for a separation purpose, they provide a more relevant coordinate system for discriminating among groups. Often, a small number of CVs are enough to provide sufficient discrimination. Like most other supervised classification methods, the LDA is prone to overfitting when only a small training dataset is used or too many CVs are employed in the model.

1.4.6 Variable selection

The essential part of molecular factor computing-based spectroscopy is the molecular filter selection. The selection procedure is done by searching a reference spectra library and designing filters based on combinations of library entries. The actual algorithm to perform such a selection procedure is adopted from the variable selection algorithms that are widely used in spectra feature selection. Instead of using the original X wavelength variables, MFC variable selection chooses the library scores that are generated through the dot product of the library spectra and training sample spectra.

1.4.6.1 Brute force variable selection

The most obvious method of variable selection is to examine all possible combinations (APC) of variables. Each time, the calibration modeling is carried out on data with the selected variables, then the model is tested on a validation dataset and the prediction error is computed. The model that results with the smallest prediction error will be chosen out of all the possible combinations. Because the number of combination will increase dramatically as the number of original variables increases, this method is called brute force variable selection. The number of possible combinations N can be calculated using the following equation: [55]

$$N = 2^{n_{Total}} - 1 \qquad \textbf{1.16}$$

where the n_{Total} is the number of original variable.

The APC method is only practical where a very small number of variables exist. In this work the dataset has around 2000 variables, rendering the APC method useless.

1.4.6.2 Variable selection by genetic algorithm

A genetic algorithm (GA)[56-58] is a nature-inspired searching method and has been proven to be especially useful for large search spaces. The GA base variable selection algorithm operates as follows: First, the complete X data (spectra) are fed as input to the GA, from which an initial set of variable subset candidates are randomly selected. Second, for each variable subset, a regression model is built and the fitness function then evaluated. The fitness function performed statistical analysis and returns a value (in this work, a prediction error) as a figure of merit of the current regression model. Third, a series of genetic operations are applied to generate a new set of variable subset candidates that are likely to lead to a better model. The GA search is directed by the fitness selection and continues toward the more optimal direction until the terminating conditions are met.

The major components in a genetic algorithm include a genetic representation of a solution to the problem, a way to create an initial population of solutions, an evaluation function rating solutions in terms of their fitness, and genetic operators that alter the genetic composition of the offspring.

1.4.6.2 Stepwise variable selection

Stepwise variable selection[59] [60]is often used where a large number of independent variables are available but only few variables are desired to be included in the discriminant model for groups separation. In stepwise selection, the first step defines an initial model by either the provided variables or a randomly chosen variable. In the second step, the variable among the reminding variables that combines with the first selected variable to improve the model the most is chosen. After each new variable has been selected, the variables previously selected are reexamined to see if each still contributes a significant amount of prediction ability to the model. The variable will be eliminated if the corresponding variable elimination improves the model performance. The stepwise variable selection process stops when a satisfactory classification result has been found or when the model can no longer be improved by stepwise selection.

Chapter Two – An Introduction to Molecular Factor Computing Based Spectroscopy and Integrated Computational Imaging

2.1 Introduction

Modern hyperspectral imaging is able to collect unprecedented amounts of information with extraordinary speed. Remote sensing with combinations of Synthetic Aperture Radar (SAR), IR, UV/visible and similar technologies is increasingly prevalent, adding to the computational burden. Reducing these volumes of data from physical fields to high-level, useful information is difficult. Part of this reduction is now being done optically. In the past, optics has served mainly to render the universe more easily visible to human observers. Now, computers are increasingly employed to make sense of the visual world in ways that humans cannot. With a new generation of optics, scientists and engineers are recasting visual scenes for interpretation exclusively by computers. To the human eye, these pictures appear distorted at best, or at worst look like visual noise, without discernable meaning. Nevertheless, to computers, such data are worth more than a thousand words. Optimizing complete vision-and-action systems for computers lies at the core of *integrated computational imaging* (ICI).[61] Computers are well-established manipulators of digitized images, and image-processing programs do it routinely on desktop machines. However, what is new is the strategy of ICI - *processing image information as it is sensed* to make it better suited for the "computer mind."

Both *spatial* and *spectral* features of samples can be encoded in ICI. When spectral images are simultaneously obtained and encoded at many different wavelengths, the process is termed hyperspectral integrated computational imaging (HICI). Molecular absorption filters can be used as mathematical factors in spectral encoding to create a factor analytic optical calibration in a high-throughput spectrometer. In this system, the molecules in the filter effectively compute the calibration function by weighing the signals received at each wavelength over a broad wavelength range. This chapter describes the principle of spectrometer designs that use molecular-computing to replace traditional principal component analysis in a computer with molecular filters tailored to produce factor scores at the detector.

A simple analogy suggests the advantage of doing as much of the processing as possible in the sensing transducer itself. Imagine two gunfighters on Main Street at sundown in the old “wild west.” The first gunfighter’s hand and revolver are controlled by his brain using image information transmitted from the retinas of his eyes. Impulses must travel from his eye to his brain, and then from his brain to his hand. The second gunfighter’s hand and revolver are controlled directly by the retinas of his eyes using nerve impulses that travel only one path instead of two. The second gunfighter’s weapon is likely to always be slightly ahead of the first's. Moreover, the second gunfighter’s brain is free to consider other, more effective strategies.

2.2 Spectral Feature Encoding with Molecular Factor Computing: *Principal Component Analysis (PCA) methods.*

Many different variations contribute to the spectrum of actual samples, typically more variations than contribute to spectra of synthetic standards prepared in the laboratory. Field samples contain variations from instrument differences such as detector noise, differences in the constituents in the sample mixtures, interactions between constituents, shifting environmental conditions that influence the spectral baseline and overall absorbance, and differences in sample preparation and presentation.[62] In spite of these compound changes taking place, there should be a limited number of independent variations taking place in the spectral data. With a bit of luck, the largest variations in the calibration set will be the changes in the spectra due to the property you wish to measure, e.g., the different concentrations of the constituents of the mixtures. The strategy in traditional principal component analysis and in molecular computing of factors is to focus on the spectral variations in the calibration set. It is possible to compute a set of "variation spectra" that correspond to the differences in the absorbances at all the wavelengths in the spectra. In molecular computing, the molecular filters are selected to maximize the integrated differences in the variation-spectra within a certain bandpass. In PCA, the so-called variation-spectra can be used in place of the raw spectral data for constructing the calibration model. Usually there are fewer common variations in these PC or MFC spectra than the total number of calibration spectra, and so the number of computations required for the calibration equations is reduced, too.

If properly constructed, the "variation spectra" can be used to reconstruct the original spectrum of a certain sample by multiplying each variation-spectrum by a unique constant scaling factor and summing the results until the new spectrum agrees with the unknown spectrum. In principle, reconstruction can be done in either traditional PCA or molecular factor computing (MFC). Each spectrum in the calibration set must have a distinct series of scaling constants for each variation because the concentrations of the constituents are dissimilar. For this reason, the fraction of each "spectrum" that must be added to reconstruct the unknown spectral data is associated with the concentration of the constituents.

In PCA, the spectra of the variations are termed eigenvectors, or loadings, spectral loadings, loading vectors, or principal components or factors, based on the means used to compute the spectra. Regardless of the means, a simple transformation will convert one into another (eigenvectors into loadings, for example). The scaling constants employed to reconstruct the individual spectra are commonly called scores. Ordinary spectroscopy and PCA chemometrics records signals with a narrow bandpass at each wavelength and then weighs the signals *a* at each wavelength λ with a coefficient *f* in a computer.

$$Score = f_1 a_{\lambda 1} + f_2 a_{\lambda 2} + f_3 a_{\lambda 3} + \ldots \qquad \textbf{2.1}$$

However, it is also possible to weigh each wavelength in a spectrum optically using the absorbance spectra of “filter” molecules. The “scores” can then simply be read by an A/D as the voltage from a detector unit by integrating the total light through the sample and filter over a broad wavelength band. In this case, while the scores are not perfectly orthogonal, they are often close enough to permit chemical analyses to be performed. For those who question the lack of orthogonality, it is worth remembering that the principal component scores calculated on spectra in a traditional computer are also not perfectly orthogonal as soon as one new sample is added to the calibration set. Yet even in these cases, chemical analyses can usually be successfully performed.

Because the computed eigenvectors were generated from the original calibration spectra, the eigenvectors must relate by some means to the concentrations of the constituents that comprise the samples. The identical loading vectors can be used to predict "unknown" samples; therefore the only difference between the spectra of samples with different constituent concentrations is the fraction of each loading vector added (the scores). Using MFC, the only difference between the spectra of samples with different constituent concentrations is the integrated detector response in the wavelength bandpass, which is equivalent to a factor score as described earlier. Varying the bandpass for different types of samples can help to spread the scores, effectively improving S/N.

By PCA or MFC, the scores are exclusive to each independent principal component (or factor) and training-set spectrum, and can be used in lieu of absorbance values in either of the typical modeling procedures (classical least-squares, CLS, or inverse least-squares, ILS). Because the description of mixture spectra is condensed from many wavelengths to a small number of scores (illustrated in Figure 2.1), spectroscopists usually apply the ILS manifestation of Beer's Law in computing concentrations because of its ability to calculate concentrations in the midst of interfering species. Of course, ILS preserves the averaging effect of CLS chemometrically by building a large number of wavelengths from the spectrum (indeed, up to the entire spectrum) into the model when calculating the eigenvectors. As a result, factor models really combine the best features of both the CLS and ILS methods simultaneously in the one series of computations. This synergy is the primary cause for the generally better statistical performance of factor models over classical models in terms of robustness and accuracy

The “secret” in these modeling methods is in the technique by which the eigenvectors are obtained. These models build the concentration predictions upon changes in the data, not absolute absorbance values (absolute absorbance values are used in the classical models). To calculate a PCA model, the spectral data must vary in some fashion. The simplest way to achieve variations in the spectra is to alter the concentrations of the constituents. When making these alterations, problems can arise with collinearity, just as can occur with ILS modeling. Collinearity is easy to understand: If the concentrations of two constituents in the calibration samples are always present in the same ratio in the samples (for example, 3:1 of chemical X to chemical Y), then the modeling process will only detect one variation instead of two. Constant ratios sometimes arise in building calibration sets when a series of dilutions are made from a single stock solution. The model “sees” all the absorbance peaks of constituent X increase or decrease together with constituent Y. As a result, only one variation is sensed: the changes in the spectrum of the sum of X and Y. For this reason, when calibrating models it is imperative that the concentrations of the individual constituents of interest be present in uncorrelated and evenly distributed ratios. The same basic principle holds whether calibration is done optically by MFC, or digitally in a computer (like PCA). Spectral data are often mean-centered prior to PCA being applied to a calibration set. Mean centering is performed by calculating the mean spectrum (i.e., average spectrum) from all of the calibration spectra, and then subtracting the mean spectrum from every calibration spectrum. Mean centering enhances slight differences between the spectra being centered, rendering the differences more easily visible. Bear in mind that computerized factor analytic methods compute the principal components based on variations in the absorbance data, and not the absolute absorbance values themselves. For that reason, any transformation that increases the likelihood of detecting differences between the calibration spectra will enhance the model. Mean-centering can be understood from the perspective of PCA calculating the eigenvectors. Because the eigenvectors correspond to the variations in the spectral data that are shared by all the calibration spectra, eliminating the mean merely removes the first, most common variation before the data are handled by the PCA code. Similarly, double-beam spectrometry can be thought of as a way of mean-centering spectral data optically, too. In MFC, a similar enhancement can be obtained by moving the limits of detector integration (the bandpass) to regions of the spectra where those same slight differences used in PCA occur. In fact, in PCA limiting a calibration to such spectral regions often improves standard errors of estimate (SEE) and standard errors of performance (SEP).

PCA is actually a process of elimination, iteratively removing each orthogonal variation from the calibration spectra sequentially to generate a group of eigenvectors (principal components) that correspond to the variations in the absorbance values that are shared among all spectra. In reality, the underlying variations are seldom actually independent, and removing them orthogonally to form the inside model space leaves a “reverse impression” of them on the outside model space. MFC filters are also never completely orthogonal for physical reasons. Sometimes this means that fewer MFC factor filters are needed than PCs in PCA, but sometimes not. Once the calibration data have been treated completely by the PCA algorithm, the data are reduced to two matrices: the eigenvectors (dimensioned like the spectra) and the scores, which act as scalar eigenvector weighing values for all the calibration spectra. The matrix expression of the model equation for the spectral data appears in the form:

$$A = SF + E \qquad \textbf{2.2}$$

where ***A*** is an ***n*** by ***p*** matrix of spectral absorbance values (as in Figure 2.2), ***S*** is an ***n*** by ***f*** matrix of score values for all of the spectra, and ***F*** is an ***f*** by ***p*** matrix of eigenvectors. The ***E*** matrix is the errors in the model’s ability to predict the calibration absorbance values and has the same dimensionality as the ***A*** matrix. In the case of eigenvector analysis, the ***EA*** matrix is often called the matrix of residual spectra. The dimensions of the matrices are representative of the data they hold; ***n*** is the number of samples (spectra), ***p*** is the number of data points (wavelengths) used for calibration, and ***f*** is the number PCA eigenvectors. Similarly, in MFC ***S*** is obtained by integrating the total light through the sample and individual MFC filters over a broad wavelength band. ***F*** are the filter spectra, and ***E*** are the errors in the model’s ability to predict the calibration absorbance values. Figure 2.3 shows how spectra in Figure 2.2 can be reconstructed from factor scores so they can be compared to spectroscopic predictions made by modeling programs like Gaussian.

MFC-computing molecules are selected by comparing the spectrum of prospective filter materials to the loadings spectra calculated by PCA. Given a set of training spectra collected at all available wavelengths, it is possible to rationally select molecular filter (MF) materials to perform PCA. PCA is designed to maximize the signals from the spectral regions with the most variability by most heavily weighing them in calibration. However, PC loadings heavily weight signals in the positive and negative direction, which cannot be done with MFs without offsetting signal gained at one wavelength with signal lost at another wavelength in the total bandpass. Because only absolute values can be represented in MFs, as many as two filters are needed for a PC, one for the positive loadings (MF1) and one for the negative loadings (MF2). The transmission spectrum (%T) of the filter material should be as similar as possible to the absolute value of the loadings spectrum being targeted. The MFC concept and process are illustrated in Figure 2.4. Using a conventional spectrometer, mixtures of liquid molecular filters can be titrated to produce the optimum PC result. A spectral library search algorithm can also be developed to generate the optimal mixture of liquid molecular filters as long as we can obtain such a candidate spectral library.

PCA and MFC share a number of important advantages over more established spectroscopic techniques. Neither PCA nor MFC require wavelength selection. Any number of wavelengths can be employed in calibration; typically, the entire spectrum is used with PCA, and large ranges with MFC. Using more wavelengths provides an averaging effect that produces a model less susceptible to spectral noise. Both PCA and MFC provide spectral data compression and permit use of inverse regression to calculate model coefficients. Each technique can be calibrated to measure constituents of interest while ignoring most interference, and can be applied to complex mixtures because only calibration information on the constituents of interest is necessary. In some cases, they can even be used to identify samples containing contaminants not present in the original calibration mixtures.

PCA and MFC each present special problems that must be considered in their use. PCA computations are slower than the molecular computing approach to estimating scores, which is effectively instantaneous. The calibration models produced by these two techniques are relatively complex, and can be difficult to interpret and understand. Optimization of models requires some knowledge of PCA and knowledge of MFC that is not yet complete. Nonlinear spectral effects in complex samples sometimes be linearized with careful choice of molecular filters in MFC, and with PCA can be modeled to trace their source. PCA vectors often do not correspond directly to constituents of interest. In MFC, filter molecules can be selected that correspond directly to sample constituents, but only for the sample constituents that are known. From time to time spectroscopy senses an effect that is merely correlated to a constituent of interest instead of originating from the constituent itself (this usually shows up as absorbance signals at unexpected wavelengths.) For this reason, analytical chemists seldom rely on a single instrumental technique to analyze a sample. Other analytical methods can be used on these samples to verify results where accidental correlation is suspected. Generally, a large number of samples are required for accurate calibration. In hyperspectral imaging, however, each pixel contains an entire spectrum so large numbers of spectra are easy to obtain. Gathering sufficient calibration samples can be problematical when one must avoid collinear constituent concentrations. However, computer methods for assisting in the collection of orthogonal samples exist to ameliorate this problem.

2.3 Spectral Feature Encoding with Digital Mirror Array

The revolution in integrated computational imaging extends beyond just lenses and light pipes. A new trend in hyperspectral imaging is to speed the visual data processing and reduce data storage requirements by downloading some of the computation to the sensing detector itself. In many cases, the detector array can perform both feature extraction (of both physical and spectral features) and encoding of these features. The codes are transmitted by the array to the computer, integrating the computation and imaging to reduce the huge data load and speed the processing. Similarly, molecular computing in a multiplex image bandpass spectrometer can accomplish hyperspectral imaging as integrated computational imaging performs feature extraction.

Current sensor system architectures detect signals from a physical stimulus, convert them to electrical signals, convert the electrical signals to digital form for processing by computers, and, finally, extract critical information from the processed signals for exploitation. Integrated Sensing and Processing (ISP), an initiative launched in the Defense Advanced Research Projects Agency[63], aims to replace this chain of processes, each optimized separately, with new methods for designing sensor systems that treat the entire system as a single end-to-end process that can be optimized globally. In the 21st century, global information dominance is necessary to protect U.S. air, space and ground assets. Sensor systems like interferometric synthetic aperture radar

(InSAR) and IR video collect unprecedented amounts of data, greater than 10^{12} pixels/day that require more than 10^{16} flops/day to process. At the same time, the "downsizing" personnel trend persists and the ratio of "pixels to pupils" is heading toward infinity. These trends combine to make training data collection, processing, downlink and distribution all problematic as the U.S. military seeks ways to reduce rapidly data from physical fields to high-level information. At the same time, computing resources are limited in size, weight, power, and cost. Application Specific Integrated Circuits (ASICs) do not really help because they solve a fixed problem in a changing sensor/target environment. ASIC design time and cost tend to be prohibitive. More flexible detection schemes like the Texas Instruments digital micromirror array (DMA) measure features, not pixels, under computer control[64]. This holistic approach boosts signal-to-noise ratio (S/N) and concentrates information the way ICI was intended to do.

The Texas Instruments DMA chip was originally designed for use in consumer televisions, home theater systems and business projectors. White light passes through a color filter wheel, causing red, green and blue light to be directed in sequence on the surface of the DMA. The micromirrors have only two positions, on and off. The switching of the mirrors, and the proportion of time they are 'on' or 'off', is synchronized according to the color illuminating them. The human visual system integrates the sequential colors and registers a full-color image in the brain.

ICI with a DMA requires only simple computing because the feature extraction and encoding is done by the detector array. In contrast to calculating Longbow classifiers, Fourier-Mellin transforms, SVDs, and PCAs with powerful digital computers, in ICI simple digital manipulations and analog detection permit low-cost, rapid identification of targets.

The same DMA technology can be employed to compute principal components optically. For example, using a lens to collimate light for a transmission grating, a spectrum can be projected across a DMA. In this example, the columns of the DMA represent wavelength, and the rows reflect a fraction of light (between zero and one) collected at each wavelength. Principal components are simply a weighed sum of light intensities across a range of wavelengths, and it is a simple matter to turn on a fraction of the mirrors in each column to represent the weighting at each wavelength. By employing this scheme, the signal observed at a single analog detector can be easily related to a principal component score, and even to an analyte concentration. Thus, a relatively complex computational function is reduced to measuring the voltage on a single detector.

Algorithms for design and operation of ICI sensor systems are being developed that allow back-end exploitation operations, such as target identification and tracking, to configure and set the operating modes of sensor elements without human intervention to ensure the most significant data are always being collected as scenarios evolve. The ISP program approach is leading to an order-of-magnitude performance improvement in detection sensitivity and target classification accuracy, with no change in computational cost, across a broad range of Department of Defense (DoD) sensor systems and networks - from surveillance to radar, sonar, optical, and other weapon guidance systems. ISP has created novel feedback strategies to administer the elements of an adaptive optical sensing system. ISP has invented new mathematical frameworks for global optimization of design and function of a number of diverse types of sensor systems. It is also realizing its software prototypes of ICI methodology in test-bed hardware systems, including missile guidance and automatic ground-target recognition modules.

Stepping away from military applications, researchers are also replacing conventional optics on microscopes and other optical instruments, and at the same time imparting extended depth of field to these devices. Other optical engineering groups are developing ICI optics to facilitate computers in sensing motion and the physical properties of remote objects. Exceeding the limits of visible light, engineers foresee construction of similar lenses that can process other segments of the electromagnetic spectrum, extending the general change in the way scientists think about sensing.

Hyperspectral ICI optical elements have been designed for collecting all of the spectra at the same time across the pixels of a complete scene. Hyperspectral data may expose camouflaged armaments in a satellite image or biological activities under a microscope, especially with the assistance of fluorescent labels that bind to particular cellular structures. For example, one ICI spectra-capturing lens generates a multicolor (multiwavelength) design in which a 30-color spectrum linked with each point in a scene is mapped onto a detector. The design is not an image at that point; it is only a disorderly collection of colors and pixels. In spite of this disorder, computerized sorting can transform the apparent visual disharmony into an image of the scene at any individual wavelength. Such hyperspectral data have become one of the most important means by which scientists analyze the physical and chemical properties of objects extending from atoms to distant planets.

More than a generation ago, scientists tried to use lenses to transcend simple imaging. DoD tried to exploit "optical correlators" that could sense threats by optically comparing reconnaissance pictures with patterns of enemy vehicles stockpiled holographically. At the time digital processing was new but unsophisticated, and the most elegant method to manage the data was to process it optically. Regrettably, the tactic was unsuccessful because optics did not afford the degree of accuracy required for detecting threats in complex and chaotic battlefield scenes. Technology has improved significantly since that time. Most noticeably, the data-analysis capabilities of computers have skyrocketed. However, there have also been important developments in mathematical tools and innovations in optics manufacture that permit more complex lenses to be prepared, such as the wavefront coding lenses and aperture arrays using light pipes. With the union of the latest computers and innovative optics, ICI is ready to reveal a universe of possibilities that have long been concealed from the human eye.

Our laboratories contain an assortment of monochromator-based spectrometers, FT instruments, and tunable lasers that are used for routine spectrometric work. These instruments function adequately in ordinary applications. However, when high sample throughput is required, such as when pharmaceutical process analytical technologies (PAT) are used to examine finished dosage forms for zero-defect process control, traditional instruments can perform poorly. These older instruments require too much time to acquire data, and produce too much raw data that must be analyzed to provide high-level information. Hyphenation of analytical methods (e.g., NIR/ARS, or NIR/acoustic-resonance spectrometry) further increases the data burden, and the time needed to acquire spectra[65]. Furthermore, traditional instrumentation can be too complex and heavy for some applications, such as remote sensing using robotic platforms on distant planets. ICI goes a long way toward solving these problems with traditional instruments. Use of MFC in the optical design of instruments reduces the total amount of data that must be analyzed as it reduces the total instrument part count. Simpler instruments are often more rugged than complex instruments, and lighter and more suitable for applications like astrobiological research.

2.4 The MFC-based Spectrometer Prototype

Figure 2.5 is a diagram of a prototype MFC spectrometer that has been constructed based on an earlier design.[66] A 12VDC, 25-watt tungsten-halogen lamp provided broadband near-IR illumination. The source was powered by a stabilized DC power supply. An optical chopper modulated the near-IR beam at 210 Hz. A lock-in amplifier provided phase-sensitive detection. A lead-sulfide detector (CalSensors, Inc.) with an active area of 1x1 mm was placed near the focal point of the near-IR beam.

In general, near-IR absorption signals grow weaker as the wavelength of the radiation decreases. Each consecutive order of overtone or combination band is approximately a factor of ten weaker. A filter pathlength of one cm was selected for operation in the near-IR range from 1400-2200 nm. In this wavelength range, most organic liquids attenuate radiation almost completely at their absorption peaks at a 1-cm pathlength (i.e., in a molecular computing context, absorption peaks place zero weight on signal values in the factor score). The majority of these liquids also deliver sizeable bands of radiation within these limits of pathlength and wavelength (i.e., in a molecular computing context, these bands place nonzero weight on signal values in the factor score). At wavelengths lower than 1400 nm, most molecules fail to absorb strongly enough to attenuate significantly the near-IR beam in a reasonable pathlength. In addition, the tungsten-halogen light source has the most radiant power at shorter wavelengths. A 1400-nm long-pass filter was installed to prevent overwhelming the detector with light containing only modest sample information on these grounds.

Molecular filters can be held on a track or on a wheel. The sample cuvette was placed immediately in front of the filters and the PbS detector directly behind them. Black baffles (not shown in the figure for simplicity) were employed to avoid stray modulated light escaping around the sample and filters. The baffles were found to be very valuable, particularly whenever the samples or the filters were strong light absorbers. In either case, even a minute quantity of stray light can lead to appreciable errors. The convex lenses were situated to produce a defocused 1:1 image of the source filament on the active area of the detector. The small defocusing was necessary because inserting the filters and samples in the beam path shifted the focal point of the filament image forward.

2.5 Near Field Scanning Optical Microscopy using Molecular Factor Computing

Our group has already constructed a constant-height near-IR near-field scanning optical microscope (NSOM) using an external cavity tunable diode laser as a light source (80 nm tuning range per diode). This NSOM used a single-mode optical fiber pulled in a special oven to a tip diameter of 100 nm. This instrument has been utilized to demonstrate the feasibility of spectrally imaging single low-density lipoprotein (LDL) particles (18-25 nm in diameter). Furthermore, electromagnetic models have been developed and proven with experimental results to explain novel near-field imaging effects. [67]

However, *NSOM scanning is a slow process* because only one point on the sample surface can be scanned at a time with the optical fiber tip. Moreover, a tunable diode laser must slowly scan a range of wavelengths at each point on the sample. Replacing the tunable diode laser in the NSOM with synchrotron light, and placing molecular filters in the synchrotron beam path before coupling the light to the single-mode optical fiber, may enable the NSOM scanning process to be speeded up substantially. This new instrument could be used to examine collagens and elastin and their distributions in mouse aortas. A number of genetic "knockout" murine models have been developed recently to mimic human atherosclerosis and abdominal aortic aneurysm (AAA)[68]. Techniques for monitoring the onset, progression, and regression of these processes in murine models could provide valuable pathophysiological insights into the disease processes. Nondestructive in vivo techniques will be needed for proteomics studies in these models. Finally, these analytical methods may be useful in assessing the effectiveness of possible treatments.

A synchrotron is a huge machine by most laboratory standards, and it produces very intense light comprising many different wavelengths. The electromagnetic radiation is considerably more intense than that from a diode in a near-IR TV remote control, a microwave oven, or dental X-ray machine covering the same wavelengths, because the synchrotron's rays of light are concentrated into very small zones. These small areas are ideal for multiwavelength microscopy in the near and far fields. The synchrotron produces light by accelerating electrons to nearly the speed of light. Magnets channel the electrons into circular paths. As the electrons turn (accelerate), photons are emitted. The mechanism of the Brookhaven National Laboratory National Synchrotron Light Source includes an electron gun, linear accelerator, a circular booster ring (to increase the speed of the electrons), two storage rings (to re-circulate electrons), and beamlines (evacuated pipes down which the infrared, UV, and X-rays are launched to the research areas where they are used for experiments).

Nanometer-sized structures are becoming increasingly important and, as a result, near-field scanning optical microscopes (NSOMs) are becoming increasingly popular analytical tools[25]. An NSOM instrument allow samples that are smaller than half the wavelength of light to be imaged by using an aperture size and an aperture-sample separation distance that is less than a wavelength of the source. Furthermore, because optical sources are utilized, NSOM instruments provide this resolution while maintaining the advantages of traditional optical microscopes including nondestructive sample analysis, as well as spectroscopic analysis. The combination of NSOM with the synchrotron and MFC should speed near-field hyperspectral imaging and research into the etiology of AAA.

Diffuse reflection near-infrared (near-IR) spectroscopy has proven to be a useful technique for identifying chemical content of biological tissues. Biological applications of near-IR spectroscopy include monitoring systemic and cerebral oxygenation and identifying plasma constituents including glucose, total protein, triglycerides, cholesterol, urea, creatinine, and uric acid. Our group has reported on the use of near-IR spectroscopy to classify human aortic atherosclerotic plaques and to identify cholesterol, HDL, and LDL in arterial wall samples.[69] Feasibility testing of NIR-NSOM with MFC to monitor collagen and elastin in the aortas of ApoE knockout mice with atherosclerosis and aneurysm formed from chronic infusion of the peptide angiotensin II has just begun. The goal of this project is to create a novel near-field microscopic probe capable of rapidly collecting molecular structural information from mouse aortas at subwavelength resolution. These data may help unravel the complex biochemistry of aneurysm formation and progression, and will certainly develop a new tool for use in other biomedical research.

2.6 Conclusion

The approaching tsunami in hyperspectral imaging is the speeding of processing and reduction of data storage requirements by downloading some of the computation to the detector array itself. In many cases, the detector array can perform both feature extraction (of both physical and spectral features) and encoding of these features. The codes are transmitted by the array to the computer, integrating the computation and imaging to reduce the huge data load and speed the processing. Molecular computing in a multiplex bandpass spectrometer can accomplish hyperspectral imaging as integrated computational imaging performs feature extraction.

The possibilities of the MFC based ICI approach are nearly endless. A synchrotron provides a bright, thin collimated beam of near-infrared and infrared light in the form of 20 ns pulses. This broadband beam of light can be readily coupled to a chalcogenide or other suitable optical fiber and used for NIR-NSOM (near-field scanning optical microscopy). Complex absorption filters can be used as mathematical factors to create a factor-analytic near-IR calibration in such a highthroughput near-IR nanospectrometer. The wavenumber (i.e., factor or principal component) selectors can be placed in the far field over the detector, simplifying construction. This new type of nanospectrometer offers simplicity, cost advantages, and enhanced throughput. The high-throughput wavelength selector is useful for ensuring the signal-to-noise ratio typically needed for chemometrics, and is vital given the extremely small aperture of a near-field probe. Using this device, collagen I and III as well as elastin can be imaged in aortas beyond the diffraction limit, enabling the mechanisms of collagen:elastin ratio increase in the genesis of aneurysms to be determined

ICI also has applications in coronary catheters, where physicians do not have much time to locate vulnerable atherosclerotic plaques. The “old western gunfight” analogy used earlier explains much of the motivation for ICI. A gunfighter who controls his revolver directly from his retina has some distinct advantages over a gunfighter who must send signals from his retina to his brain, and then from his brain to his gun hand. ICI has many military applications for the same reason – the available time to identify correctly many different targets is short. The data-analysis capabilities of computers continue to grow in accordance with Moore’s Law, and those capabilities need not be wasted on low-level operations. Moreover, significant developments in mathematical tools and innovations in optics manufacture permit more complex components to be made, such as molecular computing fiber optics and lenslet arrays. With the union of the latest computers and innovative optics, ICI is ready to reveal a universe of possibilities that have been concealed from the human eye.

Chapter Two Figures

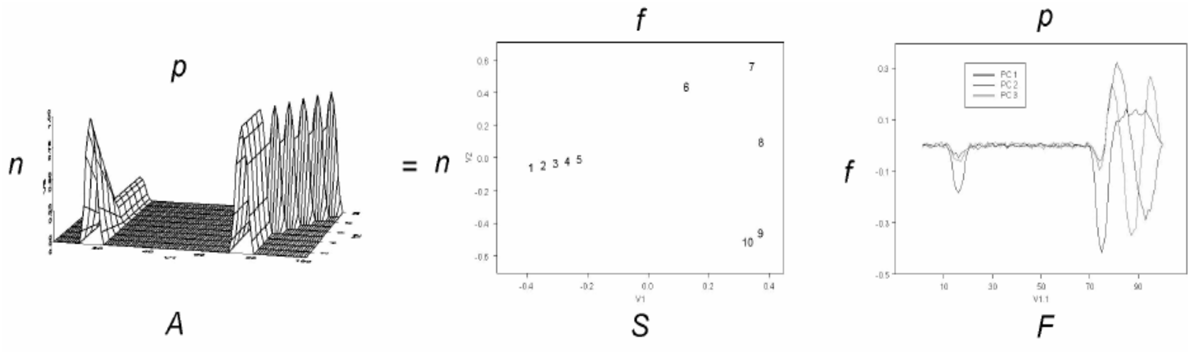


Figure 2.1 PCA decomposes the spectral data into the most common spectral variations (often called factors, eigenvectors, or loadings, and basically equivalent to multiplex bandpass molecular filters) and the corresponding scaling coefficients (scores, equivalent to the integrated bandpass detector signal in MFC). A=original spectra, S=PC scores, F=factors (loadings), n=number of spectra, p=number of data points, f=number of PCs

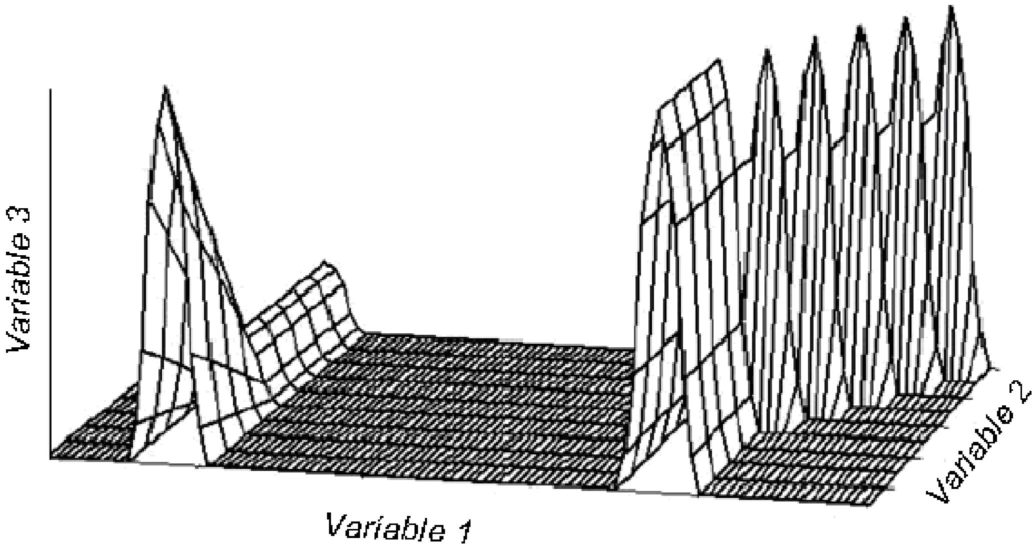


Figure 2.2 A hypothetical training set of spectra with two varying constituents. The peak on the left changes height while the peak on the right is held constant. Then the peak on the right drifts to higher energies while the intensity of the peak on the left is held constant.

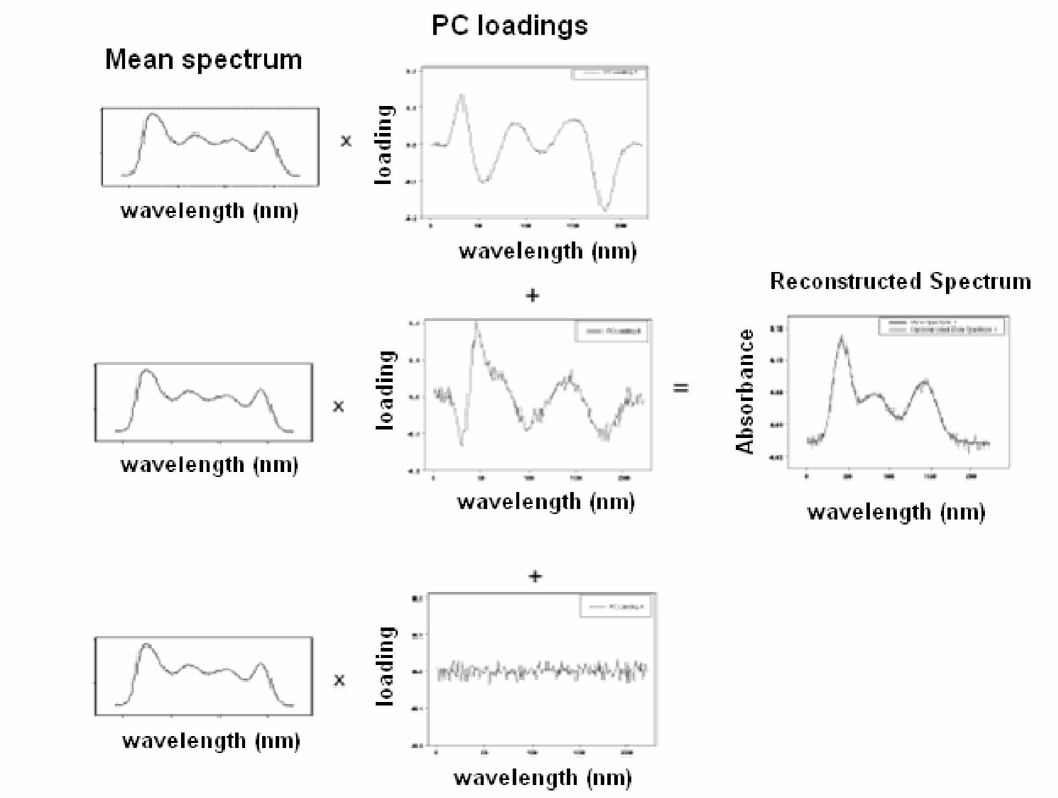


Figure 2.3 The original calibration spectra can be recreated using all of the loadings and scores. By selecting certain vectors from the loadings, the spectra of selected constituents or properties can be selectively reconstructed. Molecular filters MF1 and MF3 could be substituted for PC1 and PC3, and the scores would be the integrated detector signals observed with MF1 and MF3.

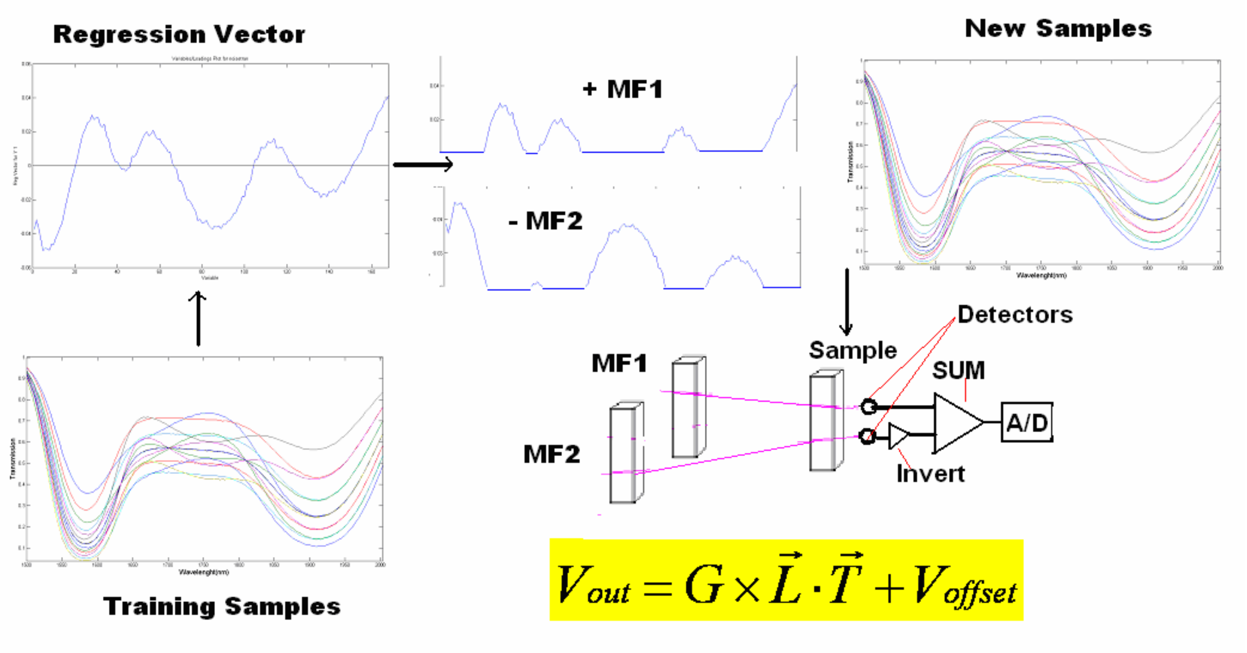


Figure 2.4 Concept illustration of the design process using MFC approach. Training spectra were collected first, followed by PCR to obtain the regression vector. The regression vector is split to the positive and negative half, and the goal is to find the molecular filter which has the spectral pattern matches the regression vector. The detected signal from negative detector is inverted and added to the detected signal from positive detector, the resulting signal should related to a property (i.e. concentration) of the sample under measurement.

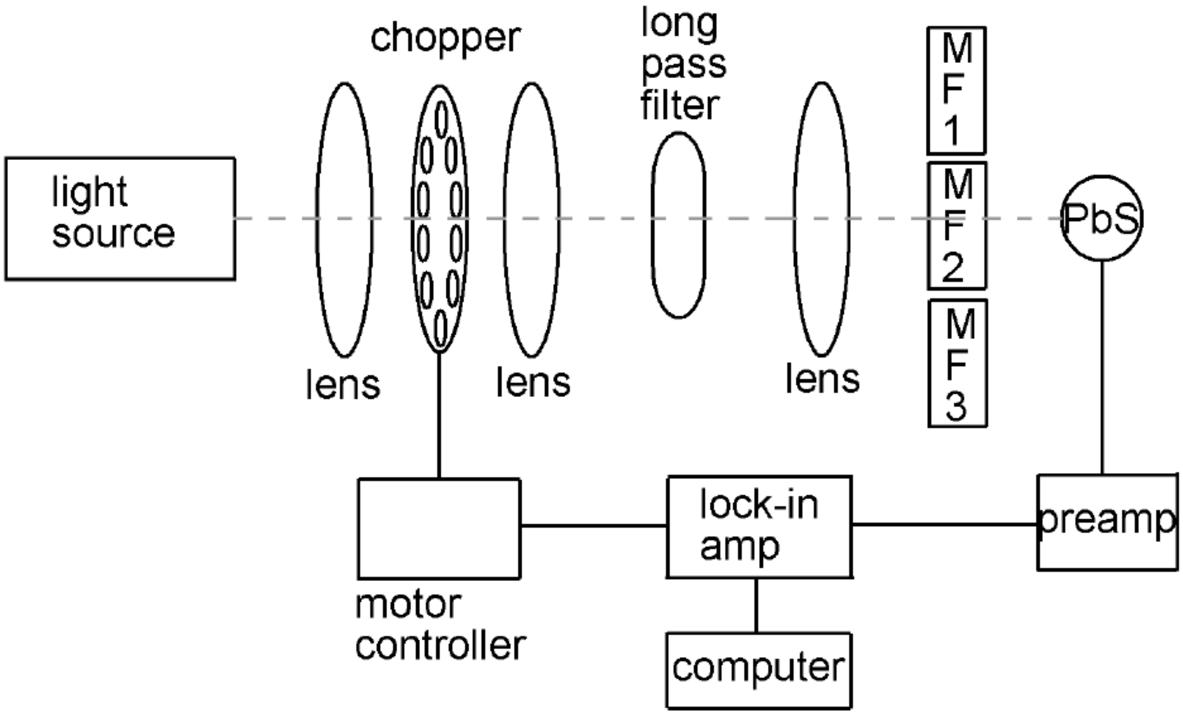


Figure 2.5 Block diagram of a prototype molecular computing near-IR spectrometer that outputs factor scores. Three molecular filters (MF1, MF2, and MF3) rotate in and out of the light path on a filter wheel.

Chapter Three -- Genetic Algorithm Based Multivariate Linear Regression for Molecular Filter Selection in Molecular Factor Computing

3.1 Introduction

On July 13 2006, Roxane Laboratories and the US Food and Drug Administration (FDA) issued a nationwide recall of a single manufacturing lot of Azathioprine tablets, 50 mg.[70] This recall was initiated due to concerns that bottles labeled as Azathioprine may contain Methotrexate, 2.5 mg tablets. A rapid, nondestructive online method of analysis and monitoring could have prevented the recall and disposal of the batch of mislabeled tablets, thus reducing the company considerable cost and avoiding a potentially serious or life-threatening health risk.

Near-infrared (NIR) spectroscopy has emerged as a widely used analytical method in process environments in the biotechnology and pharmaceutical industries.[20, 71-73] NIR spectroscopy requires very little or no sample preparation. In addition, it is a nondestructive technique. These advantages make NIR spectroscopy an ideal technique for industrial on-line process analysis. For routine application in industrial process lines, expensive and bulky laboratory-type instruments are often not the best choice. Instead, filter-based instruments are used because they have high optical throughput, offer rapid data collection and can be constructed ruggedly and at considerably lower cost than traditional dispersive or interferometric instruments.[74, 75]

Modern spectroscopy and hyperspectral imaging techniques are able to collect a huge amount of data in a short time. The sheer volume of data threatens to cause a bottleneck in process analysis and control, drug discovery and development, and other areas. Given the flood of data that are generated from spectroscopic sensors, appropriate analysis techniques need to be used in order to extract useful high-level information from fields of physical data. Multivariate calibration is a well-established tool in chemometrics for multiwavelength data such as NIR, UV-vis, and Raman spectra[8]. Conventional measurement of the chemical or physical properties of a sample using optical spectra entails constructing a predictive model based on training spectra. Two of the most commonly used methods to construct a predictive model are partial least squares (PLS) and principal component regression (PCR).[26, 48] After a predictive model is constructed, the prediction of the chemical or physical properties is carried out by multiplying the spectra of unknown sample with the regression vector. In a conventional approach, the complete data collection and analysis process for raw data can be time consuming and computationally intensive. Part of the reason for the computational demands is that the device used to acquire training data to build the model is exactly reproduced when applying the model to predict the properties of unknown samples.

One approach currently being investigated to simplify both instrumentation and computational analysis involves optical pattern encoding. This technique involves tailoring the optical spectrum of filters to encode high-level information about the samples in the sensing stage. Theoretical treatment of this methodology can be found in the literature.[11] Myrick et al. have demonstrated some practical applications of this methodology in UV-vis and NIR spectroscopy.[76-83] These applications are based on the fabrication of thin film solid-state optical filters, termed multivariate optical elements (MOEs). MOEs are designed to replicate the multivariate regression pattern by transmitting and reflecting weighed optical signals over a broad wavelength band.

A recent publication, proposed an alternative approach for spectral encoding.[84] The new spectral encoding approach designs a molecular filter-based spectrometer. As shown in Figure 3.1, a light source, few molecular filters, and a single detector are used to construct the spectrometer. Molecular absorption filters are used in the spectrometer as mathematical factors for spectral encoding to generate a factor-analytic optical calibration in a high-throughput spectrometer. The process is called molecular factor computing, or MFC. The molecules in the filter effectively represent the regression vector by weighing the signals received at each wavelength over a broad range of wavelengths. One or more molecular filters are used in the MFC-based spectrometer to produce detector signals correlated to desired sample information.

A critical part of the MFC approach is the molecular filter selection. An initial approach to select desired molecular filters is to represent a predefined principal components (PC) loading vector, and thereby attempt to maximize the analytical signal from a series of mixtures while minimizing the total signal contribution from interferences. Given a set of training spectra, collected at all available wavelengths (see Figure 3.2, left side), it is possible to rationally select molecular filter (MF) materials to perform principal component analysis (PCA) (see Figure 3.2, right side).

PCA is designed to maximize the signals from the spectral regions with the most variability by weighing them most heavily (with the loadings line in left graph in Figure 3.2) in calibration. However, PC loadings heavily weigh signals in the positive and negative direction, which cannot be integrated over the entire range of wavelengths with MFs without offsetting the signal gained at one wavelength with signal lost at another wavelength. Because only absolute values can be represented in MFs, two filters are needed to represent a PC, one for the positive loadings (MF1) and one for the negative loadings (MF2). The filter materials are selected by examining the sample spectra. In this approach, the transmission spectrum (%T) of the filter material should be as similar as possible to the absolute value of the loadings spectrum being targeted. Bandpass filters should be selected to ignore regions of the spectrum where there is no difference between the training spectra, as extra photons in those regions simply saturate the detector or add noise without providing any additional signal. The MF filters do not have to be featureless in the areas away from their peaks in the pictures above as long as bandpass filters (or prisms or gratings) are used to wipe out the %T peaks in undesired areas. For multivariate calibration problems, however, the complex spectral patterns required of molecular filters used as principal component loading vectors lead to difficulty finding appropriate molecules in the spectral library. In addition, for a complex sample matrix, more PCs are required to construct a calibration model, thus a larger number of MFs are required.

As a way to overcome some of the limitations of the filter selection approach based on principal component vector matching, a novel MFs selection algorithm is proposed that results in a requirement for fewer MFs and a simple library search for MFs. This approach uses a genetic algorithm to search the spectral library to find the MFs that minimize the prediction error for the species of interest. Instead of selecting MFs whose spectrum matches a predefined PC loading vector, the new algorithm selects MFs to generate a new coordinate system that attempts to minimize the root mean standard error of prediction (RMSEP).

There are plentiful applications of this methodology. Process Analytical Technology (PAT)[3, 4] is one field where this approach could be valuable, allowing for rapid online process monitoring. PAT aims to ensure final product quality through rapid and noninvasive measurements of critical quality of materials and processes. One typical example is rapid online monitoring for the manufacture process of pharmaceutical tablets.[19] In addition to the PAT application, MFC-based remote NIR imaging for real-time surreptitious surveillance has gained more interests.[17]

3.2 Theory

3.2.1 Molecular filters for optical computing

MFC is a form of analog computing, which has remained important in computing because of its extremely high speed. In the MFC approach to predict the chemical properties of samples, molecular filters are combined with a light source and detector to make a MFC-based spectrometer. Molecular computing of vectors by using molecular filters as transformation matrices enables spectra to be represented in any arbitrary coordinate system. The coordinate systems are selected for specific reasons (such as reducing interferences, increasing specificity, increasing robustness or increasing resistance to outliers). New coordinates can be used to reduce the dimensionality of the spectral hyperspace, to increase the orthogonality of a subspace of the coordinate system, or to simplify the mechanical/electrical/computational construction of a spectrometer. In the molecular filter selection algorithm, new coordinate systems are selected to construct a simple and robust calibration model with lower root mean standard error of prediction (RMSEP).

By directing the light through the sample then to the detector, both signal convolution and summation are implemented. The resulting detector response is proportional to the summed intensity of detected photons. For each molecular filter, only one detector response value (corresponding to a factor score) is obtained on each sample. Instead of collecting entire spectra for each sample, only few detector responses are collected for each sample, depending upon the number of molecular filters used in the system. In the MFC approach, the mechanical/electrical/computational construction of a spectrometer is significantly simplified. More importantly, MFC yields a throughput advantage and a multiplex advantage, increasing the signal to noise ratio (S/N). MFC also yields the ISP (integrated sensing and processing) advantage, which virtually eliminates the need for computer analysis after data collection.

In an MFC-based spectrometer, for each sample, the voltage output of the detector is given by Equation 3.1: [16]

$$V_{out} = G \times \vec{L} \cdot \vec{S} + V_{offset} \qquad \textbf{3.1}$$

where V_{out} is output voltage, G is amplifier gain, $\vec{L}$ is the molecular filter transmission spectrum vector, V_{offset} is voltage offset of the detector and $\vec{S}$ is the sample transmission spectrum.

With m different MFs, m V_{out} values are obtained for each sample. The prediction of a sample property is achieved by a multivariate linear regression model (MLR).

$$\hat{y} = \sum_{i=1}^{m} \beta_i V_{out\ i} \qquad \textbf{3.2}$$

where $\hat{y}$ is the estimated concentration or property, β are the regression coefficients, and m is the number of MFs used in the spectrometer. The evaluation of equation 3.2 can be accomplished using a digital computer (typical in calibration) or faster analog electronics (typical for prediction of production samples).

3.2.2 **Molecular filter selection by spectral library search**

The selection of molecular filters from the reference spectral library is carried out in a manner that minimizes the cross-validate root mean standard error of calibration (RMSECV). First, the complete transmission spectra of samples obtained from a conventional spectrometer are convolved with the emission spectrum of the light source for MFC, $\vec{I}$, and the response of the detector, $\vec{R}$, to get corrected transmission spectra of samples, $\hat{S}$.

$$\hat{S} = \vec{S} \times \vec{I} \times \vec{R} \qquad \textbf{3.3}$$

Then the corrected transmission spectra of the samples, $\hat{S}$, are multiplied (dot product) by the transmission spectra of the library, *L*, to produce simulated scores, *T,* which represent the voltage response of a detector for each sample if it had been measured .

$$T = \hat{S} \cdot L \qquad \textbf{3.4}$$

With a library that contains *l* members, and a training sample set with *n* spectra, the resulting simulated scores *T* form a $l \times n$ matrix. After the simulated library scores are obtained, a genetic algorithm was used to search the simulated score space (*l* dimensions) to find a set of optimal variables that identify the optimal molecular filters. The variable selection problem is defined within a fitness function against which each selected variable set is evaluated in order to find the best calibration model. The estimated concentration values of *n* calibration samples, $\hat{y}_i$, as in Equation 3.5:

$$\hat{y}_i = \alpha \cdot T_{(m),i} \qquad \textbf{3.5}$$

where *m* is the number of optimal variable that is determined by genetic algorithm. α are the coefficients that indicated the weights of the chosen molecular filters. Till here, the molecular has been chosen. The RMSEC is calculated according to Equation 3.6:

$$RMSEC = \left[\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n} \right]^{1/2} \qquad \textbf{3.6}$$

The RMSEC equation served as the fitness function that is optimized by genetic algorithm.

3.2.3 **Architecture of genetic algorithm (GA)**

The genetic algorithm [58] is an evolutionary computing method for solving optimization problems for which there are many possible solutions based on natural selection, the process that drives biological evolution. In GAs, the initial step is to generate a random population that consists of a predefined number of individuals and variables. At each step, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation until a particular stop criterion has been reached. Over successive generations, the population "evolves" toward an optimal solution. The genetic algorithm is typically useful for solving a variety of variable selection problems that are not well suited for standard optimization algorithms. In spectroscopy, it has already been shown that genetic algorithms can be successfully used as a feature selection technique.[56, 57]

One major risk in using the GA is overfitting. As the number of test models and variables increase, the risk of overfitting increases. Cross-validation can significantly reduce the risk of overfitting, although it does not provide complete assurance. Leave-one-out cross validation is used when the training set is less than 30, otherwise, random block cross-validation is used as the cross validation method. To further reduce the risk of overfitting, GA routine was run 50 times independently. The final model is obtained from the variables that appear most frequently. Table 3.1 lists the parameters of the genetic algorithm employed in this study.

3.3 Experimental Section

3.3.1 NIR spectra library

The chemicals chosen as MFC filters were found by searching a library of near-IR transmission spectra of 1923 compounds (John Wiley & Sons, Inc.). The library consisted of two spectra of each compound collected from two slightly overlapping NIR regions, 952-1587 nm and 1388-2630 nm. Different spectral regions in the library are used depending on the path length needed for different applications, the signal strength and the interferences present.

3.3.2 **Simulation dataset**

Simulations were run to generate transmission data to form calibration and validation datasets. Simulated datasets were used to demonstrate and validate the filter selection algorithm during the calibration and validation stage. Two simulated pure component absorbance spectra were generated by summing weighed Gaussian bands as shown in Figure3.3. The absorbance spectra of the two-component mixtures are generated by adding simulated spectra of pure components with different factors based on the concentrations. The concentration of each component in the mixtures is based on a 7-level, 2-factor central composite experiment design[23] as shown in Table 3.2. The transmission spectra are generated by log-transforming of the absorbance spectra. Gaussian distributed random numbers with a mean of zero and standard deviation of 0.1%T were added to the transmission spectra to simulate measurement noise. The wavelength range from 1400 to 2000 nm was assigned as an arbitrary wavelength range for the spectra. The resulting transmission spectra are shown in Figure 3.4. The transmission spectral dataset was split into a calibration set (15 spectra) and a validation set (15 spectra) to develop and test the algorithm.

3.3.3 **Experimental dataset I**

Spectra of 615 pharmaceutical tablets were obtained from a Multitab spectrometer (Foss NIRsystems, Silverspring, MD) and were used as experimental dataset I. These spectra dataset was originally used as a shootout dataset at IDRC 2002 (International Diffuse Reflectance Conference). All tablets were scanned in the transmittance mode. Each individual tablet was subsequently analyzed for ASSAY value, tablet weight, and tablet hardness, but only ASSAY value was used for the development of filter selection algorithm. The name of the active ingredient in these tablets was not disclosed for proprietary reasons. The reference method has an error of +/- 1.3 mg on these tablets with a nominal value of 200 mg per tablet. The data set includes tablets with a wide ASSAY range, 152 to 239 mg, for developing the calibration. All transmission spectra cover the region from 600 to 1898 nm in 2 nm increments. Because the important absorbance bands fall in 1000-1600 nm region, only the 950-1580nm region was used in this study. Data were split into a calibration set (155 spectra), and a validation set (460 spectra). The transmission spectra of the calibration set are shown in Figure 3.5.

3.3.4 **Experimental dataset II**

Dataset II is transmission spectra of ethanol in water mixtures. Ethanol was reagent grade, obtained from AAPER (Shelbyville, KY). Water was distilled in house. Quantitative mixtures of alcohol and water were prepared by volume using grade-A volumetric flasks and burettes. Twenty samples were obtained with ethanol concentration ranging from 0% to 14%. Sample solutions were placed in a quartz cuvette with a path length of 2 mm. NIR transmission spectra were collected using an Ocean Optics NIR256 temperature-regulated NIR spectrometer (Ocean Optics, Dunedin, FL) over the wavelength range of 900-2500 nm to acquire a total of 256 data points. Data analysis was limited to 1400-2200 nm to avoid the short wavelength region. Selected transmission spectra included 128 data points. The transmission spectra were convolved with the transmission spectra of a 1400 nm long-pass filter, the emission spectrum of the tungsten NIR source, and the response curve of the InGaAs photodiode detector in the MFC instrument to give a corrected representation of the response. These corrected transmission spectra were used as training spectra for MFC filter selection and multivariate analysis. The corrected spectra are presented in Figure 3.6. Due to the small training set size, leave-one-out cross validation was used as the validation method.

3.3.5 **Data analysis**

All data analysis was carried out using Matlab 7.0 (Mathworks, Inc., Natick, MA). The PLS toolbox v3.51 for Matlab (Eigenvector Research, Inc. Wenatchee, WA) was used for multivariate analysis. The genetic algorithm and direct search toolbox for Matlab were employed to develop the molecular filter selection algorithm.

3.4 Results and Discussion

3.4.1 **Conventional PCR calibration**

The PCA models were used to correlate the analyte concentration to spectra to give an estimation of regression vector, r. The number of principal components required to give sufficient prediction was determined by cross validation. In this study, PCR calibration models were constructed for both simulated data and experimental data to estimate calibration performance for comparison with the performance of the MFC approach.

For the calibration set of simulated transmission spectra (15 spectra), a four-factor PCR calibration model was constructed after evaluating the root mean standard error of calibration (RMSEC) and root mean standard error of cross validation (RMSECV) (RMSEC=0.0054, RMSECV=0.0103，r^2=0.997, F test =11.09%). The PCR model was used to predict the validation set of simulated transmission spectra (15 spectra). The root mean standard error of prediction (RMSEP) for the validation set was 0.0034. The lower RMSEP than RMSEC is due to the smaller concentration range covered by the validation set. The scaled PCR regression vector is plotted as solid line in Figure 3.7. The good performance of the PCR model based on transmission spectra suggested using transmission spectra instead of absorbance spectra is acceptable for prediction.

A four-factor PCR calibration model was built for the calibration set of transmission spectra of pharmaceutical tablets. The RMSEC was 6.47 mg and RMSECV was 6.72 mg, corresponding to 3.23% and 3.36% error relative to the mean of the calibration set, respectively. (F test=80.67%).The model was validated with the test set of transmission spectra (460 spectra), and the standard error of prediction was 6.22 mg, corresponding to 3.11% error relative to the mean of the test set.

PCR calibration was carried out with corrected transmission spectra of ethanol in water mixtures. Four principal components were required to build a calibration model with optimum predictive ability. The RMSEC was 0.359%, corresponding to 5.13% error relative to the mean of the calibration set. The four-PC model was validated by leave-one-out cross validation, and the RMSECV was 0.551%, or 7.87% relative to the mean of the calibration set. F test=26.82%. A PCR calibration was also carried out on absorbance spectra. Three principal components were required to build an optimum calibration model. The RMSEC was 0.309%, corresponding to 4.41% error relative to the mean of the calibration set. The three-PC model was validated by leave-one-out cross validation, and the RMSECV was 0.494%, or 7.06% relative to the mean of the calibration set. F test =27.88% Compared to the PCR model based on corrected transmission spectra, the PCR model based on absorbance spectra required fewer principal components and has a slightly lower RMSEC and RMSECV. Even the calibration model based on transmission data is slightly worse than the model based on absorbance, it still hold adequate predictive capability. The better model with fewer principal components obtained using absorbance spectra is due to the linearity existing between absorbance and concentration as Beer's law states.

3.4.2 **Performance of the MFC approach**

The GA-based molecular filter selection algorithm was tested first on the simulated dataset. Three molecular filters were chosen from a spectral library that consisted of 1923 components.

The molecules were ethylester-nonanoic acid, 4-ethoxy-benzaldehyde and N,N,N',N'-tetramethyl-1,4 butanediamine. Their spectra are shown in Figure 3.8. The multivariate linear regression (MLR) model using these molecular filters provided adequate calibration and prediction performance (RMSEC=0.0045, RMSEP=0.0031, r^2=0.993, F test=93.93%) for the simulated data. The GA result is actually better than the PCR model. The predictions of both the training and validation set are shown in Figure 3.9. The regression vector generated using those three molecular filters is also plotted in Figure 3.7 to provide a comparison with the regression vector generated by the PCR calibration model.

For experimental dataset I, four molecular filters were chosen to give a satisfactory MLR model. The four molecules were 1-iodo-octane, 1-ethyl-piperidine, methanesulfonic acid trifluoro-ethylester and N,N,N',N'-tetramethyl-1,4 butanediamine. Their transmission spectra are shown in Figure 3.10. The RMSEC and RMSEP were 4.57mg and 4.51mg, corresponding to 2.29% and 2.26% error relative to the mean of the calibration set, respectively. (F test =83.1%). The predictions of both the training and validation set are also shown in Figure 3.11. The regression vector generated by those four molecular filters is plotted in Figure 3.12 to give a comparison with the regression vector generated by PCR calibration model.

Four molecular filters were chosen from the spectra library to give an adequate MLR model for experimental dataset II. Water, methanol, ethanesulfonic acid, and 2,2-diethoxypropane were selected as MFs, and their transmission spectra are plotted in Figure 3.13. The RMSEC and

RMSECV are 0.229% and 0.339%, respectively. F test =80.18%.

To demonstrate that the algorithm described above is not simply useful theoretically in simulations, an actual MFC-based NIR spectrometer was constructed in the lab to test the performance of the instrument using ethanol-in-water mixtures. Details of the prototype instrument were described in a previous publication[85]. Four molecular filters chosen by the MFC selection algorithm were used in this instrument as multiplex filters. Thirty-nine ethanol-in-water samples were tested in this spectrometer. Each sample was represented by four-point factor spectra because only four molecular filters were used. The instrument was calibrated using 20 training samples, which were the same samples as those used to generate experimental dataset II. The other 19 samples were used as test samples to validate the predictive ability of the real MFC spectrometer. The RMSEC and RMSEP for ethanol in water mixture were 0.748% and 0.735%, respectively. (F test=54.02%). The predictions of both calibration and test samples were shown in Fig 3.14. This result is worse than the simulated model based on experimental dataset II (RMSEC= 0.229%, RMSECV=0.339%, r^2=0.995, F test=80.18%). The factors that contributed to the degraded predictive performance were discussed in a previous paper[85]. The regression vector of MLR model based on four molecular filters and the regression vector of PCR model based on transmission spectra are both plotted in Figure 3.15 to provide a comparison.

The performances of calibration models using PCR and MLR for different data sets are listed in Table 3.3. It is found that the MFC based MLR calibration provides better models than full spectra based PCR calibration for all the datasets

3.4.3 **Discussion**

This novel algorithm provides great flexibility for molecular filters selection. Due to the stochastic nature of the genetic algorithm, there are several different set of molecular filters that provide sufficient accurate prediction. This multi-solution approach makes it possible to choose final molecular filter by taking account of the cost, availability and toxicity of the filter compounds.

The susceptibility of MFC-based spectroscopic measurement to complex matrix interference in sample is not well understood. Theoretically, the MFC-based instrument should be able to precisely measure the specific chemical species of interest as long as the potential interferences were introduced and modeled in the training set. New testing and validation of the algorithm for a dataset that collected in more complex matrix with more predefined interfering species will be carried out to evaluate the susceptibility to interference.

Due to the radiometric nature of MFC-based spectroscopy, there exists the nonlinearity in radiometry. Multivariate statistics based chemometrics for modeling such nonlinear has not been well developed yet, so it is difficult to quantitatively estimate the error accumulated by the nonlinearity.

3.5 Conclusion

Many areas of science now generate huge volumes of data that present visualization, modeling, and interpretation challenges. Methods for selecting small but highly relevant variables to represent the original data in a reduced coordinate space are therefore receiving much attention. This research demonstrated a novel library search method using genetic algorithms that, coupled with predictive modeling methods to select molecular filters from a spectra library, permits quantitative prediction of chemical species of interest. Instead of selecting MFC filters in terms of matching the PCR regression vector that represents a fixed RMSEC, this new algorithm searches the spectral library to find MFC filters that represent the original spectra in a new coordinate system to minimize the RMSEC. The performance of MLR calibration based on the MFs selected by the algorithm was compared to PCR calibration based on transmission spectra of samples, and superior results were obtained by using MLR calibration with MFC approach. The novel algorithm has been successfully tested on both pharmaceutical tablet dataset and ethanol sensing dataset. A prototype MFC type NIR spectrometer, whose design is generated from this algorithm, has been built and tested on ethanol sensing, the good performance of this prototype instrument proved the concept.

Acknowledgement

The authors are grateful to Gary E. Ritchie at the US Phamacopoeia, Purdue Pharma, and the International Diffuse Reflectance Conference for providing the pharmaceutical tablet data set.

Chapter Three Tables

Table 3.1 Parameters of the GA in GA-MLR Algorithm

* Variables: simulated library scores S.
* Variable size: 1923 (simulated library score)
* Population size: 20 chromosomes
* Regression method: MLR
* Fitness function: cross-validated root mean square error of prediction
* Maximum number of variables selected in the same chromosome: 4
* Probability of mutation: 1%
* Crossover fraction: 50%
* Time limit: 6000 sec
* Generation limit: 100

Table 3.2 Amount of components 1 and 2 in simulated calibration and validation dataset.

Calibration set			Validation set	
Sample No.	Component 1	Component 2	Component 1	Component 2
1	0.4	1.0	0.5	1
2	0.6	1.4	0.7	1.3
3	0.6	0.6	0.7	0.7
4	0.8	0.8	0.9	0.9
5	0.8	1.2	0.9	1.1
6	1.0	1.0	1.0	1.0
7	1.0	1.0	1.0	1.0
8	1.0	1.0	1.0	1.0
9	1.2	0.8	1.1	1.1
10	1.2	1.2	1.1	0.9
11	1.4	1.4	1.3	0.7
12	1.4	0.6	1.3	1.3
13	1.6	1.0	1.5	1
14	1.0	0.4	1.0	0.5
15	1.0	1.6	1.0	1.5

Table 3.3 Performance of calibration model using PCR and MFC for different datasets.

	SEC	SECV	SEP	Num of PCs	Num of MFCs
Simulation Dataset (PCR)	0.0054	0.0103	0.0034	4	
Simulation Dataset (MFC)	0.0045	0.0065	0.0031		3
Experimental Dataset I (PCR)	6.47mg	6.72mg	6.22mg	4	
Experimental Dataset I (MFC)	4.57mg	4.57mg	4.51mg		3
Experimental Dataset II (PCR)	0.359%	0.551%	N/A	4	
Experimental Dataset II (MFC)	0.229%	0.339%	0.735%		4

Chapter Three Figures

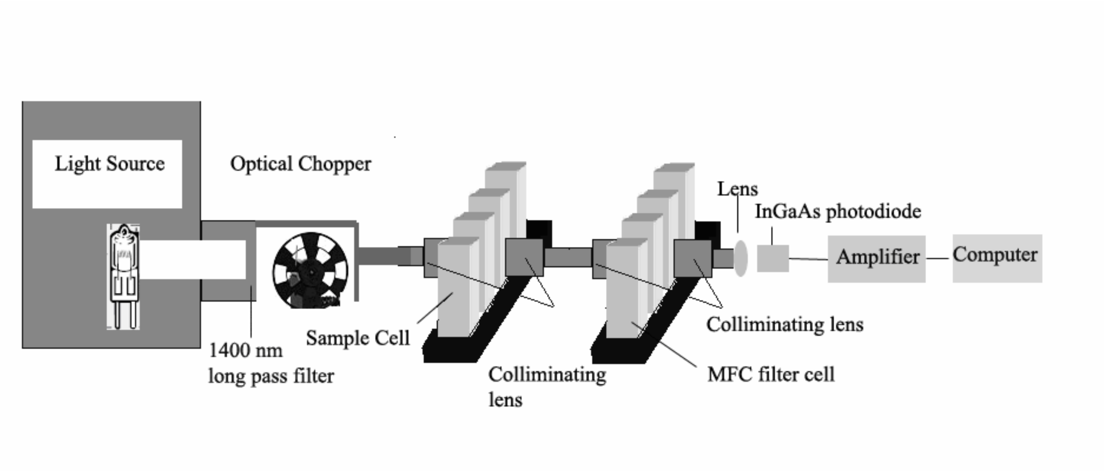


Figure 3.1 The schematic representation of MFC based high throughput spectrometer

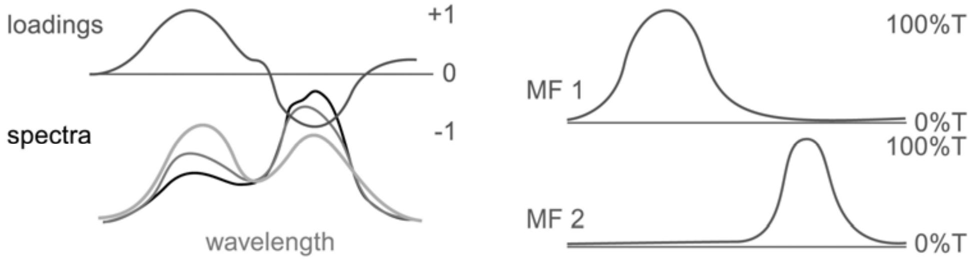


Figure 3.2 Principal components (PCs) of spectral data are formed from loadings vectors (left). The highest loadings correspond to wavelengths where the variation of interest in the sample spectra is greatest. The variations can be captured optically by selecting molecular filter (MF) substances with transmission spectra similar to the loadings. If there are positive and negative loadings in the MFC bandpass, two molecular filters must be employed for that PC to avoid ambiguity.

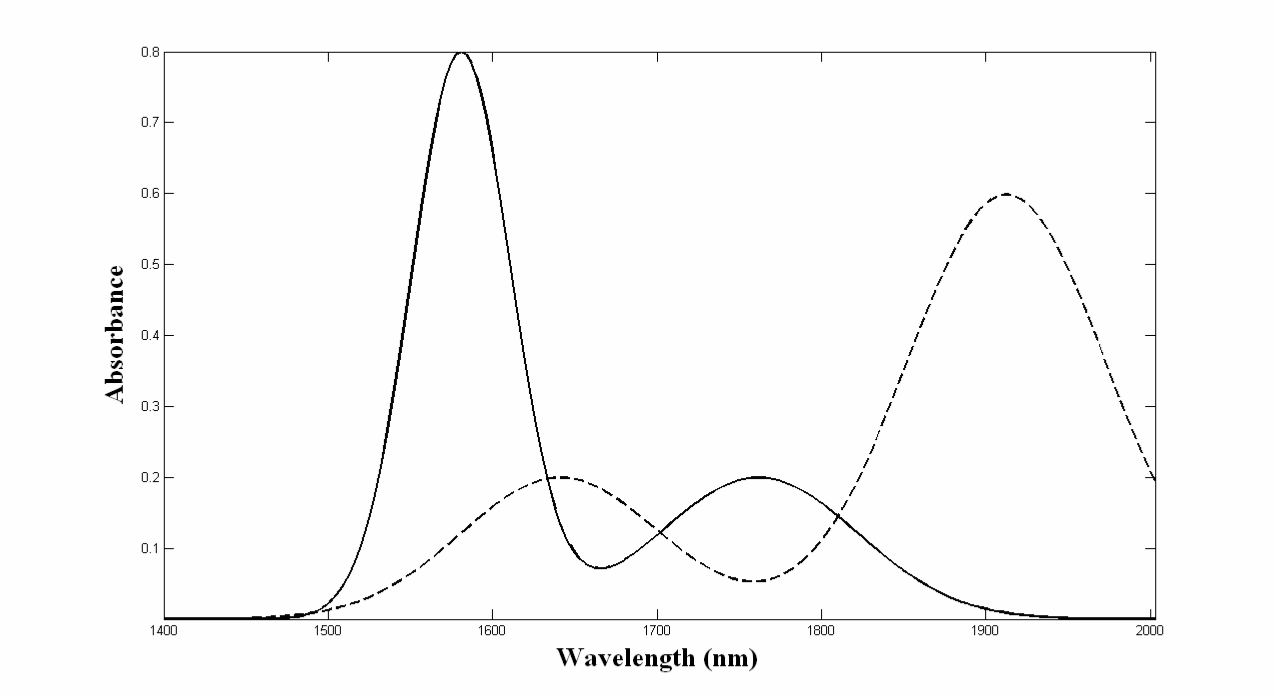


Figure 3.3 Simulated spectra of pure components: (----) Component 1, (—) Component 2.

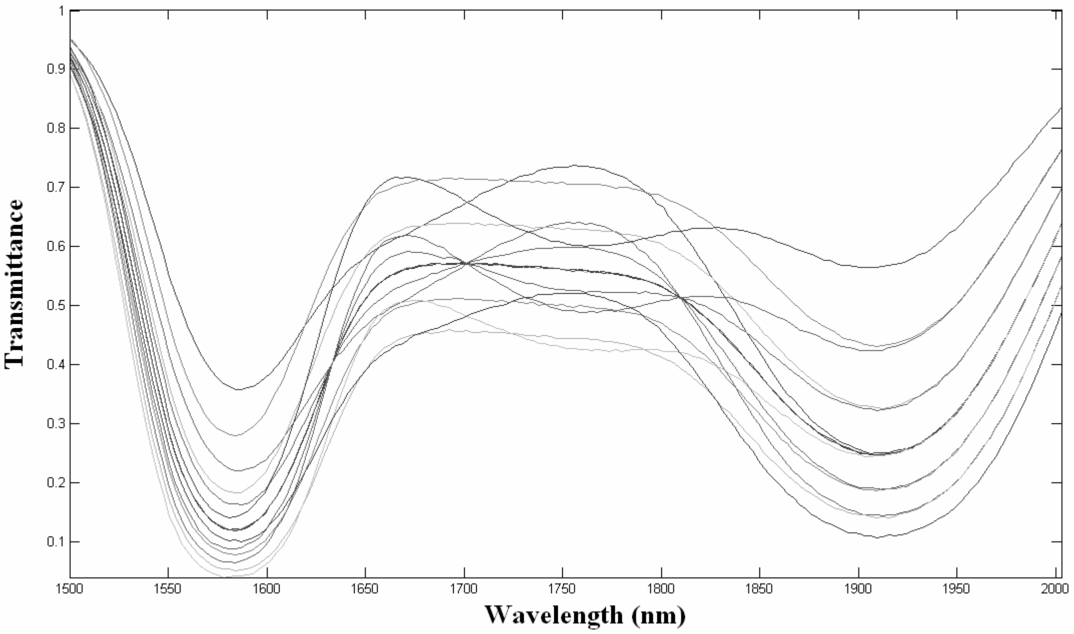


Figure 3.4 Simulated transmission spectra of calibration and validation set.

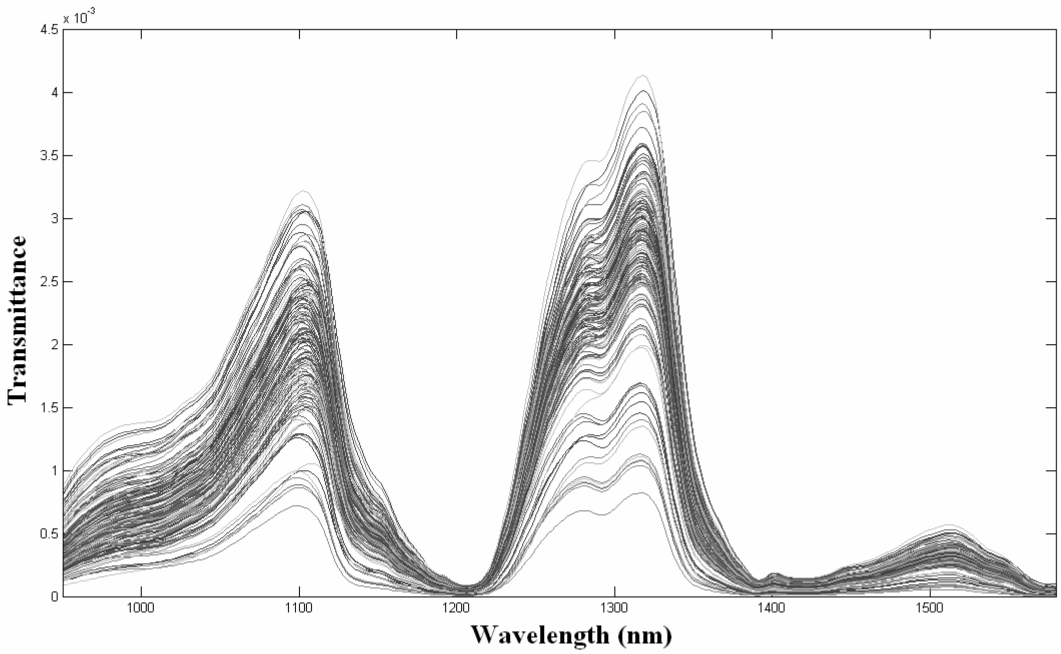


Figure 3.5 Transmission spectra of tablet samples (calibration set).

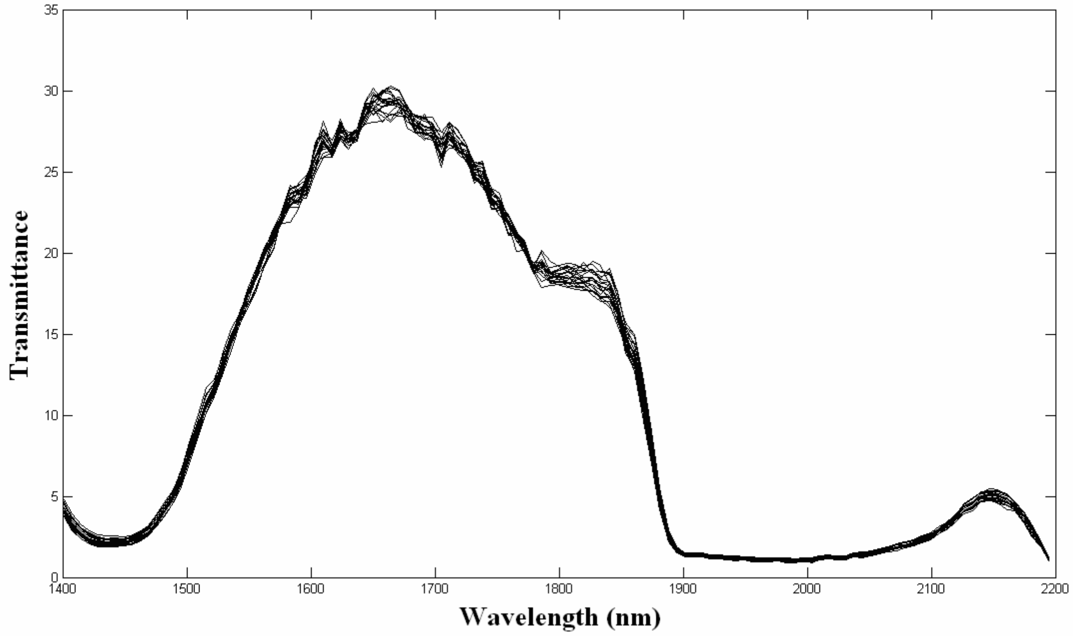


Figure 3.6 Radiometric corrected NIR transmission spectra of ethanol-in-water mixtures.

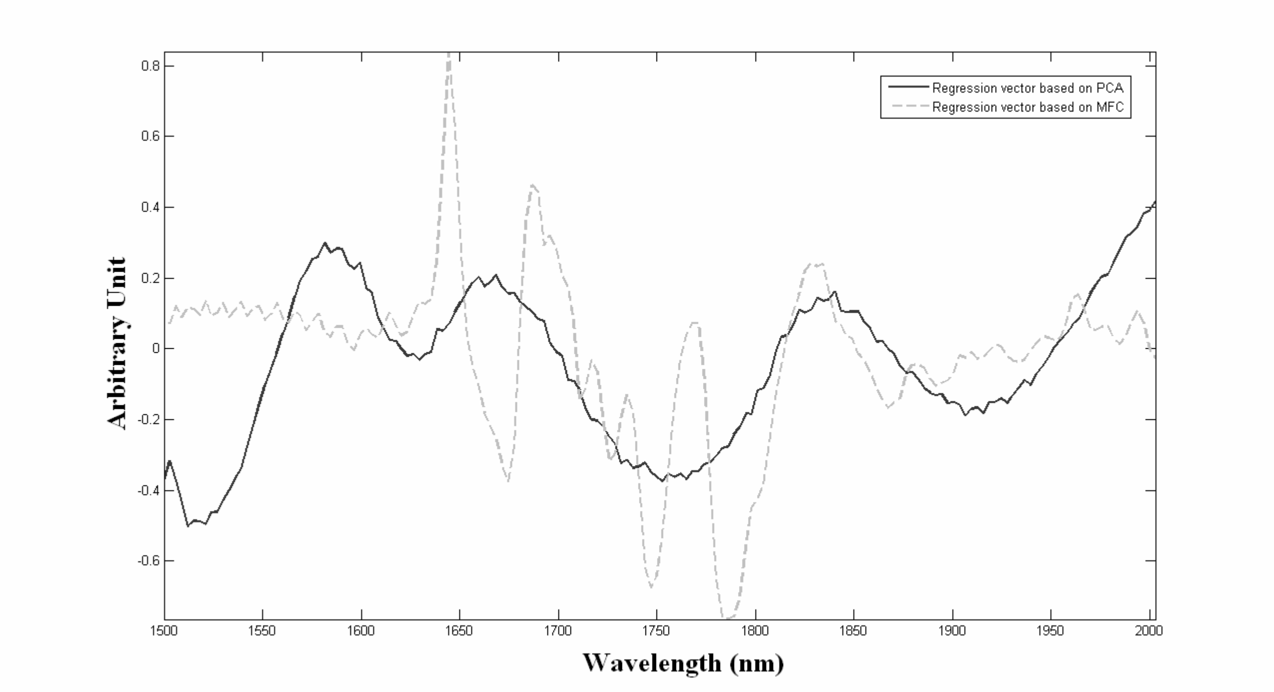


Figure 3.7 The regression vectors for simulated data set. The solid line (—) shows the PCR regression vector based on transmission spectra, dash line (----) shows the regression vector based on MLR calibration by using MFC filter.

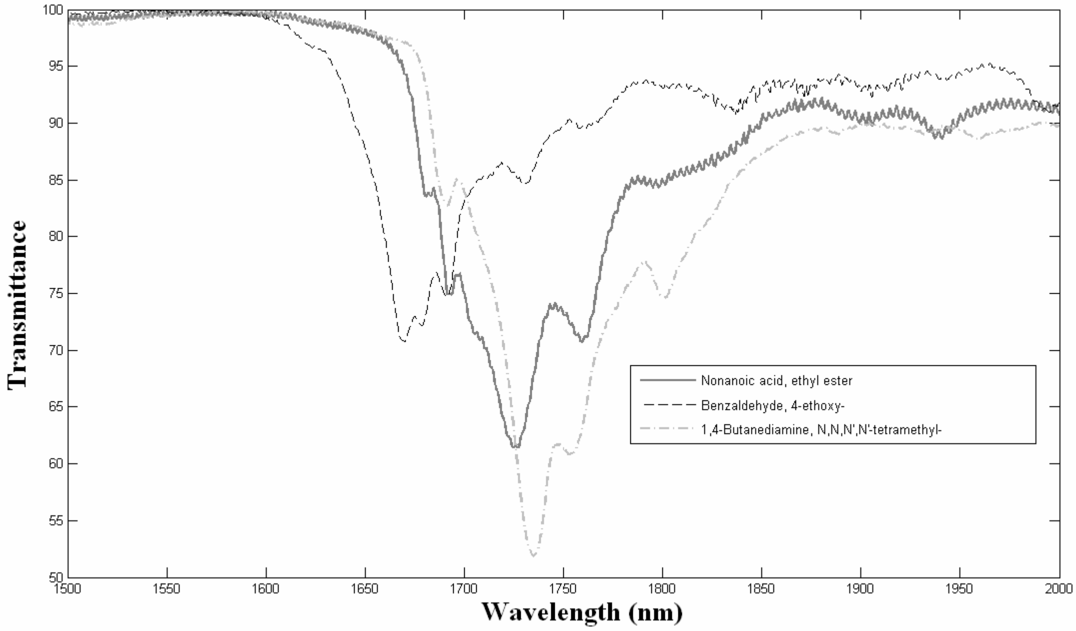


Figure 3.8 The transmission spectra of selected MFC filters for simulated data set.

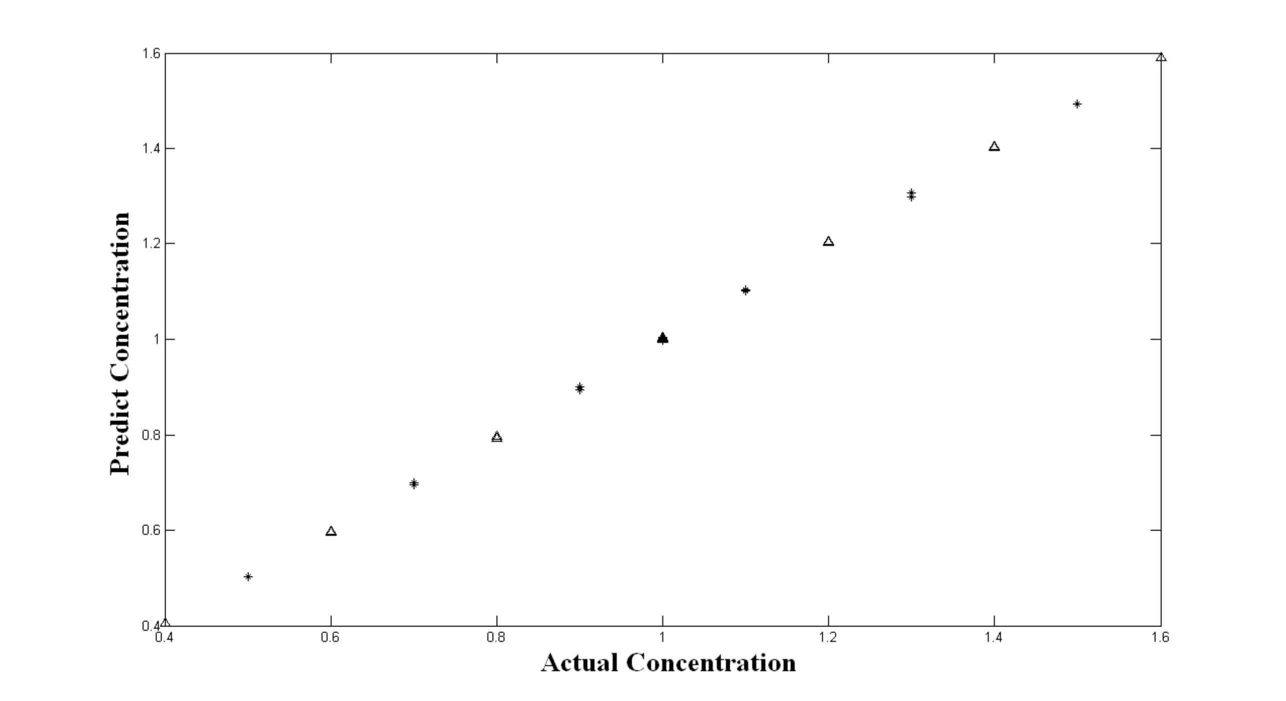


Figure 3.9 The predicted concentrations versus the actual concentrations of both simulated calibration (+) and validation set (△).

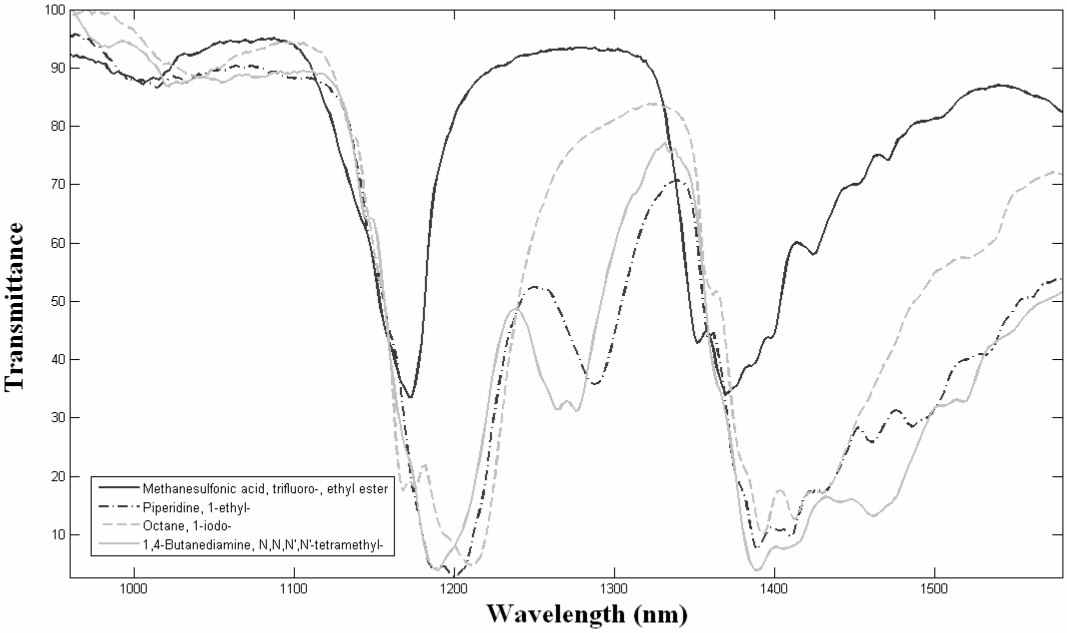


Figure 3.10 The transmission spectra of selected MFC filters for experimental data I.

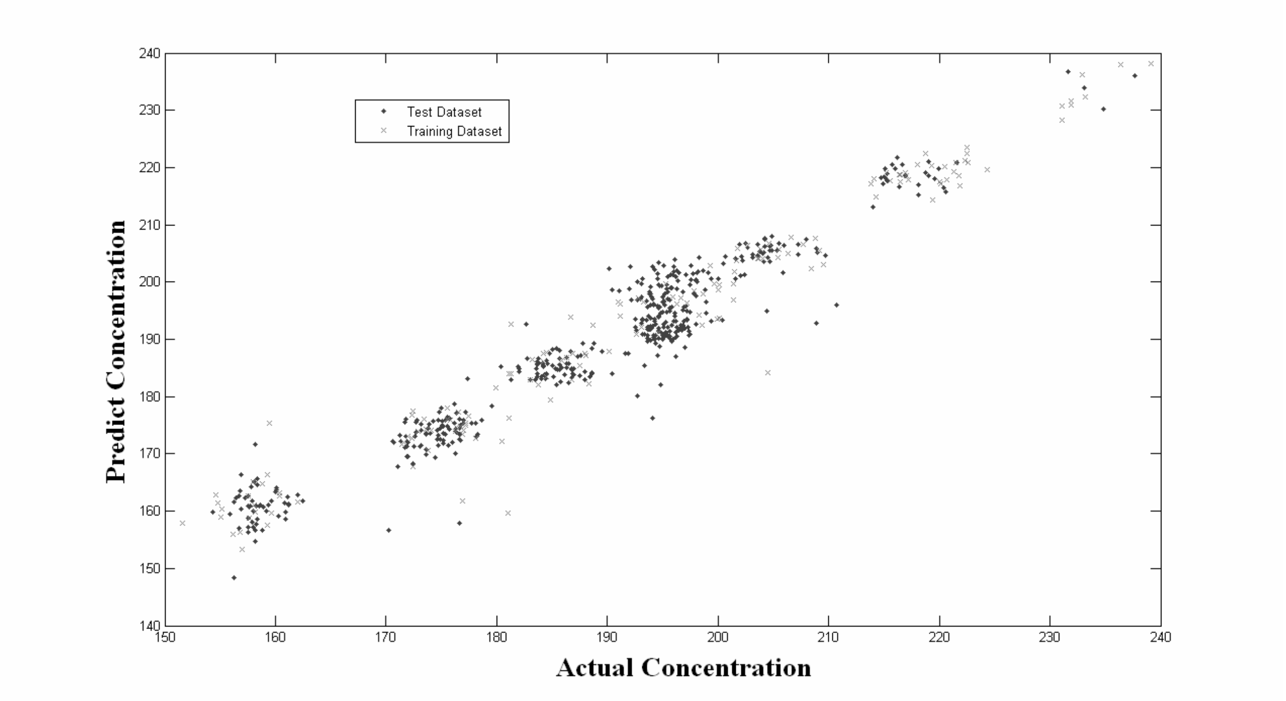


Figure 3.11 The predicted ASSAY concentrations versus the actual ASSAY concentrations of both simulated calibration (*) and test set (•).

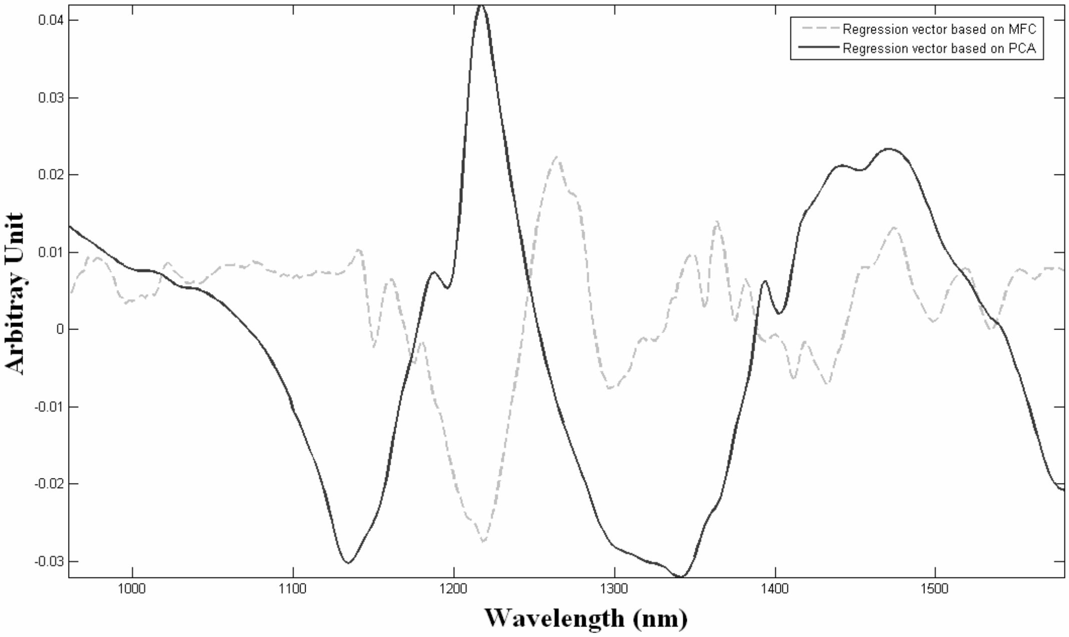


Figure 3.12 The regression vectors for experimental data I. The solid line (—) shows the PCR regression vector based on transmission spectra, dash line (----) shows the regression vector based on MLR calibration by using MFC filter.

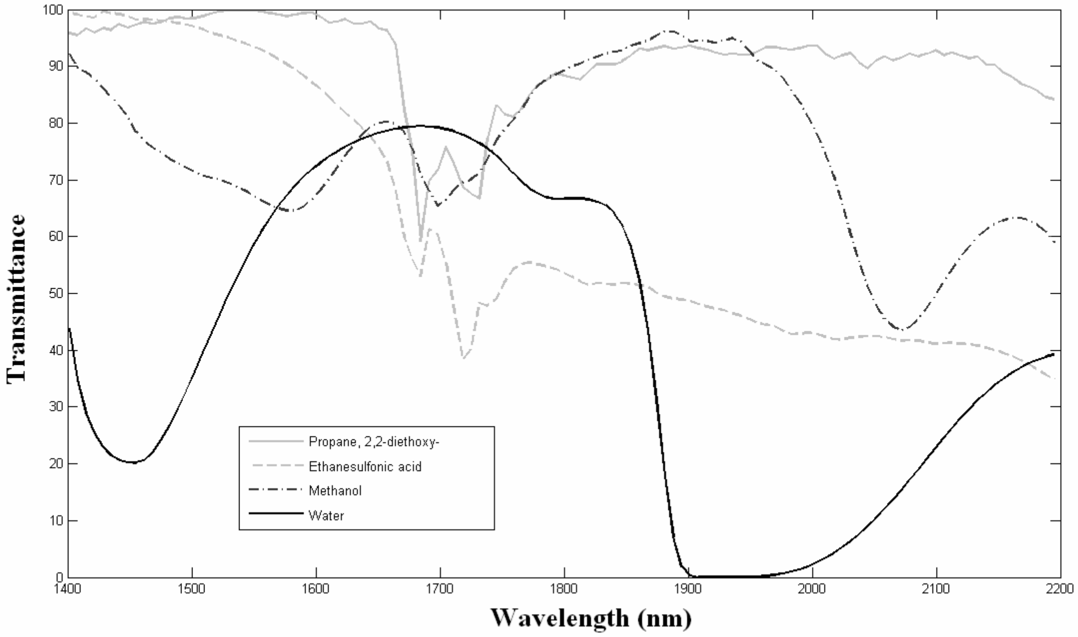


Figure 3.13 The transmission spectra of selected MFC filters for experimental data II.

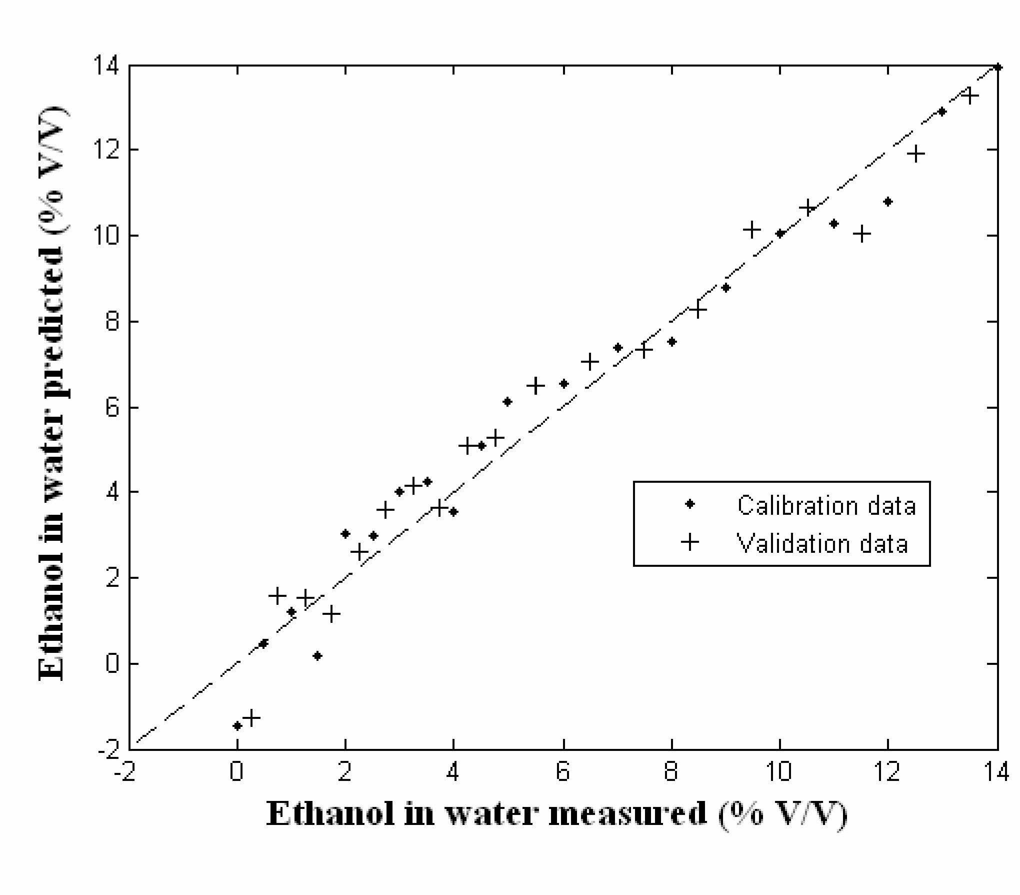


Figure 3.14 The predicted ethanol concentrations versus the actual ethanol concentrations of both simulated calibration (+) and test set (◇).

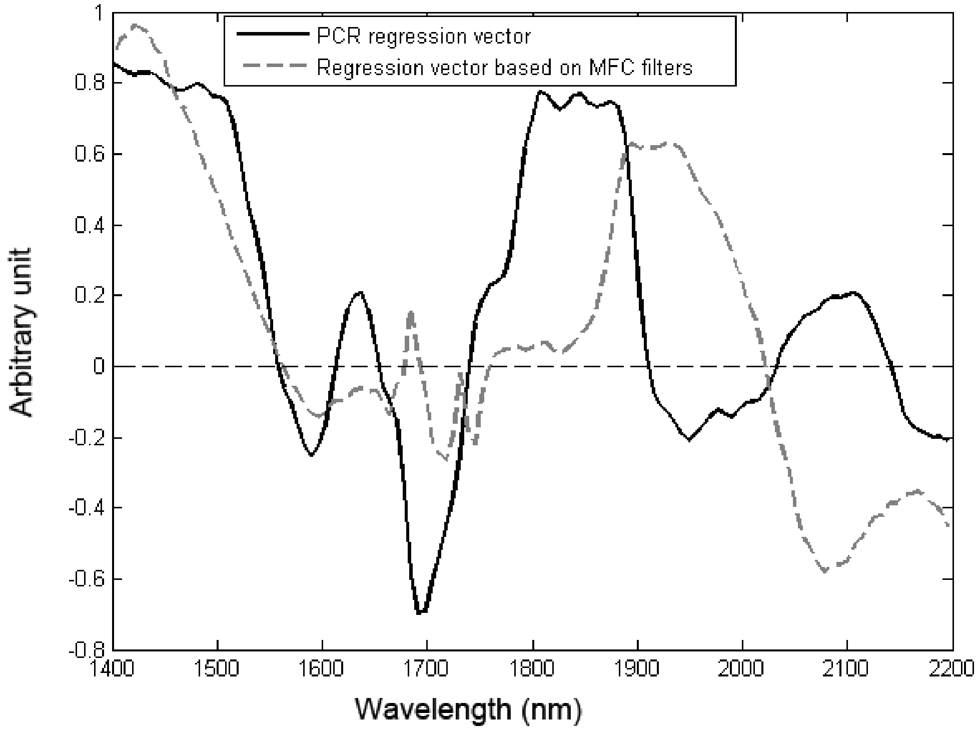


Figure 3.15 The regression vectors for experimental data II. The solid line (──) shows the PCR regression vector based on transmission spectra, dash line (----) shows the regression vector based on MLR calibration by using MFC filter.

Chapter Four – Genetic Algorithm Based Linear Discriminant Analysis for Molecular Filter Selection in Molecular Factor Computing

4.1 Introduction

The application of near infrared spectroscopy (NIRS) as a quantitative and qualitative method is rapidly increasing in pharmaceutical, biomedical and food industry.[21, 22, 69, 86] In addition to easy sample preparation, NIRS is a rapid, robust and noninvasive analytical method.[72, 73] These characteristics make NIRS technique an ideal tool for online process analysis and quality control in chemical and pharmaceutical industries.

Absorption bands in NIR region (1100nm—2500nm) are made up from overtones and combinations of fundamental molecular vibrations in mid-infrared. Unlike the sharp and specific peaks in the mid-IR spectra, the absorbance peaks in the NIR spectra are broad and heavily overlapped. Many different combination and overtone absorptions exist in NIR, thus the spectrum in the NIR region is very complex. Therefore, multivariate statistics plays a critical role in developing the calibration or classification model for the correlation of chemical properties of interest to the complex NIR spectra.

Multivariate calibration methods such as multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS) are routinely used to establish the regression vector that correlates the properties of interest[26, 48, 87]. In a classification scenario where the objective is component differentiation, linear discriminant analysis (LDA) and partial least squares discriminant analysis (PLSDA) [88]are the two major statistical modeling methods.

Modern instruments can acquire a huge amount of data in a short period time due to the rapid sampling and parallelizing of devices. Such advances in analytical instrumentation enable measurement of more properties of interest in many samples, therefore, making the investigation of more complex system in a great depth possible. However, such constantly increasing flood of data generated by modern instruments has posted a challenge for data analysis in many research areas.[89] Reduction of the raw data provided by instrument to the high-level and meaningful information is difficult. New instruments with built-in data reduction capabilities are needed in many medical and military applications where a minimizing computation time is essential.

One approach currently being investigated to perform the data reduction and simplify the spectroscopic instrument includes optical pattern encoding. Theoretical treatment of this approach can be found in literature.[90] Molecular factor computing (MFC), as one of optical encoding methods, has received much attentions due to its simple implementation and extremely high speed.[84]

A critical part of the MFC approach is the molecular filter selection. The molecular filters selection algorithm for obtaining an optimal quantitative calibration model has been addressed in a previous publication[85, 91]. This chapter focuses on the algorithm for finding optimal molecular filters, which lead to a good performance of the classification model. Thus, the aim of this study is to evaluate the practicality of the new proposed algorithm using a genetic algorithm and stepwise selection to select a small subset of the reference spectra library for the final classification model. Both simulated and experimental NIR data sets were used to test the performance of the MFC approach and the new algorithm proposed here.

4.2 Theory and Methods

4.2.1 **Data reduction and classification**

There are many different techniques for classification of data. Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two routinely used techniques for data classification and dimensionality reduction.

PCA is among the most commonly used dimension-reduction techniques in chemometrics. PCA is a bilinear projection method that determines the orthogonal space with fewest dimensions that best describes all of the variations in a variable matrix **X**. For multivariate analysis of spectra, PCA is frequently used to extract spectral features for classification of different materials.

The common route to evaluate principal components is through the singular value decomposition (SVD). For an *n by m* matrix of calibration spectra, ***X***, where n is the number of samples and *m* is the number of wavelength or variables, can be modeled by truncated SVD,

$$X = U_{(n\times a)}\Lambda_{(a\times a)}V^T_{(a\times m)} + E \qquad \textbf{4.1}$$

Λ is the diagonal matrix of the singular values $\boldsymbol{\lambda_i}$, for ***i***=1 to a, which describe the amount of variance. They are sorted by the value of $\boldsymbol{\lambda}$. The coordinates of the sample points projected onto this hyperplane are called scores ***T***, which can be expressed as the product of ***U*** and ***Λ***. The direction of each dimension in the hyperplane is its loading, which can be expressed as $\boldsymbol{V}_{(a\,x\,m)}$. The part of ***X*** that is not explained by the model is the residuals $\boldsymbol{E}_{(n\,x\,m)}$. The scores, loadings and residuals together describe the variation in ***X***.

$$X = T_{(n\times a)}V^T_{(a\times m)} + E_{(n\times m)} \qquad \textbf{4.2}$$

Often, a few principal components are enough to capture the variation, therefore reduce the dimension of the original matrices.

Preliminary research shows that LDA is a computational cheap algorithm that gives an acceptable classification [92]. LDA is aimed to maximize the ratio of between-class variance ($\boldsymbol{S_B}$) to the within-class variance ($\boldsymbol{S_W}$) in the data set by projecting the data onto a lower-dimensional vectors space, thereby guaranteeing maximal separation. The LDA can be formulated by Fisher criterion:

$$D_F(W) = \frac{W^T S_B W}{W^T S_W W} \qquad \textbf{4.3}$$

$$S_B = \sum_C N_C(\mu_C - \overline{x})(\mu_C - \overline{x})^T \qquad \textbf{4.4}$$

$$S_W = \sum_C \sum_{i\in C}(x_i - \mu_C)(x_i - \mu_C)^T \qquad \textbf{4.5}$$

$$\mu_C = \frac{1}{N_C} \sum_{i \in C} x_i \qquad \textbf{4.6}$$

$$\bar{x} = \frac{1}{N} \sum_{i} x_i \qquad \textbf{4.7}$$

where ***W*** is a linear transformation matrix, $\boldsymbol{S_B}$ is the between-class variance, and $\boldsymbol{S_W}$ is the within-class variance. ***C*** is the number of class, ***N*** is the number of population, and $\boldsymbol{N_C}$ is the number of samples in one class.

An intrinsic limitation of classical LDA is that its objective function requires the nonsingularity of the data matrices. For many multivariate data analysis applications, such as spectra-based sample classification, the data matrices under investigation can be singular due to the high-dimensional nature of the data. In high-dimensional space, the number of variables exceeds the number of samples. For example, a typical NIR spectrum (1100nm-2500nm) contains thousands of data points whole usually less than few hundred samples are collected. Such a data matrix is known as *undersampled* or *singular* matrix.

One approach used to solve the singularity problem is PCA-LDA algorithm[93], in which an intermediate dimension reduction is carried out by using PCA. However, because there is no prior information available to determine whether the small PCs are caused by noise or by highly discriminant information, the performance of PCA-LDA varies when the set of PCs selected in the PCA stage varies. In addition, PCA-LDA algorithm requires the computation of Eigen-decomposition of data matrices, which degrades the computing efficiency. For a large dataset such as hyperspectral imaging data, PCA-LDA algorithm can be very difficult to implement within a limited computing time.

4.2.2 **Molecular filters for optical computing**

MFC-based spectral pattern encoding provides a new approach to perform both data reduction and feature selection simultaneously by integrating the data processing into the sensing stage in a MFC-based spectrometer. As schematically demonstrated in Figure 4.1, molecular filters are used as multiplex filter, combined with the light source and detector to make a MFC-based spectrometer. In a MFC-based spectrometer, signal convolution and summation are both effectively performed by directing the multiplexed light through the sample to the detector. The resulting detector response is proportional to the summed intensity of detected photons. For each molecular filter, only one detector response is obtained for each sample. Instead of collecting entire spectra for each sample, only a few detector responses are recorded for each sample, depending upon the number of molecular filters used in the system. With MFC approach, mechanical/electrical/computational construction of a spectrometer is significantly simplified. Because the dimension of new data matrices generated from MFC-based spectrometer is very small, the singularity problem does not exist. More importantly, the MFC-based spectrometer yields a throughput advantage and a multiplex advantage, thereby increasing the signal-to-noise ratio (S/N). MFC also yields the ISP (integrated sensing and processing) advantage, which virtually eliminates the need for computer analysis after data collection.

In a MFC based spectrometer, for each sample, a voltage output of the detector is given by Equation 4.8:

$$V_{out} = G \times \vec{L} \cdot \vec{T} + V_{offset} \qquad \textbf{4.8}$$

where V_{out} is output voltage, G is amplifier gain, $\vec{L}$ is the molecular filter transmission spectrum vector, V_{offset} is voltage offset of detector and $\vec{T}$ is the sample transmission spectrum.

With *m* different MFs, $m\ V_{out}$ are obtained for each sample. The classification of samples is accomplished by constructing a LDA model using the training data set. Based on the LDA model, the class of new samples can be predicted by directly computing classification scores.

$$S_i = C_i + W_{i1} \times V_1 + W_{i2} \times V_2 + ... + W_{im} \times V_m \qquad \textbf{4.9}$$

where the subscript *i* denotes the respective group; the subscripts *1, 2, ..., m* denote the m V_{out} (*m* denotes the number of MFs used in a spectrometer); $\boldsymbol{C_i}$ is a constant for the i^{th} group, $\boldsymbol{W_{ij}}$ is the weight for the j^{th} V_{out} in the computation of a classification score for the i^{th} group; $\boldsymbol{V_j}$ is the observed value of the j^{th} V_{out}. $\boldsymbol{S_i}$ is the resultant classification score.

4.3 Algorithm

4.3.1 Selected molecular filters for optical computing

In the MFC approach to predict the chemical properties of samples, molecular filters used as multiplex filter, combined with the light source and detector comprise a MFC-based spectrometer. Molecular computing of vectors by using molecular filters as transformation matrices enables spectra to be represented in any arbitrary coordinate system. The coordinate systems are selected for specific reasons. New coordinates can be used to reduce the dimensionality of the spectral hyperspace, to increase the orthogonality of a subspace of the coordinate system, or to simplify the mechanical/electrical/computational construction of a spectrometer. In the molecular filter selection algorithm described in this chapter, the new coordinate systems are selected to construct a simple and robust classification model.

4.3.2 **Molecular filters selection by spectral library search**

The selection of molecular filters from a reference spectra library is carried out in a manner that maximizes the separation among groups (dispersion of the centroids of the groups). The classification model is obtained by maximizing the ratio of the among-object-group dispersion over the pooled within-object-group dispersion. A schematic overview of the algorithm for selecting MFs from a large spectral library (1921 spectra) is illustrated in Figure 4.2. Three main steps of the algorithm are library scores generation, genetic algorithm based LDA, and stepwise variable selection.

4.3.2.1 Library scores generation

First, transmission spectra of samples obtained from a conventional spectrometer were convoluted with the emission spectra of light source, $\vec{I}$ and the sensitivity of detector, $\vec{R}$, to get a correct transmission spectra of samples, $\hat{T}$.

$$\hat{T} = \vec{S} \times \vec{T} \times \vec{R} \qquad \textbf{4.10}$$

Then the corrected transmission spectra of samples, $\hat{T}$, were multiplied (dot product) by transmission spectra of library, *L*, to give a simulated score, *S,* which represent the voltage responses of detector.

$$S = \hat{T} \cdot L \qquad \textbf{4.11}$$

With a library containing *l* components, and a training sample set with *n* spectra, the resultant simulated scores *S* are a $n \times l$ matrix.

4.3.2.2 Genetic algorithm based linear discriminant analysis

After the simulated library scores were obtained, genetic algorithm based linear discriminant analysis (GALDA) was employed to search the simulated scores space (*l* dimensions) to find a set of optimal variables as molecular filter candidates.

A genetic algorithm (GA) is a heuristic search process based on natural genetic selection.[58] It has proven to be very useful on large search space (i.e. spectra feature extraction).[33, 40, 56, 57, 94] In GA, the initial step is to generate a random population consisting of a predefined number of individuals and variables. At each step, genetic operators (reproduction, crossover and mutation) are applied to individuals in the pool to create the new population after the fitness function (the ratio of the Mahalanobis distance within group and Mahalanobis distance between groups) is evaluated. The genetic algorithm selects individuals from the current population to be parents and uses them to produce the children for the next generation until a particular stop criterion, such as a predefined fitness value or maximum computing time, has been reached. Over successive generations, the population "evolves" toward an optimal solution. The genetic algorithm is typically useful for solving a variety of variable selection problems that are not well suited for standard optimization algorithms.

It must be taken into account that GA based LDA could lead to an overfitted classification model, since the variable-to-sample ratio of the data matrices is high in this study. So the GALDA is used here as a variable screening tool instead of the final variable selection. The GALDA runs 1000 times, and 200 variables out of 1921 are selected as molecular filter candidates based on the frequency of being selected. After the GALDA step, to decide the set of variables to be included in the final model, a stepwise variable selection approach is used. The stepwise LDA is used to select the optimal variable set out of 200 variable candidates.

4.3.2.3 Stepwise variables selection

Stepwise variable selection[60] is often used where a large number of dependent variables are available but only a few variables are desired to be included in a discriminant model for groups separation. The initial model is defined by either the provided starting variables or randomly chosen variables. In every step new models are generated by including every single variable that is not in the model, and by excluding every single variable that is in the previous model. Then the resulting model performance is tested by cross-validation in the training dataset, and if the model performance is improved, then the corresponding variable is in- or excluded. The stepwise variable selection process stopped when a satisfactory classification result has been found or when the model can no longer be improved by stepwise selection.

The validation of the selected variables was done by both an internal leave-block-out cross validation, in which a number of samples were leave out each time then the model was evaluated on the reminded samples, and by an external validation, in which the model was evaluated using an external data set.

4.4 Experimental Section

4.4.1 NIR spectral library

The chemicals severed as MFC filters were found by searching a library of near-IR transmission spectra (John Wiley & Sons, Inc.) The near-IR library of 1923 liquid compounds contains substituted aromatic compounds, aliphatic alkanes, alcohols, and other carbonyl groups, most of which exhibit remarkable absorbance in NIR region. The library consisted of two spectra of each compound collected on slightly overlapping NIR region, 952-1587 nm and 1388-2630 nm. Depending upon the applications, different region of spectra in the library were used.

4.4.2 Simulation dataset

Simulations were run to generate simulated transmission data to form the training validation and the test dataset. Data generated through the simulations were then used to test the filter selection algorithm. Each simulated absorbance spectrum was synthesized by summing four randomly weighed Guassian bands. For each group, fifty spectra were generated, resulting in 150 spectra in total for 3 groups. The transmission spectra were generated by log-transfer of the absorbance spectra of mixtures. Normally distributed random number with a mean of zero and standard deviation of 0.1%T were added to transmission spectra as simulated measurement noise. The wavelength range from 1500 to 2000nm was assigned as arbitrary wavelength to the spectra. The resulting transmission spectra are shown in Figure 4.3. Transmission spectra dataset were randomly split into calibration set (120 spectra) and test set (30 spectra) to test the algorithm.

4.4.3 **Experimental dataset**

Diffusion reflectance spectra of different compounds under the aliquots of rabbit red blood cell (RBC) solutions were collected from 1100-2500 nm using a scanning monochromator (Bran+Luebbe InfraAlyzer 500, Germany) with an external reflectance probe. The samples were pure cholesterol (Sigma, St. Louis, Mo), collagen (Type I, Sigma) and elastin (Sigma) in cylindrical recesses (1 cm diameter, 2 mm deep). The path length through RBC solutions was varied from 0.1 to 1.0 mm to generate the spectral variation. After close examination of the spectra, the spectral region 1150-1300 nm showed significantly different degrees of reflectance. The reflectance spectra and the second derivate spectra are shown in figure 4.4 and figure 4.5, respectively. So it was decided to use this spectral band in the MFC chemical selection routines. Molecular factor scores were simulated by calculating the dot products between transmission spectra from the NIR library (1150-1300 nm) and the raw NIR reflectance spectra (1150-1300 nm) collected using the scanning monochromator.

4.4.4 **Data analysis**

All data analysis was carried out by using Matlab 7.0 (Mathworks, Inc., Natick, MA). The PLS toolbox v3.51 for Matlab (Eigenvector Research, Inc. Wenatchee, WA) was used for multivariate analysis. The genetic algorithm toolbox and GALDA toolbox[56, 57] for Matlab were employed to develop the molecular filter selection algorithm.

4.5 Results and Discussion

A log transformation of the reflectance *R* or transmittance *T* to absorbance *A* linearizes the reflectance or transmittance measurement in near-IR spectroscopy, which is important in quantitative measurements. However, in classification applications, only qualitative information is concerned, it is not absolutely necessary to use absorbance as the data representation. Using reflectance or transmittance data is appropriate for MFC based spectroscopy because the usage of absorbance as spectra representation is difficult due to the radiometric nature of MFC based spectroscopy.

The performances of both the PCA-LDA models and the MFC approach were compared by examining the classification accuracy as defined in the following equation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad \textbf{4.12}$$

where *TP* is True Positive, *TN* is True Negative, *FP* is False Positive, *FN* is False Negative.

4.5.1 **Classification performance with full spectra using PCA-LDA**

In this study, PCA-LDA models were constructed for both the simulated data and the experimental data to estimate classification performance for comparison with the performance of MFC approach.

PCA was first applied in the simulated transmission spectra (150 spectra). Eight principal components are included to capture 93.5% of the variation. LDA was then carried out in dimension reduced dataset (150 × 8). For simulation data, which consists of three different groups, an LDA model includes two eigenvectors that define a two-dimensional canonical space. The projection of high-dimensional PC space to two-dimensional canonical space is done by multiplication of the data matrix with the LDA eigenvectors. The classification result is demonstrated in Figure 4.6. It is found that three groups are well separated, although a small cluster overlap exists between group 1 and 3. Leave-block-out cross validation was used to validate the LDA model. Mahalanobis distance is used to calculate the probability that a leave-out sample belong to one of the three groups. A leave-out sample that has the smallest Mahalanobis distance to the group center is assigned to that group accordingly. The cross validation classification results for simulation dataset are listed in Table 4.2.

For experimental dataset, LDA is applied after PCA. The same modeling method is used as it does in simulation dataset. The classification result shown in figure 4.7 demonstrated that the samples can be clearly separated with full spectra data using PCA-LDA. The cross validation classification results are listed in Table 4.2.

4.5.2 Classification performance with MFC approach

The molecular filter selection algorithm described in the algorithm section was implemented on both simulated dataset and experimental dataset. The classification results based on selected MFs are shown here to give a comparison of PCA-LDA approach. Seven chemicals were chosen from the reference NIR spectra library as molecular filters. These molecular filters were methyl-oxirane, 3-bromo-1-propene, 3-methyl-benzenamine, 1-cyclopropyl-ethanone, 2,5-dimethyl-pyrazine, ethenyl-benzene, cyclooctane. Their spectra are shown in Figure 4.8. The LDA model using these molecular filters provided good classification among the groups in the simulated data as demonstrated in Figure 4.9. The classification result shown in Figure 4.9 is better than PCA-LDA model. Leave-block-out cross validation was used to validate the LDA model, and the classification results are listed in Table 4.2

For the experimental dataset, eight molecular filters were chosen to give a satisfactory LDA model. The eight molecular filters are phenyl-hydrazine (PH), 2-dimethylbutyric acid (DMBA), 2,5-dihydrofuran (DHF), ethyl iodoacetate (EIA), tetramethylurea (TMU), thiopheneethanol (TE), 2,2-diethoxypropane (DEP), dicyclohexyl phthalate (DCP). Dicyclohexyl phthalate (DCP) was in powder form and dissolved in CCl_4 (0.5 g/mL). Their transmission spectra are shown in Figure 4.10. The classification result shown in Figure 4.11 demonstrated that, in principle, MFC approach using eight chemical filters should be able to achieve similar classification results.

To demonstrate that the algorithm described above is not just a theoretical simulation, an actual MFC-based NIR spectrometer was constructed in lab to test the performance of the prototype instrument in differentiating cholesterol, collagen and elastin through red blood cell solutions. Figure 4.1 illustrates the schematic diagram of the prototype instrument. Details of the prototype instrument were described in a previous publication.[95] Eight molecular filters chosen by the MFC selection algorithm were employed in the instrument as multiplex filters. Three different components were immersed in red blood cell solution for analysis. The distance between the fiber optic termination and the target surface was varied from 0 and 0.5 mm to introduce the path length variation. Eighteen measurements of each component were analyzed resulting in 54 total spectra. Each sample was represented by eight-point score spectra because eight molecular filters were used.

The LDA classification result based on eight-point score spectra is shown in figure 4.12. A slightly difference is found when the results of actual experiments are compared with the simulation study. The factors that likely contribute to the disparity between the actual experiment and simulation study were addressed in a previous publication.[95] Due to the small sample size, a leave-one-out cross validation was carried out, and the classification error rates are listed in Table 4.3.

4.6 Conclusion

This research demonstrated a novel library search method using genetic algorithms that, coupled with multivariate modeling methods to select molecular filters from a spectra library, permits qualitative classification of samples with different chemical species of interest.

The performance of LDA classification based on the MFs selected by the algorithm was compared to PCA-LDA classification based on full spectra of samples. Similar results were obtained by both methods; however, the MFC approach yields more advantages that include higher S/N and data reduction. This novel molecular filter selection algorithm has been successfully tested on both simulation and experimental datasets. A prototype MFC type NIR spectrometer whose design is generated from this algorithm has been built and tested on discriminating cholesterol, collagen, and elastin samples through red blood cell solutions. Good classification accuracy provided by this method proved the success of the algorithm.

Chapter Four Tables

Table 4.1 Parameters of the GALDA

* Variables: simulated library scores *S*.
* Variable size: 1921 (simulated library score)
* Population size: 30 chromosomes
* Model method: LDA
* Maximum number of variables selected in the same chromosome: 8
* Probability of mutation: 1%
* Crossover fraction: 50%
* Number of runs: 1000

Table 4.2 Comparison of the classification performance using PCA-LDA and MFC-LDA for different dataset.

Accuracy	Group 1	Group 2	Group 3	Num of PCs	Num of MFs
Simulation Dataset (PCA-LDA)	96%	100%	96%	8	
Simulation Dataset (MFC-LDA)	100%	100%	100%	N/A	7
Accuracy	Cholesterol	Elastin	Collagen	Num of PCs	Num of MFs
Experimental Dataset (PCA-LDA)	100%	100%	100%	10	N/A
Experimental Dataset (MFC-LDA)	100%	100%	100%	N/A	8

Table 4.3 Classification performance of experimental MFC data acquired by the prototype MFC spectrometer.

Cholesterol	Elastin	Collagen
94.4%	90.7%	75.9%

Chapter Four Figures

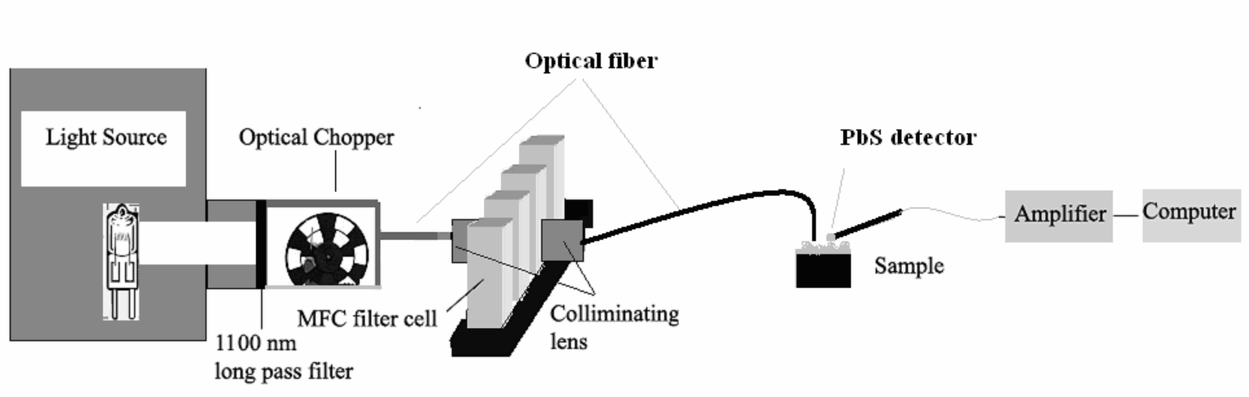


Figure 4.1 Schematic diagram of the MFC NIR spectrometer (diffusion reflectance mode).

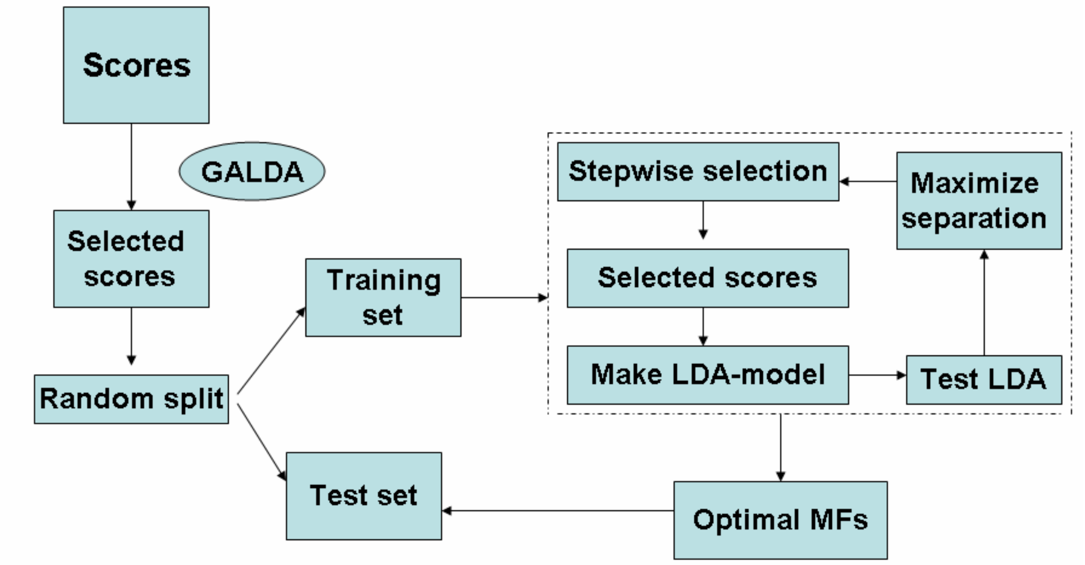


Figure 4.2 Schematic view of all the steps involved in the molecular filters selection algorithm.

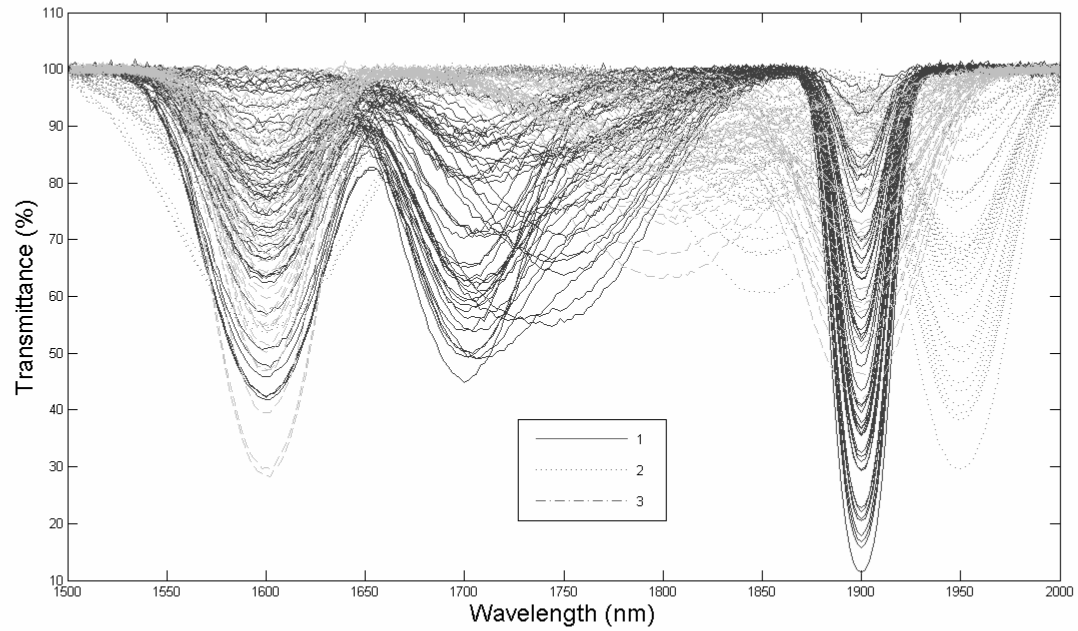


Figure 4.3 Transmittance spectra of simulated spectra. The simulated spectra include 150 spectra made of 3 different groups.

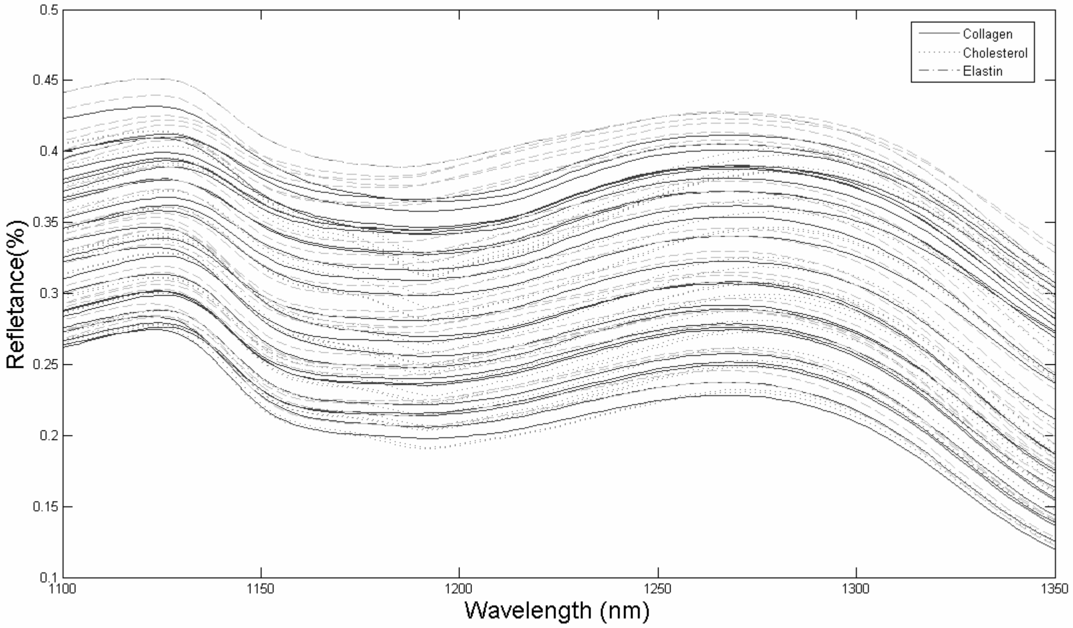


Figure 4.4 Reflectance spectra of cholesterol, collagen and elastin through red blood cell solutions.

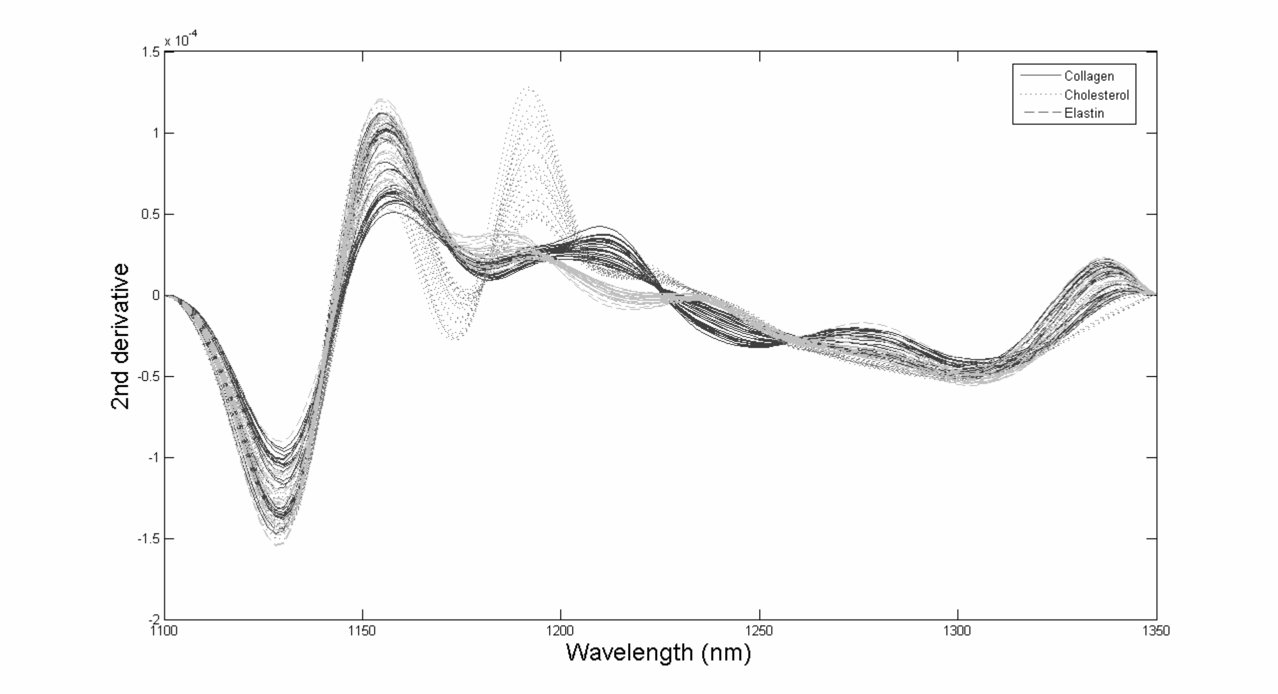


Figure 4.5 Second-derivate spectra of cholesterol, collagen and elastin through red blood cell solutions. The spectral region covers 1150-1350 nm.

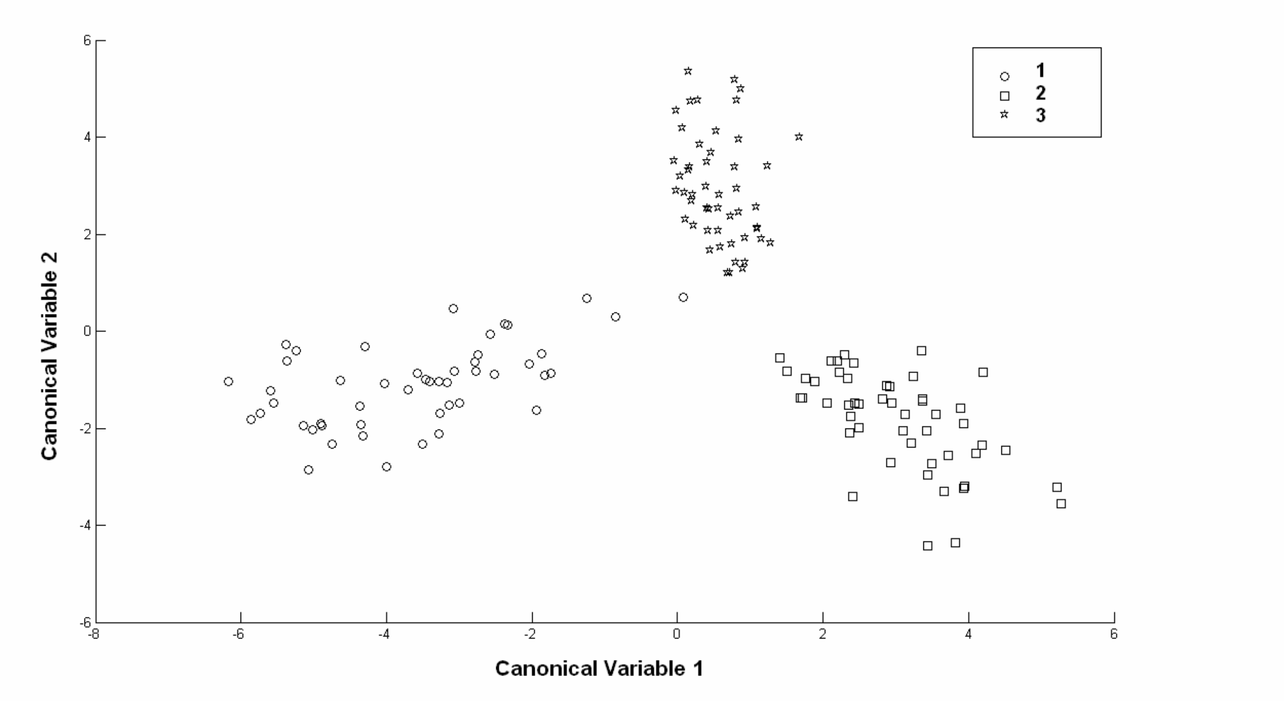


Figure 4.6 Canonical variables from simulated transmittance full spectra. PCA was first applied to the spectra, eight principal components were included, and LDA was followed.

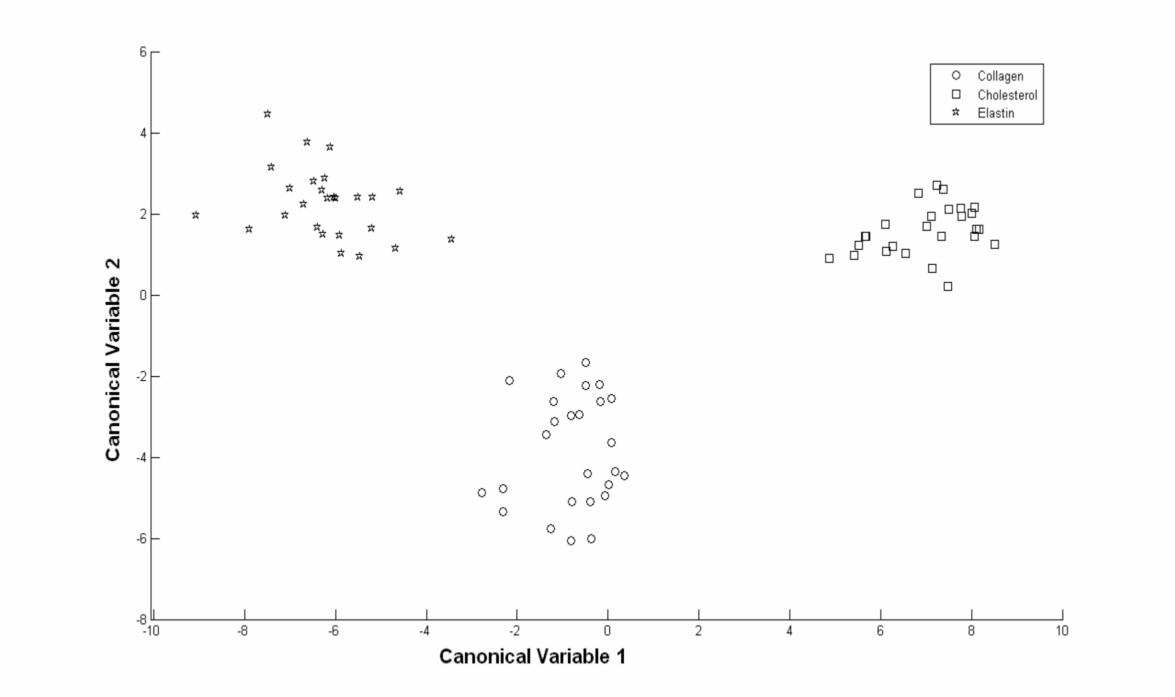


Figure 4.7 Canonical variables from the near-infrared full spectra of the sample targets in the 1150-1350 nm range.

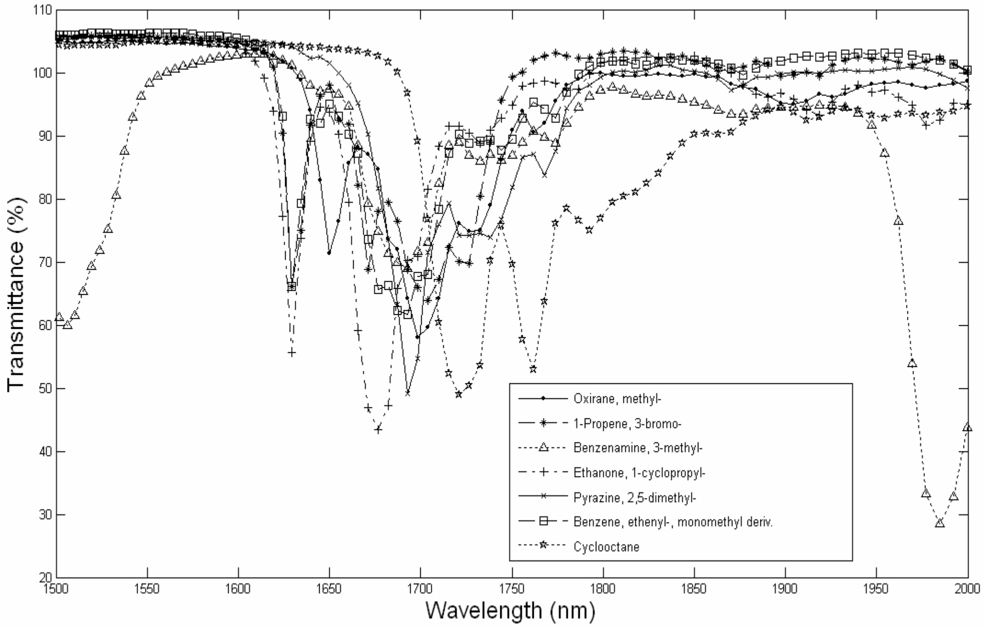


Figure 4.8 Transmittance spectra of selected molecular filters for classification of simulated spectra data.

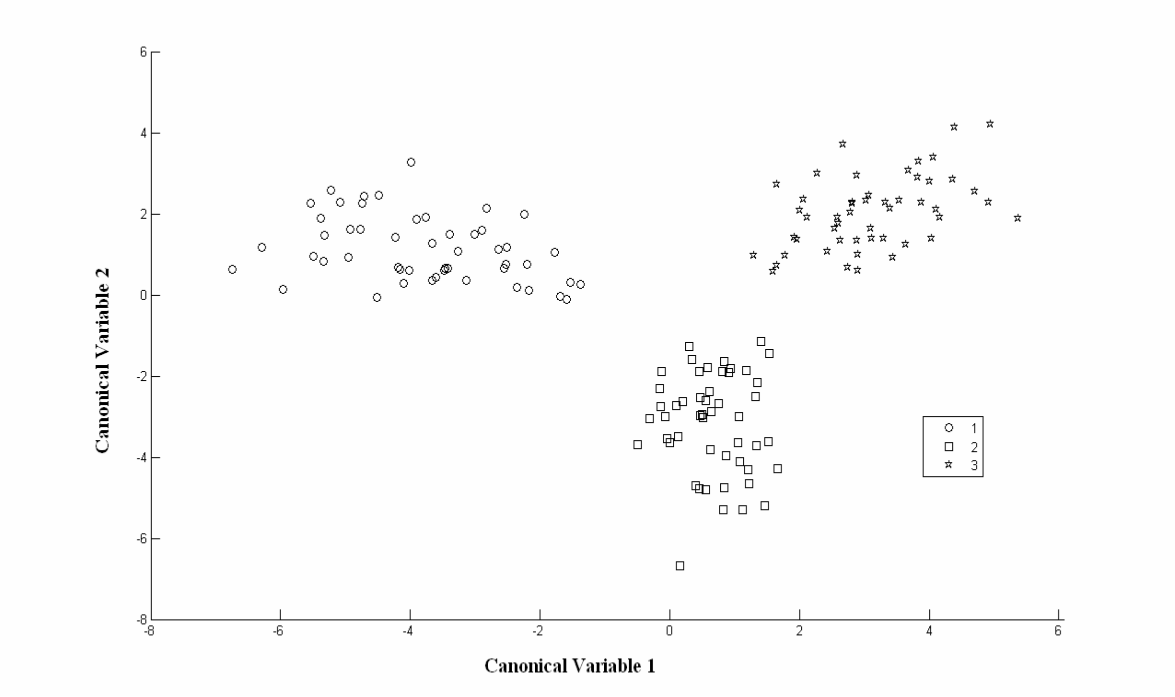


Figure 4.9 Canonical variables from selected seven MFC score of simulated spectra.

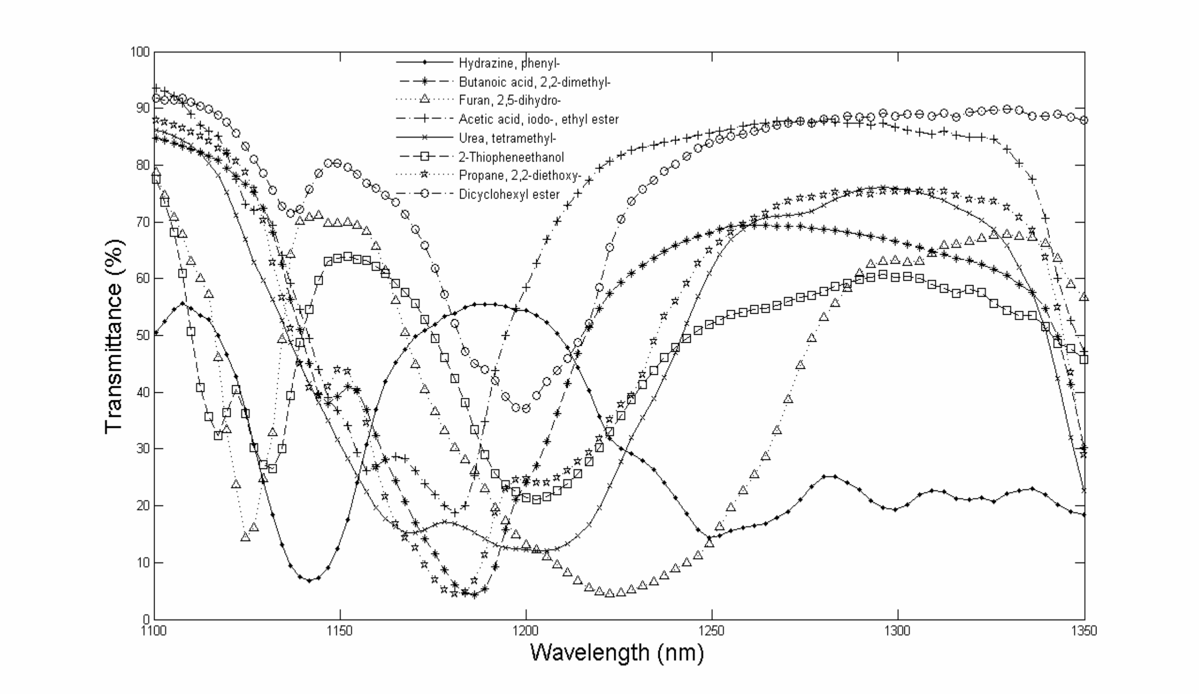


Figure 4.10 Transmittance spectra of selected molecular filters (MFs) for classification of cholesterol, elstin, and collagen through red blood cell solutions.

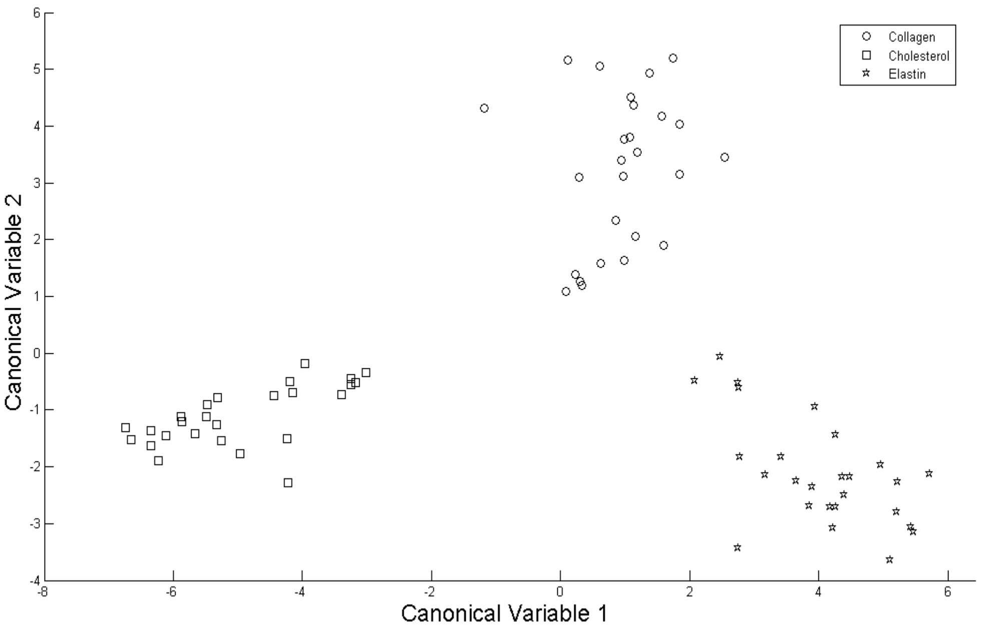


Figure 4.11 Canonical variables from the simulated MFC data of the eight chemical optical filters.

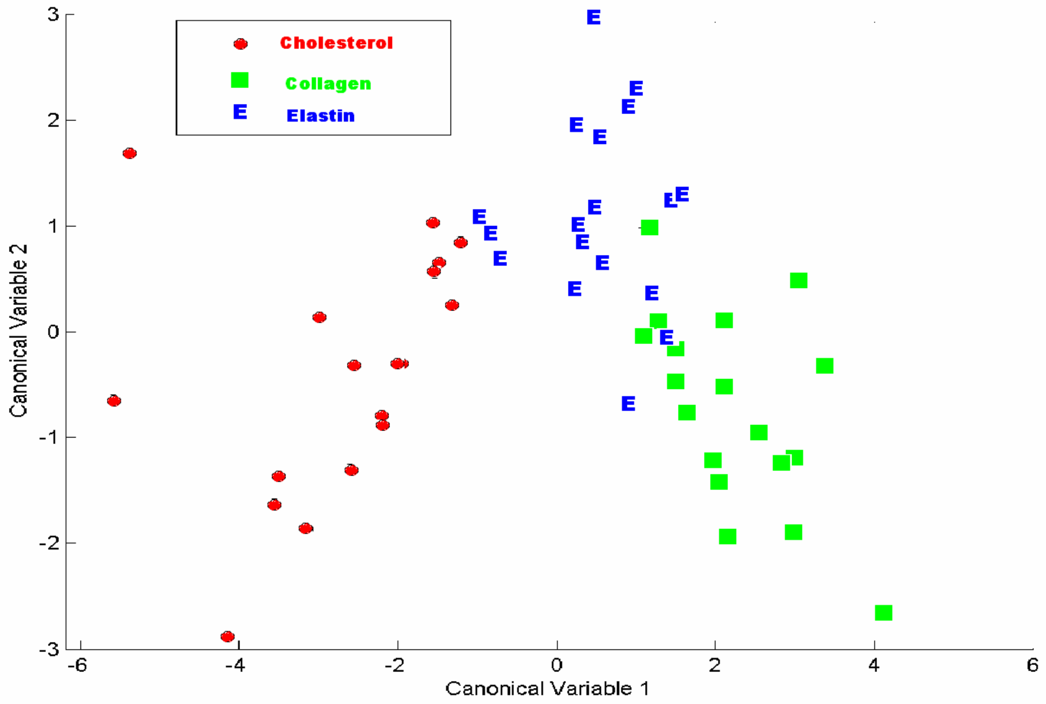


Figure 4.12 Canonical variables from the experimental MFC data using the eight selected chemical optical filters. (The data published in Urbas et. Al. [95] is used for generating this figure).

Chapter Five – Molecular Factor Computing for Predictive Spectroscopy

5.1 Introduction

Near infrared spectroscopy (NIR) has become an important process analytical method for simultaneous multicomponent chemical analysis. NIR has found many applications in process environments and in measurements in the biotechnology and pharmaceutical industries[20, 71-73, 96, 97] where NIR spectroscopy provides online, nondestructive and noninvasive sensing. In September of 2004, the US FDA released a Guidance for Industry, PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance.[98, 99] This guidance is designed to facilitate innovation in, process development and quality assurance. Process Analytical Technology (PAT) will help in better design, monitor and control of pharmaceutical manufacturing process by integrating multivariate modeling, sensors design and process optimization with the goal of ensuring final product quality.[100]

Industrial environments are usually less friendly to analytical instrumentation than research laboratories. Filter instruments are usually much more stable and rugged than their dispersive or interferometric counterparts, making them ideally suited for the harsh conditions found in industrial environments.[66, 74] Multivariate calibration is a well-established tool in chemometrics for analysis of NIR, UV-vis, and Raman spectra. Conventional measurement of chemical or physical properties from spectra is carried out by constructing a predictive model. [26, 48, 101] Two of the most commonly used methods to construct a predictive model are partial least squares

(PLS) and principal component regression (PCR). In a conventional spectrometer with typical chemometrics, data collection and processing of raw data can be time consuming and computationally expensive, especially when spatial relationships (image data) are required. Methods for selecting small but highly relevant variables to represent the original data in a reduced coordinate space and methods for integrated sensing and processing (ISP) are therefore receiving much attention.[56, 57]

ISP aims to design and optimize sensing systems that integrate the traditionally independent units of sensing, signal processing, communication and targeting. By employing ISP, computational complexity within traditional sensing system has been substantially reduced through determining efficient low-dimensional representations of those sensing problems that were originally posed in high-dimensional settings by traditional sensing architecture. Successful ISP is expected to yield entirely new ways of designing and operating sensor systems.[89]

One approach currently being investigated to simplify both instrumentation and computational analysis involves optical pattern encoding.[78] This technique involves tailoring the optical spectrum of filters to encode high level information about the samples in sensing stage. Theoretical treatment of this methodology can be found in the literature.[79, 102] Myrick et al. have demonstrated some practical applications of this methodology in UV-visible and NIR spectroscopy.[76, 77, 80-83, 103, 104] Encoding applications were based on the fabrication of thin film solid-state optical filters, termed multivariate optical elements (MOEs). MOEs were designed to replicate the multivariate regression pattern by transmitting and reflecting weighed optical signals over a broad wavelength band.

Recent publications from our laboratory have offered an alternative approach for spectral encoding .[84, 105] Molecular absorption filters can be used as mathematical factors in spectral encoding to generate a factor-analytic optical calibration in a high-throughput spectrometer, which we term molecular factor computing (MFC). The molecules in the filter effectively compute the calibration function by weighing the signals received at each wavelength over a broad range of wavelengths. One or more molecular filters are used in MFC-based spectrometer to produce detector signals correlated to desired sample information. Advantages of this new approach over conventional spectroscopy include significantly reducing the computational demand (the integrated sensing and processing, or ISP, advantage), shorter data collection and analysis time with higher signal-to-noise ratio (S/N) (especially for imaging spectrometry, through the Fellgett advantage), higher optical throughput (the Jacquinot advantage), and more rugged instrumentation with a considerably lower cost.

This chapter describes the instrumentation and application of a molecular factor computing-based spectrometer. Such a spectrometer may be particularly useful in applications where real-time video analyses of remote sensing data are required. In such cases, molecular filters placed in front of near-IR cameras would produce images in which the intensities were proportional to the factor scores, without the need for additional computation. Ethanol in water mixtures were selected as training and validation samples to design molecular filters that would test the concept of MFC-based spectroscopy. Ethanol is used in liquid pharmaceuticals to enhance solubility, for example. Ethanol is also sometimes abused in the general population. Sensing alcohol in the environment is necessary in such an application to evaluate the effectiveness of pharmacotherapy or other therapies for alcohol abuse.

5.2 Materials and Methods

5.2.1 Traditional NIR training spectra collection

The ethanol was reagent grade, obtained from AAPER (Shelbyville, KY). The water was distilled in house. Quantitative mixtures of alcohol and water were prepared by volume using grade-A volumetric flasks and burettes. Twenty samples were prepared with ethanol concentrations ranging from 0% to 14%. Sample solutions were placed in a quartz cuvette with a path length of 1 mm. Conventional near-infrared transmission spectra for comparison with MFC were collected using a dispersive spectrometer (Ocean Optics NIR256 temperature-regulated NIR, Dunedin, FL) over the wavelength range of 900-2500 nm to acquire a total of 256 data points per spectrum. Data analysis was limited to 1400-2200 nm to avoid the short wavelength region, which was not used for the MFC tests. As a result, the selected transmission spectra included 128 data points. The sample temperature remained constant at 25 ^{o}C during the data collection period. Each recorded spectrum was the average of 10 scans, with the total integration time ca. one second. The transmission spectra are shown in Figure 1a. To calculate the required composition of MFC filters for ethanol determinations, full spectra of ethanol/water mixtures over the wavelength range of interest must be available. For maximum accuracy, these spectra must represent the optical characteristics of the MFC spectrometer, not the conventional instrument. As a result, the transmission spectra of the dispersive spectrometer were convolved with the transmission spectra of a 1400-nm long pass filter, the emission spectrum of the tungsten NIR source, and the response curve of the InGaAs photodiode in the prototype instrument to give a corrected representation of the MFC instrument response. These corrected transmission spectra were used as training spectra for MFC filters selection and multivariate analysis. The corrected spectra are presented in Figure 5.1b.

5.2.2 **MFC-based high throughput NIR spectrometer**

A graphic representation of the instrumental setup is given in Figure 5.2. A 12V, 100W tungsten-halogen broadband source (model 621, McPherson Inc., Chelmsford, MA) with 1400-nm long pass filter (Thorlabs, Newton, NJ) was used as the source of broadband NIR light. The tungsten-halogen light source has more intense radiation in the shorter NIR wavelength region. To avoid saturating the detector with short wavelength NIR radiation that contains little chemical information about the samples, the 1400 nm long pass filter was used to block the short wavelength radiation. The source beam was modulated with an optical chopper (Model SR540, Stanford Research Systems Inc., Sunnyvale, CA) at a frequency of 280 Hz. The light beam was focused onto an InGaAs photodiode (Fermionics Opto-Technology, Simi Valley, CA) through a convex lens after passing through the molecular filter cuvette and sample cuvette. A step-indexed sliding cuvette tray was constructed in-house that permitted manual selection of cuvettes in the beam path. All cuvettes used for holding the liquid MFC filters were 2 mm path length optical glass. The sample cuvette had 1 mm path length. A two-factor spectrum from a sample consisted of four data points because the positive and negative factor loadings were represented by separate molecular filter mixtures. Thirty-nine ethanol-in-water mixtures were scanned with the MFC-based spectrometer. Twenty samples were used to calibrate the instrument and build a multivariate linear regression prediction model, and the remaining samples were used to validate the predictive ability of the model. To avoid possible false responses due to instrument drift, samples were measured in a random order. The sample temperature was held constant at 25 ^{o}C during the data collection period. A 3-second integration was employed at each MFC filter.

5.2.3 **Data analysis**

All data analysis was carried out using Matlab 7.0 (Mathworks, Inc., Natick, MA). The PLS toolbox v3.51 for Matlab (Eigenvector Research, Inc. Wenatchee, WA) was used for multivariate analysis. A genetic algorithm and direct search toolbox for Matlab were used to perform the NIR library search to generate combinations of liquids for use as MFC filters.

5.2.4 **Theory**

As illustrated in Figure 5.2, using the MFC approach, traditional bulky multi-channel wavelength selection devices such as gratings and moving mirrors are replaced with simple MFC filters. Only a light source, detector and MFC filters are needed to construct a minimal MFC-based spectrometer. The weighed combination of spectral responses from the filters is designed to match the regression vector from transmission spectra-based factor methods like PCR or PLS calibration. Because a multivariate regression vector can be positive or negative while all transmission spectra of MFC filters are positive, two distinct MFC filters are employed to represent accurately the multivariate regression vector. Depending on the complexity of the regression vector and availability of MFC filter materials, an exact match of the regression vector to the filter might be very difficult. Fortunately, an exact match is not absolutely necessary, for reasons that are addressed in MFC filters selection. For each MFC filter, the signal produced at the detector is a dot product of the filter transmission spectrum and the sample transmission spectrum, with a signal offset $\boldsymbol{v_{Offset}}$ in practice[76].

$$v_{out} = G \times \vec{s} \cdot \vec{f} + v_{offset} \qquad \textbf{5.1}$$

$\boldsymbol{v_{out}}$ is the output voltage, $\boldsymbol{G}$ is the constant amplifier gain, , $\boldsymbol{f}$ represents the MFC filter spectrum vector, and $\boldsymbol{s}$ represents the corrected sample spectrum vector.

For m samples and n filters, $\boldsymbol{V_{out}}$ (m by n) is output voltage matrix.

$$V_{out} = G \times SF^{T} + V_{offset} \qquad \textbf{5.2}$$

where $\boldsymbol{F}$ (n by k) is the transmission spectra matrix of MFC filters, and $\boldsymbol{S}$ (m by k) is transmission spectra matrix of samples.

The vector of concentration values, $\boldsymbol{Y}$ (m by 1), of the training samples are predicted by multivariate linear regression (MLR) according to Equation 5.3:

$$\hat{Y} = V_{out}C + E = G \times SF^{T}C + offset \qquad \textbf{5.3}$$

where ***C*** (*n* by *1*) are the regression coefficients, ***E*** is a scalar, *m* is the number of training samples, and *n* is the number of MFC filters.

After MFC filters were selected and the regression coefficients ***R*** obtained,

$$R = F^{T}C \qquad \textbf{5.4}$$

this ***R*** works in a similar fashion to PCR loadings.

$$\hat{y}_i = G \times S_i F^{T}C + offset = G \times S_i R + offset \qquad \textbf{5.5}$$

For *m* training samples, the root-mean-square error of calibration (RMSEC)[76] is

$$RMSEC = \left[\sum_{i=1}^{m} \frac{(\hat{y}_i - y_i)^2}{m}\right]^{1/2} = \left[\sum_{i=1}^{m} \frac{(G \times S_i R + offset - y_i)^2}{m}\right]^{1/2} \qquad \textbf{5.6}$$

The minimum RMSEC is reached by searching a NIR spectral library to select the best molecules for MFC filters. ***G*** and ***offset*** are the parameters adjusted after the MFC filters have been chosen. While one could select MFC filter molecules to match a regression vector that provides a fixed RMSEC specified *a priori*, searching the NIR library to find a combination of MFC filters that minimizes the RMSEC is usually more desirable. A perfect spectral match may require a large number of different filters molecules or filter molecules that are not available in the library.

5.2.5 **Spectral region selection**

Theoretically, the MFC-based spectrometer approach should function in any spectral region where molecular filters are available. For this research, the NIR spectral region was used because NIR spectrometry is a widely employed PAT and ethanol has a significant absorbance between two water absorbance bands in the NIR region between 1400 and 2200 nm.

5.2.6 **Radiometric correction**

The multivariate prediction of analyte concentration using MFC is inherently radiometric in nature. Radiometric measurement is based on a detector response that is directly related to sample transmission instead of absorbance. Of course, sample concentration is linearly related to absorbance when Beer's law holds and transmission is logarithmically related to sample concentration. In a low absorbance regime, transmission relates to concentration approximately linearly, however, in a higher absorbance regime, the nonlinear relationship between concentration and transmission predominates. To model both regimes in transmission mode, extra principal components or latent variables have to be used in a linear multivariate calibration model. [46]

Before using transmission spectra to perform a library search for MFC filter constituent selection, the transmission spectra were corrected for unique optical characteristics of the MFC spectrometer. Data provided by manufacturers' test sheets were used to form the correction factors. In the experimental MFC system, the radiometric correction was performed by convolving the transmission spectra with the emission spectrum of the source lamp, the transmission spectra of the 1400 nm long pass filter, and the response curve of the InGaAs photodiode in the prototype instrument. Thus, the corrected transmission spectra represented an unbiased detector response as a function of wavelength. The corrected spectra in Figure 5.1b revealed that the transmission of the spectrometer is not completely cut off at 1400 nm. The transmission of the actual 1400-nm long pass filter employed was approximately 25% at 1400 nm. However, the transmission was much lower at shorter wavelengths and was less than 1% at 1370 nm. Because the variation of the sample spectra from 1370 nm to 1400 nm was small, the effects of the slightly wider bandpass on prediction of sample composition were negligible.

5.2.7 MFC Filter Selection

The chemicals chosen as MFC filters were found by searching a library of near-IR transmission spectra containing 1923 compounds (John Wiley & Sons, Inc.). The library consisted of two spectra of each compound collected over slightly overlapping regions, 952-1587 nm and 1388-2630 nm. Because the coverage of the MFC system is 1400-2200 nm, only the spectra from 1388-2630 nm were used in the library search. Molecular factor scores were calculated from the product of the transmission spectra from the NIR spectral library and the corrected transmission spectra of ethanol / water mixtures:

$$U_{m \times l} = S_{m \times k} L^{T}_{l \times k} \qquad \textbf{5.7}$$

where $\boldsymbol{U}$ is the score matrix, $\boldsymbol{L}$ is the transmission spectra of the NIR library, $\boldsymbol{S}$ is the corrected transmission spectra of training samples, l is number of compounds in the library (l=1923), m is number of training spectra (m=20), and k is the number of wavelength values in the spectra (k=128).

A modified genetic algorithm [106] was used to search the score space to find four filters that yielded a predictive model with the lowest root mean square error of cross validation (RMSECV). The RMSECV function was used as the fitness function of the genetic algorithm. A genetic algorithm (GA) is a search procedure employed in computing to find actual or approximate solutions to optimization and search problems. Genetic algorithms are classified as global search heuristics. Genetic algorithms form a specific class of evolutionary algorithms that are based on methods motivated by evolutionary biology such as inheritance, mutation, selection, and crossover (also termed recombination).

Genetic algorithms are executed as a computer simulation in which a population of conceptual symbols (termed chromosomes, or the genotype or the genome) of possible solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves in the direction of superior solutions. Usually, solutions are symbolized in binary, but other symbol encodings are also feasible. The evolution usually begins from a population of randomly generated individuals and occurs in generations. In each generation, the fitness of each individual in the population is assessed, multiple individuals are randomly selected from the existing population (based on their fitness), and adapted (recombined and perhaps mutated) to create a new population. The new population is then employed in the subsequent iteration of the algorithm. A typical genetic algorithm needs two items to be specified: (a) a genetic representation of the solution domain, and (b) a fitness function to assess the solution domain. A fitness function is a specific form of objective function that quantifies the optimality of a solution (i.e., a chromosome) in a genetic algorithm in order that that individual chromosome may be ranked against every one of the other chromosomes. Optimal chromosomes, or at least chromosomes that are *more* optimal, are permitted to breed and combine their datasets by numerous techniques, leading to a new generation that will (with luck) be improved. An ideal fitness function connects closely with the algorithm's aim, and still can be computed rapidly. Speed of calculation is vital, because a conventional genetic algorithm must be iterated lots of times in order to yield a practical result for a nontrivial problem.

The genetic algorithm library search was performed 50 consecutive times. Due to the indefinite nature of the genetic algorithm, each time the search routine produced a somewhat different MFC filter combination, but roughly the same RMSECV. Four common chemicals were selected as molecular filters: water, methanol, ethanesulfonic acid, and 2,2-diethoxypropane. The transmission spectra of these chemicals are shown in Figure 5.3. These four MFC filters gave an adequate calibration model (r^2 =0.995, RMSEC=0.229%, RMSECV=0.339%, p=0.05 by f test). This result is slightly better than a corresponding PCR calibration model based on corrected transmission spectra (r^2=0.993, RMSEC=0.359%, RMSECV=0.551%, p=0.05 by f test). The PCR regression vector and simulated regression vector based on MFC filters are both presented in Figure 5.4. It is evident in Figure 5.4 that these two regression vectors do not match. The search for a regression vector by genetic algorithm is intended to reach a minimum on a multidimensional response surface. The PCR regression vector is one of many such minima, and it can be visualized as a point in a reduced orthogonal p-factor space that describes a linear relationship between the spectra and concentration. Due to the stochastic nature of the genetic algorithm, it is possible to obtain several essentially equivalent solutions to the optimization problem. Therefore, it is not surprising that the regression vector generated by MFs did not match a predefined PCR regression vector. Such a pattern match is unnecessary. Indeed, the fact that multiple solutions exist makes it easier to find molecular filters that are stable and compatible with other molecules in the filter system.

5.3 Results and Discussion

5.3.1 Multivariate analysis of absorbance and transmission spectra

In order to compare the results of MFC measurements with the results of multivariate analysis using a conventional spectrometer, PCR was performed on the same training set used for selection of MFC filters. First, the PCR calibration was performed with corrected transmission spectra. The optimum predictive model was defined as the model with lowest RMSECV by leave-one-out cross validation. Four principal components were required to build a calibration model with optimum predictive ability. Theoretically, two principal components should be sufficient to model the ethanol/water mixtures. The extra principal components were included due to the nonlinear response between the transmission spectra and concentration. The RMSEC was 0.359%, corresponding to 2.56% error relative to the range of the calibration set. The four PCs model was validated by leave-one-out cross validation, and the RMSECV was 0.551%, or 3.93% relative to the range of the calibration set. Next, a PCR calibration was carried out on absorbance spectra, which were calculated from original transmission spectra (using A=1/logT). Three principal components were required to build an optimum calibration model. The RMSEC was 0.309%, corresponding to 2.20% error relative to the range of the calibration set. The three PCs model was validated by leave-one-out cross validation, and the RMSECV was 0.494%, or 3.53% relative to the mean of the calibration set. Compared to the PCR model based on corrected transmission spectra, the PCR model based on absorbance spectra required fewer principal components and had a slightly lower RMSEC and RMSECV. The better performance of the model based on absorbance spectra is expected because of the linear response between absorbance and concentration.

5.3.2 **Expectation from simulation**

As described in the MFC filters selection section, the simulation study for the MFC filters predicted a RMSEC of 0.229% and RMSECV of 0.339% with corrected transmission spectra. . The result showed that the PCR model is not necessarily the best model. MFC filters outperform the traditional scanning PCR model in terms of RMSECV. The simulation result in Figure 5.5 shows a plot of the predicted ethanol concentrations versus the actual ethanol concentrations using a MLR model based on 4 MFC filters and a 4-component PCR model based on corrected transmission spectra.

5.3.3 **Determination of ethanol with the MFC approach**

The voltage output from the detector was recorded for each of 39 samples through each MFC filter. The samples were split into two groups for cross validation, and 20 samples were used to calibrate the MFC-based instrument, while the other 19 samples were used as the validation dataset. The 20 calibration samples were different from those samples used as training samples for the MFC filters selection, but were prepared at the same nominal concentrations. Additional calibration was necessary because the correction factors used with the training spectra were all obtained from manufacturer's test datasheets and set-ups, and might be different once assembled in the prototype instrument. The optimal correlation between detector output voltage and ethanol concentration were obtained by following equation.

$$\hat{Y} = \begin{bmatrix} v_{out[1,1]}v_{out[1,2]}v_{out[1,3]}v_{out[1,4]} \\ \dots\dots \\ \dots\dots \\ v_{out[m,1]}v_{out[m,2]}v_{out[m,3]}v_{out[m,4]} \end{bmatrix} \begin{bmatrix} -28567 \\ -14368 \\ 27997 \\ 21164 \end{bmatrix} - 34 \qquad \mathbf{5.8}$$

where $\hat{Y}$ was the predicted ethanol concentration, and v_{out} was the voltage output of each sample for each MFC filter. The RMSEC of the model was 0.748%, and the RMSEP by data splitting was 0.735%. Figure 5.6 shows a plot of predicted ethanol concentrations versus actual ethanol concentrations of all 39 samples.

5.3.4 **Discussion**

The estimated RMSEP (0.735%) of the MFC-based measurement was not as good as the RMSEP (0.339%) predicted by the simulation. Still, the actual MFC result shows that the MFC instrument is able to produce a useful numerical concentration result. The difference between the simulated instrument and the actual instrument results was due to several factors:

1. Sampling noise arose from the positioning of molecular filter cuvettes and/or sample cuvettes that did not exist in the simulation.

2. The transmission spectra in the NIR library were obtained with a path length of 2.5 mm, while cuvettes with a path length of 2 mm were used as MFC filters in the prototype instrument. The difference in the profile of transmission spectra due to the different path length likely increased prediction error.

3. Instrumental limitations prevented obtaining the exact transmission spectrum of the 1400 nm long pass filter, the emission spectrum of the light source, and the detector response curve in the prototype MFC-based instrument to correct the training transmission spectra. Alternative correction factors were obtained from manufacturers' datasheets. Better results might be expected if each individual optical component in the prototype instrument were carefully calibrated.

4. Although an optical chopper and lock-in-amplifier were used to reduce noise and thermal drift, the MFC-based prototype instrument was shown to have a significant instrument drift. Simple studies with the light source (e.g., 1400 nm long pass filter in place but without MFC chemicals or sample cell present) exhibited signal drift as high as 4% relative over 20 minutes, which was roughly the time required to scan all 39 samples in the MFC instrument. This significant drift could contribute to the high RMSEP. Future studies will utilize a double-beam

design to eliminate this drift.

5. Using the genetic algorithm-based MFC filters selection algorithm, only the predictive ability of the MLR model was considered in the fitness function. The sensitivity of each individual MFC filter to changes in ethanol concentration was not taken into account. PCR is a regression method based on orthogonal principal components that maximize variance. However, MLR only aims to minimize the sum of the squared errors, and variance maximization for dependent variables is not taken into account. Therefore, the genetic algorithm-based MFC filters selection could select MFC filters with high prediction ability but low sensitivity, which results a hypothetical low RMSEP in the simulation study that is difficult to achieve with real, physical filters. (A new search algorithm that takes both prediction and sensitivity into account is currently being investigated.)

In addition to the multivariate regression model for ethanol concentration, an estimate of the detection limit for binary mixtures of ethanol and water was also calculated. The estimate was based on an extension of the BEST metric for sub-cluster detection with sample populations that has been described previously. [51, 107] The experimental MFC data were then analyzed to estimate the limits of detection of each component in binary mixtures of two components. This was performed by translating the sample population mean of 1% ethanol in water sample towards pure water sample population's mean until the two clusters could not be differentiated using the BEST subcluster detection algorithm. The estimate of the detection limit for ethanol in water determined by this procedure is 0.26%. The dynamic range for ethanol detection by MFC is a factor of 57. The extended BEST metric provided lower errors than traditional regression approaches because it took both changes in sample cluster location as well as scale into account. However, to achieve its better results the extended BEST requires multiple replicates of the same sample, which can be impractical in real-life remote sensing applications.

In order to assess of the long-term stability of molecular filters, the molecular filters were directly exposed to the near-IR light beam for 10 hs. For each of those four molecular filters, the signal was continually monitored and variations in signal level of 4% were observed in this study (the same range as the variation in the light source intensity). The molecular filters were also sealed in cuvettes over two-month period, and there appeared no visible degradation of these molecular filters over that time. It is worth noting that, for some other molecular filters that were not used in this study, severe degradation of MFs can be observed. Thus, it is necessary to compile a spectral library using only stable molecules for MFC.

The susceptibility of MFC-based spectroscopic measurement to complex matrix interference in samples is not well understood. Theoretically, the MFC-based instrument should be able to precisely measure the specific chemical species of interest as long as the potential interferences were introduced and modeled in the training set. Future research will include determination of ethanol containing other alcohols as interferences that are not in the training set to evaluate the susceptibility of MFC to this sort of interference.

5.4 Conclusion

A prototype MFC-based spectrometer was designed, constructed, and tested for the analysis of ethanol-in-water mixtures. The concept of molecular factor computing was demonstrated. The results obtained from an MFC-based measurement were compared to PCR calibration based on conventional scanning spectrometry. Although the actual results from MFC-based prediction in the first prototype were slightly worse than from conventional PCR prediction, the MFC simulation study suggested that a better prediction model could be built based on MFC. A double-beam MFC instrument under construction may achieve the superior results predicted by the simulation. Advantages of the MFC approach over conventional spectroscopy include significantly reducing the computational demand (the integrated sensing and processing, or ISP, advantage), shorter data collection and analysis time with higher signal-to-noise ratio (S/N) (especially for imaging spectrometry, through the Fellgett advantage), higher optical throughput (the Jacquinot advantage), and more rugged instrumentation with a considerably lower cost. The high optical throughput of an MFC system could offer improved analytical ability in systems with a weak signal.

Problems with reproducibility in positioning of filter cuvettes and samples cuvettes increased measurement noise in the MFC-based prototype spectrometer. The effect will be reduced by using aperture control and through better design of slides for holding filters and samples.

A new library search algorithm should be developed to select the optimal MFC filters. Prediction ability and sensitivity of MFC filters both should be taken into account in the fitness function of genetic algorithm-based searches.

The number of potential filter materials is huge. Solutions and solid-state mixtures could both be used as molecular filters. The use of organic solvents as MFC filters introduces some ruggedness problems for process analysis. To simplify the instrument and improve the system stability, solid-state MFC filters constructed from materials such as polymers may offer a good alternative to liquid filters [74].

MFC offers users a simpler ISP instrument with significant reduction of computational complexity and processing time at the cost of some experimental flexibility. In other words, MFC-based instruments are not general-purpose research tools. Instead, the MFC approach is for practical measurement in the real world where fast results are needed and achieved by integrating the processing into the sensing stage.

In addition to applications of this technique as a process analytical technology (PAT), MFC-based remote NIR imaging for real-time surveillance has gained interest. A MFC-based NIR imaging system for remote ethanol sensing is currently under construction in our laboratory. The range of possible applications is likely to expand when imaging systems are available.

Chapter Five Figures

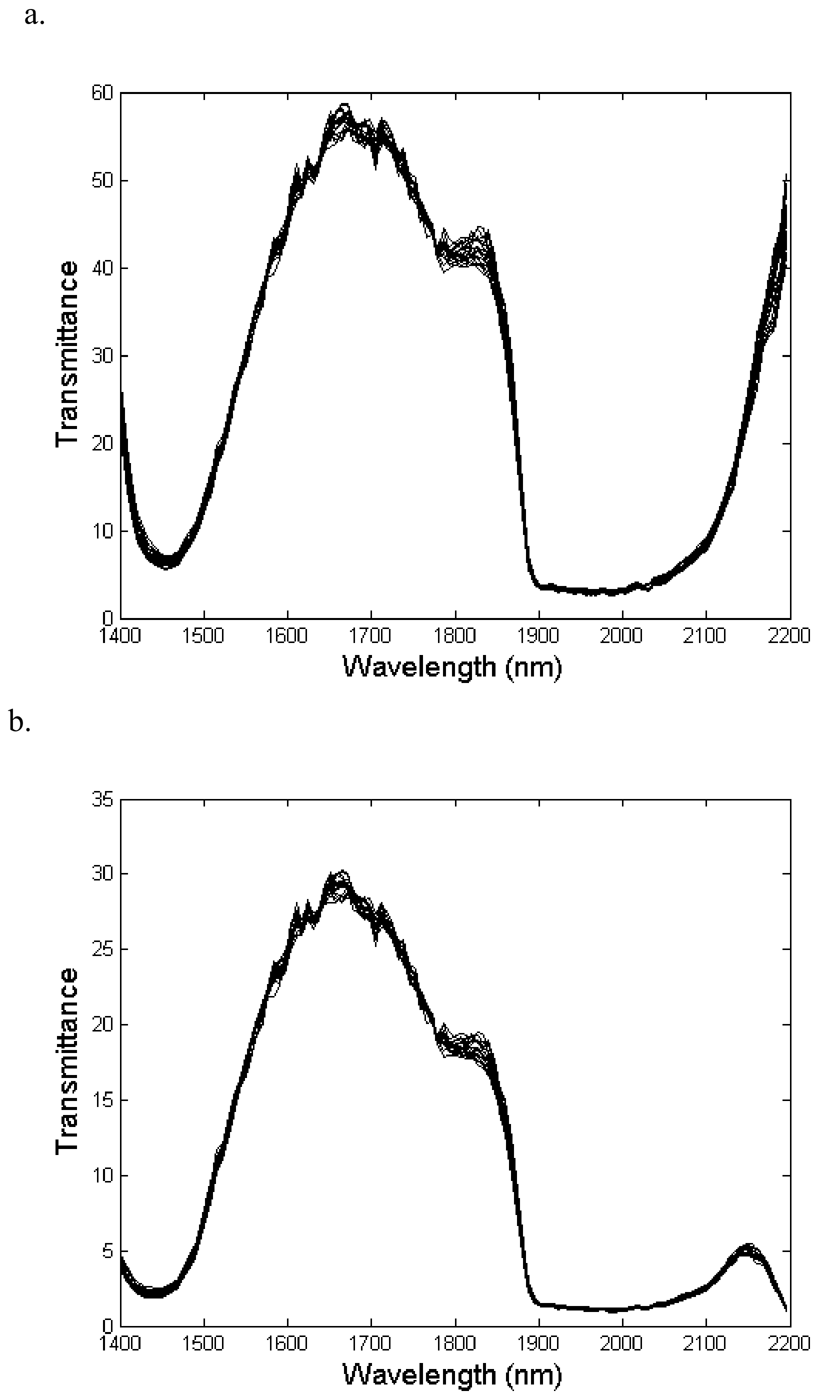


Figure 5.1 a. Raw, uncorrected transmission spectra of 20 ethanol / water mixtures acquired on a conventional dispersive NIR spectrometer. b. Corrected spectral response function. These data are based on the transmission spectra in Figure 1a, convolved with following radiometric vectors: radiance spectrum of tungsten light source, the transmission spectrum of 1400 nm long pass filter, and the response curve of the InGaAs photodiode.

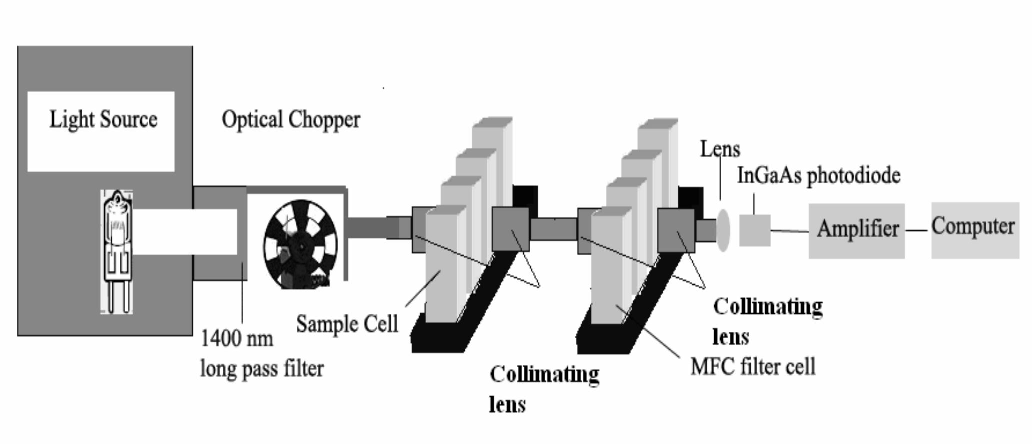


Figure 5.2 A graphical representation of the MFC-based high throughput spectrometer.

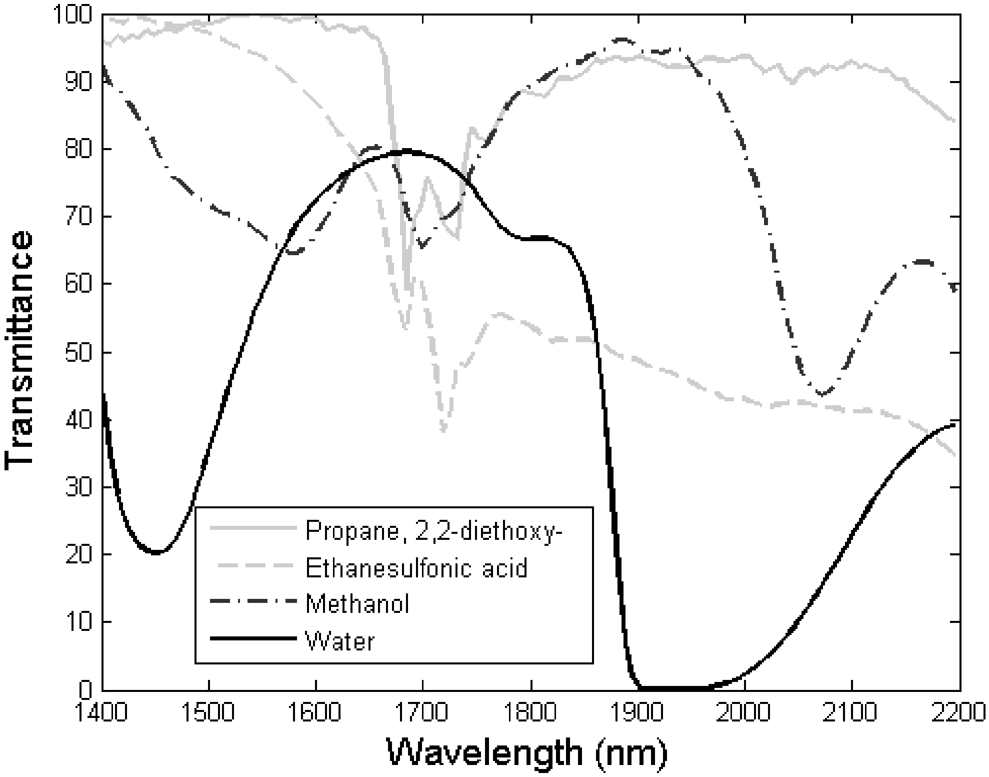


Figure 5.3 The transmission spectra of the selected MFC filters.

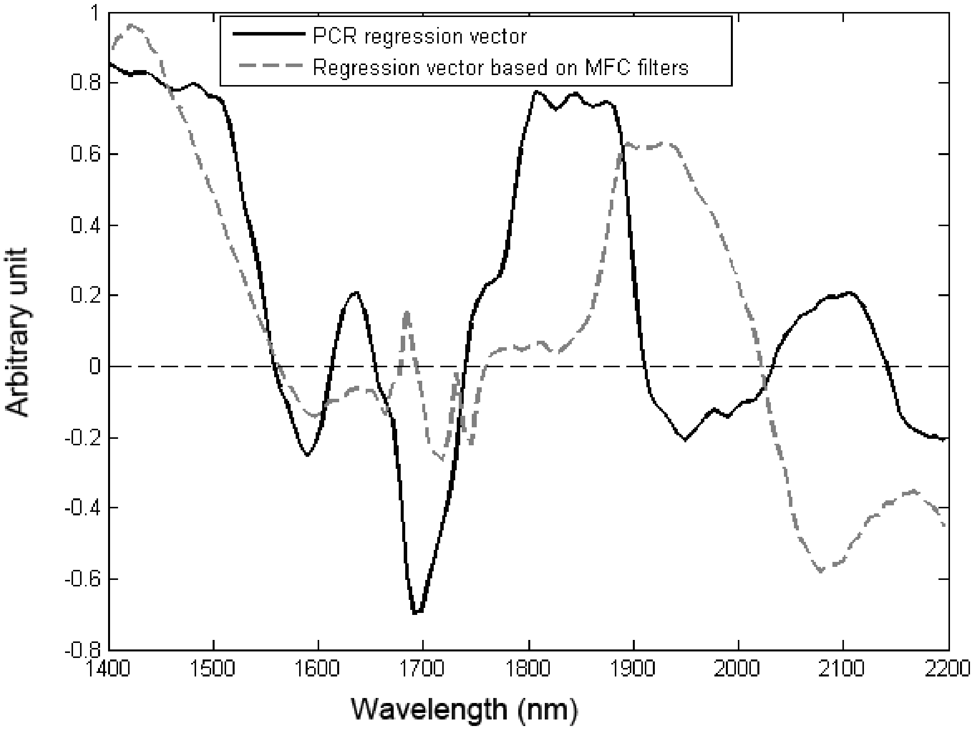


Figure 5.4 The regression vectors versus wavelength. The solid line shows the PCR regression vector, and the dashed line shows the regression vector based on MLR calibration of the MFC filter.

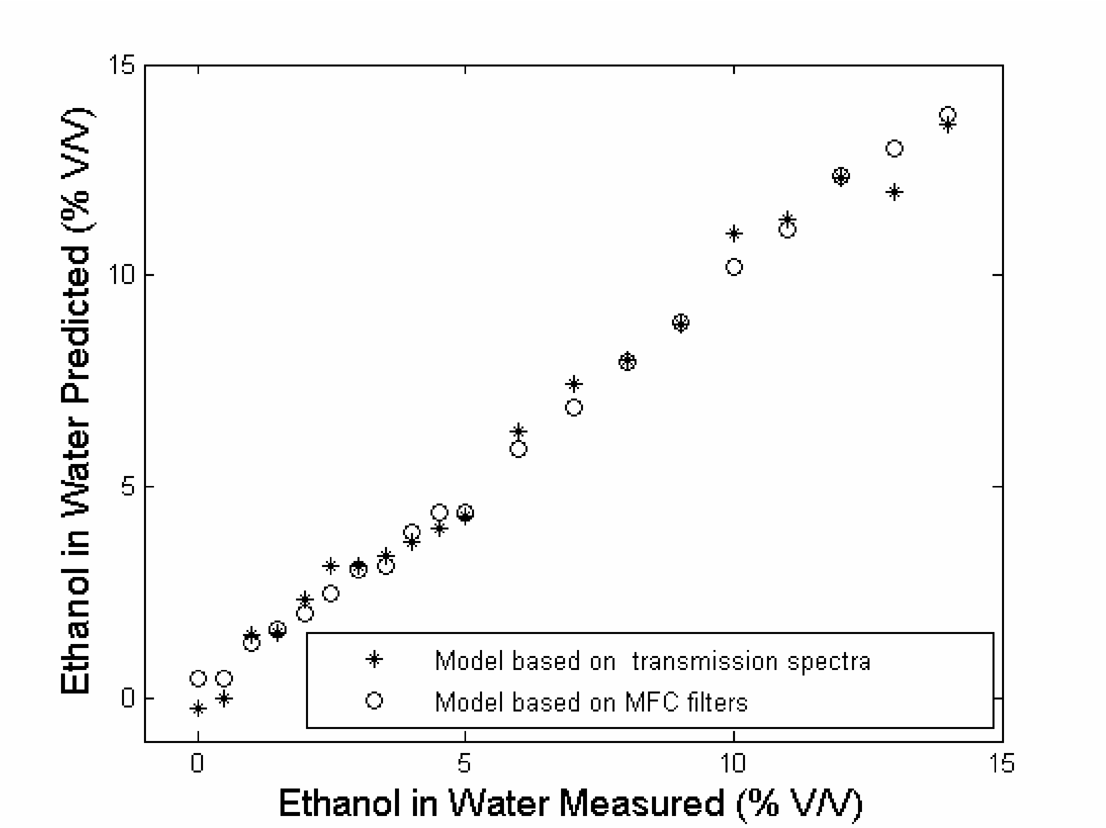


Figure 5.5 A plot of the predicted ethanol concentrations versus the actual ethanol concentrations using a MLR model based on 4 simulated MFC filters and a PCR model based on corrected transmission spectra. Stars: PCR model based on corrected transmission spectra, RMSEC=0.359%, RMSECV= 0.551%. Circles: MLR model based on 4 simulated MFC filters, RMSEC=0.229%, RMSECV=0.339%.

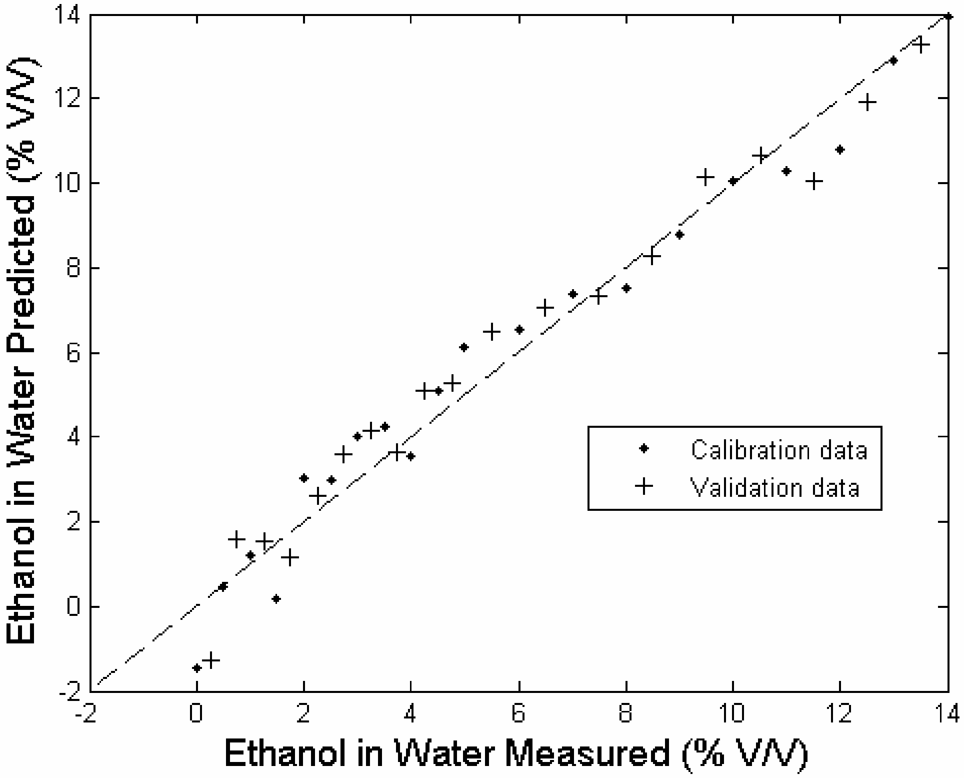


Figure 5.6 A plot of the predicted ethanol concentrations versus the actual ethanol concentrations of all 39 samples. Diamonds: calibration samples, r^2=0.968, RMSEC=0.748%. Crosses: validation samples, RMSEP=0.735%. Significant at p=0.05 by f test.

Bin Dai, Aaron A. Urbas, Craig C. Douglas, Robert A. Lodder. *Pharmaceutical Research.*
2007; Accepted Dec. 2006

Chapter Six -- Conclusion and Future Work

This dissertation demonstrates that with a carefully designed system, the promise of molecular factor computing-based spectroscopy can be fulfilled. The technical challenge for MFC based spectroscopy is the development of the method for selecting optimal molecular filters. Two library searching algorithms are developed and tested to meet the challenge. Transmission spectra of a library of reference compounds are multiplied by transmission spectra of training samples to obtain simulated library scores. The algorithm treats the simulated library scores as variables and searches the library-scores space to select a few variables that represent the sample spectra in a new coordinate system. The new coordinate system is either chosen to minimize the standard error of sample prediction for the chemical properties of interest if a quantitative characterization is desired, or chosen to maximize the class separation if a qualitative characterization is pursued.

A prototype instrument is constructed and tested for its feasibility in PAT application using ethanol-in-water mixtures. The design of this MFC PAT sensor is guided by the result of the filter generating algorithm in a series of simulation studies. The successful results obtained from the instrumental testing prove the concept of MFC-based spectroscopy.

Future work will test the MFC-filter generating algorithm with more complex sample matrices, including samples collected from actual manufacturing process lines. The hardware enhancements will include instrument simplification and system stability improvement to improve instrument ruggedness for PAT applications.   Solid-state MFC filters constructed from materials such as polymers may offer a better alternative to the liquid filters employed in the current system.

Appendix A Hardware Design of MFC-Based Near Infrared Analyzer

A.1 The concept of MFC Spectrometer

A schematic of the spectrometer is demonstrated in Fig. A.1. The spectrometer includes a light source, collimating optic, optical chopper, chemical filters and detector.

The chemical filters are specifically designed to perform optical computing for measurement of samples. Details of the formulation of chemical filters can be found in chapter 3 and chapter 4. The potential advantages of this unique spectrometer are described in next section.

A.2 Advantages of the MFC-based Spectrometer

The hardware design of MFC-based spectrometer is fundamentally different from the conventional and commonly used spectrometers like diffraction grating-based monochromators or Fourier-transform (FT) spectrometers. The advantages of the proposed spectrometer are discussed in the following paragraphs.

A.2.1 Integrated Sensing and Processing (ISP) advantage

The essential concept of the MFC-based spectrometer is the integration of a calibration or classification model into the spectrometer itself. The chemical filters are chosen specifically for the purpose of replacing vector computing in computer with optical computing, which is typically done in the detector of the spectrometer itself. Since only very few data points (detector responses) are collected for each sample, both amount of data and the computing demands for data processing are substantially reduced. The successful application of this concept might eventually eliminate the necessity of the data processing after the data have been collected.

A.2.2 Simple, compact, rugged

As shown in Figure A.1, the schematic of the hardware design offers a simpler and cheaper way to construct a spectrometer compared with the design of a conventional Fourier transform or grating spectrometer. The MCF-based spectrometer includes significantly fewer optical and electronic parts. The MFC-based spectrometer also does not require a high-precision motion control system as other spectrometers do; therefore, the MFC based spectrometer is more suitable in hostile environment such as manufacturing.

A.2.3 High optical throughput and multiplex advantage

The MFC-based spectrometer is multiplex filter type instrument. An MFC based spectrometer is not like a grating type or interference filter type instrument that allows only a very narrow band of light to pass through the sample. An MFC spectrometer gains huge optical throughput by allowing broadband multiplex light to pass through the sample. In addition, as a filter type spectrometer, the MFC instrument provides more geometrical throughput than the grating type monochoromators. The MFC-based spectrometer gains ~100 times higher throughput than a grating type instrument. The multiplex advantage (Fellgett advantage) is obvious because an MFC-based spectrometer simultaneously measures all incident wavelengths all of the time.

The geometric throughput improvement provided by MFC can be described by a few simple equations.

The optical throughput Θ is defined as: [108]

$$\Theta = A_{aperture} \frac{A_{fill}}{f2} \qquad \textbf{A.1}$$

where f is the focal length of the focusing lens, A_{fill} is the illuminated area in the sample or wavelength selector component. $A_{aperture}$ is the area of limiting aperture, for a circular aperture with the radius of r,

$$A_{aperture} = \pi \cdot r^2 \qquad \textbf{A.2}$$

Combining Eq. (B.1) and Eq. (B.2), the geometric throughput of a filter type spectrometer can be expressed as

$$\Theta_{filter} = \pi \frac{r^2}{f^2} A_{fill} \qquad \textbf{A.3}$$

For comparison, the throughput of a Fourier transform spectrometer is defined[109] as

$$\Theta_{FT} = \pi \frac{A_{fill}}{R} \qquad \textbf{A.4}$$

In a diffraction grating spectrometer, the throughput is given by:

$$\Theta_G = \frac{l \cdot A_{fill}}{f \cdot R} \qquad \textbf{A.5}$$

where l is the height of the optical slit and R is the spectral resolving power. The requirement of higher spectral resolving power reduces the optical throughput.

A.2.4 High signal-to-noise ratio (S/N)

In many diffusion reflectance applications, the detected light is weak, and the measured S/N is usually limited by detector noise. In this case, the S/N is proportional to the three factors, the optical throughput Θ, the transmission efficiency of the optical system ζ, and the square root of the number of data collected N.

$$S / N \propto \Theta \times \xi \times \sqrt{N} \qquad \textbf{A.6}$$

As described in A.2.3, the throughput advantage (Jacquinot advantage) optimizes Θ, and the multiplex advantage (Fellgett advantage) optimizes N. As a result, the S/N of an MFC-based spectrometer is optimized.

A.3. Development of Prototypes

In this study, two different prototypes were developed. The first one was tested by measuring the concentration of methanol in water mixtures using randomly selected chemical filters. The main purpose for building the first instrument was a feasibility study and proof-of-concept. The second prototype was then built and used in the ethanol sensing experiment. The second prototype also served as the instrument for differentiating the cholesterol, collagen, and elastin through the blood cell solution.

A.3.1 The first prototype instrument

Light source: A 12-V, 25-W tungsten halogen lamp (GE, Cleveland, OH) was used to provide the near-infrared (NIR) radiation, powered by a stabilized 12-V DC power supply. A 250W heat lamp (GE) was used as an alternative light source, directly powered by 110V AC power supply.

Optical chopper: An optical chopper provided a convenient way to modulate the signal. Direct amplification of a low-frequency or DC signal is troublesome particularly when an instrument exhibits amplifier drift (or detector drift) and flicker noise. Often, this 1/f noise is much larger than the types of noise that predominate at higher frequency. For this reason, low frequency or DC signals from a detector were often converted to higher frequency, thereby reducing the 1/f noise. After amplification, the modulated signal was freed from 1/f noise by a low-pass filter, producing an amplified AC signal suitable for output. Of course, both signal filtering and demodulation could be completed in a computer after the signal was collected by using the Fast Fourier Transform (FFT). The optical chopper system included a motor speed controller, motor, and chopper wheel. A reference signal from the chopper was collected by adding a photointerrupter (GP1L57, SHAPE Corp., Japan). The chopper frequency was set at 210 Hz.

Detector circuit and data acquisition graphic user interface (DAQ GUI): A lead sulfide (PbS) detector (Hamamatsu Corp., Japan) with an active area of 2 x 5 mm was placed at the focal point of the NIR beam. An AC-coupled preamplifier removed dark current in the detector and amplified the chopper-modulated signal. A 1 kHz low-pass filter smoothed the output signal from the preamplifier to eliminate the high frequency noise. The modulated and low-pass filtered signal was then delivered to an external PC sound card (Creative Labs Sound Blaster MP3 +).

The sound card served as an analog-to-digital (A/D) converter with a 16-bit A/D and a 48 KHz sampling rate, and transformed the AC signal to a PC-compatible digital format. The circuit design of a preamplifier and 1 kHz low-pass filter is based on McClure's papers. [110, 111] Figure A.2 shows the screen printing of the Matlab-based graphical user interface (GUI) that controlled the instrument and performed the data acquisition. After the digitalized AC signal was transferred to a computer through a USB port, the GUI software integrated the AC signal using a Fast Fourier Transform (FFT), and then converted the signal to its Root Mean Square (RMS) equivalent DC voltage value. The magnitude of this DC voltage represented the intensity of the transmitted light through the molecular filter and sample. This voltage was also proportional to the MFC (PC) score. The raw score data were saved and loaded into Matlab for further processing.

A.3.2 Second prototype instrument

This instrument was a modified version of first instrument. The difference between these two instruments was that the major components were purchased commercially instead of being constructed in the lab. The picture in figure A.3 shows the actual instrument.

Light Source: A 12V, 100W tungsten-halogen broadband source (model 621, McPherson Inc., Chelmsford, MA) with 1400-nm long pass filter (Thorlabs, Newton, NJ) was used as the source of broadband NIR light. The tungsten-halogen light source has more intense radiation in the shorter NIR wavelength region. To avoid saturating the detector with short wavelength NIR radiation that contains little chemical information about the samples, the 1400 nm long pass filter was used to block the short wavelength radiation.

Optical Chopper: The source beam was modulated with an optical chopper (Model SR540, Stanford Research Systems Inc., Sunnyvale, CA) at a frequency of 280 Hz.

Detector: The light beam was focused onto an InGaAs photodiode (Fermionics Opto-Technology, Simi Valley, CA) through a convex lens after passing through the molecular filter cuvette and sample cuvette. The latest prototype used the fiber-optics to replace the lens-optics, resulting in a more rugged instrument.

Cuvette holder: A step-indexed sliding cuvette tray was constructed in-house that permitted manual selection of cuvettes in the beam path. All cuvettes used for holding the liquid MFC filters were 2 mm path length optical glass. A picture of the cuvette holder is shown in Figure A.4

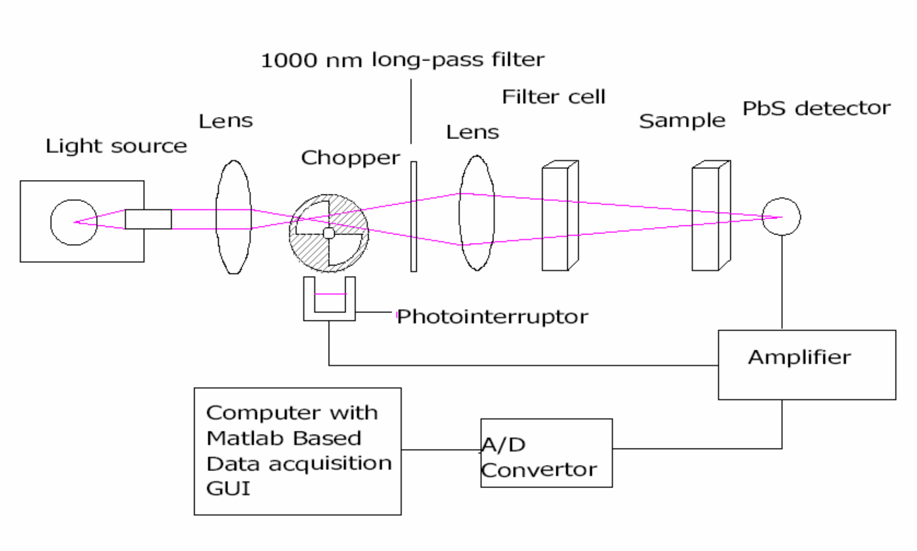
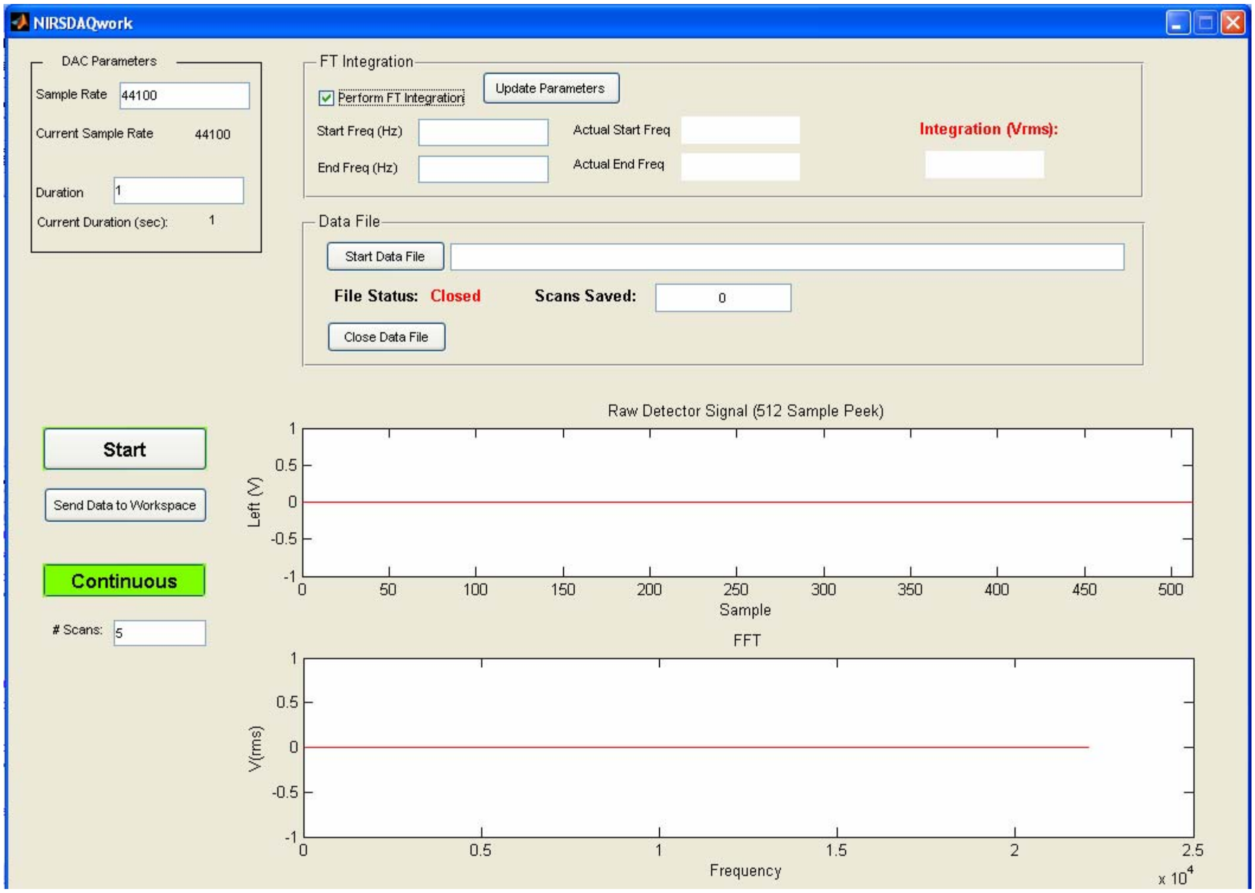


Figure A.1 Instrumental schematic of MFC-based NIR spectrometer. Shown is a setup for transmission measurement. A 12-V 50-W tungsten halogen lamp provides the near-infrared (NIR) radiation. A photo-interrupter is mounted beside the optical chopper to collect the reference signal for correcting the chopper drift**.**



NIRSDAQwork

DAC Parameters

Sample Rate: 44100

Current Sample Rate: 44100

Duration: 1

Current Duration (sec): 1

FT Integration

☑ Perform FT Integration

Start Freq (Hz):

End Freq (Hz):

Actual Start Freq:

Actual End Freq:

Integration (Vrms):

Data File

File Status: Closed

Scans Saved: 0

Scans: 5

Figure A.2 The DAQ graphic user interface (GUI) used in the first prototype instrument.

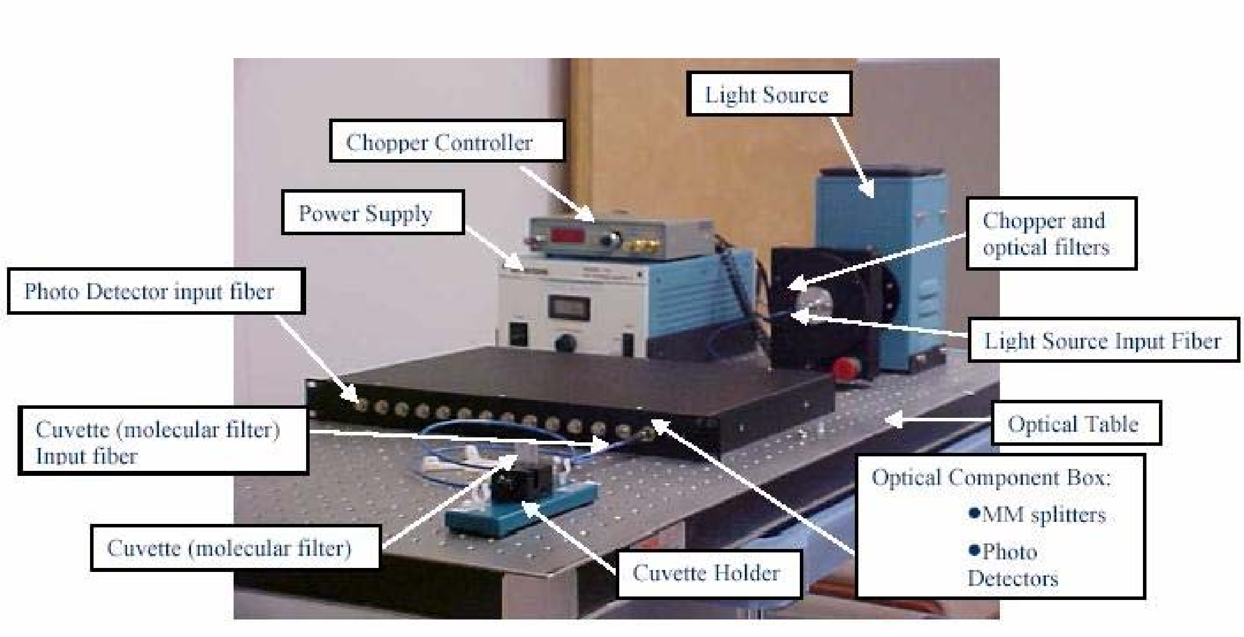


Figure A.3 Photograph of the second prototype instrument.

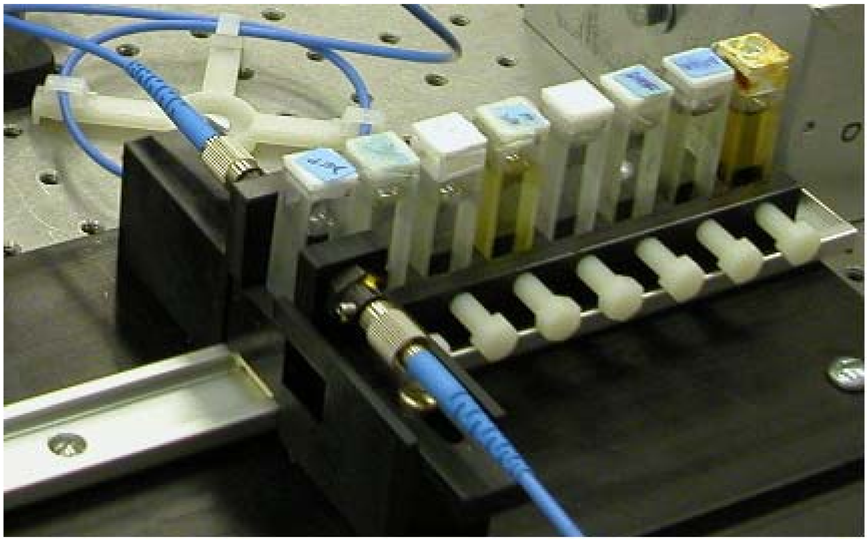


Figure A.4 Photograph of the cuvette holder with molecular filters. This picture shows that 10mm cuvettes are placed in the holder. In MFC ethanol experiment, 2mm cuvettes were used.

Appendix B Matlab Program List

B.1 Genetic Algorithm Based Multivariate Linear Regression Algorithm for MFC Filters Selection

```
% GAScript.m
% following m-script for GA based variables selection routine. Sample
% spectra and library spectra are provided. Copyright@ Bin Dai, Aaron
% Urbas, Robert Lodder Analytical Spectroscopy Research Group. University
% of Kentucky.
load transspecslow; % load the library spectra
load waveshighnm;   % load the wavelength of library spectra
load correctspec;   % load radiometric corrected sample spetra
wavelength=linspace(1400, 2200,117);
% wavelength for raw spectum of ethanol-water mixtures
[newspecs]=resamplespecs(transspecslow(1:1000,:),waveshighnm,wavelength);
% resample the library spectra to the wavelength range as sample spectra
score=correctspec*newspecs';
% generate score matrix by multiply sample transmission spectra with
% library spectra
score=score';
calibrate_conc=[0:0.5:8,9:14];  % concentration of training sample
numSelectedVars=4;   % limit the number of Molecular Filters to 4
numSamples=20;   % training sample size=20
% following settings are parameters for Genetic Algorithm
options = gaoptimset('CreationFcn',{@varselect_gacreate,numVariables},...
'CrossoverFcn',@varselect_crossoverscattered,...
'MutationFcn',@varselect_mutationuniform,...
'PopulationSize',100,...
'PopInitRange',[1;numVariables],...
'FitnessLimit', 0.002,...
'StallGenLimit',100,...
'StallTimeLimit',600,...
'TimeLimit',6000,...
'Generations',100,...
'CrossoverFraction',0.5,...
'Display','iter',...
'PlotFcn',@gaplotbestf);
% creat GA fitness function minimize the SEP by crossvalidation
FitnessFcn = {@varselect_gafit,score,calibrate_conc,numSamples};
% run GA to select the variables and compute the SEP (errorrate)
```

```
[selectedVars, errorRate] = ga(FitnessFcn,numSelectedVars,options);

function [sep,avesee,fpercnt,rpercnt,preds] = xvalid(tnspec,conc,...
   printstats,includeconst)

% Xvalid.m  Cross-validation MLR
% Copyright@ Aaron Urbas,Bin Dai  University of Kentucky.

 if(nargin < 2)
   error('XVALID requires at least two input arguments');
elseif(nargin < 3)
   printstats = 1;
   includeconst = 1;
elseif(nargin < 4)
   includeconst = 1;
end

% Check that data and conc have compatible dimensions
[n,p] = size(tnspec);
[ny,py] = size(conc);
if n ~= ny,
   error('The number of rows in conc must equal the number of rows in tnspec');
end
if py ~= 1,
   error('conc must be a vector');
end

% Create empty output variables and statistics and add constant column to
% data if needed
[nrows,ncols] = size(tnspec);
if(includeconst ~= 0)
   tnspec = [ones(nrows,1),tnspec];
end
sees = zeros(nrows,1);
res = zeros(nrows,1);
preds = zeros(nrows,1);
rsqrs = zeros(nrows,1);

% Cross-validation
for i=1:nrows
   train = tnspec([1:(i-1),(i+1):nrows],:);
   tconc = conc([1:(i-1),(i+1):nrows]);
   [b,bint,r,rint,stats] = regress(tconc,train);
   sees(i) = sqrt((r'*r)/(nrows-ncols-1));
   rsqrs(i) = stats(1);
   preds(i) = tnspec(i,:)*b;
   res(i) = conc(i)-preds(i);
end
 % Find cross-validation statistics
avesee = mean(sees);
conctemp = conc;
restemp = res;
maxconcind = find(conctemp==max(conctemp));
if(length(maxconcind) < 2)
   restemp = restemp(find(conctemp~=max(conctemp)));
   conctemp = conctemp(find(conctemp~=max(conctemp)));
end
minconcind = find(conctemp==min(conctemp));
if(length(minconcind) < 2)
   restemp = restemp(find(conctemp~=min(conctemp)));
end
sep = sqrt((restemp'*restemp)/(length(restemp)-1));
f = avesee^2/sep^2;
if f > 1
   f = 1/f;
end
fpercnt = fcdf(f,nrows,nrows)*2*100;
t = mean(rsqrs)/std(rsqrs);
rpercnt = (1-tcdf(t,(length(rsqrs)-1)))*2*100;
rsep = sep/(max(conc)-min(conc));

% If not returning results, display some statistics
if(printstats ~= 0)
   disp(' ');
   disp(['Min Conc               = ' num2str(min(conc))]);
   disp(['Max Conc               = ' num2str(max(conc))]);
   disp(['Average-SEE            = ' num2str(avesee)]);
   disp(['SEP                   = ' num2str(sep)]);
   disp(['Rel SEP (Over Conc Range) = ' num2str(rsep)]);
   disp(['F Percent (SEE = SEP)    = ' num2str(fpercnt)]);
   disp(['R^2 Percent (R^2 = 0)    = ' num2str(rpercnt)]);
   disp(' ');
```

```
end
```

```
function errorRate = varselect_gafit(variableIndices,...
   data,groups,numSamples)
 % varselect_gafit defines the fitness function=SEP
[pcs,pv,lds] = mypca(data(:,variableIndices),2,0);
[sep,avesee,fpercnt,rpercnt,preds] = xvalid(pcs,groups,0,1);
errorRate = sep;
```

```
function [pcs,propvar,loads,S] = mypca(data,precode,printstats)

% Check input arguments and set precode and printstats if not supplied
if(nargin < 1)
   error('pca requires at least one argument');
elseif(nargin < 2)
   precode = 2;
   printstats = 1;
elseif(nargin < 3)
   printstats = 1;
end

% Do preprocessing, precode: (1) mean centering, (2) Z-scoring
[nrows,ncols] = size(data);
if(precode == 1)
   data = data - repmat(mean(data),nrows,1);
elseif(precode == 2)
   data = data - repmat(mean(data),nrows,1);
   data = data ./ repmat(std(data),nrows,1);
end

% Transpose data if necessary for a more efficient SVD
flipflag = 0;
if(ncols > nrows)
   data = data';
   flipflag = 1;
end

% Find SVD
[U,S,V] = svd(data,0);

% Find propvar, pcs and loadings from SVD results
propvar = diag(S*S);
propvar = propvar/sum(propvar);
if(flipflag == 0)
   pcs = U;
   sinv = diag(ones(size(S,1),1)./diag(S));
   loads = (V*sinv)';
else
   pcs = V(:,1:(nrows-1));
   S = S(1:size(S,1)-1,1:size(S,1)-1);
   sinv = diag(ones(size(S,1),1)./diag(S));
   loads = (U(:,1:(nrows-1))*sinv)';
   propvar = propvar(1:length(propvar)-1);
end

% Print pca results if requested
if(printstats ~= 0)
   disp(' ');
   disp('Parameter Table:');
   disp('  ROOTNUM      ROOT        PROPVAR        CUMVAR');
   disp('  -------      ----        -------        ------');
   disp(sprintf('%5.0f %15.5g %15.5g %15.5g \n',[(1:length(propvar))',...
      diag(S),propvar,cumsum(propvar)]'));
   disp(' ');
end
```

```
function [newspecs] = resamplespecs(specs,origwaves,newwaves,p)

% Check input arguments and assign default smoothing parameter, 'p', if
% necessary
if(nargin < 3)
   error('resamplespecs requires at least three arguments')
elseif(nargin < 4)
   p = 0.999999;
end
 % Check for extrapolation outside of original wavelength range
origmin = min(origwaves);
origmax = max(origwaves);
newmin = min(newwaves);
newmax = max(newwaves);
if((newmin < origmin) || (newmax > origmax))
   error('New x data is outside range of original x data, extrapolation beyond
end points is not allowed.')
end
 % Fit a cubic smoothing spline to the data and resample the spectrum
[newspecs,p] = csaps(origwaves,specs,p,newwaves);
```

B.2 Genetic Algorithm Based Linear Discriminant Analysis Algorithm for MFC Filters Selection

```
function [keepers,coefs,mdistscv,gmdists,relpdists] =
mfcmahalpicoptgrps(scrs,groups,maxkeeps,trnsets,tstsets,start)

% Input List:
% scrs:  The scores candidate
% groups: The assigned class of spectra
% maxkeeps:  The maximum number of variables to keep
% trnset: training set
% tstset: test set
% start:  Preselected candidates
%
% output List:
% keepers:  The index of selected variables
% coefs: Coefficients of selected variables
% mdistscv: corss validate mahalanobis distance
% gmdists: mahalanobis distance between groups
% relpdists: the ratio of mahalanobis distance between group and within group


if(nargin < 3)
   error('Not enough input arguments');
end
[ns,nx] = size(scrs);
% Convert group to indices 1,...,g and separate names
[groupnum, gnames] = grp2idx(groups);
ng = max(groupnum);

% Check input arguments and set defaults
switch(nargin)
   case 3  % Generate LOO training and test sets
      trnsets = zeros(ns,ns-1);
      tstsets = (1:ns)';
      for i=1:ns
         trnsets(i,:)=[1:(k-1),(k+1):ns];
      end
      start = [];
   case 4 % Generate test sets as the compliments of the training sets provided
      tstsets = zeros(size(trnsets,1),ns-size(trnsets,2));
      for i=1:size(trnsets,1)
         tstsets(i,:) = setxor(1:ns,trnsets(i,:));
      end
      start = [];
   case 5
      if(size(trnsets,1) ~= size(tstsets,1))
         error('TRNSETS and TSTSETS must have equal numbers of rows');
      end
      start = [];
   otherwise
      if(nargin ~= 6)
         error('Incorrect number of input parameters');
      end
end

ntrnsets = size(trnsets,1);
ntsts = length(tstsets(:));
ntst = length(tstsets(1,:));
tstgrps = tstsets';
tstgrps = groupnum(tstsets(:));

relpdists = zeros(maxkeeps,nchoosek(ng,2));
keepers = zeros(0); % holds variables selected for model
candids = 1:nx; % holds candidate list, initially all variables

% Include variable preselected for the model if needed
if(~isempty(start))
   keepers = start; % variables preselec for model
   candids = setxor(candids,keepers);
end

disp(['Mahalanobis Distances Between Groups Means:']);
lastmingmdist = 0;
numvars = 1;
while(numvars <= maxkeeps)
   mingmdists = zeros(length(candids),1);
   nobetter = 0;
   for j = 1:length(candids)
      mdists = zeros(ntsts,1);
      tstcnt = 1;
      for k = 1:ntrnsets
         train = scrs(trnsets(k,:),[keepers,candids(j)]);
         traing = groupnum(trnsets(k,:));
         test = scrs(tstsets(k,:),[keepers,candids(j)]) -
repmat(mean(train),ntst,1);
         [d,p,stats] = manova1(train,traing);
         if(d < 1)
            d = 1;
         end
         testcanon = test*stats.eigenvec;
         trainmean = grpstats(stats.canon(:,1:d), traing);
         for i=1:ntst
            mdists(tstcnt) = sum((testcanon(i,1:d) -
trainmean(groupnum(tstsets(k,i)),:)).^2);
            tstcnt = tstcnt+1;
         end
      end
      [d,p,stats] = manova1(scrs(:,[keepers,candids(j)]),groupnum);
      % Compute Mahalanobis distances between group means
      mmdist = grpstats(mdists,tstgrps);
      reldists = zeros(length(stats.gmdist),1);
      %gmdist = stats.gmdist;
      gmeans = grpstats(stats.canon(:,1:d),groupnum);
      gmdist = zeros(ng,ng);
      for k=1:ng
         for m=1:ng
```

```
            if(k==m)
               continue;
            end
            gmdist(k,m) = sqrt(sum((gmeans(k,:)-gmeans(m,:)).^2));
         end
      end
      cnt = 1;
      for k=1:ng
         for m = k+1:ng
            reldists(cnt) = gmdist(k,m)/(mmdist(k)+mmdist(m));
            cnt = cnt+1;
         end
      end
      mingmdists(j) = min(reldists);
   end
   bestcandid = find(mingmdists == max(mingmdists));
   if(length(bestcandid)>1)
      bestcandid = bestcandid(1);
   end

   mdists = zeros(ntsts,1);
   tstcnt = 1;
   for k = 1:ntrnsets
      train = scrs(trnsets(k,:),[keepers,candids(bestcandid)]);
      traing = groupnum(trnsets(k,:));
      test = scrs(tstsets(k,:),[keepers,candids(bestcandid)]) -
repmat(mean(train),ntst,1);
      [d,p,stats] = manova1(train,traing);
      if(d < 1)
         d = 1;
      end
      testcanon = test*stats.eigenvec;
      trainmean = grpstats(stats.canon(:,1:d), traing);
      for i=1:ntst
         mdists(tstcnt) = sum((testcanon(i,1:d) -
trainmean(groupnum(tstsets(k,i)),:)).^2);
         tstcnt = tstcnt+1;
      end
   end

   % Compute Mahalanobis distances between group means
   [d,p,stats] = manova1(scrs(:,[keepers,candids(bestcandid)]),groupnum);
   mmdist = grpstats(mdists,tstgrps);
   %gmdist = stats.gmdist;
   gmeans = grpstats(stats.canon(:,1:d),groupnum);
   gmdist = zeros(ng,ng);
   for k=1:ng
      for m=1:ng
         if(k==m)
            continue;
         end
         gmdist(k,m) = sqrt(sum((gmeans(k,:)-gmeans(m,:)).^2));
      end
   end
   cnt = 1;
   for k=1:ng
      for m = k+1:ng
         relpdists(numvars,cnt) = gmdist(k,m)/(mmdist(k)+mmdist(m));
         cnt = cnt+1;
      end
   end
   mingmdistsbest = min(relpdists(numvars,:));
   if(mingmdistsbest > lastmingmdist)
      keepers = [keepers,candids(bestcandid)];
      candids =
candids([1:(bestcandid-1),(bestcandid+1):length(candids)]);
      lastmingmdist = mingmdistsbest;
      disp(relpdists(numvars,:));
   else
      nobetter = 1;
   end

   if(numvars > 1)
      %mingmdistsbest = %
sum(sqrt(relpdists(numvars,:)))*(min(relpdists(numvars,:)).^2);
      mingmdistsbest = min(relpdists(numvars,:));
      [mingmdists] =
mfcmahalremovalgrps(scrs(:,keepers),groupnum,trnsets,tstsets);
      if(max(mingmdists) > mingmdistsbest)
         remvar = find(mingmdists==max(mingmdists));
         if(length(remvar)>1)
            remvar = remvar(1);
         end
         if(remvar == numvars)
            keepers = keepers([1:(numvars-1)]);
            break;
         else
            throwback = keepers(remvar);
            if(nobetter == 1)
               keepers = keepers([1:(remvar-1),(remvar+1):numvars-1]);
               numvars = numvars-1;
            else
               keepers = keepers([1:(remvar-1),(remvar+1):numvars]);
            end
            candids = [candids,throwback];
         end
      else
         if(nobetter==1)
            break;
         else
            numvars = numvars+1;
         end
      end
   else
      numvars = numvars+1;
   end
end
```

```
[dfull, pfull, statsfull] = manova1(scrs(:,keepers),groups);
coefs = statsfull.eigenvec(:,1:dfull);
%gmdists = statsfull.gmdist;
gmeans = grpstats(statsfull.canon(:,1:d),groups);
gmdists = zeros(ng,ng);
for k=1:ng
   for m=1:ng
      if(k==m)
         continue;
      end
      gmdists(k,m) = sqrt(sum((gmeans(k,:)-gmeans(m,:)).^2));
   end
end

mdistscv = zeros(ntsts,1);
tstcnt = 1;
for k = 1:ntrnsets
   train = scrs(trnsets(k,:),keepers);
   traing = groupnum(trnsets(k,:));
   test = scrs(tstsets(k,:),keepers) - repmat(mean(train),ntst,1);
   [d,p,stats] = manova1(train,traing);
   if(d < 1)
      d = 1;
   end
   testcanon = test*stats.eigenvec;
   trainmean = grpstats(stats.canon(:,1:d), traing);
   ntst = length(tstsets(k,:));
   for i=1:ntst
      mdistscv(tstcnt) = sum((testcanon(i,1:d) -
trainmean(groupnum(tstsets(k,i)),:)).^2);
      tstcnt = tstcnt+1;
   end
end

% Display table header
%if(printstats)
%   disp(' ');
%   disp('Model:');
%   disp('   VAR   COEF         CVCOEF         CVSTD        SEP
AVESEE       F-PERCENT');
%   disp('   ---   ----         ------         -----       ---------');
%end
```

References

1. Raju, G. K. White paper: New Opportunities in Pharmaceutical Manufacturing. http://www.pharmamanufacturing.com/whitepapers/2004/118.html

2. Abboud, L.; Hensley, S., New prescription for drug makers: update the plants. *The Wall Street Journal* March 9, 2003.

3. Pharmaceutical cGMPs for the 21st Century - A Risk-Based Approach. http://www.fda.gov/cder/gmp/gmp2004/GMP_finalreport2004.htm#_Toc84065734

4. PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. In 2004.

5. Watts, D. C.; Clark, J. E., PAT: Driving the Future of Pharmaceutical Quality. *Journal of Process Analytical Technology* **2006,** 3, (6), 6-9.

6. Ciurczak, E., Near-Infrared Spectroscopy: Why It Is Still the Number One Technique in PAT. *Journal of Process Analytical Technology* **2006,** 3, (1), 19-21.

7. Brereton, R. G., Chemometrics and PAT. *Journal of Process Analytical Technology* **2005,** 2, (3), 8-11.

8. Miller, C. E., Chemical Principles of Near-Infrared Technology. In *Near-Infrared Technology in Agriculture and Food Industries*, Williams, P.; Norris, K., Eds. St Paul, MN, 2001; pp 19-38S.

9. De Braekeleer, K.; De Juan, A.; Sanchez, F., Deterimination of the end point of a chemical synthesis using on-line measured min-infrared spectra. *Appl. Spectrosc.* **2000,** 54, 601-607.

10. Baylor, L. C.; O'Rourke, P. E., UV-Vis for On-line Analysis. In *Process Analytical*

Technology, Bakeev, K. A., Ed. Blackwell Publishing: 2005; pp 170-186.

11. Ngai, S. S. H. Multi-scale analysis and simulation of powder blending in pharmaceutical manufacturing. Massachusetts Institute of Technology, 2005.

12. McGill, C.; Nordon, A.; A, L., D., Potential applications of process analytical analysis by NMR spectrometry. *Journal of Process Analytical Chemistry* 6, (1).

13. Nordon, A.; McGill, C. A.; Littlejohn, D., Process NMR spectrometry. *Analyst* **2001,** 126, 260-272.

14. Dell'Orco, P.; Brum, J.; Matsuoka, R.; Badlani, M., Monitoring Process-scale Reactions using API Mass Spectrometry. *Anal. Chem.* **1999,** 71, 5165-5170.

15. Jestel, N. L., Process Raman Spectroscopy. In *Process Analytical Technology*, Bakeev, K. A., Ed. Blackwell Publishing: 2005; Vol. 133-170.

16. Lodder, R. A.; Selby, M.; Hieftje, G. A., Detection of Capsule Tampering by Near-Infrared Reflectance Analysis. *Anal Chem* **1987,** 59, (15), 1921-1930.

17. Kamat, M. S.; Lodder, R. A.; DeLuca, P. P., Near-infrared spectroscopic determination of residual moisture in lyophilized sucrose through intact glass vials. *Pharm Res* **1989,** 6, (11), 961-5.

18. Galante, L. J.; Brinkley, M. A.; Lodder, R. A., Bacterial monitoring in vials using a spectrophotometric assimilation method. *Pharm Res* **1992,** 9, (3), 357-64.

19. Galante, L. J.; Brinkley, M. A.; Drennen, J. K.; Lodder, R. A., Near-infrared spectrometry of microorganisms in liquid pharmaceuticals. *Anal Chem* **1990,** 62, (23), 2514-21.

20. El-Hagrasy, A. S.; Morris, H. R.; D'Amico, F.; Lodder, R. A.; Drennen, J. K., 3rd,

Near-infrared spectroscopy and imaging for the monitoring of powder blend homogeneity. *J Pharm Sci* **2001,** 90, (9), 1298-307.

21. Buice, R. G., Jr.; Gold, T. B.; Lodder, R. A.; Digenis, G. A., Determination of moisture in intact gelatin capsules by near-infrared spectrophotometry. *Pharm Res* **1995,** 12, (1), 161-3.

22. Buice, R. G., Jr.; Cassis, L. A.; Lodder, R. A., Near-IR and IR imaging in lipid metabolism and obesity. *Cell Mol Biol* **1998,** 44, (1), 53-64.

23. Lewis, E. N.; Schoppelrei, J. W.; Lee, E.; Kidder, L. H., Near-Infrared Chemical Imaging as a Process Analytical Tool. In *Process Analytical Technology*, Bakeev, K. A., Ed. Blackwell Publishing: 2005; pp 187-225.

24. Lewis, E. N.; Lee, E.; Kidder, L. H.; Schoppelrei, J., Near Infrared Chemical Imaging Microscopy of Pharmaceutical Products. *Microscopy and Microanalysis* **2004,** 10, 1294-1295.

25. Simpson, M. B., Near-Infrared Spectroscopy for Process Analytical Chemistry: Theory, Technology and Implementation. In *Process Analytical Technology*, Bakeev, K. A., Ed. Blackwell Publishing: 2005; pp 39-90.

26. Martens, H.; Naes, T., *Multivariate Calibration*. John Wiley and Sons: New York, 1989.

27. Wold, S., Closs-validatory estimation of the Number of Components in Factor Analysis and Prinicpal Component Models. *Technometrics* **1978,** 20, 397-406.

28. Massie, D. R.; Norris, K. H., The Spectral Refletance and Transmittance properties of Grains in the Visible and Nar-Infrared. *Trans. Am. Sco. Agric. Eng* **1965,** 8, 598.

29. Hrusschka, W. R.; Norris, K. H., Least Squares CurveFitting of Near-Infrared Spectra Predicts Protein and Moisture Content in Ground Wheat. *Appl. Spectrosc.* **1982,** 36, 261-265.

30. Gerlach, R. W.; Kowalski, B. R.; Wold, H. O. A., Partial Least Squares Path Modeling with Latent Variables. *Anal. Chim. Acta* **1979,** 112, 417-421.

31. Tauler, R.; Kowalski, B. R.; Fleming, S., Multivariate Curve Resoltion Applied to Spectral Data from Multiple Runs of an Industrial Process. *Anal. Chem.* **1993,** 65, 2040-2047.

32. Miller, C. E., Chemometrics and NIR: A Match Made in Haven. *Am.Pharm. Rev.* **1999,** 2, (2), 41-48.

33. Miller, C. E., Chemometrics for On-Line Spectroscopy Application: Theory and Practice. *J. Chemom.* **2000,** 14, 513-528.

34. Box, G. E. P.; Hunter, W. G.; Hunter, J. S., *Statistics for Experimenters: An Introdcution to Design, Data Analysis and Model Building*. John Wiley & Sons: New York, 1978.

35. Eriksson, L.; Johansson, E.; Kettaneh-wold, N.; C.Wikstorm; Wold, S., *Design of Experiments-Principles and Application*. Learnways AB: 2000.

36. Meyers, R. H.; Montgomery, D. C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley-Interscience: New York, 1995.

37. L. Erikson, E. J.; Wikström, C., Mixture Design - Design Generation, PLS Analysis and Model Usage. *Chemom. Intell. Lab. Syst.* **1998,** 43, 1-24.

38. Lundstedt, T., Experimental Design and Optimization. *Chemom. Intell. Lab. Syst.* **1998,** 42, 3-40.

39. Mobley, P. R.; Kowalski, B. R.; Workman, J. J.; Bro, R., Review of Chemometrics Applied to Spectroscopy:1985-1995. *Appl. Spec. Rev.* **1996,** 31, 347-368.

40. Geladi, P.; MacDougall, D.; Martens, H., Linearization and Scatter-Correction for

Near-Infrared Reflectance Spectra of Meat. *Appl. Spectrosc.* **1985,** 39, 491-500.

41. Isaksson, T.; Naes, T., The Effect of Multiplicative Scatter Correction and Linearity Improvement on NIR Spectroscopy. *Appl. Spectrosc.* **1988,** 42, 1273-1284.

42. Brown, C. W.; Obremski, C. W., Multicomponent Quantitative Analysis. *Appl. Spec. Rev.* **1984,** 20, 373-418.

43. Sutter, J. M.; Kalavias, J. M.; Lang, P. M., Which Principal Components to Utilize for Principal Component Regression. *J. Chemom.* **1992,** 1, 19-31.

44. Jolliffe, I. T., *Principal Component Analysis*. Spriinger-Verlag: New York, 1986.

45. Lorber, A.; Wangen, L. E.; Kowalski, B. K., A Theoretical Foundation for the PLS Algorithm. *J. Chemom.* **1987,** 1, 1-17.

46. Geladi, P.; Kowalski, B., Partial least-squares regression: A tutorial. *Analytica Chimica Acta* **1986,** 185, 1-17.

47. Wise, B. W.; Kowalski, B. R., *Process Chemometrics in Process Analytical Chemistry*. Blackie Academic: London, 1995.

48. Martens, H.; Martens, M., *Multivariate Analysis of Quality*. John Wiley and Sons: New York, 2001.

49. Jong, S. d., SIMPLS: an Alternative Approach to Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **1993,** 18, 251-263.

50. Manne, R., Analysis of Two PLS Algorithm for Multivariate Calibration. *Chemom. Intell. Lab. Syst.* **1987,** 2, 187-197.

51. Zou, Y., et al., Making Your Best Case - near-Ir Spectral Identification of Soil. *analtical*

chemistry **1993,** 65, (9), A434-A439.

52. Fukunaga, K., *Introduction to Statistical Pattern Recognition*. Academic Press: San Diego, 1990.

53. Mark, H. L.; Tunnell, D., Qualitative near-infrared reflectance analysis using Mahalanobis distances. *Anal Chem* **1985,** 57, (7), 1449-1456.

54. CANDOLFI, A.; DE, M. R.; MASSART, D. L.; HAILEY, P. A.; HARRINGTON, A. C., Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA. *Journal of pharmaceutical and biomedical analysis* **1999,** 19, (923-935).

55. Krzanowski, W. J., *Principles of Multivariate Analysis - A User's Perspective*. Oxford University Press: Oxford, 1988.

56. Leardi, R., Genetic algorithm-PLS as a tool for wavelength selection in spectral data sets. *Data Handling in Science and Technology* **2003,** 23, 169-196.

57. Leardi, R., Application of genetic algorithm-PLS for feature selection in spectral data sets. *Journal of Chemometrics* **2000,** 14, (5-6), 643-655.

58. Holland, J. H., Adaptation in Natural and ArtificialSystems. **1992**.

59. Brook, R. J.; Arnold, C., *Applied regression analysis and experimental design*. Marcel Dekker Inc: New York, 1985.

60. Rencher, A. C., Methods of Multivariate Analysis. **2002**.

61. Weiss, P., New lenses create distorted images for digital enhancement. *Science News* 2003, p 200.

62. Thermogalactic http://www.galactic.com/

63. DARPA http://www.darpa.mi

64. DeVerse, R. A.; Hammaker, R. M.; Fateley, W. G., Realization of the Hadamard Multiplex Advantage Using a Programmable Optical Mask in a Dispersive Flat-Field Near-Infrared Spectrometer. *Applied Spectroscopy* **2000,** 54, (12), 1751-1758.

65. Buice, R. G.; Lodder, R. A., Determination of Cholesterol Using a Novel Magnetohydrodynamic Acoustic-Resonance Near-IR Spectrometer. *Appl. Spectrosc.* **1993,** 47, 887-890.

66. Fong, A.; Hieftje, G. M., Near-IR Multiplex Bandpass Spectrometer Using Liquid Molecular Filters. *Applied*

Spectroscopy **1995,** 49, (4), 493-498.

67. Symons, W. C.; Whites, K. W.; Lodder, R. A., Theoretical and Experimental Characterization of a Near-Field Scanning Microwave Microscope (NSMM). *IEEE Transactions on Microwave Theory and Techniques* **2003,** 51, (1), 91-99.

68. Urbas, A.; Manning, M. W.; Daugherty, A.; Cassis, L. A.; Lodder, R. A., Near-Infrared Spectrometry of Abdominal Aortic Aneurysm in the ApoE -/- Mouse. *Anal. Chem.* **2003, 75**, (14), 3650-3655.

69. Cassis, L. A.; Lodder, R. A., Near-IR imaging of atheromas in living arterial tissue. *Anal Chem* **1993,** 65, (9), 1247-56.

70. http://www.fda.gov/medwatch/safety/2006/safety06.htm#azathioprin. In 2006.

71. Dempsey, R. J.; Davis, D. G.; Buice, R. G.; Lodder, R. A., Biological and medical applications of near-infrared spectroscopy. *Applied Spectroscopy* **1996,** 50, 18A-34A.

72. Drennen, J. K.; Lodder, R. A., Nondestructive near-infrared analysis of intact tablets for determination of degradation products. *J Pharm Sci* **1990,** 79, (7), 622-7.

73. Urbas, A.; Manning, M. W.; Daugherty, A.; Cassis, L. A.; Lodder, R. A., Near-infrared spectrometry of abdominal aortic aneurysm in the ApoE-/- mouse. *Anal Chem* **2003,** 75, (14), 3318-23.

74. Fischer, M. R.; Hieftje, G. M., Near-IR multiplex bandpass spectrometer utilizing polymer filters. *Applied Spectroscopy* **1996,** 50, (10), 1246-1252.

75. Fong, A.; Hieftje, G. M., Near-IR multiplex bandpass spectrometer utilizing liquid molecular filters. *Applied Spectroscopy* **1995,** 49, (4), 493-498.

76. Soyemi, O. O.; Haibach, F. G.; Gemperline, P. J.; Myrick, M. L., Nonlinear optimization algorithm for multivariate optical element design. *Applied Spectroscopy* **2002,** 56, (4), 477-487.

77. Soyemi, O. O.; Haibach, F. G.; Gemperline, P. J.; Myrick, M. L., Design of angle-tolerant multivariate optical elements for chemical imaging. *Applied Optics* **2002,** 41, (10), 1936-1941.

78. Soyemi, O.; Eastwood, D.; Zhang, L.; Li, H.; Karunamuni, J.; Gemperline, P.; Synowicki, R. A.; Myrick, M. L., Design and testing of a multivariate optical element: The first demonstration of multivariate optical computing for predictive spectroscopy. *Analytical Chemistry* **2001,** 73, (6), 1069-1079.

79. Prakash, A. M. C.; Stellman, C. M.; Booksh, K. S., Optical regression: a method for improving quantitative precision of multivariate prediction with single channel spectrometers. *Chemometrics and Intelligent Laboratory Systems* **1999,** 46, (2), 265-274.

80. Myrick, M. L.; Soyemi, O. O.; Schiza, M. V.; Farr, J. R.; Haibach, F.; Greer, A.; Li, H.;

Priore, R., Application of multivariate optical computing to simple near-infrared point measurements. *Proceedings of SPIE-The International Society for Optical Engineering* **2002,** 4574, 208-215.

81. Myrick, M. L.; Soyemi, O. O.; Haibach, F.; Zhang, L.; Greer, A.; Li, H.; Priore, R.; Schiza, M. V.; Farr, J. R., Application of multivariate optical computing to near-infrared imaging. *Proceedings of SPIE-The International Society for Optical Engineering* **2002,** 4577, 148-157.

82. Myrick, M. L.; Soyemi, O.; Li, H.; Zhang, L.; Eastwood, D., Spectral tolerance determination for multivariate optical element design. *Fresenius' Journal of Analytical Chemistry* **2001,** 369, (3-4), 351-355.

83. Myrick, M. L.; Soyemi, O.; Karunamuni, J.; Eastwood, D.; Li, H.; Zhang, L.; Greer, A. E.; Gemperline, P., A single-element all-optical approach to chemometric prediction. *Vibrational Spectroscopy* **2002,** 28, (1), 73-81.

84. Cassis, L. A.; Dai, B.; Urbas, A.; Lodder, R. A., In vivo applications of a molecular computing-based high-throughput NIR spectrometer. *Proc. SPIE-Int. Soc. Opt. Eng.* **2004,** 5329, (239-253).

85. Dai, B.; Urbas, A. A.; Douglas, C. C.; Lodder, R. A., Molecualr Factor Computing based Spectroscopy for Predictive Spectroscopy. *Pharmaceutical Research* **2007,** Paper Accepted.

86. Cassis, L. A.; Yates, J.; Symons, W. C.; Lodder, R. A., Cardiovascular near-infrared imaging. *Journal of Near Infrared Spectroscopy* **1998,** 6, (1-4), A21-A25.

87. Dayal, B. S.; MacGregor, J. F., Improved PLS algorithms. *Journal of Chemometrics* **1997,** 11, (1), 73-85.

88. Barker, M.; Rayens, W. S., A partial least squares paradigm for discrimination. *Journal of Chemometrics* **2003,** 17, 166-173.

89. Schwartz, C. Integrated Sensing and Processing. http://www.darpa.mil/dso/thrust/math/isp.htm

90. Nelson, M. P.; Aust, J. F.; Dobrowolski, J. A.; Verly, P. G.; Myrick, M. L., Multivariate optical computation for predictive spectroscopy. *Analytical Chemistry* **1998,** 70, (1), 73-82.

91. Dai, B.; Urbas, A. A.; Douglas, C. C.; Lodde, R. A., An Algorithm for Molecular Filters Selection in Molecular Factor Computing using Genetic Algorithm Based Multivariate Linear Regression. *Applied Spectroscopy* **2007,** Paper submitted.

92. Lavine, B. K.; C.E.Davidson; A.J.Moores; P.R.Griffiths, Raman Spectroscopy and Genetic Algorithm for the Classification of Wood Types. *Appl. Spectrosc.* **2001,** 55, (8), 960-966.

93. Jarvis, R. M.; Goodacre, R., Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. *BIOINFORMATICS* **2004,** 21, (7), 860-868.

94. Esteban-Diez, I.; Gonzalez, J. M.; Gomez-Camara, D.; Millan, C. P., Multivariate calibration of near infrared spectra by orthogonal WAVElet correction using a genetic algorithm. *analytica cheimica acta* **2005**, (555), 84-95.

95. Urbas, A. A.; Dai, B.; Ali Shamsaie; Lodder, R. A., Molecular Factor Computing Near-Infrared Spectroscopy for Differentiating Cholesterol, Collagen and Elastin through Red Blood Cell Solutions. *analtical chemistry* **2006**, Submitted.

96. Meza, J. C. S. C. P.; Caraballo, W.; Conde, C.; T. Li; Morris, K. R.; Romanach, R. J., On Line Non-Destructive Determination of Drug Content in Moving Tablets Using Near Infrared

Spectroscopy. *Journal of Process Analytical Technology* **2005,** 2, (5), 8-14.

97. Ridder, T. D.; Hendee, S. P.; Brown, C. D., Noninvasive Alcohol Testing Using Diffuse Reflectance Near-Infrared Spectroscopy. *Applied Spectroscopy* **2005,** 59, 181-189.

98. Guidance for Industry PAT - A Framework for Innovative Pharmaceutical. In Manufacturing and Quality Assurance, U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation (CDER), Research (CDER),Center for Veterinary, Medicine (CVM), Office of Regulatory Affairs (ORA): September 2004.

99. Hussain, A. S., Process Analytical Technology: A First Step in a Journey Towards the Desired State. *Journal of Process Analytical Technology* **2005,** 2, (1), 8-13.

100. Byrn, S. R.; Liang, J. K.; Bates, S.; Newman, A. W., PAT - Process Understanding and Control of Active Pharmaceutical Ingredients. *Journal of Process Analytical Technology* **2006,** 3, (6), 14-19.

101. Beebe, K. R.; Kowalski, B. R., Introduction to multivariate calibration & analysis. *Anal Chem* **1987,** 59, 1007A-1017A.

102. Bialkowski, S. E., Species discrimination and quantitative estimation using incoherent linear optical signal processing of emission signals. *Analytical Chemistry* **1986,** 58, (12), 2561-2563.

103. Haibach, F. G.; Greer, A. E.; Schiza, M. V.; Priore, R. J.; Soysmi, O. O.; Myrick, M. L., On-line reoptimization of filter designs for multivariate optical elements. *Applied optics* **2003,** 42, (10), 1833-1838.

104. Haibach, F. G.; Myrick, M. L., Precision in multivariate optical computing. *Applied*

optics **2004,** 43, (10), 2130-2140.

105. Cassis, L. A.; Urbas, A.; lodder, R. A., Hyperspectral Intergrated Computational Imaging. *Anal. Bioanal. Chem.* **2005,** 868, 382.

106. Huang, E.; S. H. Cheng, H.; Dressman, J. P.; Tsou, M.-H.; Horng, C.-F.; Iversen, A. B. E. S.; Liao, M.; Chen, C.-M.; West, M.; Nevins, J. R.; Huang, A. T., Gene Expression Predictors of Breast Cancer Outcomes. *Lancet* **2003, 361**, 1590-1596.

107. Lodder, R. A.; Hieftje, G. A., Detection of Subpopulations in Near-Infrared Reflectance Analysis. *Applied Spectroscopy* **1988,** 42, (8), 1500-1512.

108. MCluney, W. R., *Introduction to Radiometry and Photometry*. Artech House: Massachusetts, 1994.

109. Saptari, V. A., *Fourier Transform Spectroscopy Instrumentation Engineering*. SPIE Press: Washington, 2003.

110. Morimoto; McClure, W. F.; Stanfield, D. L., Hand-Held NIR Spectrometry: Part I: An Instrument Based upon Gap-Second Derivative Theory. *Applied Spectroscopy* **2002,** 55, (2), 182 - 189.

111. Morimoto; McClure, W. F.; Stanfield, D. L., Hand-held NIR spectrometry. Part II: an economical no-moving parts spectrometer for measuring chlorophyll and moisture. *Applied Spectroscopy* **2002,** 56, (6), 720-724.

Vita

Bin Dai
Born: May 21, 1977
Xiapu, Fujian Province, CHINA

EDUCATION

- Tongji University, Bachelor of Science, July 1999

PROFESSIONAL

- Ph.D. Summer Intern, the Procter & Gamble Company, Summer, 2006
- Research Assistant, University of Kentucky Department of Chemistry, 2002-2007
- Teaching Assistant, University of Kentucky Department of Chemistry, 2002-2003

HONORS AND AWARDS

- 1995-1998 People Scholarship, Tongji University, Shanghai, China

PUBLICATIONS

(1) Bin Dai, Aaron Urbas, Robert A. Lodder. Genetic Algorithm Based Linear Discriminant Analysis for Molecular Filters Selection in Molecular Factor Computing. *Applied Spectroscopy.* Submitted 2007.

(2) Bin Dai, Aaron Urbas, Robert A. Lodder. Genetic Algorithm Based Multivariate Linear Regression for Molecular Filters Selection in Molecular Factor Computing. *Applied Spectroscopy*. Submitted 2007.

(3) Bin Dai, Aaron Urbas, Robert A. Lodder. Molecular Factor Computing for Predictive Spectroscopy. *Pharmaceutical Research.* Accepted Nov. 2006.

(4) Aaron Urbas, Bin Dai, Ali Shamsaie, Robert A. Lodder. Molecular Factor Computing Near-Infrared Spectroscopy for Differentiating Cholesterol, Collagen and Elastin through Red Blood Cell Solutions. *Anal. Chem* Submitted 2007

(5) Bin Dai, Aaron Urbas, Robert A Lodder. Prospects for implantable sensors powered by near infrared rechargeable batteries. *NIR news* Vol. 17 No. 1 (**2006**) 14-18.

(6) Bin Dai, Robert A. Lodder. Parallel Hyperspectral Integrated Computational Imaging. *Conference Proceedings 2004 International Symposium on Distributed Computing and Applications to Business, Engineering and Science.* 256-261.

(7) Lisa A. Cassis, Bin Dai, Aaron A. Urbas, Robert A. Lodder, In Vivo Applications of A Molecular Computing-based High-throughput NIR Spectrometer. *Proc. SPIE-Int. Soc. Opt. Eng*. **2004**, 5329. 239-253.

(8) Yanchun Wang, Yufei Yuan, Binglin Ding, Peijiao Wang, Xiangfeng Sun, Bin Dai**,** Shuping Wu, Zhiqin Jian. Study on Electron Transfer Mechanism of Nucleic Acid Precursors and Their Modified Structure *Chinese Journal of Organic Chemistry* Vol.2, **2001**,No.11, 890-897

PRESENTATIONS

(1) Molecular Factor Computer-based High Throughput NIR Spectrometer. *The Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy.* Mar. 7-12, 2004, Chicago, IL.

(2) Parallel Hyperspectral Integrated Computational Imaging. *2004 International Symposium on Distributed computing and Applications to Business, Engineering and Science.* Sep. 11-17, 2004, Wuhan, China.

(3) Cholesterol, Elastin and Collagen Classification by Hyperspectral Integrated Computational Imaging. *The Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy.* Feb. 28-Mar. 4, 2005, Orlando, FL.

(4) Molecular Factor Computing Based Remote Ethanol Sensing. *The Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy.* Feb. 28-Mar. 4, 2005, Orlando, FL.

(5) A Non-invasive Near IR Integrated Alcohol Sensor System. *NIAAA Biosensors Annual Review Meeting.* July 21-22, 2005, Bethesda, Maryland.

(6) In Vivo Multimodal Remote Sensing of Ethanol Concentrations. *The Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy.* Mar. 12-17, 2006, Orlando, FL.

(7) Molecular Factor Computing Near-Infrared Spectroscopy for Differentiating Cholesterol, Collagen and Elastin through Red Blood Cell Solutions. *The Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy.* Mar. 12-17, 2006, Orlando, FL