University of Kentucky
UKnowledge

University of Kentucky Doctoral Dissertations | Graduate School

2009

BOOTSTRAP ENHANCED N-DIMENSIONAL DEFORMATION OF SPACE WITH ACOUSTIC RESONANCE SPECTROSCOPY

David John Link
University of Kentucky, dlink4884@gmail.com

Right click to open a feedback form in a new tab to let us know how this document benefits you.

Recommended Citation

Link, David John, "BOOTSTRAP ENHANCED N-DIMENSIONAL DEFORMATION OF SPACE WITH ACOUSTIC RESONANCE SPECTROSCOPY" (2009). *University of Kentucky Doctoral Dissertations*. 728.
https://uknowledge.uky.edu/gradschool_diss/728

This Dissertation is brought to you for free and open access by the Graduate School at UKnowledge. It has been accepted for inclusion in University of Kentucky Doctoral Dissertations by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

ABSTRACT OF DISSERTATION

David John Link

The Graduate School

University of Kentucky

2009

BOOTSTRAP ENHANCED N-DIMENSIONAL DEFORMATION OF SPACE WITH ACOUSTIC RESONANCE SPECTROSCOPY

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
David John Link
Lexington, Kentucky

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

2009

Copyright © David John Link 2009

ABSTRACT OF DISSERTATION

BOOTSTRAP ENHANCED N-DIMENSIONAL DEFORMATION OF SPACE WITH ACOUSTIC RESONANCE SPECTROSCOPY

Acoustic methods can often be used with limited or no sample preparations making them ideal for rapid process analytical technologies (PATs). This dissertation focuses on the possible use of acoustic resonance spectroscopy as a PAT in the pharmaceutical industry. Current good manufacturing processes (cGMP) need new technologies that have the ability to perform quality assurance testing on all products. ARS is a rapid and non destructive method that has been used to perform qualitative studies but has a major drawback when it comes to quantitative studies. Acoustic methods create highly non linear correlations which usually results in high level computations and chemometrics.

Quantification studies including powder contamination levels, hydration amounts and active pharmaceutical ingredient (API) concentrations have been used to test the hypothesis that bootstrap enhanced n-dimensional deformation of space (BENDS) could be used to overcome the highly non linear correlations that occur with acoustic resonance spectroscopy (ARS) eliminating a major drawback with ARS to further promote the device as a possible process analytical technology (PAT) in the pharmaceutical industry. BENDS is an algorithm that has been created to calculate a reduced linear calibration model from highly non linear relationships with ARS spectra. ARS has been shown to correctly identify pharmaceutical tablets and with the incorporation of BENDS, determine the hydration amount of aspirin tablets, D-galactose contamination levels of D-tagatose powders and the D-tagatose concentrations in resveratrol/D-tagatose combinatory tablets.

KEYWORDS: acoustic resonance spectroscopy, non linear calibration, chemometrics, process analytical technology, multivariate analysis

David J. Link

Student's Signature

23 June 2009

Date

BOOTSTRAP ENHANCED N-DIMENSIONAL DEFORMATION OF SPACE WITH ACOUSTIC RESONANCE SPECTROSCOPY

By

David John Link

Robert A. Lodder

Director of Dissertation

Robert Grossman

Director of Graduate Studies

23 June 2009

Date

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

David John Link

The Graduate School

University of Kentucky

2009

BOOTSTRAP ENHANCED N-DIMENSIONAL DEFORMATION OF SPACE WITH ACOUSTIC RESONANCE SPECTROSCOPY

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
David John Link
Lexington, Kentucky

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

2009

Copyright © David John Link 2009

ACKNOWLEDGMENTS

I would like to extend my upmost gratitude and appreciation to all those that have helped me in the preparation and execution of my dissertation. I would like to give a special thank you to my advisor, Dr. Robert Lodder, for teaching me that research is only as good as your ability to sell it. I would also like to give an endearing thank you to my wife, Hollie; without your support through the most trying and stressful times, I would have never made it as far as I have. The list below is by no means exhaustive but I would like to give a warm thank you to each and every person below.

My Graduate Committee, University of Kentucky

Dr. Robert Grossman, University of Kentucky

Dr. Jay Baltisberger, Berea College

Dr. Matthew Saderholm, Berea College

Thaddeus Hannel, University of Kentucky

Chemistry Office, University of Kentucky

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SECTION ONE – INTRODUCTION AND BACKGROUND

Chapter One: Acoustic Resonance Spectroscopy

According to the Oxford English Dictionary (OED), sound is defined as the vibration transmitted through a medium of matter with frequencies that can be heard with the human ear [1]. In spectroscopy, sound is part of the larger umbrella term, acoustics, which encompasses sound, ultrasound and infrasound. Acoustic methods include the generation, propagation, resonation and acquisition of mechanical waves and vibrations. The reason for using acoustics is to use a nondestructive and rapid method that is specific to many different physical and chemical properties. Acoustic velocity [2], ultrasonic attenuation [3], acoustic reflection [4] and acoustic emission [5] are different types of acoustic properties that are used in infer other analytical properties of interest.

The human ear and brain essentially form an acoustic spectrometer that deciphers different physical properties. A trained ear can decipher the frequencies present in complex waveforms and tell the difference between different frequencies played in succession. Studies have even demonstrated that individuals can acoustically distinguish the shape of different vibrating plates [6]. The different experiments performed on blindfolded individuals included specifying shapes, dimensions and materials of different objects that were struck by a pendulum. Another study tested blindfolded participants listening to partially occluded sound passing through a doorway [7]. The individuals were asked to listen to a doorway-like structure with obstructions of different apertures. They were surveyed to determine whether they could infer the hole to be large enough for the individual to pass through the doorway. Judgments were found to be relatively accurate, which suggests that individuals can hear surfaces that obstruct apertures like doorways. Acoustic resonance spectroscopy (ARS) coupled to a computer works in a similar manner to analyze samples.

ARS Instrumentation

Most published works in acoustics have been in the ultrasonic region and their instrumentation has dealt with propagation through a medium and not a resonance effect. The ARS has come a long way since its conception in 1988, when researchers designed a V shaped quartz rod instrument that utilized ultrasonic waves to obtain signatures of micro liter volumes of different liquids [8]. The instrument now has the ability to use a

larger region of the acoustic spectrum, including sonic and ultrasonic [9]. Since the conception of the ARS, it has evolved and has been used to differentiate wood species, pharmaceutical tablets, determine burn rates and determine dissolution rates of tablets [10-13]. In 2007, *Analytical Chemistry* featured the past and current work of the quartz-rod ARS discussing the potential of acoustics in the analytical chemistry and engineering fields [14].

The ARS is designed to create a fingerprint for different samples by constructive and destructive interferences. Figure 1.1 is a schematic of the quartz rod ARS and illustrates the path of the sound through the quartz rod. A function generator is depicted as the source (A), though any device that is capable of outputting sound in voltage form could be used (e.g., a CD-player, MP3 player or sound card). White noise is generated and the voltage is converted into a sound wave by a piezoelectric transducer disc (B), which is coupled to the quartz rod. The sound is shown as a blue sinusoidal wave (C), and resonates along the quartz rod, where two key interactions occur. A portion of the energy (red) is introduced into the sample and interacts in a specific manner dependent on the sample, and another portion of the energy (blue) continues unaltered through the quartz rod. The two energies still have the same frequency, but they will have most likely show changes in their phase and amplitude. The two waves recombine after the sample (D) and constructive or destructive interference occurs, depending on the phase shift and amplitude change due to the sample. The altered combined energy (purple) is converted to an electrical voltage by another piezoelectric disc at the end of the quartz rod (E). The voltage is then recorded onto a computer by a sound card (F). The sample is coupled to the quartz rod at constant pressure, which is monitored by a pressure transducer that also acts as the sample holder. Rubber grommets are used to secure the quartz rod to a stable stand, minimizing acoustic coupling of the rod to the surroundings. Broadband "white" noise is used to obtain a full spectrum; however, most sound cards only operate between 20 and 22050 Hz. The waveform that is sent to the computer is a time based signal of the interactions of white noise with the sample. A Fast Fourier transform (FFT) is performed on the waveform to transform the time-based signal into the more useful frequency spectrum (see figure 1.2).

The AR spectrometer had drawbacks which were causing low consistency with the measurements and so three major alterations were completed. Originally piezoelectric strips were used which have low sensitivity and were replaced with piezoelectric discs (figure 1B and figure 1E). An amplifier was needed to obtain a valuable signal which in turn incorporated unwanted noise. The piezoelectric discs have a higher sensitivity and also created a louder excitation from the source. The discs increased the signal to noise of the instrument which is discussed later in the experiments. The amplifier was no longer needed which further reduced the noise. Another change made to the AR spectrometer was including a second receiving piezoelectric disc not coupled to the quartz rod. The two receiving piezoelectric discs were wired through two channels of a stereo input. Originally when scanning was in progress the entire lab had to be quiet which was inevitably impossible. The second receiving disc creates a background reference that can be subtracted from the coupled signal. The background reference can also pick up on background electrical noise that can fluctuate over time and alter the readings. Electrical noise was not completely removed by subtracting the background and therefore a third change was made to the AR spectrometer. All three of the piezoelectric discs were shielded using electrical tape and aluminum casings. The shielding caused the background electrical noise to remain consistently low.

Chapter One Figures

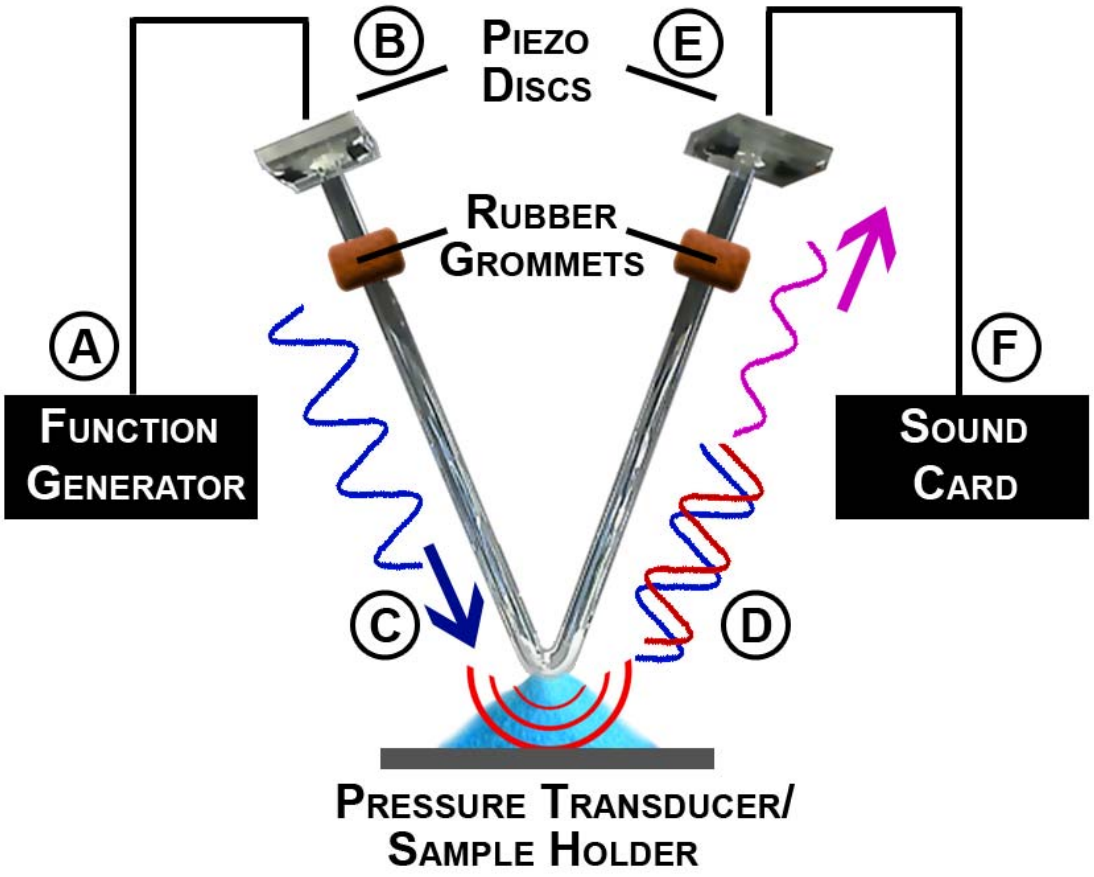


Figure 1.1: Schematic of the quartz rod AR spectrometer. A function generator is depicted as the source (A). White noise is generated and the voltage is converted into a sound wave by a piezoelectric disc (B) which is coupled to the quartz rod. The sound resonates down the quartz rod which is shown as a blue sinusoidal wave (C) and two key interactions occur. A portion of the energy (red) is introduced into the sample and interacts in a specific manner dependent of the sample and another portion of the energy (blue) continues unaltered through the quartz rod. The two waves recombine after the sample (D) and constructive or destructive interference occurs depending on the phase shift due to the sample. The altered combined energy (purple) is converted to an electrical voltage by another piezoelectric disc at the end of the quartz rod (E). The voltage is then recorded onto a computer by a sound card (F).

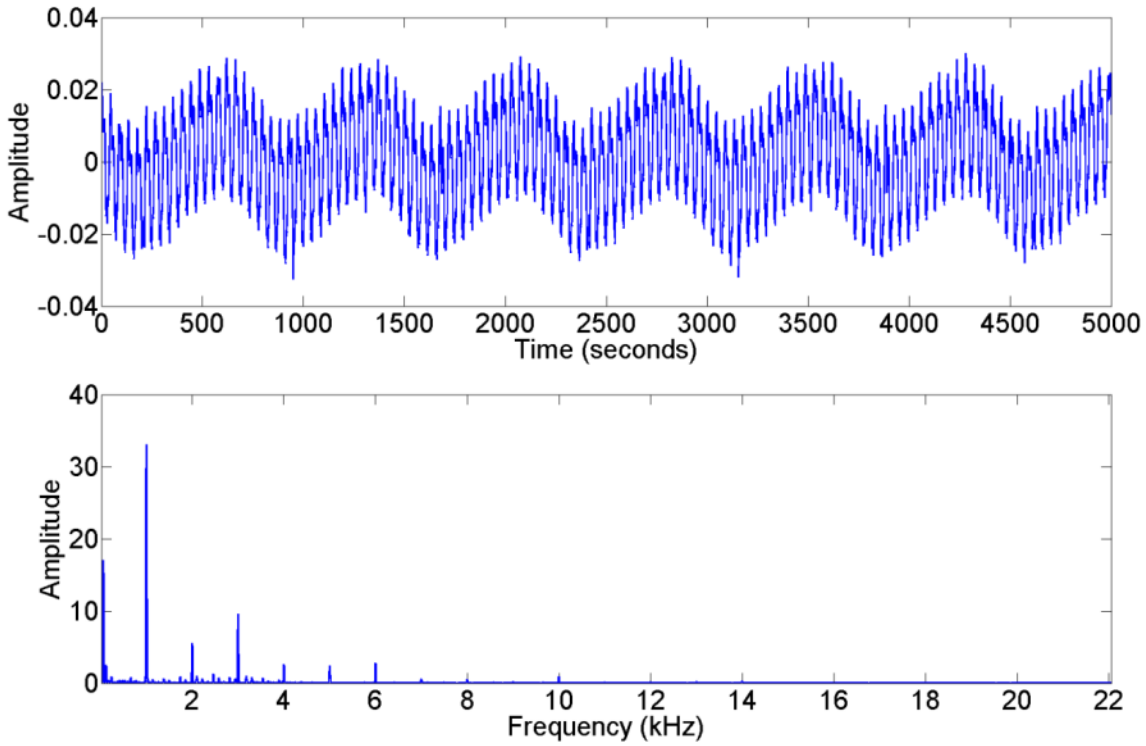


Figure 1.2: Fast Fourier Transform. Illustration of the time based waveform (top) and the frequency based spectrum (bottom). The frequency based spectrum is calculating using a fast Fourier transform of the time based waveform.

Copyright Statement

Copyright © David Link 2009

Chapter Two: Chemometrics

Multivariate calibration (MVC) methods are a vital part of many applications in analytical chemistry. Many different MVC methods are often being used in many industrial applications to relate easily measured spectra to parameters of interest. The food and pharmaceutical industries are two major places where a fast and reliable method is desired for constant quality control monitoring. Current quality control methods are often tedious, expensive and time-intensive which causes them to be performed off the production line. For example, a pharmaceutical manufacturer is mixing compounds to produce a drug and they need to know the purity of the product coming out before they can continue to the next stage of production. The manufacturer must first take a certain amount of randomly selected batches known as batch sampling, and test those with current methods to test the purity. The company now has two choices, either wait for the answer to come back from the lab or continue production while the testing is going on; either way they are taking a chance and risking money. A real time monitoring system involving optical sensors (i.e. near infrared) or acoustic sensors (i.e. acoustic resonance) could be used with an MVC method to determine the parameter of interest.

Multiple Linear Regression

MLR (also known as inverse least squares) is a straightforward extension of simple univariate linear regression and is used to generate a quantitative relationship between a group of predictor variables and a response:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon \qquad \textbf{2.1}$$

where y is the response, x_i are predictors, β_i are regression coefficients, and ε is the residual. Note that the additional terms, such as powers (x_i^k) or cross-terms ($x_i x_j$), can be included and the model remains linear even though the function may not be a straight line. For a data set of known responses and predictors, the linear model can be expressed in matrix form according to equation 5.10:

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon} \qquad \textbf{2.2}$$

where $\mathbf{y}$ ($m\times1$) is the column vector of responses, $\mathbf{X}$ ($m\times n$) is the matrix of predictors, $\mathbf{b}$ ($n\times1$) is the unknown column vector of regression coefficients, and $\boldsymbol{\varepsilon}$ ($m\times1$) is the column vector of residuals. For a unique solution to exist, m must be larger than n. Equation 5.11 provides the least squares estimate of the regression coefficients ($\hat{\mathbf{b}}$):

$$\hat{\mathbf{b}} = \left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y} \qquad \textbf{2.3}$$

where $\mathbf{X}^\mathbf{T}$ denotes the transpose of $\mathbf{X}$, and $\left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-1}$ denotes the inverse of $\mathbf{X}^\mathbf{T}\mathbf{X}$. This solution is only attainable when $\mathbf{X}$ is of full rank. The equation is useful theoretically but in practice has poor numerical properties and more robust methods are generally used.

For a typical spectroscopic data set we are often confronted with the situation of having significantly more variables than samples ($m<n$). One solution to this dilemma is the selection of a subset of variables to constrain n to be smaller than m. This approach is used frequently in practice but is not without drawbacks. In spectroscopic data, for instance, the absorbance values in a spectrum at multiple wavelengths tend to vary together with changing constituent concentration. This effect is known as collinearity,

and causes instability in the mathematical solution. Another drawback is that the removal of variables from the model discards potentially useful information. Finally, determination of an optimal subset of variables, especially when thousands may be available, presents considerable difficulties. An alternative to the problem of too many variables involves factor-based approaches for dimensionality reduction. One such approach is principal component analysis, a discussion of which follows.

Principal Component Regression

Both near infrared spectroscopy (NIRS) and ARS both involve what is called an "ill posed problem," where there are more variables than observations [15]. Traditional calibration methods like ordinary least squares regression breaks down in ill posed problems which is why MVC techniques are used to reduce the data. Principal component regression (PCR) and partial least squares regression (PLS) are the two most common MVC methods in the literature [11][16-23]. PCR and PLS are both valuable MVC methods and are both incorporated with a linear assumption of the correlation [24-25]. PCR is calculated using singular vector decomposition (SVD) which is a data reduction technique that transforms the data for regression. Once SVD is performed, the new variables are no longer under the ill posed problems. The number of observations is now greater than the number of variables and least squares regression can be performed on the data.

PCR is a regression method that is based on simple properties of linear algebra. Principal component analysis (PCA) essentially is a factorization known as SVD [26]. Consider an n by k matrix $\boldsymbol{X}$ and let $t = \min\{n,k\}$. The SVD of $\boldsymbol{X}$ is the factorization

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{P}^{\top} \qquad \textbf{2.4}$$

where $\boldsymbol{U}$ is an n by t orthogonal matrix, $\boldsymbol{P}$ is a k by t orthogonal matrix and $\boldsymbol{D} = \text{diag}\{d_1,\ldots,d_t\}$ is a t by t diagonal matrix with elements $d_1 \geq \ldots \geq 0$. The $r \leq t$ non-zero values d_i are the singular values of $\boldsymbol{X}$. The last $t - r$ zero d_i may be discarded along with the last $t - r$ columns of $\boldsymbol{U}$ and $\boldsymbol{P}$, giving the SVD on reduced form which is done for PCR. The matrix $\boldsymbol{P}$ is sometimes referred to as the loadings matrix with its values being called the loadings. The matrix $\boldsymbol{T}$ where

$$\boldsymbol{T = UD} \qquad \textbf{2.5}$$

is called the scores matrix and its columns $t_1,\ldots,t_t$ are called the PC scores. The PC scores are a reduced form of the variance in the matrix $\boldsymbol{X}$. PCA, in effect, takes a cloud of data points, rotates and projects it onto a space of lower dimension, selecting the directions in the data space with maximum variability, or equivalently high information. For example, a spectral block $\boldsymbol{X}$ contains a lot of redundant information, because the absorbance for adjacent frequencies is highly correlated, and because features stemming from a given analyte are spread out over a range of different frequencies. PCA can provide a few factors (scores) that represent the different variances in the data. A regression can then be performed on the PC scores which is the last step of PCR (see page 34 for least squares linear regression). As with all regression methods over fitting must always be accounted for by cross validation techniques.

Partial Least Squares Regression

A problem can occur with PCR when there is high variability in $\boldsymbol{X}$ due to some other factor than the analytical property being studied. The PC scores that may statistically look the best may be due to these interferences rather than the analytical property. PLS is better suited to deal with this problem by forming variables that are relevant to $\boldsymbol{y}$ (the analytical property of interest). Assuming the same data as described above for PCR where $\boldsymbol{X}$ is an n by k centered data matrix and y is an n by 1 centered data vector. The PLS algorithm starts with the initialization $j = 1$, $\boldsymbol{X}_1 = \boldsymbol{X}$ and $y_1 = y$. The algorithm then proceeds through the following steps to find the first g latent variables:

1. Let $w_j = \boldsymbol{X}_j^{\top} y_j / \left\| \boldsymbol{X}_j^{\top} y_j \right\|$.
2. Let $t_j = \boldsymbol{X}_j w_j$.
3. Let $\hat{c}_j = t_j^{\top} y_j / t_j^{\top} y_j$.
4. Let $p_j = \boldsymbol{X}_j^{\top} t_j / t_j^{\top} t_j$.
5. Let $\boldsymbol{X}_{j+1} = \boldsymbol{X}_j - t_j p_j^{\top}$ and $y_{j+1} = y_j - t_j \hat{c}_j$.
6. Stop if $j = g$; otherwise let $j = j + 1$ and return to Step 1.

Now form the two k by g matrices $\boldsymbol{W}$, $\boldsymbol{P}$ and the n by g matrix $\boldsymbol{T}$ with columns w_j, p_j and t_j respectively, and form a column vector $\hat{c}(g \times 1)$ with elements $\hat{c}_j$. Let

$$\hat{X} = TP^{\top} = \sum_{j=1}^{g} t_j p_j^{\top} \qquad \textbf{2.6}$$

and

$$\hat{y} = T\hat{c} = XW(P^{\top}W)^{-1}\hat{c} \qquad \textbf{2.7}$$

which are the predicted values of $\boldsymbol{X}$ and y, respectively. The matrix W is orthogonal, and T has orthogonal columns.

In PLS, we seek the direction in the space of X, which yields the biggest covariance between X and y. This direction is given by a unit vector w, and is such that large variations in x-values are accompanied by large variations in the corresponding y-values. The unit vector $w_1(k \times 1)$ is thus formed by standardizing the covariance matrix for X and y. The n by 1 score vector t_1 is formed as a linear combination of the columns of X with weights w_1. As explained above, the relative weights are given by the covariances between y and each of the columns of X, and t_1 may be understood as the best linear combination of the columns of X for the purpose of predicting y. The latent vectors t_j are also called scores, similar to the terminology for PCA.

The regression coefficient $\hat{c}_1$ is calculated by ordinary linear regression of y on t_1. The k by 1 vector p_1 is the transpose of the vector of regression coefficients obtained from simple linear regressions of the columns of X on t_1. The n by k vector $X_2 = X - t_1 p_1^{\top}$ represents the residuals after regressing X on t_1, and correspondingly, $y_2 = y_1 - t_1\hat{c}_1$ are the residuals after regressing y on t_1. Step 5 ensures that the t_j-vectors become orthogonal and thus ensures that the multiple regression of y on T can be calculated one column at a time, as done in Step 3. After the first run through Steps 1-5, the procedure is repeated using the residuals X_2 and y_2. The algorithm then finds the best linear combination of the columns of X_2 for the purpose of predicting y_2, thus picking up any further structure in the connection between X and y not accounted for by t_1. This is repeated on and on, such that each run of the algorithm in principle reveals more and

more information about the connection between X and y. Just as for PCR the information accounted for by each step usually becomes less and less for each step taken.

After the g runs have been completed, the following relations hold:

$$X = TP^{\top} + X_{g+1} \qquad \textbf{2.8}$$

$$y = T\hat{c} + y_{g+1} \qquad \textbf{2.9}$$

The number of scores g should be chosen such that X_{g+1} contains no further information about y_{g+1}. In the extreme case where $X_j^{\top} y_j$ becomes zero, the algorithm is stopped prematurely. Further scores should be extracted as long as each new variable contributes significantly to the description of y.

Bootstrap Technique

In every instance of statistical analysis a data set **x** is used to calculate some statistic $t(\mathbf{x})$ in order to make an approximation of some quantity of interest. For demonstration purposes, the data in box 1 are plasma glucose concentrations

100	105	110	111	115	120	128	129	138	157	162	188

Box 1

for twelve women, 21 years of age or older of Pima Indian heritage; the scores are an ordered random sample from a larger data set of 768 women [27]. If the data in box 1 is **x** then $t(\mathbf{x})$ could be something simple like their mean, $\bar{\mathbf{x}}$. The common next step in statistics is to ask the question of how accurate is $t(\mathbf{x})$. Since we are dealing with the mean we simply look at the standard deviation or standard error,

$$se(x) = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \right)^{1/2} \qquad \textbf{2.10}$$

The results on our data set **x** would be reported as $\bar{\mathbf{x}} = 130 \pm 26.6$, which can easily be applied to a normal Gaussian distribution and a confidence interval can be calculated.

Unfortunately most statistical measures do not simple equations such as the mean and standard deviation. Consider PCR and PLS as described in chapter 1, it takes pages to simply explain the statistical calculations and how would you find their errors according to a distribution. Bootstrapping was created for this exact reason, to provide a technique

of finding a bootstrap error for any statistical measure proportional to the true population [28]. To demonstrate bootstrapping a more complex statistic than the ordinary mean is needed, but to keep it simple $t(\mathbf{x})$ can be the 25% trimmed mean, $\bar{\mathbf{x}}\{0.25\}$. A similar explanation of bootstrapping can be found by the creator of bootstrapping, Bradley Efron in a paper published in *Science* in 1991 [29]. The 25% trimmed mean is defined as the average of the middle 50% of the data. The data is ordered and the lower and upper 25% of the data is excluded and the remaining data is averaged.

$$\bar{\mathbf{x}}\{0.25\} = \frac{111 + 115 + 120 + 128 + 129 + 138}{6} \qquad \textbf{2.11}$$

The equation is simple for this case, though there is not a universal equation for this method in part due to if the number of values in **x** is not divisible by four, interpolation is required. For our demonstration data set, the $\bar{\mathbf{x}}\{0.25\} = 124$. The next step again is to find out how accurate $t(\mathbf{x})$ is but the standard error equation is only for the ordinary mean. In place of simple equations, bootstrapping uses computer power to obtain a numerical estimate of the standard error.

The bootstrap algorithm randomly samples from data set **x** in replacement of the original data **x**. Bootstrap data and statistical measure will be denoted $\mathbf{x}^*$ and $t^*(\mathbf{x})$, respectively. Each new data, $\mathbf{x}^*$ has the same number elements of **x** but consists of values randomly pulled from **x**. Values can be repeated because the number of bootstrap sets created is a simulation of the true population's distribution. Assume *B* bootstrap sets are taken, and now the statistical measure of each $\mathbf{x}^*$ is taken, in this case the $t^*(\mathbf{x})$ is the $\bar{\mathbf{x}}\{0.25\}$. The empirical standard deviation of the *B* bootstrap trimmed means is the bootstrap estimate of the standard deviation for the trimmed means. In other words, the standard deviation of all the bootstrap trimmed means is taken. Since the bootstrap population is analogous to the true population then the standard deviation of the bootstrap trimmed means is a representation of the error estimate of the original data, **x** trimmed mean. Table 1 gives the bootstrap error for the demonstration data at different values of *B* which can be compared with a true standard error of twelve randomly sampled values from the full dataset of 768 women of 8.66. Figure 9 is a visual representation of the bootstrap method adapted from afore mentioned paper in *Science* [29].

Bootstrapping is used with BENDS by weighting the final manifold by the standard deviation of the bootstrap manifolds. The inverse of the standard deviation of the manifolds at each instance in the spectral data is normalized and used as weights as described in the BENDS algorithm section below.

Chapter Two Tables

Table 2.1 Bootstrap estimates of error for the trimmed mean.

Bootstrap Replicates (*B*)	Bootstrap Estimate of Error(±)
50	7.47
100	8.56
200	8.61
500	8.75
1000	8.85

Chapter Two Figures

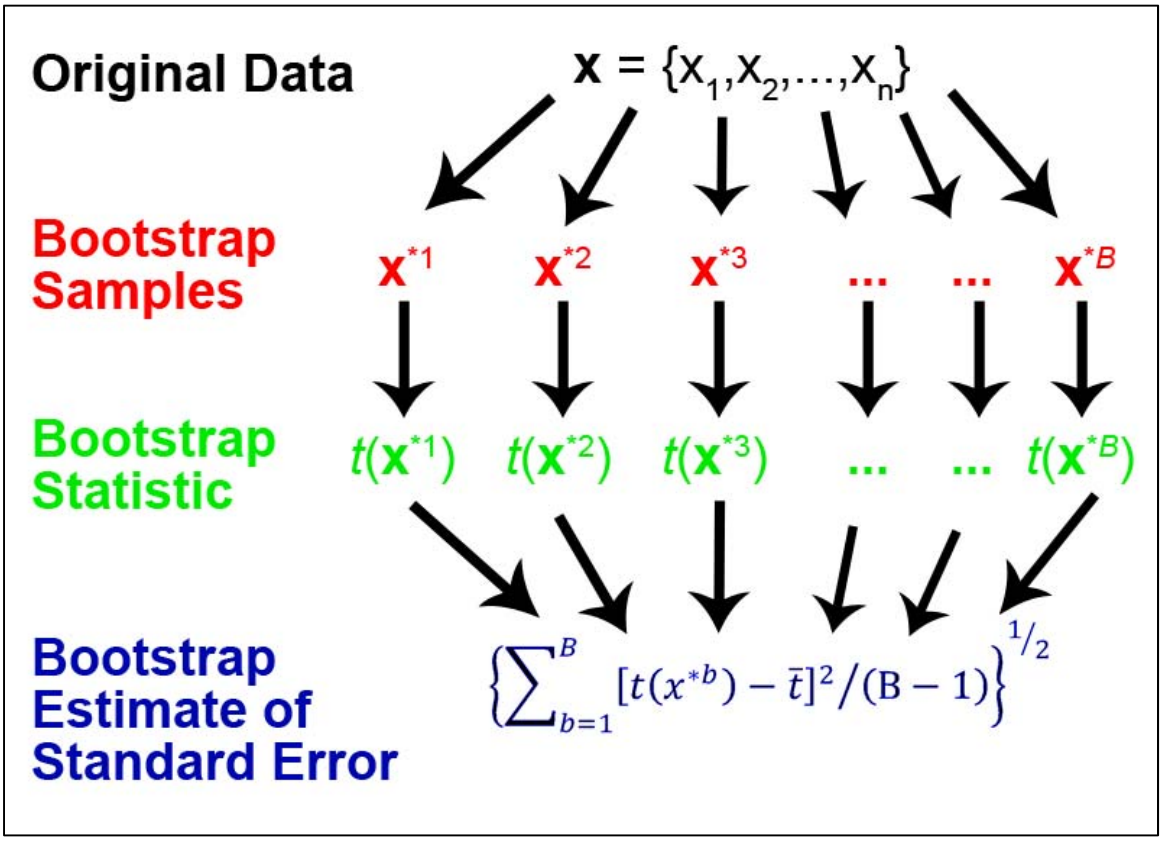


Figure 2.1: Statistical bootstrapping demonstration. The original data (black) is randomly sampled to B bootstrap sample sets (red) and the statistic *t*(x*) is performed (green) on each bootstrap sample set. The bootstrap estimate of the standard deviation *t*(x*) is performed by calculated the standard deviation off all the *t*(x*) (blue). Figure adapted from the paper in *Science* by Bradley Efron [29].

Copyright Statement

Copyright © David Link 2009

SECTION TWO – ACOUSTIC RESONANCE SPECTROSCOPY

Chapter Three: Integrated Sensing and Processing - Acoustic Resonance Spectrometry (ISP-ARS) in Differentiating D-Tagatose and Other Toll Manufactured Drugs

Introduction

Over 200,000 Americans die each year due to complications from type 2 diabetes [30]. Type 2 diabetes is a chronic disease where insulin in the body is no longer being used effectively. Also, a high level of glucose in the blood occurs due to low insulin production. Hyperglycemia leads to kidney, blood vessel, nervous and heart disorders and when not diagnosed or monitored can lead to death. Between nineteen to twenty million U.S. citizens have been diagnosed with type 2 diabetes according to current statistics [31]. If the current model stands, by 2030 over 366 million people in the world will have type 2 diabetes [32]. Type 2 diabetes and its associated complications cost the U.S. $132 billion in 2002 [33].

D-tagatose (D-tag) is currently being tested to treat type 2 diabetes in a global phase 3 clinical trial. D-tag occurs naturally in heated dairy products and is a hexose sweetener. It has a sweetness level of nearly 92% of the commonly used table sugar. D-tag has not been shown to increase insulin production; however, it does cause moderate weight loss and decreases glycemic response according to clinical trials. The mechanism of action of D-tag is based on enzymatic activity taking place in liver [34-35]. The production of experimental drugs is often contracted to manufacturers where different dosage levels and placebo drugs that are required to be visually indistinguishable from the actual drug are needed. Process analytical techniques need to be investigated to guarantee the quality, quantity and identity for newly developed drugs.

The manufacturing needs of big pharmaceutical companies are often given out to toll-manufacturers in order to meet their financial and production quotas. The manufacturing companies are often producing multiple drugs nearly simultaneously that are visually similar. The manufacturing companies may be contracted by many different outside companies to produce similar looking drugs as well. A rapid and nondestructive method for tablet verification is crucial step that could reduce or even eliminate tablets from accidentally becoming contaminated or confused. A strategy for placing each tablet

identification and verification system prior to shipping the product is essential. Process analytical technologies (PAT) incorporated into the manufacturing line should be able to complete verification in real-time. There are millions of tablets that are recalled because there are currently no guaranteed methods to remove all the problems of contamination and mislabeling. A great example was reported at the end of 2006 where eleven million bottles of acetaminophen was recalled by the Perrigo Company because metal wire was later found in the tablets [36]. Current good manufacturing processes (cGMP) are at their limits according to the FDA who also stated better risk-based approaches should be investigated to insure the safety of pharmaceutical products [37]. PATs could ensure prevention of large recalls because they are designed to find the problems before they occur.

Integrated sensing and processing (ISP) acoustic resonance spectroscopy (ARS) is a rapid and nondestructive analytical method. Unlike many optical methods, a major benefit of all acoustic methods is their ability to penetrate different types of opaque packages. Many clinical trial drugs and placebos are required to be hidden from the users creating a situation where acoustic methods have a clear advantage. A system of ISP-ARS sensors would act as a PAT and could analyze every tablet manufactured, creating a situation where only the tablets below the quality guidelines would need to be removed. A controlling system such as a dynamic data-driven application system (DDDAS) would adjust processing conditions and chemical compositions based on the data from the ISP-ARS sensors [38-39]. DDDAS would allow for the integration of real-time information to predict and model an event or measurement. The predictions become more dependable when dynamic rather than static information is being constantly included into a model. For example, if weather predictions were calculated with information collected statically from different sensors then any prediction made would be obsolete immediately following its conception due to the rapidly changing nature of weather. Imagine the same situation with many sensors collecting data in real-time where each change would be integrated into the prediction creating a continuous flow of information. In this way, DDDASs have the ability to guide their measurement processes and focus their resources, much as forecasts guide US Air Force 53rd Weather Reconnaissance Squadron ("Hurricane Hunter") aircraft away from calm seas and into the eyes of hurricanes to

concentrate their data collection. The information collected makes possible advance warning of hurricanes and increases the accuracy of hurricane predictions and warnings by as much as 30 percent [40].

Fourier transform acoustic resonance spectroscopy (FTARS) is performed on data in order to formulate the ISP acoustic waveforms. FTARS is used as a predictor for the frequencies and regions that will be used for the creation of the ISP waveforms for a specific group of samples. The process in which the information is transferred from the training FTARS data to the predictive models of the ISP waveforms parallels a DDDAS because there is continuous monitoring and adjusting of the ISP waveform through retraining. To demonstration, in pharmaceutical manufacturing, FTAR spectra of active pharmaceutical ingredients (API) are calculated from the ISP waveforms. Once new tablets are formulated and manufactured, ISP-ARS is used to identify each tablet. If a sample is unidentifiable by the current calibration then the sample will be scanned using FTARS and its identity and ISP information is incorporated into the global calibration (see figure 3.1).

FTARS is well established and has been shown to differentiate drugs, [41] powders, [45-47] liquids, [48-50] as well as predict dissolution rate in otherwise identical samples [42]. FTARS is nondestructive and complete scans can be made in seconds, therefore it is a prime candidate for use as a PAT. However, FTARS relies on intensive computer processing following data collection due to the amount of information gained in each scan. An ARS spectrum recorded over the interval of 20 Hz to 20 kHz with a sample rate of 44.1 kHz for one second generates a substantial amount of data (1 s x 44.1 kHz = 44100 data points). Chemometric analysis of multiple FTARS data sets can become computationally demanding and could limit the production rate of tablets, especially if 100% tablet inspection is considered. ISP-ARS reduces the computational burden of FTARS because it directly produces the analyte identity as an output.

Theory

In FTARS, white noise comprising a mixture of all frequencies over a specified range is used as excitation for scanning a calibration set of samples. If no *a priori* information is

given about the samples then sample classification is made via undirected (unsupervised) data mining. For this qualitative tablet identification experiment, multivariate techniques are employed to group the calibration data into specific classes. The specific classes are then used to build a predictive model on which ISP-ARS is based. To begin, a training set of data is scanned over the entire frequency range using FTARS techniques. PCA is employed to separate the samples into classes. PCA is a multivariate analysis technique that reduces the amount of data in large sets. PCA has been previously applied to FTARS and other spectroscopic data to differentiate samples [42][44-45][51-52]. In PCA a new set of data, the principal components (PCs), are generated from the acoustic frequencies such that the first PC contains the most variation of the original data, the second PC the next highest variation orthogonal to the first, and so on until the total sample variation is explained. If there is a significant amount of correlation present in the original data then the number of useful PCs is small [53]. The PCs that denote the greatest variation among the calibration set tablets are used to create the ISP acoustic waveforms. The loadings (coefficients) of the PCs are used to indicate the frequency regions that have the greatest effect on each PC (figure 3.2). In ISP-ARS, the acoustic waveforms are created from those frequency regions where the greatest sample variation was observed, and that had the largest loading coefficients. The PC loadings, however, are weighted in both the positive and negative direction, and each contains useful data. Thus, loadings over the frequency region corresponding to the highest variation in the data must be found in both the positive and negative directions. Separate acoustic waveforms must be created for the positive and negative loading data. If the data are not separated then frequency components from the positive data domain may offset the components from the negative when the entire waveform is integrated during the detection process. The same is true for each specific PC loading that is used. Frequency components from one loading may overlap with components of another.

In many cases, a single PC is sufficient for a tablet analysis. But suppose that in practice it is found that the top three PCs separate the tablet calibration data sufficiently. The corresponding loadings for PCs 1-3 must be broken into positive and negative pieces, and an ISP acoustic waveform constructed from three PCs would have six segments that

would be played sequentially and integrated into six distinct detector values. Cluster analysis of the detector voltage data would complete the classification of a sample.

One method of classifying the output voltages is the Bootstrap Error-adjusted Single-sample Technique (BEST). The BEST method of sample classification calculates the distance between data clusters in multidimensional standard deviations (MSD) [54]. When the distance of a spectrum from a cluster is less than three MSDs, the unknown spectrum is considered to be of the same sample as the cluster. ISP acoustic waveforms can be generated from many samples, and an MP3 player can be used to hold an entire database of ISP excitation waveforms. This makes ISP-ARS a great choice for PAT as pharmaceutical manufacturers could calculate an ISP acoustic waveform from FTARS data to determine many sample properties and characteristics in their production line.

Experimental

Tablet Preparation. Tablets of different over-the-counter pharmaceutical drugs were obtained for scanning by the ARS. Tablets included: vitamin C (Spring Valley, 1000 mg), vitamin B-12 (Spring Valley, 2000 mcg), acetaminophen (Equate, 325 mg), aspirin (TopCare, 325 mg), ibuprofen (Equate, 200 mg) and D-tagatose (Spherix Inc, 300 mg). The tablets were scanned intact with no special preparation.

ARS Data Collection. Four tablets each of the pharmaceutical drugs were scanned along with a blank (a scan of the empty base-plate at equal pressure as the tablets), in triplicate and in random order. Each tablet was placed on a scale (Model 3120, Health O Meter, Bridgeview, IL, USA) and adjusted to a pressure of 150g so that contact between the sample and the quartz rod of the ARS was maintained and constant throughout scanning. After each scan, the scale was reset and the tablets repositioned. White noise in the frequency range of 0 to 3.1 MHz was generated using a function generator (Stanford Research Systems, Sunnyvale, CA, USA). The sound card used to capture the data (Model No. SB0490, Creative Labs) had a range of 20 Hz to 22 kHz and the card contained an anti-aliasing filter that prevented problems from excitation outside the frequency range of the function generator. All data processing was done in Matlab 7.0.1 (The Mathworks Company, Natick, MA, USA). All sound was captured for 5 seconds

with a sample rate of 44.100 kHz. An FFT at size 44100 was performed on the sound files to convert the data from the time domain into the frequency domain. The mean of the three replicate measures was taken. Frequency domain data were z-scored in intensity and principal axis transformation was performed on the data before cluster analysis. Loadings from the first three PCs were used to find the frequencies that contributed the most to the total variance between sample types. The positive and negative loadings were separated and sorted by principal component number in descending order. The frequencies corresponding to the largest 10, 100 and 1000 loading values were used to create the excitation signal for ISP-ARS. The excitation signal consisted of 18 frequency ensembles in sequence, one second of each. The first three frequency ensembles were from the positive loadings of PC one through three using only ten frequencies. Sequence four through six were created from the negative loadings of PC one through three using ten frequencies. The order was then repeated for 100 and 1000 frequencies to give the total of 18 ensembles (figure 3.3). The average detector voltage signal of each frequency mixture became a single dimension in the classification process. Because the excitation was performed using the frequency ensemble created from the PC loadings, the detector voltage was directly proportional to the PC scores, and MANOVA was used for classification of tablets in the ISP-space.

Results and Discussion

Three PCs representing 76% of the total variance of the data set was used to classify the tablets. Of each 10, 100 and 1000 frequencies, three orthogonal excitations (one for each PC) were employed for both the positive loadings and negative loadings obtained from PC analysis. The three orthogonal excitations were visualized in a three dimensional scatter plot (figure 3.4). Similar samples can be visualized as clusters in a three dimensional scatter plot with dissimilar samples clustering in different regions hyperspace. When ISP waveforms were constructed from PC loadings the resultant ISP voltages observed at the detector were functionally equivalent to the PC scores. Projecting the three integrated detector voltage signals scanned from a sample onto a three-dimensional scatter plot illustrated the group in which the sample belonged. Additional scans of the same type of tablets contained the information needed to draw

probability density contour plots encompassing the regions where spectral points of more samples of the same material were likely to be found. This approach of digitally calculating PCs initially to form an excitation waveform (effectively an analog computing alternative to the more typical digital analysis after spectra have been collected) allows for a rapid data acquisition and determination of probability densities for classification [55].

ISP-ARS vs FTARS Clusters. Figure 3.4 illustrates the cluster patterns using conventional ARS with PCA. The PCs that captured the largest variations between spectra were plotted against each other in an XYZ type scatter plot. The figure depicts the separation between the different types of tablets. Figure 3.5 and Figure 3.6 illustrate similar plots as Figure 3.4, but rather than calculating the PCs from full acoustic spectra, the XYZ axes represent the observed detector voltages from the ISP-AR excitations. Figure 3.5 represents the voltages acquired from the ten frequencies with the positive loadings contributing to the largest variation. Figure 3.6 represents the voltages acquired from the ten frequencies with the negative loadings contributing to the largest variation. All clusters from both methods contain 4 sample points, and the ellipses represent one standard deviation level in each direction. Adding frequencies sometimes improved the cross validated separation between tablets, but sometimes did not (see Table 3.1).

Comparisons between the positive loadings (figure 3.5) and the negative loadings (figure 3.6) indicate that each excitation was important on different tablet types. Note that while the positive loading excitations do not separate the blank rod from tagatose and acetaminophen, the negative loading excitations do separate them. Employing the positive and negative loadings together in the analysis allows the benefits of both to separate the different types of tablets. Canonical Variables (CV) were calculated from the voltages obtained from the ten frequencies of both the positive and negative loads to produce figure 3.7. The ellipses in figure 3.7 depict the BEST three standard deviation contour level.

Table 3.1 reports the mean inter-cluster and intra-cluster BEST MSD for the different tablets. The mean intra-cluster MSDs from cross validation are reported on the diagonal in bold face type. ISP-ARS represents a slight improvement over full spectrum FFT-

ARS, but some intra-cluster MSDs are increased with more frequencies in the excitation process, while others are being decreased (over fit) with more frequencies in the excitation process. Neither FFTARS nor ISP-ARS have all intra-cluster MSDs below three standard deviations, probably due to the low number of tablets scanned in each group (four). However, ISP-ARS does lead to larger inter-cluster MSDs than FFTARS. For example, for FFTARS the largest inter-cluster distance is 176.903 between ibuprofen and the blank rod, and that same inter-cluster distance with ISP-ARS is 236.694 (with 10 frequency, positive and negative loadings excitation). Because there are such large inter-cluster MSDs with ISP-ARS, it would be possible to increase the MSD distance cutoff for classification to the largest intra-cluster MSD, as long as it is much smaller than the smallest inter-cluster MSD, and still maintain accurate classification.

ISP-ARS Classifications. MANOVA was used for cross validation classification where each tablet was classified to clusters three standard deviations or less away. Tables 3.2, 3.3 and 3.4 report classification, accuracy, precision and recall when using 10 frequencies, 100 frequencies and 1000 frequencies respectively in the ISP waveform to represent the loadings. These statistics were calculated as follows:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{3}$$

Where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative. These statistics are represented in percentages and each table includes the averages for the method as a whole.

Comparisons using 10, 100 and 1000 frequencies to represent the loadings depict rather similar results; with 10 frequencies having the best accuracy (using only 10 frequencies

gives the most correct classifications, while 1000 frequencies classified the least amount correctly). However, the largest percentages in each table are the precision measures. No matter how many frequencies were used, the precision was 100 percent because there were no false positives. The data suggest that it is possible to over fit some tablet separations using 100 or 1000 frequencies to represent the loadings, and when faced with a need to differentiate two types of tablets, one should specifically optimize the number of different frequencies selected for the desired differentiation.

It is perhaps not surprising that it is possible to use too many frequencies in an ISP waveform. The largest 10 factor loadings become the 10 frequencies with the largest amplitude in the ISP waveform. The largest 100 factor loadings become the largest 100 frequencies in the ISP waveform, and this set of 100 includes the largest 10 by definition. The point is that there can be some rather weak signals in the largest 100 or 1000 frequencies, and these weaker signals are more easily overwhelmed by noise. Detector bandwidth must be increased to accommodate more frequencies, increasing the chance of picking up extraneous noise at the additional frequencies.

Speed and Versatility of ISP-ARS. ISP-ARS is a large improvement in speed and efficiency when compared to the traditional FT-ARS. Both methods have the benefits of no sample preparation needed for the tablets and acquisition times in only a few seconds; however ISP-ARS does not need heavy computation. ARS is attributed to for its non-destructive ability to analyze different materials and with the ISP accompaniment, the method only needs a calibration set. An ISP waveform can be constructed, compressed to an MP3 format to save space and an entire web based database can be created to house the information. Researchers and manufacturers could be linked together via the internet to continually add new drugs to the database in a matter of seconds with current internet speeds. With the addition of ISP a system could easily be automated to grab information from the web database and simply read the voltage at the detector for results.

Conclusion

Integrated sensing processing acoustic resonance spectroscopy has been explored as a rapid and non-destructive method to differentiate D-tagatose tablets (an experimental toll-

manufactured drug) from different tablets including aspirin, acetaminophen, vitamin C, vitamin B and ibuprofen. With an experiment-specific ISP waveform, the classification is far more rapid than with conventional ARS. Simpler ISP waveforms using fewer frequencies to represent the factor loadings that separate the tablets may outperform more complex waveforms using more frequencies. By encoding waveforms on an MP3 player, ISP-ARS could become a method to quickly identify different unlabeled tablets with a similar appearance created in a contract-manufacturing environment.

Chapter Three Tables

Table 3.1 BEST Distances from Cluster Analysis. The inter-cluster and intra-cluster (by cross validation) MSD using the BEST for the different tablets. The intra-cluster MSDs are reported on the diagonal in bold face type. The different tablet names are abbreviated for easier viewing.

FFT-ARS							
	Vit C	Blank	Asp	Ibu	Tag	Vit B	Acet
Vit C	**3.967**	16.798	10.946	25.647	41.853	2.028	5.041
Blank	-	**5.240**	95.268	176.903	30.358	59.677	79.483
Asp	-	-	**2.369**	6.622	10.599	3.603	2.022
Ibu	-	-	-	**2.806**	6.729	4.157	2.223
Tag	-	-	-	-	**1.520**	5.135	2.673
Vit B	-	-	-	-	-	**2.427**	4.985
Acet	-	-	-	-	-	-	**2.415**
ISP-ARS 10 Frequencies							
	Blank	Asp	Ibu	Acet	Tag	Vit B	Vit C
Blank	**3.234**	292.776	236.694	586.349	122.096	90.723	71.394
Asp	-	**2.180**	109.896	56.738	28.600	23.160	42.752
Ibu	-	-	**1.448**	39.902	28.230	18.336	42.202
Acet	-	-	-	**1.685**	8.667	33.079	37.650
Tag	-	-	-	-	**2.411**	50.455	47.551
Vit B	-	-	-	-	-	**2.056**	36.598
Vit C	-	-	-	-	-	-	**5.312**
ISP-ARS 100 Frequencies							
	Blank	Asp	Ibu	Acet	Tag	Vit B	Vit C
Blank	**1.875**	49.635	56.821	142.394	165.027	35.000	34.604
Asp	-	**3.489**	12.338	81.420	34.759	36.403	58.583
Ibu	-	-	**1.548**	49.770	19.618	21.577	39.843
Acet	-	-	-	**5.121**	7.886	20.616	14.339
Tag	-	-	-	-	**2.441**	23.366	14.283
Vit B	-	-	-	-	-	**1.582**	29.977
Vit C	-	-	-	-	-	-	**2.423**
ISP-ARS 1000 Frequencies							
	Blank	Asp	Ibu	Acet	Tag	Vit B	Vit C
Blank	**4.248**	190.660	106.067	125.373	150.548	79.002	69.626
Asp	-	**8.608**	19.628	71.496	100.572	36.624	88.467

Ibu	-	-	**2.458**	78.743	90.821	35.275	84.459
Acet	-	-	-	**2.242**	44.920	29.948	32.958
Tag	-	-	-	-	**3.942**	49.826	35.770
Vit B	-	-	-	-	-	**2.381**	62.774
Vit C	-	-	-	-	-	-	**2.786**

Table 3.2 Summary statistics for ISP-ARS utilizing both the ten frequencies with the greatest change according to the positive factor loadings and the 10 frequencies with the greatest change according to the negative loadings. MANOVA was used for the classification and each tablet was classified to any group within three standard deviations in hyperspace.

Group	Correct Classification	Accuracy (%)	Precision (%)	Recall (%)
Blank	3	96.43	100.00	75.00
Asp	4	100.00	100.00	100.00
Ibu	4	100.00	100.00	100.00
Pain	4	100.00	100.00	100.00
Tag	4	100.00	100.00	100.00
VitB	4	100.00	100.00	100.00
VitC	2	92.86	100.00	50.00
AVERAGE	3.57	98.47	100.00	89.29

Table 3.3 Summary statistics for ISP-ARS utilizing both the 100 frequencies with the greatest change according to the positive loadings and the 100 frequencies with the greatest change according to the negative loadings. MANOVA was used for the classification and each tablet was classified to any group within three standard deviations in hyperspace.

Group	Correct Classification	Accuracy (%)	Precision (%)	Recall (%)
Blank	4	100.00	100.00	100.00
Asp	2	92.86	100.00	50.00
Ibu	4	100.00	100.00	100.00
Pain	2	92.86	100.00	50.00
Tag	3	96.43	100.00	75.00
VitB	4	100.00	100.00	100.00
VitC	4	100.00	100.00	100.00
AVERAGE	3.29	97.45	100.00	82.14

Table 3.4 Summary statistics for ISP-ARS utilizing both the 1000 frequencies with the greatest change according to the positive loadings and the 1000 frequencies with the greatest change according to the negative loadings. MANOVA was used for the classification and each tablet was classified to any group within three standard deviations in hyperspace.

Group	Correct Classification	Accuracy (%)	Precision (%)	Recall (%)
Blank	2	92.86	100.00	50.00
Asp	3	96.43	100.00	75.00
Ibu	3	96.43	100.00	75.00
Pain	3	96.43	100.00	75.00
Tag	2	92.86	100.00	50.00
VitB	3	96.43	100.00	75.00
VitC	3	96.43	100.00	75.00
AVERAGE	2.71	95.41	100.00	67.86

Chapter Three Figures

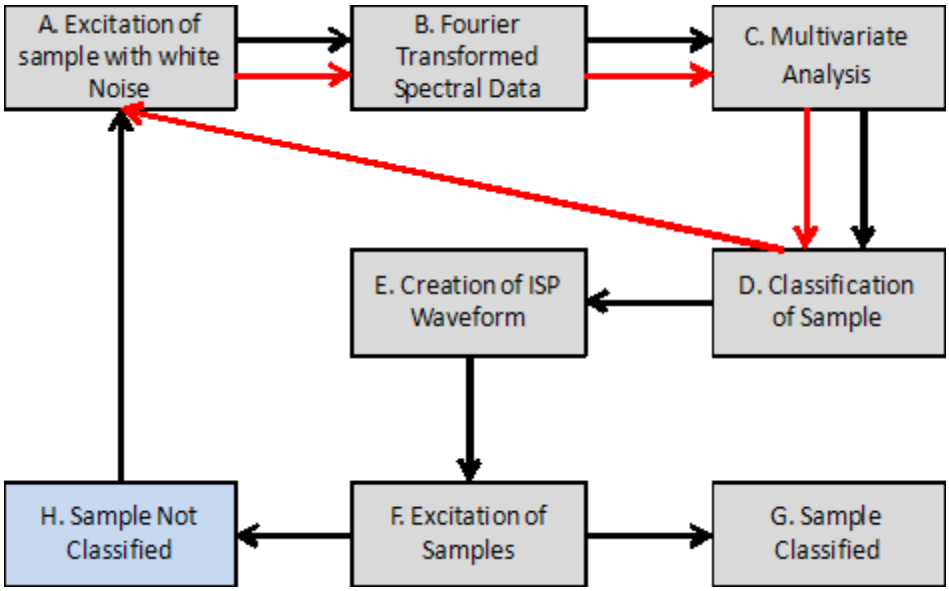


Figure 3.1 Block diagram of ISP-ARS process. The red arrows indicate the traditional FTARS cycle. In traditional FTARS, samples are scanned and classified according to their inter-cluster distances found via multivariate analysis (A-D). This process is repeated for each sample scanned. With ISP-ARS, the FTARS data are used to calculate factor loadings and an ISP acoustic waveform is constructed to represent these loadings (E). Once the ISP waveform is constructed, the traditional FTARS operation cycle is not needed. Samples scanned with the ISP waveform are classified according to their detector voltages (F-G). If a sample cannot be classified (H), then FTARS is employed for recalibration and a new ISP acoustic waveform is constructed that includes the new unknown. As samples change the ISP waveform can evolve with the new data.

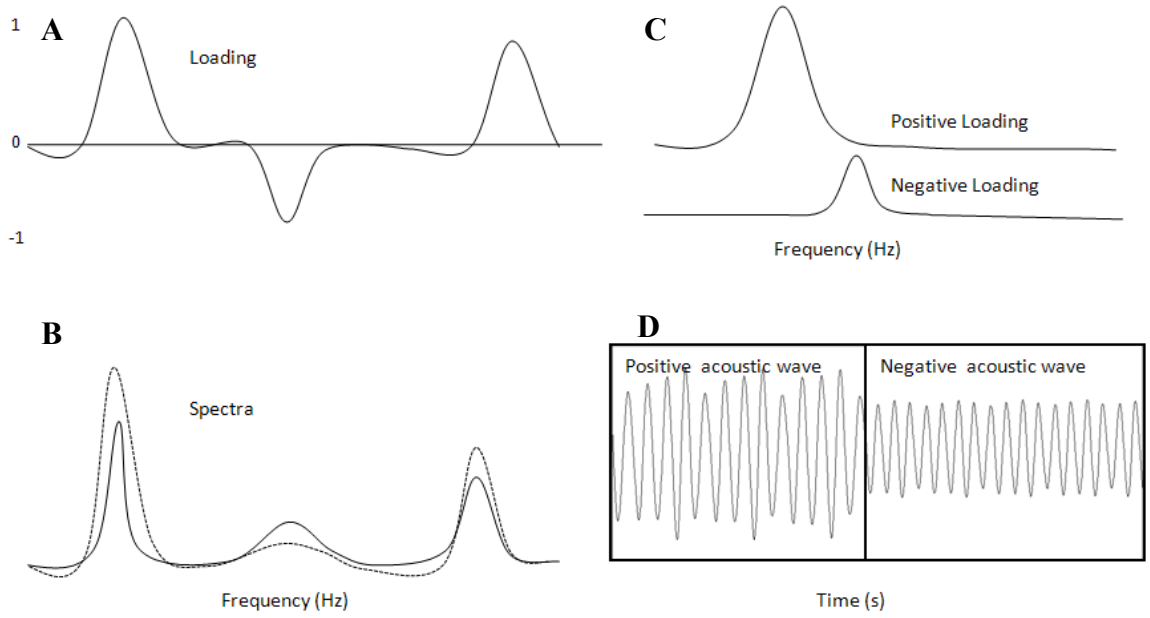


Figure 3.2 ISP Flowchart. In a series of samples (B) the highest variation in the frequency range of interest can be viewed in the highest loadings (A). Selection of the greatest frequencies from the positive and negative component of the loadings (C) can be used to construct the ISP waveform (D). To avoid cancellation of integrated signal from the positive and negative loading frequencies, an acoustic waveform must be constructed separately for the positive and negative loadings and transmitted independently through the sample.

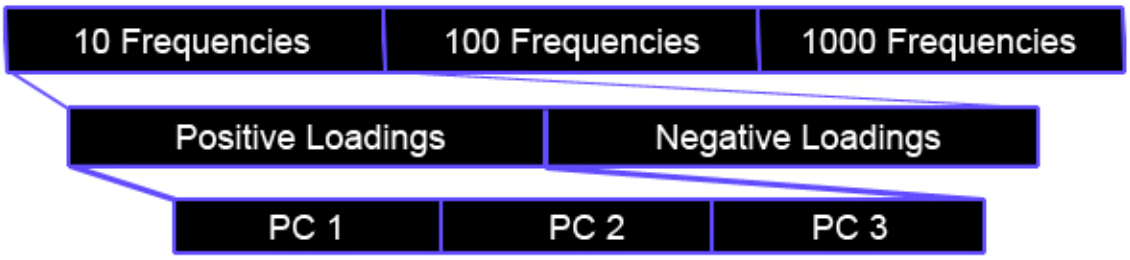


Figure 3.3 ISP waveform composition. An 18-second excitation sequence was contructed to enable the three ISP experiments to be conducted simultaneously. The first three frequency ensembles were from the positive loadings of PC one through three using only ten frequencies. Sequence four through six were created from the negative loadings of PC one through three using ten frequencies. The order was then repeated for 100 and 1000 frequencies, enabling three different calibration and prediction experiments (testing calibration using 10, 100, and 1000 frequencies) to be conducted with the same tablets at the same time.

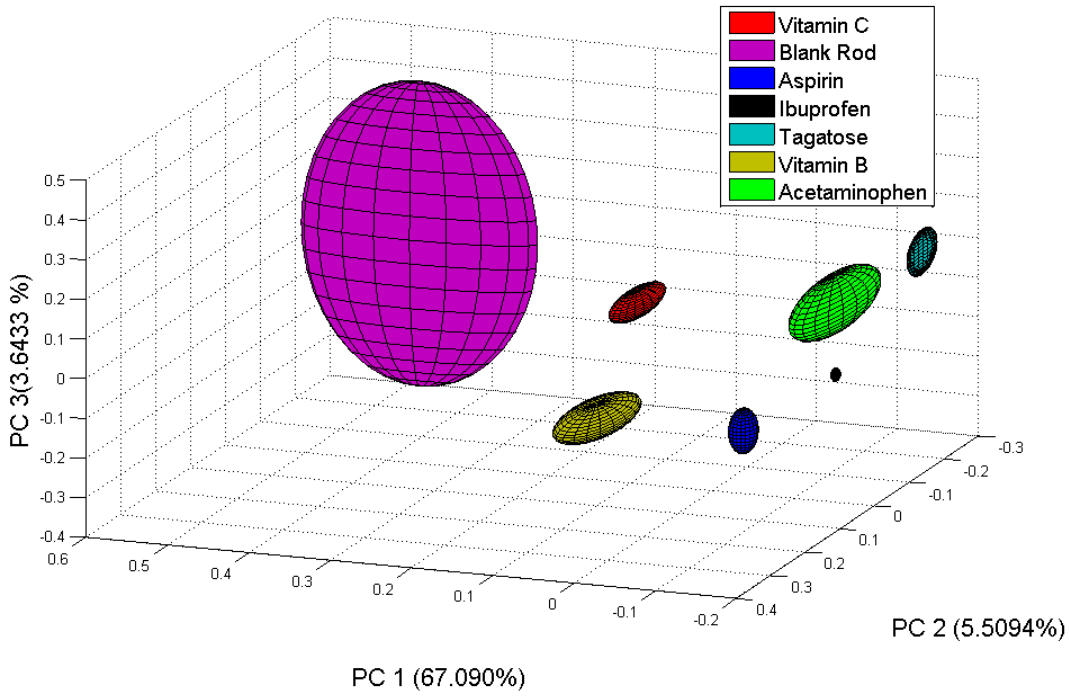


Figure 3.4 Principal component standard deviation plot. The PCs that captured the largest variations between spectra were plotted against each other in an XYZ type scatter plot. The ellipses depict a one standard deviation contour level for each tablet type.

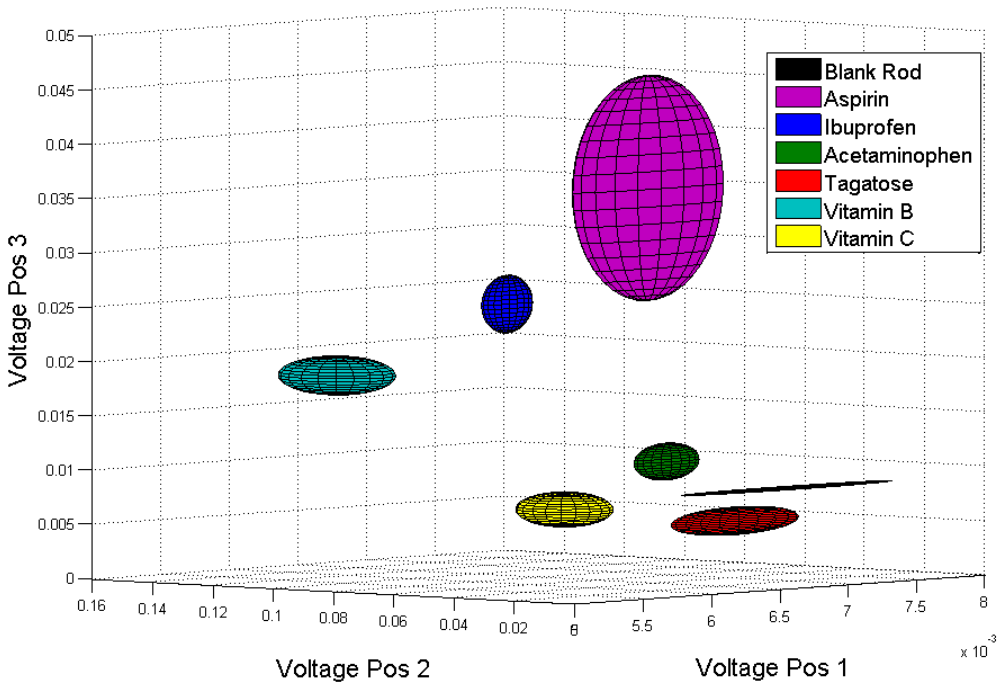


Figure 3.5 ISP voltage standard deviation plot (positive). The coordinate axes represent the detector voltages from the ISP-AR spectra. This figure represents the voltages acquired from the ten frequencies with the positive loadings contributing to the largest variation in the FTARS scans. The ellipses here depict the one standard deviation contour level for each tablet type.

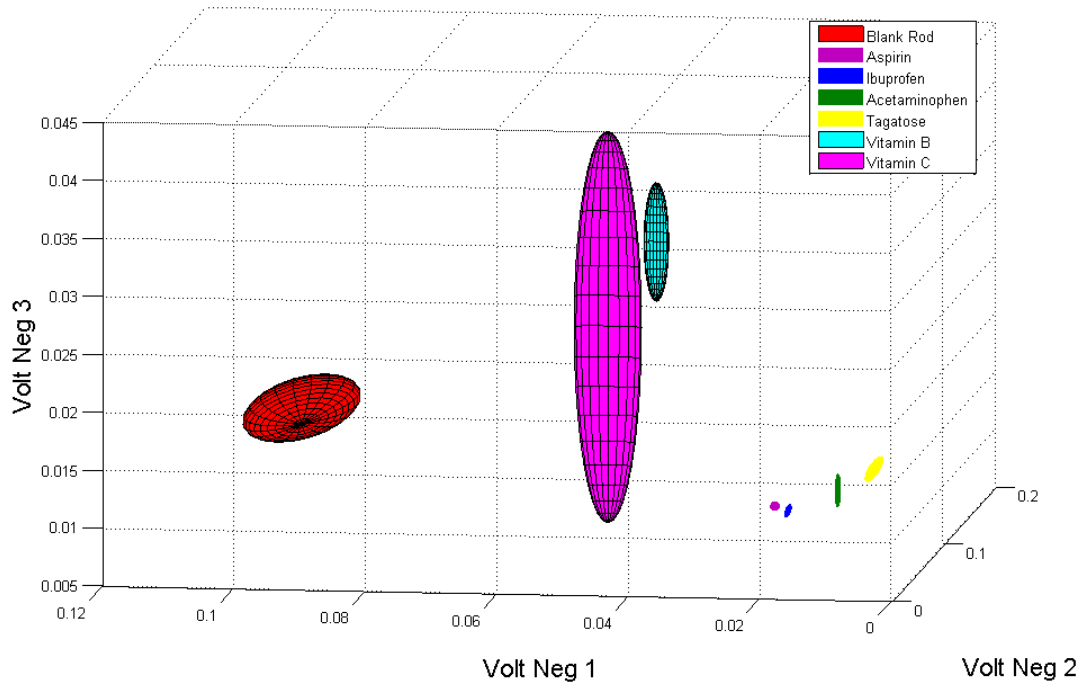


Figure 3.6 ISP voltage standard deviation plot (negative). The coordinate axes depict the detector voltages from the ISP-AR spectra. This figure represents the voltages acquired from the ten frequencies with the negative loadings contributing to the largest variation in the FTARS scans. The ellipses here depict the one standard deviation contour level for each tablet type.

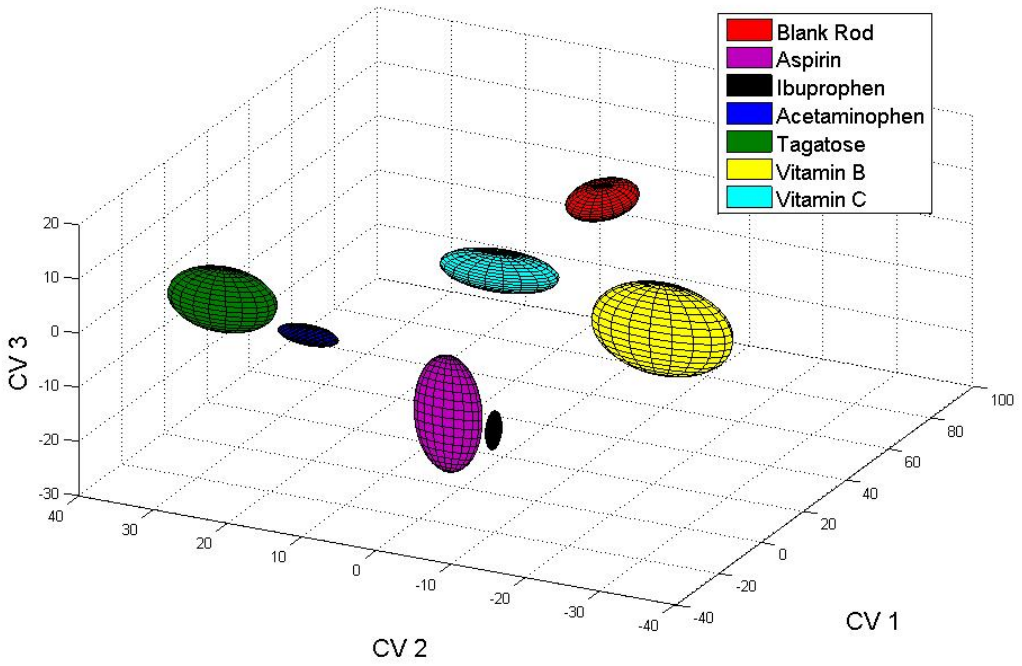


Figure 3.7 Canonical variable standard deviation plot. CV from the voltages obtained from the ten frequencies of both the positive and negative loadings. The ellipses here depict the three standard deviation contour level for each tablet type.

Copyright Statement

Copyright © SpringerLink 2008

Hannel, T. S.; Link, D. J.; Lodder, R. A. *Journal of Pharmaceutical Innovations*. **2008**, 3(3), 154-160.

Chapter Four: Incorrect or Defective Pill Detection Using a Dynamic Data-Driven Application System Paradigm

1. Introduction

Administration of incorrect medications by professional caregivers is estimated in 1997 to have killed as many as 44,000 to 98,000 Americans after prescriptions were filled [56]. These numbers are likely to be underestimates due to unreported deaths. To put this number in perspective, use of incorrect medication is the eighth leading cause of death in the United States and actually kills more people in a given year than traffic accidents, breast cancer, or AIDS. The situation is no better in 2007.

A secondary issue is defective tablets coming off a pharmaceutical production line or mistaken packaging. Many errors are readily visible and are caught immediately. However, not all are detected and the defective or mislabeled tablets reach the marketplace. In Section 2, we discuss the advantages of using a real-time dynamic approach instead of using static data. In Section 3, we discuss why catching errors at the pharmaceutical production and packaging areas is essential to reducing recalls and should be part of process analytical technologies. In Section 4, we describe an integrated acoustic sensing and processing device. A handheld version can also be used to identify medications before a caregiver delivers them to individuals. We also describe a cyber physical system (CPS) to detect incorrect or defective tablets. In Section 5, we provide simple results based on a prototype system that has been built and tested in a limited manner. In Section 6, we provide conclusions and briefly describe what needs to done next.

2. Dynamic versus Static Data

A data driven system allows for the implementation of real-time data to model or predict a measurement or event. By incorporating data dynamically rather than statically, the predictions and measurements become more reliable. Consider weather forecasting. If predictions are made based on static data collected from sparsely distributed sensors, then rapidly changing conditions often make a prediction obsolete shortly after it is made. A more reliable forecasting system continuously incorporates real-time changes from many

sensors into its predictions so that the forecast is always built around current conditions. As the conditions change, so does the forecast, in real-time. Data driven applications have the ability to guide their measurement processes and refocus their resources, much as forecasts guide US Air Force 53rd Weather Reconnaissance Squadron aircraft away from calm seas and into the eyes of hurricanes to concentrate their data collection. The information collected makes possible advance warning of hurricanes and increases the accuracy of hurricane predictions and warnings by as much as 30 percent [40].

3. Catching Mistakes at the Source

Numerous large pharmaceutical manufacturers outsource their small-scale manufacturing needs as a way of reducing cost or meeting their production deadlines. A contract manufacturer may make several kinds of pills that are similar in appearance at almost the same time, e.g., testing various dosages and placebos for clinical trials. A contract manufacturer may also produce pills for multiple companies. One way to reduce the possibility that pills may inadvertently become confused or contaminated is to employ a rapid and nondestructive means of verifying tablet identity. Such systems for identifying contaminated or mislabeled products must be strategically placed to prevent problems with pills before they are shipped. PAT on the production line should have the ability to work in real-time. Currently there are no foolproof methods to eliminate mislabeling or contamination. As a result, millions of pills are recalled in some years. For example, in November 2006, 11 million bottles of contaminated acetaminophen were voluntarily recalled by the Perrigo Company of Allegan, Michigan due to contamination of the tablets with metal wire [36]. The FDA admits that cGMP have reached their limits and better "science-based" approaches are needed to insure product safety [37]. PATs are designed to prevent large recalls by detecting problems before they occur.

4. An Integrated Sensing and Processing Approach

Integrated sensing and processing acoustic resonance spectroscopy (ISP-ARS) is a novel approach to conventional acoustic spectroscopic techniques. In ISP-ARS, an ISP acoustic waveform is created such that it comprises only the distinguishing spectral details associated with an analyte in question. FTARS is used to develop ISP acoustic waveforms employed in differentiating different drugs. ISP-ARS is fast and non-

destructive. Acoustic methods are able to deeply penetrate many types of opaque packaging, in contrast to near-infrared and other optical methods. The ability to penetrate many types of packaging can be a distinct advantage in preparation of clinical trial lots, where drugs and placebos must be blinded from users. As a PAT, a series of ISP-ARS sensors could potentially scan every pill produced by a manufacturer, enabling the removal of only those pills that did not meet quality standards.

A dynamic system should control a manufacturer's product line based on measurements from a series of ISP-ARS sensors, adjusting process conditions and ingredients in real time based on actual process measurements [39][57]. ISP-ARS reduces the time required for processing that is normally observed with full spectrum FTARS. An ISP acoustic waveform is the result of chemometric analysis of the FTARS spectrum. By weighting the frequency changes according to their individual component scores, an acoustic waveform can be made that excites only those frequencies important to the analyte under observation. The ISP output is a voltage that can be read immediately and corresponds to only the analyte under investigation. Creation of the ISP acoustic wave begins with the chemometric analysis of the initial FTARS data. Therefore, FTARS itself makes a prediction about what will work as an ISP acoustic waveform for a given set of samples. This training process can be viewed as a cyber physical system when the performance of the ISP waveform is continuously monitored and the ISP waveform is continuously adjusted through retraining.

5. Preliminary Results

ISP acoustic waveforms composed of 10, 100, and 1000 frequencies were used to identify several toll manufactured drugs. The pills used in this study were aspirin, acetaminophen, D-tagatose, ibuprofen, vitamin C, and vitamin B. It was found that only the top 10 frequencies were required to properly classify each pill used in this study. Intra-cluster distances were calculated to be less than 3 MSD for each pill type. The average accuracy of prediction was 98.47, 97.45 and 95.41 percent for the 10, 100 and 1000 frequency component acoustic waveforms respectively.

6. Conclusion

We have described a prototype cyber physical system for use in identifying defective or mislabeled pills. Integrated sensing processing acoustic resonance spectroscopy has the ability to differentiate between different types of pills in contract manufacturing and bedside applications. The results are preliminary and much more research and development will be necessary in order to produce systems that can be deployed on pharmaceutical manufacturing lines. A handheld version that can be networked needs to be refined so that caregivers can correctly identify all pills before giving them to patients.

7. Acknowledgments

This research was funded in part by grants from the National Science Foundation (EIA-0219627, ACI-0305466, ACI-0324876, OISE-0405349, and CNS-0540178), Kentucky Science and Engineering Foundation (148-502-05-154), and the National Institute on Alcohol Abuse and Alcoholism (N01AA 33003).

Copyright Statement

Copyright © International Multi-Conference of Engineers and Computer Scientists

Douglas, C; Hannel, T; Link, D; Lodder, R; Haase, G. *International Multi-Conference of Engineers and Computer Scientists*, **2009**, ISBN: 978-988-17012-2-0.

SECTION THREE– BOOTSTRAP ENHANCED N-DIMENSIONAL DEFORMATION OF SPACE (BENDS)

Bootstrap enhanced n-dimensional deformation of space (BENDS) is an algorithm that has been created in order to overcome highly non linear correlations in quantitative analytical chemistry. ARS was selected as the instrument of choice for testing BENDS due to the non linear correlations that occur regularly in AR experiments. The reasons behind the non linear trends in ARS and the details of the BENDS algorithm are described in section three. BENDS mathematically contorts, pushes and unravels the non linear trend in order to resolve the unrecognizable non linear correlations and output a reduced, linear calibration model. Using well known mathematics and basic algorithms used in other areas of science, BENDS can be used with ARS in order to quantify many different mediums including powders and tablet.

Chapter Five: ARS with BENDS to Determine Aspirin Hydration

Introduction

There are many factors that affect the quality of pharmaceutical products. Things like drying, mixing and stability of APIs must be monitored at each stage in the manufacturing process to insure maximum quality. cGMPs rely on end-point testing of a few samples from a batch to assure quality in the entire batch. When products of low quality are found, the entire batch must be restarted. Even worse, in some cases low quality pharmaceuticals are released to the public and must be recalled at a high cost. For example, in 2003, Ivax Pharmaceuticals of Miami, FL recalled over 4000 one hundred-count bottles of aspirin and codeine phosphate tablets because of a stability problem with salicylic acid, a hydrolysis product of aspirin [58]. PAT can be put into place to prevent large recalls by detecting problems before they occur. PAT is a shift to process understanding and control. An ideal PAT for in-line processes must be nondestructive and have the ability to make accurate and rapid measurements. ARS possesses many qualities that make it a useful PAT for monitoring pharmaceutical processes.

According to the OED, sound is defined as the vibration transmitted through a medium of matter with frequencies that can be heard with the human ear [1]. In spectroscopy, sound is part of the larger umbrella term, acoustics, which encompasses sound, ultrasound and infrasound. Acoustic methods include the generation, propagation, resonation and acquisition of mechanical waves and vibrations. The reason for using acoustics is to use a nondestructive and rapid method that is specific to many different physical and chemical properties. Acoustic velocity [2], ultrasonic attenuation [3], acoustic reflection [4] and acoustic emission [5] are different types of acoustic properties that are used in infer other analytical properties of interest.

The human ear and brain essentially form an acoustic spectrometer that deciphers different physical properties. A trained ear can decipher the frequencies present in complex waveforms and tell the difference between different frequencies played in succession. Studies have even demonstrated that individuals can acoustically distinguish the shape of different vibrating plates [6]. The different experiments performed on blind-

folded individuals included specifying shapes, dimensions and materials of different objects that were struck by a pendulum. Another study tested blindfolded participants listening to partially occluded sound passing through a doorway [7]. The individuals were asked to listen to a doorway-like structure with obstructions of different apertures. They were surveyed to determine whether they could infer the hole to be large enough for the individual to pass through the doorway. Judgments were found to be relatively accurate, which suggests that individuals can hear surfaces that obstruct apertures like doorways. ARS coupled to a computer works in a similar manner to analyze samples.

ARS Instrumentation

Most published works in acoustics have been in the ultrasonic region and their instrumentation has dealt with propagation through a medium and not a resonance effect. The ARS has come a long way since its conception in 1988, when researchers designed a V shaped quartz rod instrument that utilized ultrasonic waves to obtain signatures of micro liter volumes of different liquids [8]. The instrument now has the ability to use a larger region of the acoustic spectrum, including sonic and ultrasonic [9]. Since the conception of the ARS, it has been used to differentiate wood species, pharmaceutical tablets, determine burn rates and determine dissolution rates of tablets [10-13]. In 2007, *Analytical Chemistry* featured the past and current work of the quartz-rod ARS discussing the potential of acoustics in the analytical chemistry and engineering fields [14].

The ARS is designed to create a fingerprint for different samples by constructive and destructive interferences. Figure 5.1 is a schematic of the quartz rod ARS and illustrates the path of the sound through the quartz rod. A function generator is depicted as the source (A), though any device that is capable of outputting sound in voltage form could be used (e.g., a CD-player, MP3 player or sound card). White noise is generated and the voltage is converted into a sound wave by a piezoelectric transducer disc (B), which is coupled to the quartz rod. The sound is shown as a blue sinusoidal wave (C), and resonates along the quartz rod, where two key interactions occur. A portion of the energy (red) is introduced into the sample and interacts in a specific manner dependent on the sample, and another portion of the energy (blue) continues unaltered through the quartz rod. The two energies still have the same frequency, but they will have most likely show

changes in their phase and amplitude. The two waves recombine after the sample (D) and constructive or destructive interference occurs, depending on the phase shift and amplitude change due to the sample. The altered combined energy (purple) is converted to an electrical voltage by another piezoelectric disc at the end of the quartz rod (E). The voltage is then recorded onto a computer by a sound card (F). The sample is coupled to the quartz rod at constant pressure, which is monitored by a pressure transducer that also acts as the sample holder. Rubber grommets are used to secure the quartz rod to a stable stand, minimizing acoustic coupling of the rod to the surroundings. Broadband "white" noise is used to obtain a full spectrum; however, most sound cards only operate between 20 and 22050 Hz. The waveform that is sent to the computer is a time based signal of the interactions of white noise with the sample. A FFT is performed on the waveform to transform the time-based signal into the more useful frequency spectrum.

Nonlinear Responses with ARS

ARS creates highly nonlinear responses in quantitative measurements [13]. Recall the recombination of energies that occurs after the sample in figure 5.1. Likely differences between the two energies include a phase shift and amplitude change due to acoustic absorption and interaction with the sample. The phase shift is caused in part by the change in the velocity of sound through different media, and in part by variations in sample dimensions. The speed of sound through air is 343 meters per second, while through water the speed of sound is 1482 meters per second [59]. Because the frequency that is passing through both the quartz rod and sample are the same, the change in acoustic velocity will cause the two waveforms to be out of phase. The degree of the phase shift is one variable used to uniquely identify the sample. When making quantitative measurements, it is possible for the two waveforms to pass in and out of phase multiple times over a concentration range. Figure 5.2 is a depiction of two waves out of phase (blue and green), interfering with one another. Notice the result of the phase shift on the amplitude of the combined wave (red). The amplitude grows as the two waves come into phase and then decreases as they start moving out of phase again (see figure 5.3).

A change in hydration of a tablet creates a corresponding change in the acoustic velocity of a sample. As frequency increases, a given shift in acoustic velocity in the sample contributes a larger phase shift between the sample and reference waveforms. As a result, variation in acoustic velocities in tablets with variation in tablet hydration is apt to take a high frequency peak through multiple cycles of constructive and destructive interference, while taking a low frequency peak through less than one complete cycle of constructive and destructive interference. If one frequency does not complete a full cycle of phase shift (frequency A in figure 5.4) while another frequency does (frequency B in figure 5.4), the resulting plot of change of amplitude at frequency A versus the amplitude at frequency B will be non linear. A full spectrum of frequencies will be far more complex, creating a highly nonlinear multidimensional response that renders conventional linear multivariate calibration methods nearly useless.

Nonlinear Fitting

Many nonlinear fitting techniques have been developed with progressively increased efficiency of computation over the years. A number of these were evaluated for use in BENDS, and the most valuable ones are described below along with their positive and negative aspects. In the end, cubic smoothing splines were selected to define the coordinate system for BENDS due to the speed and versatility of the available spline code.

Polynomial Fitting. Polynomial fits are calculated with a priori knowledge of the behavior of the curve. The model follows the degree that the user specifies, creating a fundamental drawback. . Highly nonlinear situations are often ones in which a model is not well known, and therefore polynomial fitting either fails or requires a guess and check method for many different degrees. For example, in figure 5.5, a plot of an unknown polynomial is shown with different degrees of polynomial fitting. Notice that in this case one could look at the curve and know how to plan the fitting; however, none of the curves fit the data very well.

Principal Curves. Principal curves were originally used to reposition misplaced magnets that kept a particle beam focused [60]. The algorithm has been adapted to identify

outlines in satellite images [61], to use as clustering techniques [62], to use in ecological studies [63] and sonification [64]. The basic idea is to use an iterative process that projects the points onto a curve and uses a least squares method to determine goodness of fit. The approach has many advantages when using small data sets and for image analysis. The approach is very time intensive even on these smaller data sets, and therefore would be impossible to implement on a large scale data set such as one created by the ARS. A very good applet that describes the algorithm and how it works can be found on the University of Montreal's website [65].

Cubic Spline Interpolation. The origin of cubic spline interpolation came from the need to draw smooth curves through a number of points in engineering. The idea was to place weights on the design and use a strip of flexible wood to bend around the weights creating a relatively smooth curve. Now, cubic splines are a mathematical tool that uses the same principle. The points are numerical data points and the weights are coefficients on a cubic curve used to interpolate the data [65]. The coefficients allow the curve to pass through the points with a smooth and continuous path. The cubic spline provides a piecewise approach to interpolation in which the spline is calculated in pieces and those pieces are fit together to make a continuous function.

A piecewise function $S(x)$ is fit to the form

$$S(x) = \begin{cases} s_1(x) & \text{if} \quad x_1 \le x < x_2 \\ s_2(x) & \text{if} \quad x_2 \le x < x_3 \\ \vdots & \\ s_{n-1}(x) & \text{if} \quad x_{n-1} \le x < x_n \end{cases} \tag{5.1}$$

where s_i is a third degree polynomial defined by

$$S_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i \tag{5.2}$$

for $i = 1, 2, \ldots, n - 1$. The first and second derivatives of these $n - 1$ equations are used later to create continuity in the function, and they are,

$$S_i'(x) = 3a_i(x - x_i)^2 + 2b_i(x - x_i) + c_i \tag{5.3}$$

$$S_i''(x) = 6a_i(x - x_i) + 2b_i \tag{5.4}$$

for $i = 1, 2, \ldots, n - 1$. Traditionally a cubic spline would then have four constraints placed on the curve. (1) The piecewise function $S(x)$ passes through all data points, (2)

$S(x)$ is continuous on the interval $[x_1, x_n]$, (3) the first derivative is continuous on the interval $[x_1, x_n]$ and (4) the second derivative is continuous on the interval $[x_1, x_n]$. These constraints are described below; however, in the cubic smoothing spline constraint (1) is not used and is instead replaced with a minimizing parameter to deal with noise. The minimizing function is discussed later.

The piecewise function $S(x)$ will pass through all the data points,

$$S(x_i) = y_i \tag{5.5}$$

for $i = 1, 2, \ldots, n-1$. Because $x_i \in [x_i, x_{i+1}], S(x_i) = s_i(x_i)$ and equation 5.2 can b used to produce

$$\begin{gathered} y_i = s_i(x_i) \\ y_i = a_i(x_i - x_i)^3 + b_i(x_i - x_i)^2 + c_i(x_i - x_i) + d_i \\ y_i = d_i \end{gathered} \tag{5.6}$$

for each $i = 1, 2, \ldots, n-1$. After applying the other three constraints and simplifying, the other weighting coefficients are found to be

$$\begin{gathered} a_i = \frac{M_{i+1} - M_i}{6h} \\ b_i = \frac{M_i}{2} \\ c_i = \frac{y_{i+1} - y_i}{h} - \left(\frac{M_{i+1} - 2M_i}{6}\right) h \\ d_i = y_i \end{gathered} \tag{5.7}$$

where $M_i = s_i''(x_i)$ and $h = x_{i+1} - x_i$. The systems of equations in 6.7 are converted and displayed in matrix form for easy calculation of the coefficients. The resulting matrix equation is

$$\begin{bmatrix} 1 & 4 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 4 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ M_3 \\ \vdots \\ M_{n-3} \\ M_{n-2} \\ M_{n-1} \\ M_n \end{bmatrix} = \frac{6}{h^2} \begin{bmatrix} y_1 - 2y_2 + y_3 \\ y_2 - 2y_3 + y_4 \\ y_3 - 2y_4 + y_5 \\ \vdots \\ y_{n-4} - 2y_{n-3} + y_{n-2} \\ y_{n-3} - 2y_{n-2} + y_{n-1} \\ y_{n-2} - 2y_{n-1} + y_n \end{bmatrix} \tag{5.8}$$

The system of equations has $n - 2$ rows and n columns, which makes it under-determined. In order to generate a unique cubic spline, two other conditions must be imposed upon the system.

Natural splines, parabolic runout splines and cubic runout splines are three types of splines that impose the other conditions. Natural splines are the most common because they simplify the problem while creating a spline that simulates the original wood spline fitting. While natural splines force the second derivative at the end points to be zero, the other two types force the end point second derivative to play key roles in the shape of the curve. The resulting matrix equation using natural splines is

$$\begin{bmatrix} 1 & 4 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 4 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ M_2 \\ M_3 \\ \vdots \\ M_{n-3} \\ M_{n-2} \\ M_{n-1} \\ 0 \end{bmatrix} = \frac{6}{h^2} \begin{bmatrix} y_1 - 2y_2 + y_3 \\ y_2 - 2y_3 + y_4 \\ y_3 - 2y_4 + y_5 \\ \vdots \\ y_{n-4} - 2y_{n-3} + y_{n-2} \\ y_{n-3} - 2y_{n-2} + y_{n-1} \\ y_{n-2} - 2y_{n-1} + y_n \end{bmatrix} \tag{5.9}$$

Cubic splines can be calculated very easily and the calculation time is minimal compared to other nonlinear fitting methods. The data shown in figure 5.5 were used to create a comparison between polynomial fitting and using a cubic spline to interpolate the data. Figure 5.6 illustrates how the cubic spline describes a continuous smooth function that passes through all points.

Other advantages of cubic splines are the natural weighting of different data points and the ability to create a smoothing parameter so that the curve can be corrected for error such as regions that incur additional noise or highly variable regions. The cubic smoothing spline inserts a *m*-value that is used to create the cubic smoothing spline *S,* minimizing

$$P \sum_{i=1}^{n} w_i |y_i - S(x_i)|^2 + (1 - m) \int \lambda(t) |S''(t)|^2 dt \tag{5.10}$$

for $i = 1, 2, \ldots, n$. Here, $|z|^2$ stands for the sum of the squares of all the entries of z and the integral is over the smallest interval containing all the entries of x. The weights are

inserted as *w* and the smoothing parameter is *m*. λ is the piecewise constant weight function. Figure 5.7 shows the effect on manipulating the *m*-value.

Bootstrap Method

In statistical analysis a data set x is used to calculate some statistic t(x) in order to make an approximation of some quantity of interest. For demonstration purposes, the data in table 5.1 are plasma glucose concentrations for twelve women, 21 years of age or older of Pima Indian heritage; the scores are an ordered random sample from a larger data set of 768 women [22].

If the data in table 5.1 are x, then t(x) could be something simple like their mean, x. The common next question is "how accurate is t(x)?" Because we are dealing with the mean, we simply look at the standard deviation or standard error,

$$se(x) = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right)^{1/2} \tag{5.11}$$

The results on the glucose data set x would be reported as $\bar{\mathbf{x}} = 130 \pm 26.6$, which can easily be applied to a normal Gaussian distribution and a confidence interval calculated. Unfortunately, many statistical measures do not have simple equations relating expectation and variability in the same manner as the mean and standard deviation. Consider multivariate methods, where it can take pages to simply explain the statistical calculations. Bootstrapping was created for this exact reason, to provide a general technique for finding a bootstrap estimate of the error for any statistical measure estimated from a population [23]. To bootstrapping a slightly more complex statistic, consider the 25% trimmed mean, $\bar{\mathbf{x}}\{0.25\}$. A similar explanation of the procedure can be found from the creator of bootstrapping, Bradley Efron, in a paper published in *Science* in 1991 [24]. The 25% trimmed mean is defined as the average of the middle 50% of the data. The data are ordered, the lower and upper 25% of the data are excluded, and the remaining data are averaged.

$$\bar{\mathbf{x}}\{0.25\} = \frac{111 + 115 + 120 + 128 + 129 + 138}{6} \tag{5.12}$$

The equation is simple for this particular case, although there is not a universal equation for this method in part because if the number of values in x is not divisible by four, interpolation is required to select 25% of the data. For the glucose demonstration data set, the $\bar{\mathbf{x}}\{0.25\} = 124$. The next step is to estimate the accuracy of $t(x)$, but the standard error equation is designed only for the ordinary mean. In place of simple equations, bootstrapping uses computer power to obtain a numerical estimate of the standard error.

The bootstrap algorithm randomly samples from the data set x with replacement from the original data, x. In the following discussion, bootstrap data and statistical measure are denoted x^* and $t^*(x)$, respectively. Each new bootstrap sample, x^*, has the same number of elements as x but consists of values randomly pulled from x, with replacement from x. Values are repeated because the bootstrap sets created simulate the true population's distribution. Assume B bootstrap sets are created, and the statistical measure of each x^* is calculated; in this case, the $t^*(x)$ is $\bar{\mathbf{x}}\{0.25\}$. The empirical standard deviation of the B bootstrap trimmed means is the bootstrap estimate of the standard deviation for the trimmed means. Because the bootstrap population is analogous to the true population, the standard deviation of the bootstrap trimmed means is a representation of the error estimate of the original statistic, $\bar{\mathbf{x}}\{0.25\}$. Table 5.2 gives the bootstrap error for the demonstration data at different values of B, which can be compared with a true standard error of twelve randomly sampled values from the full dataset of 768 women (7.66). Figure 5.8 is a visual representation of the bootstrap method adapted from afore mentioned paper in *Science* [24].

Bootstrapping is used with BENDS by weighting the final manifold by the standard deviation of the bootstrap manifolds. The inverse of the standard deviation of the manifolds at each instance in the spectral data is normalized and used as a weight as described in the BENDS algorithm section below.

BENDS Algorithm

Assume a matrix, $\boldsymbol{X}$ of M by N dimensions (e.g., ARS spectra) with predictors in a vector, $\boldsymbol{Y}$ of length M (i.e. concentrations). The columns of $\boldsymbol{X}$ are sorted with respect to strictly increasing $\boldsymbol{Y}$ (i.e. $\boldsymbol{Y}_1 > \boldsymbol{Y}_2 > \ldots > \boldsymbol{Y}_M$). Create two hundred bootstrap sample data sets $\boldsymbol{X}^*$

in which a cubic smoothing spline is calculated as described above. A simple arc is depicted in figure 5.9, which will be used throughout this section as the algorithm is described. Figure 5.9 illustrates 25 bootstrap curves that are used to create the final spline as described below.

Interpolation for the spline manifold is done by the average interval of $\boldsymbol{Y}$,

$$I = \frac{\sum_{i=1}^{M} \boldsymbol{Y}_{i+1} - \boldsymbol{Y}_i}{M} \tag{5.13}$$

The interpolation values, $\boldsymbol{YY^*}$ are calculated by creating a vector of values from the minimum of $\boldsymbol{Y}$ minus I to the maximum of $\boldsymbol{Y}$ plus I in intervals of one tenth of I,

$$\boldsymbol{YY}^* = [\boldsymbol{Y}_1 - I \quad \boldsymbol{YY}^*_{i-1} + 0.1I \quad \cdots \quad \boldsymbol{YY}^*_M + I\,] \tag{5.14}$$

The interpolated values of the manifold, $\boldsymbol{XX^*}$ for each $\boldsymbol{X^*}$ are averaged and a standard deviation, $\boldsymbol{S}_X$ is found. Each value in $\mathbf{S}_X$ that corresponds to a value of $\boldsymbol{Y}$ is extracted, where $\boldsymbol{S}_Y = \boldsymbol{S}_X$ when the identifier $\boldsymbol{YY}^* = \boldsymbol{Y}$. $\mathbf{S}_Y$ is then used to weight the final manifold,

$$\boldsymbol{W} = \left[\frac{\boldsymbol{S}_Y}{\max(\boldsymbol{S}_Y)}\right]^{-1} \tag{5.15}$$

Notice that the inverse of the standard deviations is taken in order to down-weight the areas of the manifold that have higher bootstrap estimates of error. The new manifold is calculated with a cubic smoothing spline using $\boldsymbol{W}$ as the bootstrap weight values. The manifold is interpolated with the same interpolation values as in equation 5.14. The resulting manifold $\boldsymbol{XX}$ is the bootstrap enhanced manifold. The bootstrap enhanced manifold of our demonstration data is shown in figure 5.10.

Distance Along the Curve. Data reduction is a major component of all MVC procedures, and BENDS is no exception. In order to represent the multivariate data in a reduced form, the distance along the manifold curve to each spectral point is calculated (equivalent to a principal component score, but in a curved coordinate system). In order to calculate the distance along the curve, arc length must be found. If a real function f(x) exists such that $f(x)$ and $f'(x)$ are continuous on $[a, b]$, then the arc length s between $x = a$ and $x = b$ is found the formula:

$$s = \int_b^a \sqrt{1 + [f'(x)]^2}\,dx \tag{5.16}$$

BENDS utilizes a cubic spline that comprises piecewise cubic functions and therefore finding arc length is more difficult. The current implementation of BENDS determines the distance along the curve using the distance formula. For a point (x_1, x_2, …,x_n) and a point (y_1, y_2, … , y_n) the Euclidean distance (d) between these two points is defined as:

$$d_{x,y} = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2} \tag{5.17}$$

The data points are projected orthogonally onto the manifold to locate their position on the curve (see figure 5.11). The sum of the distances along the manifold to each new spectral point provides a rough estimate of the arc length. For a set of spectral points (p_1, p_2, p_i, …,p_M) in the manifold $\boldsymbol{XX}$ where each point has the dimensions (x_1, x_2, x_j, …, x_N) the distance along the curve, D_i to any spectral point, i is defined as:

$$D_i = \sum_{i=1}^{i} \left\{ \sum_{j=1}^{N} \left(\left| p_{i,j} - p_{i+1,j} \right|^2 \right)^{1/2} \right\} \tag{5.18}$$

In the same manner that the expression for arc length is derived, the distances begin to approach the true arc length as the number of segments summed approaches infinity. The manifold can be interpolated to create spectral points on the manifold in between measured spectral points, increasing the number of segments to sum. The default interpolation has been set to a factor of ten points more than the number of original data points for this reason, and also to estimate possible points not scanned during an actual calibration. Note that the distance along the curve is only calculated to the values of $\boldsymbol{XX}$ that correspond to values of $\boldsymbol{Y}$.

Least Squares (LS) Regression. To complete the BENDS calibration process, a best fit line is calculated through $\boldsymbol{Y}$ and its corresponding distances, D. Ordinary least squares regression can be used with the linear model,

$$D = m\boldsymbol{Y} + b \tag{5.19}$$

where m is the slope and b is the intercept of the line, as long as only one BEND manifold is needed to calibrate the instrument for the data. The least squares method

finds the line that minimizes the squares of the deviations of the data points from the line. The slope, *m* is calculated by the determinant,

$$m = \begin{vmatrix} \sum_{i=1}^{M} (\boldsymbol{Y}_i D_i) & \sum_{i=1}^{M} \boldsymbol{Y}_i \\ \sum_{i=1}^{M} D_i & N \end{vmatrix} \div Q \tag{5.20}$$

and the intercept, *b* is calculated by the determinant,

$$b = \begin{vmatrix} \sum_{i=1}^{M} (\boldsymbol{Y}_i^2) & \sum_{i=1}^{M} (\boldsymbol{Y}_i D_i) \\ \sum_{i=1}^{M} \boldsymbol{Y}_i & \sum_{i=1}^{M} D_i \end{vmatrix} \div Q \tag{5.21}$$

where *Q* is,

$$Q = \begin{vmatrix} \sum_{i=1}^{M} (\boldsymbol{Y}_i^2) & \sum_{i=1}^{M} \boldsymbol{Y}_i \\ \sum_{i=1}^{M} \boldsymbol{Y}_i & N \end{vmatrix} \tag{5.22}$$

The best-fit LS line is then employed as the linear predictive model for new unknown samples. Leave-one-out cross validation is used to assess the fit. Imagine the spectral point that is being left out as a new unknown sample; the BENDS algorithm is used to predict each sample that is left out using the other samples still left in the model. Figure 5.12 is a plot of the best fit line through the model sample distances along the curve. The validation samples are superimposed on the line.

Leave-one-out cross validation. A leave-one-out (LOO) cross validation is used to estimate the predictability of a data set by a certain model without needing to obtain new spectral points, but instead using the already acquired calibration data set. The algorithm loops through the ***X*** and ***Y,*** successively, leaving one observation out at each iteration, predicting that one observation using the remaining observations, and testing the model's ability to predict the identity of each data point left out. Each of the *M* observations in *X* is left out and tested with the remainder of ***X***. The first time through the loop, X_1 is left

out and the bootstrap manifold is created as described above, but with a subset of X, $\boldsymbol{tX}$ where $\boldsymbol{X}_1 \notin t\boldsymbol{X}$. The BENDS algorithm is completed as described above yielding a best-fit line. The left out point, X_1 is projected into hyperspace with the manifold and the Euclidean distance is calculated to each interpolated point on the curve (equation 5.17). The closest point on the interpolated manifold is determined by the minimum of these Euclidean distances. The distance along the manifold to the point determined to be the closest to the point left out is calculated as described in equation (5.18). The distance D is used along with m and b from the calibration to determine the predicted value, p_1 on the best-fit line. This LOO procedure is performed for each $\boldsymbol{X}$ and the standard error of prediction (SEP) is measured,

$$\left(\frac{\sum_{i=1}^{n}(\boldsymbol{Y}_i - p_i)^2}{N-1}\right)^{1/2} \tag{5.23}$$

The SEP value is used to determine the predictive utility of the model.

Data Mining with BENDS

BENDS is an MVC method and therefore deals with large data sets where some data mining methodologies would be beneficial. A few different types of data mining have been tried with BENDS with varying results, all with benefits and drawbacks. Acoustic data could have anywhere from 8-22 thousand variables (dimensions in hyperspace), and therefore an efficient and non-supervised method of selecting the important frequencies is desirable.

All possible combinations (APC). Originally, it was thought that taking every possible combination of the variables, BENDing them, and finding the best correlating variables might be effective for frequency selection. Note that the order of the selected variables does not matter, i.e., whether variable 1 is assigned dimension 1 or 3 does not change the outcome. The total number of combinations possible, C of r variables in a set of variables, n is the binomial coefficient,

$$\mathrm{C} = \binom{n}{r} = \frac{n!}{r(n-r)!} \tag{5.24}$$

On simple datasets with only a few independent variables, the APC strategy could easily be implemented (e.g., $n \leq 20$ and $r \leq 3$). The upper limit of combinations for a simple data set would require the BENDS to be performed more than 1000 times. A realistic data set with 8000 variables would require the number of iterations of BENDS to approach 85 billion, which is simply not practical.

Sort According to Standard Deviation. The second method of data mining for BENDS was to utilize the standard deviation (see equation 5.11). The standard deviations of all observations, independent of their identities were taken and the variables were sorted according to strictly decreasing standard deviations. For instance, given the data in table 5.3 the standard deviation is calculated for each variable independently across all samples regardless of their identity. The standard deviations are represented in table 5.4. The dataset is now reordered so that the first column of the data represents the column with the highest standard deviation, and the last column represents the lowest standard deviation (table 5.5).

Once the data are reordered, the variables with the variance are grouped to the left while those with the least variance are grouped to the right. The assumption is that the data change for chemical and physical reasons other than noise. Higher variability in the spectral dataset at certain frequencies usually connotes independent variables that are changing the most with changes in sample composition or identity. BENDS would then be performed on the high variability portion of the dataset while the rest of the frequencies were left out. There are some major drawbacks to the frequency selection by variance technique. In a real application, such as acoustic resonance spectroscopy, most of the response at each frequency is due to differences or similarities in the different samples; however, the responses might not be due to the analytical property that is being studied. It could be beneficial to select the portions of the data changing over the different samples and leave out the portions that are similar because the similar portions will not have a large standard deviation. Another drawback is the possibility that the samples may have a very small change with respect to the analytical property being studied, and therefore sample signals could be buried in the resorted data just as much as they were in the original ordered data.

The standard deviation approach was not incorporated into the final BENDS algorithm, although it could be if a priori knowledge were available of the different responses that are occurring in the data to be acquired. For example, if a selected portion of the spectrum is being used that exhibits a known non linear functional relationship with respect to the specific analytical property being studied (and not to any interference that may be present in the samples) then this magnitude of variance technique could be beneficial.

Interval BENDS. The most efficient and reliable data mining technique used with BENDS has been the interval technique. The idea behind interval BENDS is to set up a moving window of size B (also referred to as bin size) and move the window progressively across the data in steps of t. Using the same data as above in table 5.3, a bin size of two could be set with a step size of two. The first window of BENDS would be performed on variables one to two and then the window would move over two variables and then BEND variables three to four. If a bin size of two and a step size of one were desired, the first window BEND would be performed on variables one to two as before, but the window would only move over one variable, using variables two to three in the next window. The window would progress to variables three to four in the next jump.

The larger data sets that are acquired via ARS can be efficiently scanned with steadily decreasing intervals. A large bin size (e.g., 500) with a large step size (e.g., 500) would be set at first to select the best 500-variable region. The bin size and step size can then be incrementally reduced to a level more commensurate with the peak width (e.g., B = 5; t = 1). The efficiency of this technique is much faster than the all possible combinations approach; however, some precision is lost in finding the best possible region. The region selected out of the 500 data points may not contain the five data point region that has the greatest correlation because the five data point region is in another interval of the data that may have more uncorrelated regions around it. Because it is impossible to test all possible combinations of frequencies from a time standpoint, this potential drawback of the interval technique does not outweigh the benefits and strengths of the approach.

The purpose of this investigation is to test the hypothesis that the distances along an interpolated nonlinear manifold calculated through observed variables in order of monotonically increasing analytical property will correlate in a linear fashion to the analytical property of interest, thus transforming and reducing a nonlinear data set into a more manageable linear model. The nonlinear manifold will be optimized with the bootstrap method to estimate the error of the manifold and adjust it accordingly. The model should outperform traditional linear methods with respect to the linear correlation (r-squared) and the predictive parameter (standard error of prediction).

Results and Discussion

Bootstrap Enhanced N-dimensional Deformation of Space

The BENDS algorithm was used to find the region of the spectra that correlated to tablet hydration. The large bin section was a 500 data point region from 7.714 to 9.093 kHz. Note that the spectra in this region were highly variable, though it was not the only section that showed higher than average variability. Using standard deviations at each frequency for data mining is not a particularly effective method of isolating regions with information correlated to the analytical property of interest. The other regions could have increased noise, extra overtones, signal contributions from manufacturing differences between tablets, or they could simply be contaminated by some other factor not yet perceived in the acoustic resonance spectra. The large bin size did perform comparably with the other bin sizes (see table 5.6). The relative SEP values ranged from 10.41% to 6.760% for the 500-bin manifold and the 5-bin manifold, respectively. Table 5.6 provides the statistics calculated with BENDS and a leave-one-out cross validation of the data. The optimum *m*-value for most of the analysis is 1.00, meaning that the data were highly nonlinear and the BEND needed to pass through every point directly.

The linear regression of percent hydration versus distance along the curve is shown for the 5 data point bin size in figure 5.14A. The regression shows strong linearity created by BENDS with an R-square of 0.9834. A scatter plot of the actual percent hydration versus predicted percent hydration with a leave-one-out cross validation is next to the regression in figure 5.14B. A diagonal line with a slope of 1 is included to depict a perfectly

predictable model. The majority of the data points lie on the diagonal line giving support a strongly predictable model with a relative SEP of 6.760%.

Principal Component Regression

A comparative analysis using PCR instead of BENDS was performed on the spectra, with the results that might be expected for a highly nonlinear calibration. PC scores were calculated using a singular value decomposition of the data matrix. An ordinary least squares regression was calculated from the PC scores and the corresponding percent hydration values. The results were poor, with an R-square of 0.7785 and a relative SEP of 23.69% (see table 5.6). Figure 5.15 depicts the predictability of the data as a linear model using PCR. Figure 5.15 is a plot of percent hydration versus hydration values predicted with PCR. A diagonal line with a slope of 1 is drawn through the predictions to illustrate where the true values are located. The m-values for the entire bin sizes are above 0.9, indicating that the BEND through the data needed to pass very closely to every point in order to create a valuable trend, and that the S/N was therefore good. BENDS was able to create an effective linear model in which the percent hydration could be predicted.

Experimental Section

Tablet Preparation

Aspirin tablets (TopCare, 325 mg) were obtained for scanning by the ARS. Twenty-five tablets were divided into five different hydration groups, maintained 0, 4, 8, 16 and 24 hours in a hydrator. Hydration ranged from 1.81 to 4.85 percent by mass. The tablets were placed in the hydrator at different times so that scanning of the tablets could occur at the same time. The wet mass of each tablet was taken before scanning in triplicate with the ARS to determine moisture content. The tablets were stored in 20-ml vials with air-tight lids in order to keep the tablets from drying out in between scans. Percent hydration was calculated as follows:

$$\frac{mass_{wet} - mass_{dry}}{mass_{wet}} \times 100\% \qquad (23)$$

ARS Data Collection

All hydrated pills were scanned along with blanks (a scan of the empty base-plate at a pressure equal to that used on the tablets), in triplicate and in random order. Each tablet was placed on a scale (Model 3120, Health O Meter, Bridgeview, IL, USA) and adjusted to a pressure of 150g so that contact between the sample and the quartz rod of the ARS was held constant throughout scanning. After each scan, the scale was reset and the tablets repositioned. White noise in the frequency range of 0 to 3.1 MHz was generated using a function generator (Stanford Research Systems, Sunnyvale, CA, USA) and used to excite the tablets through the quartz rod of the AR spectrometer. The sound card used to capture the data (Model No. SB0490, Creative Labs) had a range of 20 Hz to 22 kHz. All data processing was done in Matlab 7.0.1 (The Mathworks Company, Natick, MA, USA). All sound was captured for five seconds with a sample rate of 44.100 kHz.

Data Processing

A FFT was performed on the sound files to convert the data from the time domain into the frequency domain. The mean of the three replicate measures was taken and smoothed to reduce noise (see figure 5.13). Interval BENDS (iBENDS) was performed with varying bin sizes, starting at 500 and increasing in steps of 100, to find a region with moderate correlation., iBENDS was then performed on the correlating data with bin sizes of (5, 10, 20, . . . , 100) in steps of 1. PCA was also performed for comparison with the iBENDS result.

Cross Validation. The predictability of the algorithm was tested with a leave-one-out cross validation, where each spectral point was used as an unknown while the remainder of the data were used as a calibration set. The SEP was calculated to characterize the ability of the data set to predict unknowns within its range (see equation 5.23). Relative SEP was calculated by dividing the SEP by the range of the water concentrations (% hydrations).

Conclusion

Acoustic resonance spectrometry did indeed produce a nonlinear relationship between frequency intensities and percent hydration. Bootstrap enhanced n-dimensional deformation of space provided a metric to develop a linear model from the nonlinear relationships with acoustic spectra of hydrated aspirin tablets. BENDS outperformed the conventional PCR method in both accuracy and precision. The BENDS model only required a five data point region with each hydration level, which is a positive for the method when attempting to use ARS as a PAT. The study gave strong evidence that BENDS can be applied to highly nonlinear correlations in both simulated and actual data sets. The next step for BENDS research is to find more ARS applications with nonlinear correlations, attempt to apply BENDS to other instrumentation that creates a nonlinear MVC problem and also to develop an algorithm that applies ISP with BENDS.

ISP-ARS has the benefit of performing all the high end MVC mathematics directly at the sensor by tailoring the excitation. The idea would be to create an excitation waveform that traces the nonlinear model in such a way that the sum of the voltage at the detector is directly proportional to the distance along the curve. ISP-ARS with BENDS would create method that is more rapid, uses simpler and more robust instrumentation, and does not require a broadband excitation.

Chapter five Tables

Table 5.1 Plasma glucose concentrations.

100	105	110	111	115	120	128	129	138	157	162	188

Table 5.2 Bootstrap estimates of error for the trimmed mean.

Bootstrap Replicates (*B*)	Bootstrap Estimate of Error(±)
50	7.47
100	8.56
200	8.61
500	8.75
1000	8.85

Table 5.3 Sample data for demonstrating standard deviation.

Identity	Variable 1	Variable 2	Variable 3	Variable 4
1	0.233	0.111	0.433	1.233
2	0.234	0.125	0.523	1.221
3	0.231	0.923	0.599	1.527
4	0.235	0.422	0.483	1.588

Table 5.4 Standard deviations of sample data.

Identity	Variable 1	Variable 2	Variable 3	Variable 4
Stand. Dev.	0.002	0.380	0.070	0.192

Table 5.5 Sample data reordered according to standard deviation.

Identity	Variable 2	Variable 4	Variable 3	Variable 1
1	0.111	1.233	0.433	0.233
2	0.125	1.221	0.523	0.234
3	0.923	1.527	0.599	0.231
4	0.422	1.588	0.483	0.235
Stand. Dev.	0.380	0.192	0.070	0.002

Table 5.6 Results of iBENDS analysis and PCA performed on the hydration data. The data were binned according to bin size below and the data window was moved in steps of 50 for bin size of 500 and in steps of 1 for bin sizes of 5 to 100. The frequency region shown for bin size of 500 was used when calculating BENDS with the other bin sizes. The last row represents the PCA results.

Bin Size	m-value	Freq Range (kHz)	R^2	Relative SEP
5	0.99	7.996 - 8.007	0.9834	0.0676
10	1.00	7.833 - 7.858	0.9750	0.0619
20	1.00	8.776 - 8.828	0.9547	0.0818
30	1.00	8.773 - 8.853	0.9630	0.0822
40	1.00	8.754 - 8.861	0.9622	0.0847
50	1.00	8.729 - 8.864	0.9711	0.0781
60	1.00	8.702 - 8.864	0.9619	0.0868
70	1.00	8.743 - 8.933	0.9733	0.0889
80	1.00	8.751 - 8.969	0.9784	0.0936
90	1.00	8.746 - 8.991	0.9777	0.0934
100	1.00	8.743 - 9.016	0.9772	0.0933
500	0.94	7.714 - 9.093	0.9623	0.1041
PCA	---	---	0.7785	0.2369

Chapter five Figures

Figure 5.1 Schematic of the quartz rod AR spectrometer. A function generator is depicted as the source (A). White noise is generated and the voltage is converted into a sound wave by a piezoelectric disc (B) which is coupled to the quartz rod. The sound resonates down the quartz rod which is shown as a blue sinusoidal wave (C) and two key interactions occur. A portion of the energy (red) is introduced into the sample and interacts in a specific manner dependent of the sample and another portion of the energy (blue) continues unaltered through the quartz rod. The two waves recombine after the sample (D) and constructive or destructive interference occurs depending on the phase shift due to the sample. The altered combined energy (purple) is converted to an electrical voltage by another piezoelectric disc at the end of the quartz rod (E). The voltage is then recorded onto a computer by a sound card (F).

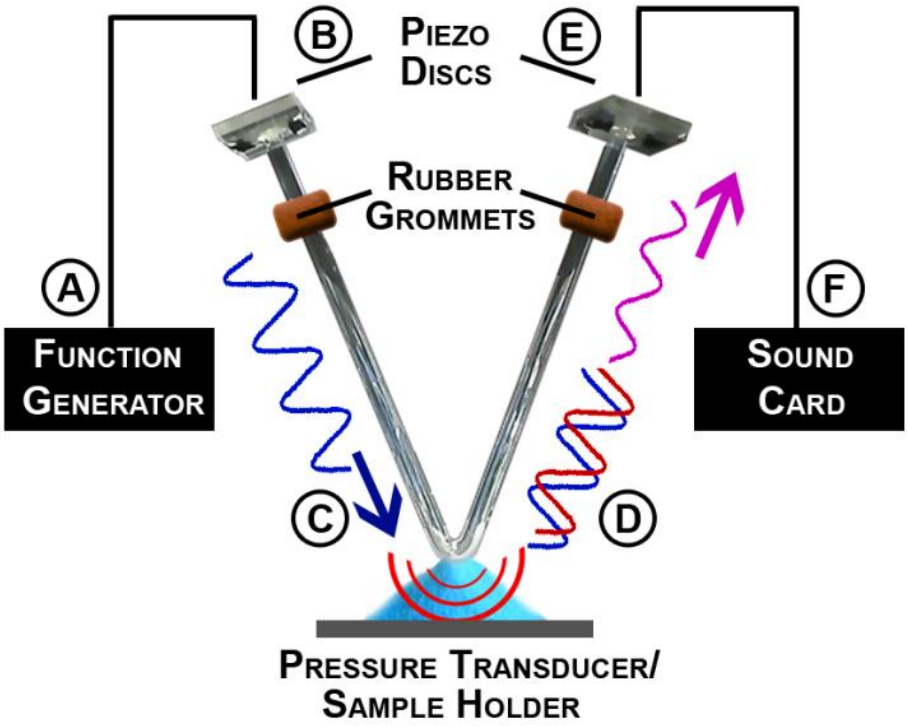


Figure 5.2 Constructive interference of two waves of the same frequency and amplitude, but one delayed behind the other (green and blue lines). The solid red line is the sum of the green and blue lines, indicating the observed amplitude will be greater than that of both smaller waves.

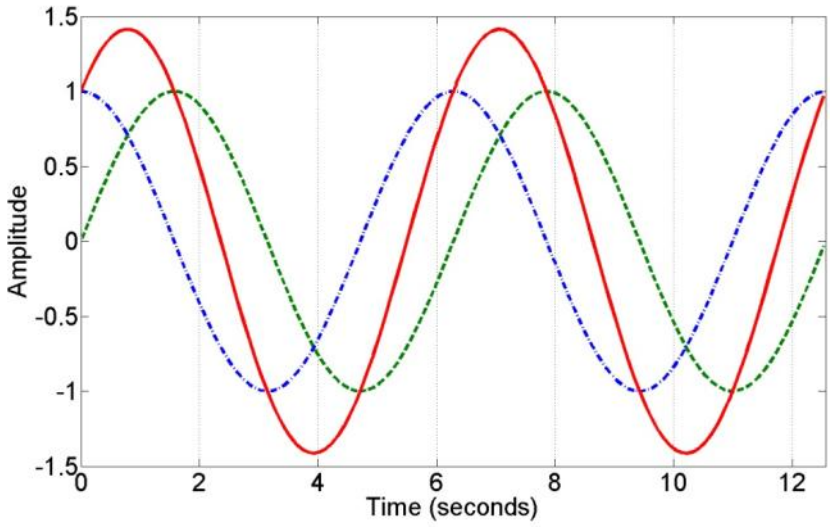


Figure 5.3 Destructive interference of two waves of the same frequency and amplitude, but one delayed behind the other (green and blue lines). The solid red line is the sum of the green and blue lines, indicating the observed amplitude will be less than both smaller waves.

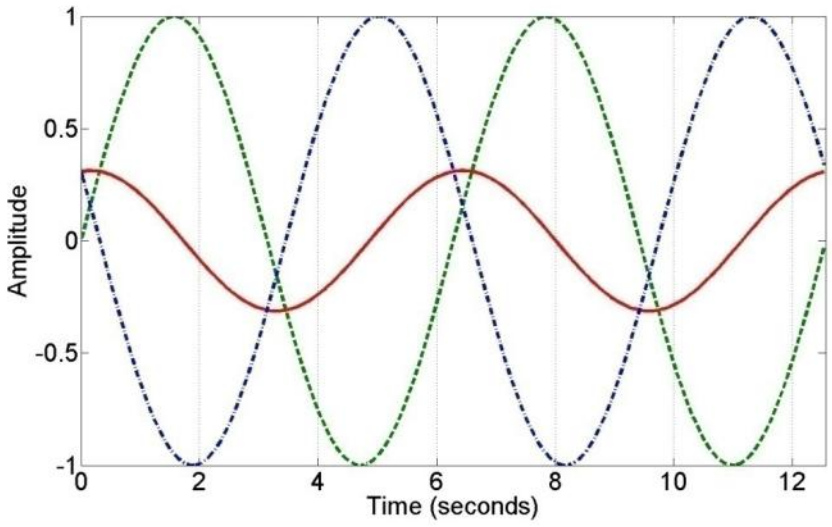


Figure 5.4 Origins of non-linear responses in acoustic resonance spectroscopy. Frequency A is changing linearly with increased concentrations while intensity B is reacting nonlinearly. Their combined contribution traces a 1-D curve that is shown in the lower chart of intensity A versus intensity B.

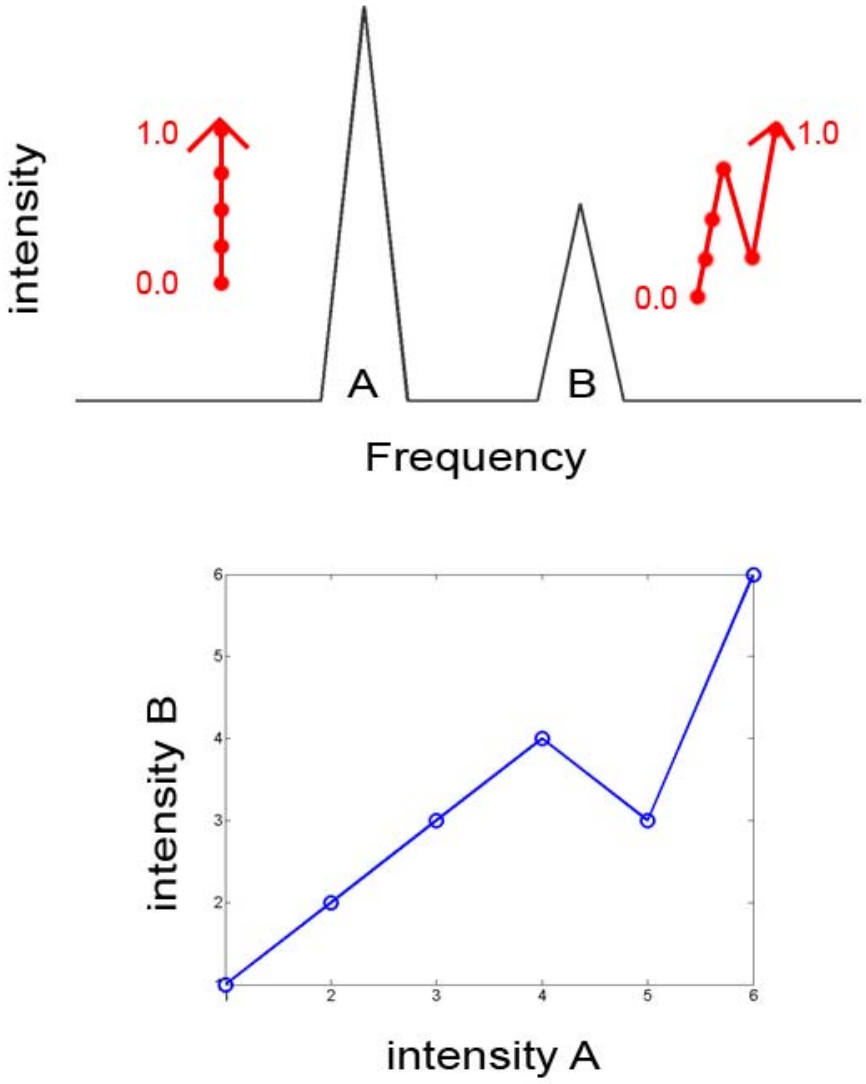


Figure 5.5 Depiction of polynomial fitting. The blue circles represent the data points of an analytical property of interest plotted versus some measured observation. The lines represent polynomial fitting to the third degree (purple), fifth degree (black), seventh degree (green) and ninth degree (blue).

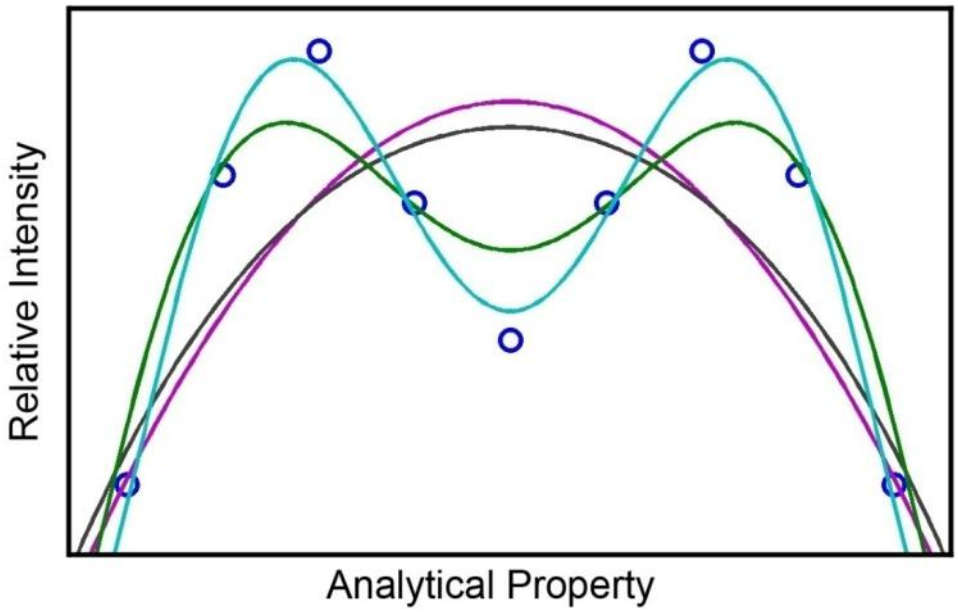


Figure 5.6 Comparison of cubic spline versus polynomial fitting. The blue circles represent the data points of an analytical property of interest plotted versus some measured observation. The red line represents a cubic spline while the other lines represent polynomial fitting to the third degree (purple), fifth degree (black), seventh degree (green) and ninth degree (blue).

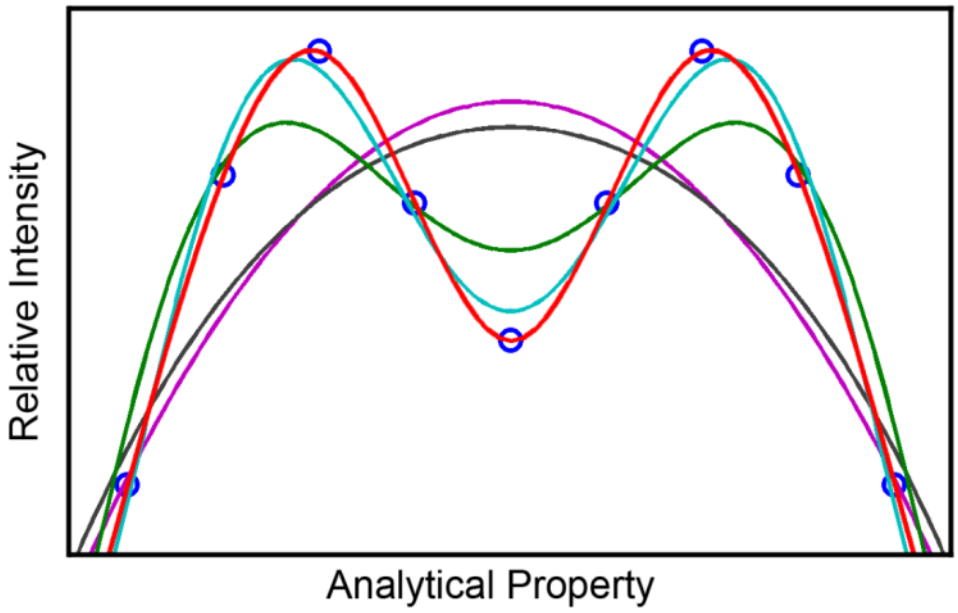


Figure 5.7 m-value in splining. The spline algorithm includes a variable *m*-value that ranges from 0 to 1, which forces the spline to pass through all points as it approaches 1. The different *m*-values shown are (*A*) *m* equal to 0, which approximates a least-squares linear fit, (B) *m* equal to 0.001, (C) *m* equal to 0.5, and (D) *m* equal to 1.0, which forces the spline to pass through all points.

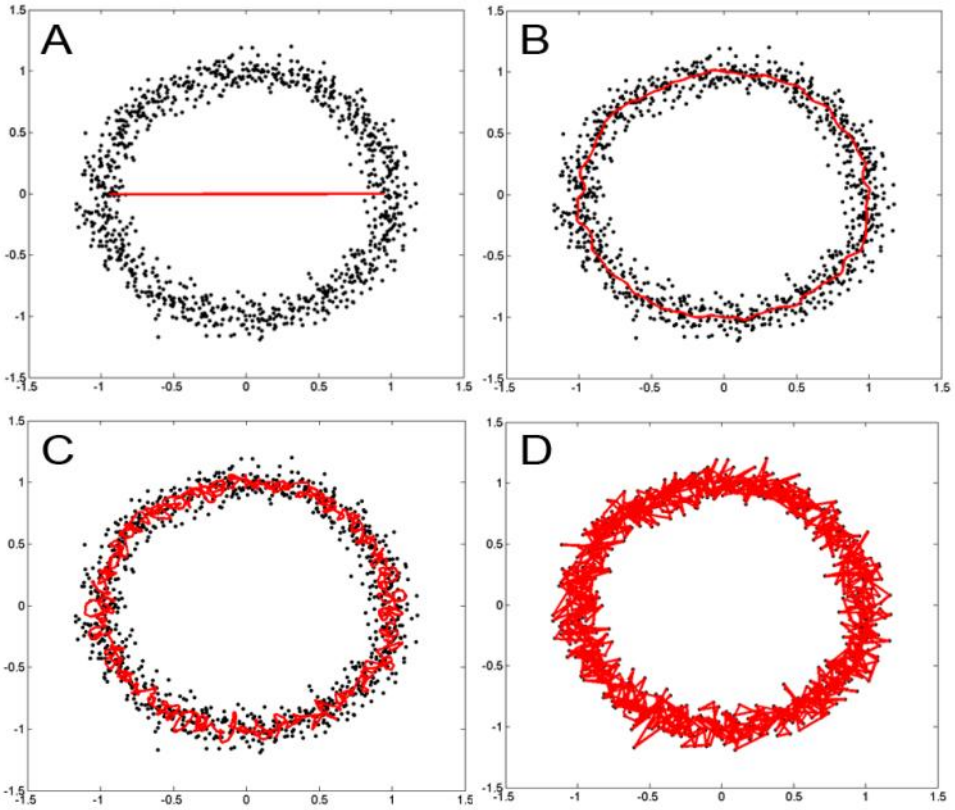


Figure 5.8 Statistical bootstrapping demonstration. The original data (black) are randomly sampled to B bootstrap sample sets (red) and the statistic t*(x) is calculated (green) for each bootstrap sample set. The bootstrap estimate of the standard deviation of t(x*) is performed by calculating the standard deviation off all the t(x*) (blue). Figure adapted from the paper in Science by Bradley Efron [24].

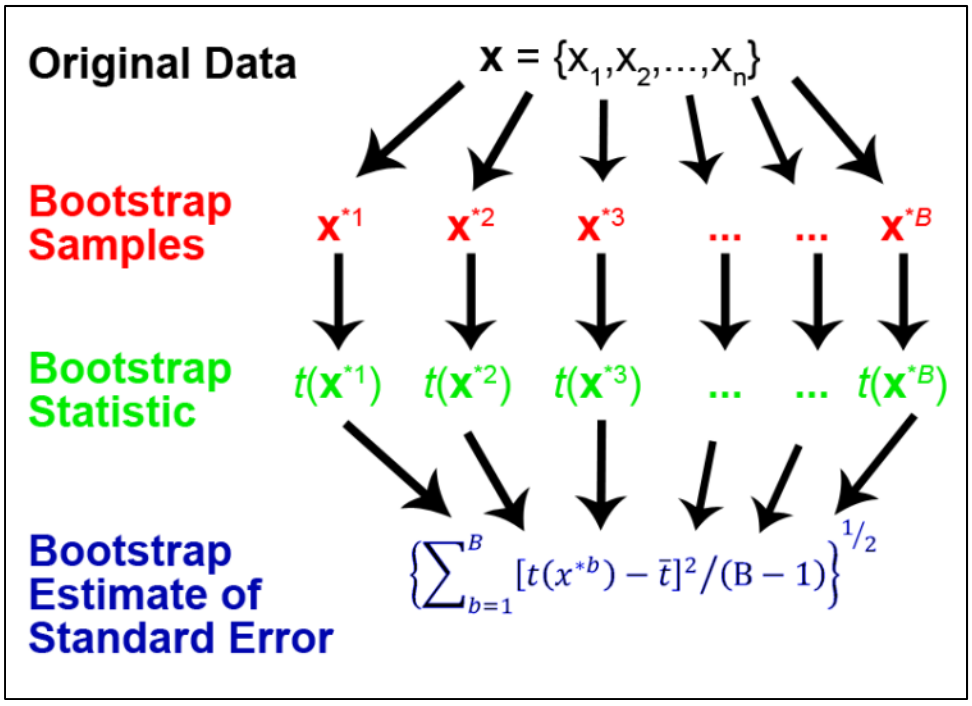


Figure 5.9 Bootstrap cubic smoothing splines. The plot depicts the 25 bootstrap manifolds (blue lines) through the ordered data (red circles).

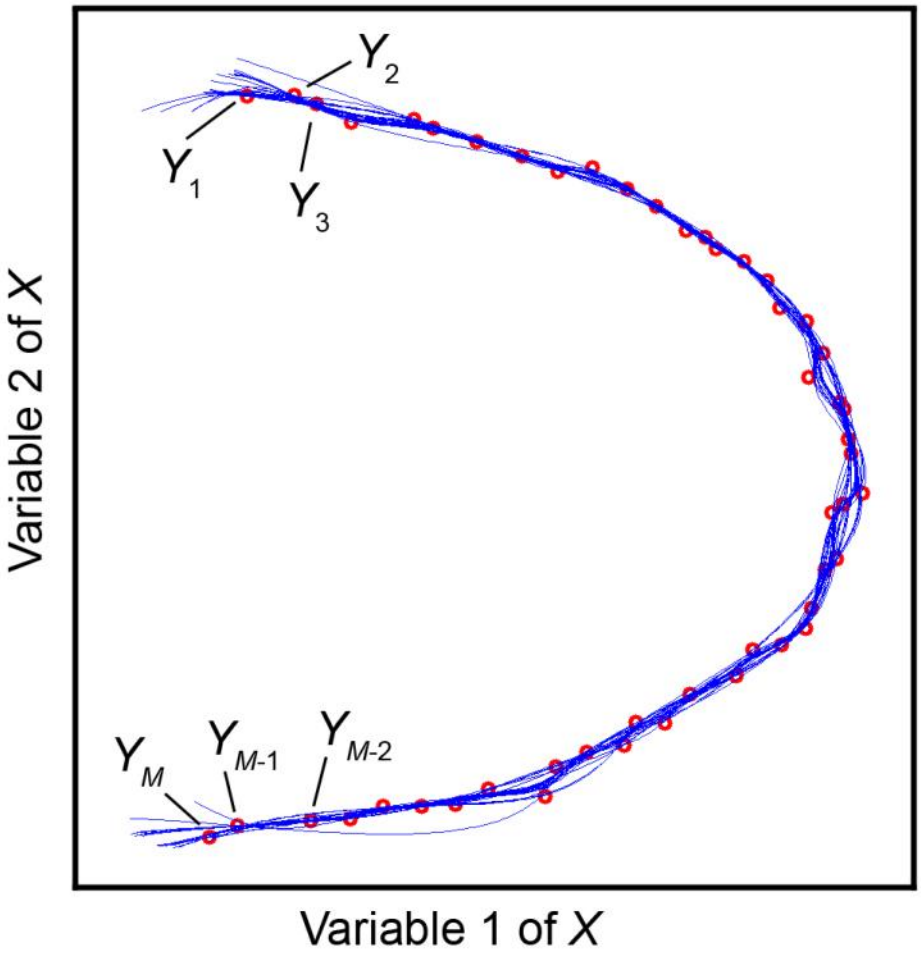


Figure 5.10 Bootstrap enhanced manifold. The data (red circles) are used to calculate bootstrap enhanced manifold (blue)

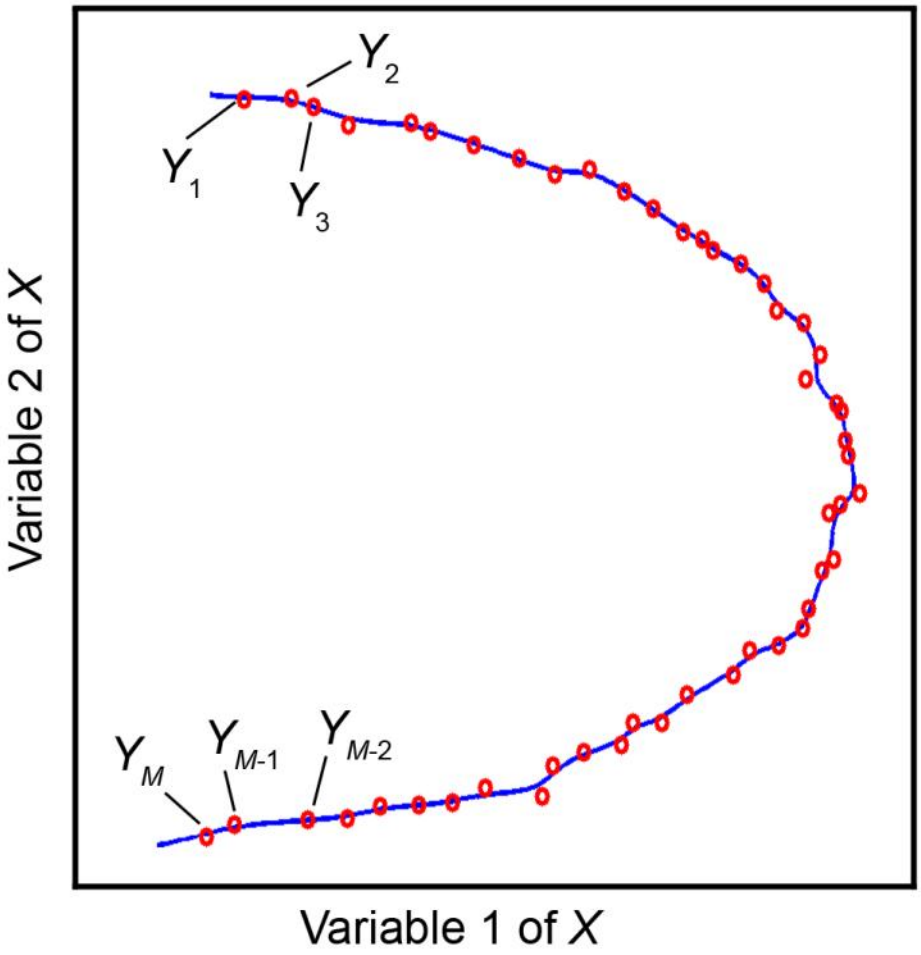


Figure 5.11 Projection of data points onto curve. The plot is a zoomed version of figure 5.10. The data points (red dots) are projected onto the bootstrap enhanced manifold (blue line) orthogonally in order to find their position on the curve. The projection is done by finding the interpolated point on the manifold that is closest to the actual spectral data point in hyperspace.

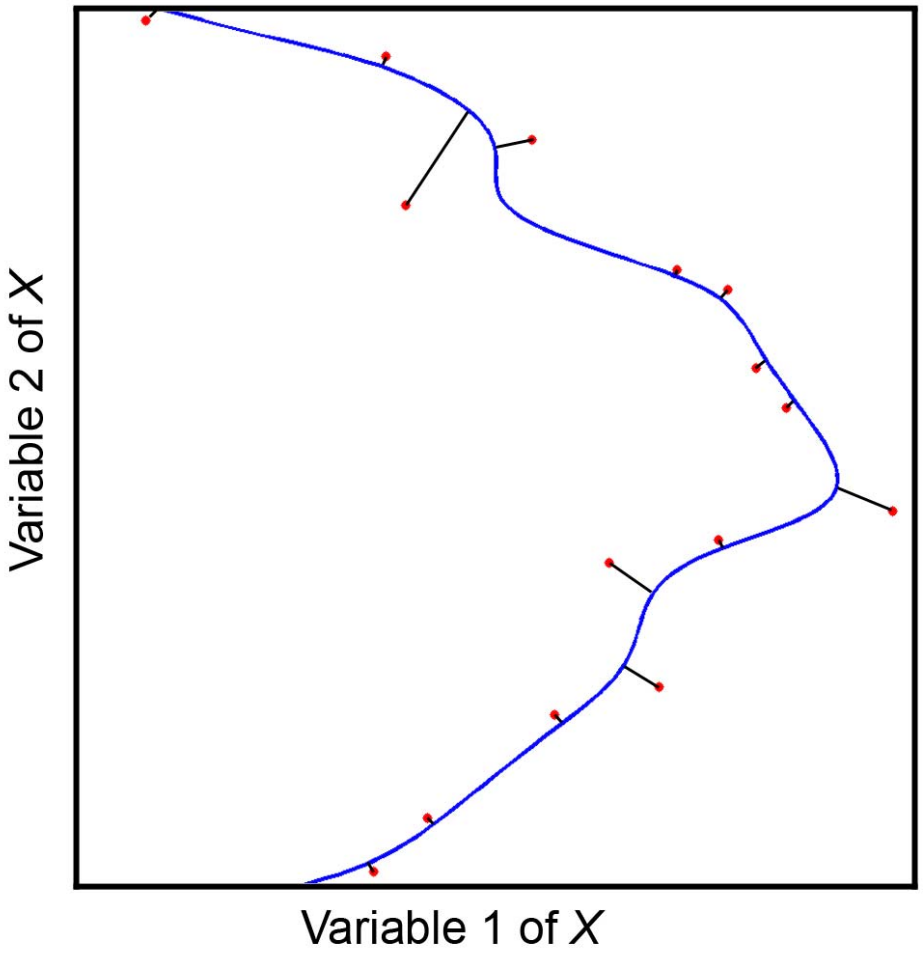


Figure 5.12 Best fit line. The distances along the curve were regressed with respect to *Y* resulting in an r^2 =0.9996. The correlation is very linear, especially compared to the original data which is seen in Figure 5.9.

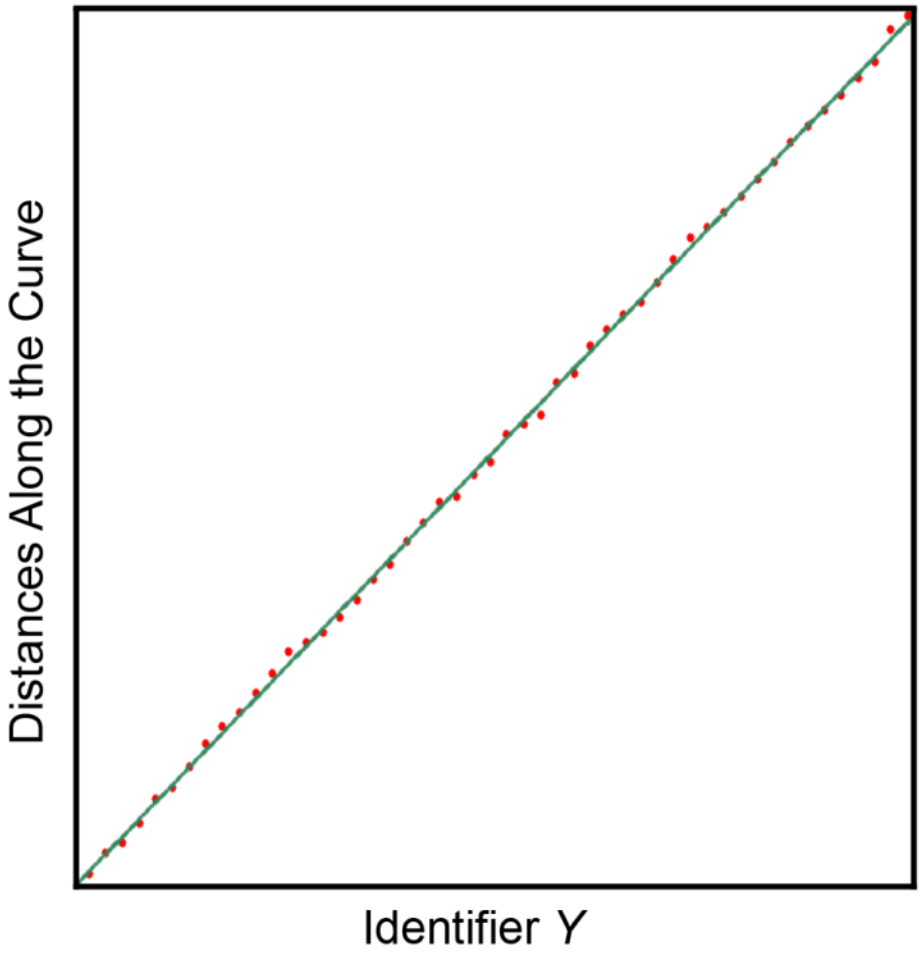


Figure 5.13 Acoustic spectra of hydrated aspirin tablets. The data are smoothed to reduce noise, and the spectra displayed here are smoothed beyond what was used in the analysis to provide a clearer picture graphically. A band of frequencies around 8 kHz differentiate the levels of hydration in the tablets best.

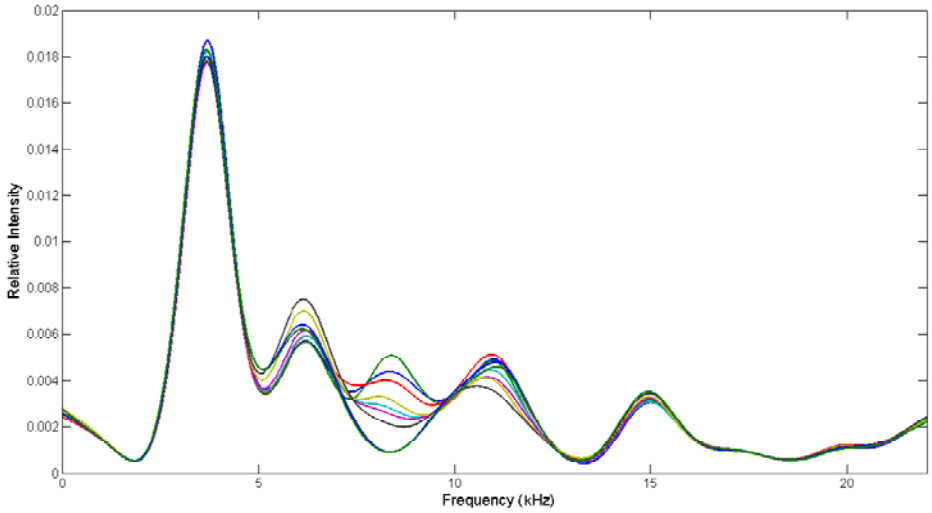


Figure 5.14 A linear model created with BENDS is superior to principal components. A bin size of 5 data points was used to create the linear model. The statistics for the regression and the cross validation are shown in table 5.6. Plot A depicts the distance along the curve to each point with a least squares linear fit through the distances corresponding to those points. Plot B are the actual hydrations versus the predicted values using leave-one-out cross validation with a diagonal line drawn through those points.

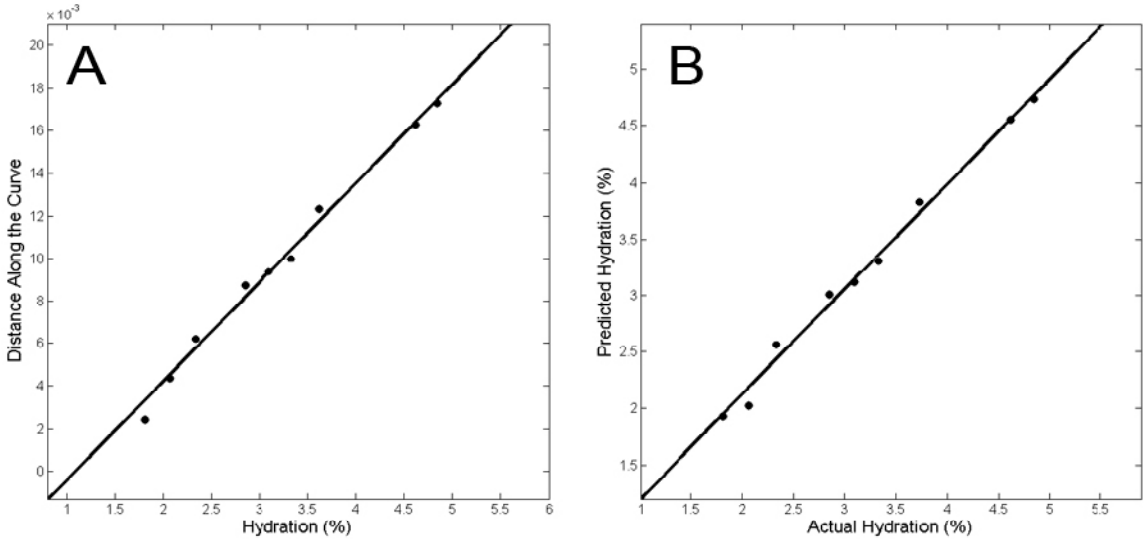
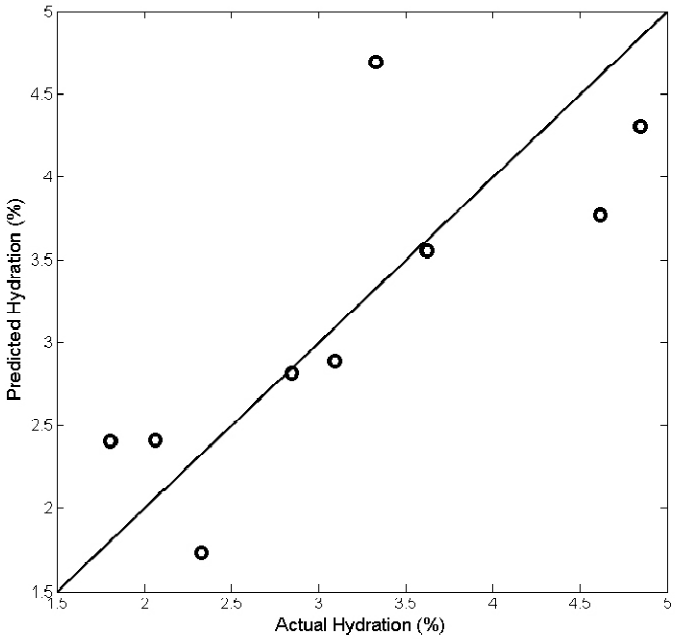


Figure 5.15 Prediction using PCA on acoustic data of hydrated tablets is less effective than the BENDS. The plot depicts the actual percent hydration versus the predicted values using a leave-one-out cross validation. The regression and cross validation statistics are represented in table 5.15.



Copyright Statement

Copyright © *Algorithms*

Link, D. J.; Hannel, T. S.; Lodder, R. A. *Algorithms*. (submitted 27 May 2009).

Chapter six – ARS with BENDS to Quantify D-tagatose Concentrations in Resveratrol Tablets

Introduction

Type 2 diabetes is a major problem in the United States and is responsible for nearly 225,000 deaths a year [30]. Type 2 diabetes is often referred to as adult-onset diabetes because it usually occurs in adults when the body stops producing enough insulin or cannot use it effectively. However, type 2 diabetes is also known to occur in children but is more difficult to diagnose. Complications arising from diabetes may consist of coronary artery, eye, blood vessel, kidney and nervous disorders. There are considerable economic consequences associated with diabetes. For example, the WHO estimates that $558 billion of Chinese national income will be lost due to heart disease, stroke and diabetes over the 10 year period of 2006 – 2015 [32]. The American Heart Association suggests that weight loss, exercise, and healthy eating can help reduce the risk factors associated with diabetes [67]. With obesity in the United Sates reaching epidemic proportions, it not likely that diet and exercise will be enough to reduce type 2 diabetes diagnoses [67]. Metabolic syndrome is another leading health problem in the United States with an estimated 50 million Americans currently diagnosed [68]. Metabolic syndrome is comprised of risk factors that lead to both heart disease and type 2 diabetes [68]. Heart disease is the biggest contributor to deaths in the United States attributing to the deaths of over 650,000 Americans in 2005 [69]. Clearly there is a need for the development of drugs that help reduce the risks and symptoms of diseases such as diabetes and metabolic syndrome.

Resveratrol (RSV; 3,5,4'-trihydroxystilbene) is an antioxidant commonly found in the skins of red grapes that is thought to increase cardiovascular health and has recently been shown to improve health and lifespan of mice [70]. RSV is also thought to possess insulin-like effects and has been reported to have a variety of pharmacotherapy effects [71-74]. RSV is of interest in treating type 2 diabetes because of its insulin-like effects. Also RSV has not been shown to overwork pancreatic β-cells [75]. D-tag is being developed as a treatment for obesity and type 2 diabetes. D-tag is an epimer of fructose and a natural component of heated milk products. Studies have shown D-tag to produce

antihyperglycemic properties as well as stimulate weight loss [34-35]. D-tag also reduces serum LDL cholesterol as well as the extent of atherosclerosis and may increase serum HDL cholesterol [77]. Pterostilbene shares many of the beneficial effects of resveratrol and D-Tagatose. It has been found to be antihyperlipidaemic increasing cardiac health and reducing "bad" cholesterol in rats [76].

D-tag, RSV and pterostilbene compounds are a recent a push for natural occurring compounds to assist in metabolic syndrome related health issues. This paper discusses the analytical procedures used to quantify D-tag levels in D-tag/RSV tablets because of their possible combined abilities against metabolic syndrome and type 2 diabetes. Future work will include formulations and tests involving a series of RSV-related stilbenes starting with pterostilbene. Correct formulations must be found to exploit the greatest potential of these naturally occurring compounds.

Production of experimental drugs such as D-tag and RSV in clinical trials must be closely monitored to ensure accurate dosage and content uniformity. If a synergistic effect exists between two or more active ingredients, then the ratio of APIs will need to be closely monitored. Frequently, drugs in phase 1 and phase 2 clinical trials are produced in small experimental batches in facilities far different from those that will eventually be used for mass production. Analytical science-based monitoring techniques should be in place to monitor as close to 100% as possible of such pharmaceutical products. This is especially true where small lots of many different experimental drugs are being prepared along with placebos. Identical experimental drugs can be mislabeled, and a simple nondestructive test capable of differentiating active from placebo could be useful.

PAT represents a shift from classical GMP to process understanding and control. An ideal PAT for in-line processes should be nondestructive and possess the ability to make accurate and rapid measurements [78]. PAT processes can streamline change from empirical standards such as cGMP to more scientific standards for manufacturing quality control. NIRS is a well-established method for qualitative and quantitative analysis and has been investigated as a PAT for online prediction of APIs [79[[80[[81[[82[[83[[84[. FTARS and ISP-ARS are new PAT techniques that have been applied to several model pharmaceutical systems [10-11]. FT-ARS has previously been investigated for qualitative

analysis [9][12][43][49-50]. A new approach currently being investigated is ISP-ARS. Using ISP-ARS, both the instrumentation and the computational analysis of ARS can be simplified. ISP-ARS is accomplished by tailoring the acoustic excitation spectrum to encode high-level information about the samples directly in the detector during the sample scanning process. Potential advantages of this method over others include smaller data volumes, rapid analysis, simpler and more robust instrumentation, and higher sample throughput. One weakness of ARS that also disables ISP is the presence of non linear correlations [13]. In this paper the BENDS algorithm is used to transform the non-linear correlation of ARS to a more simple and linear model that could potentially be coupled with ISP.

Theory

The AR spectrometer relies on structural properties of the tablets. Acoustic velocity is the major variable affecting the recorded AR spectrum. Acoustic velocity depends on the tablets' density, bulk modulus and shear modulus [12]. The quartz rod ARS has two acoustic paths that recombine to give the resulting acoustic wave that is recorded at the detector. One path is altered by the tablet while the other is the unaltered resonance of the quartz rod. The change in the acoustic wave is due to the tablet's specific acoustic velocity and absorption of the excitation. Once the two waveforms recombine, constructive and destructive interferences occur resulting in a highly unique spectrum to be recorded at the detector. One problem that can occur is the ability of the two waves to shift in and out of phase with one another and result in a nonlinear correlation in quantitative studies [85]. BENDS is an algorithm designed to combat the nonlinear correlation and transform the model into a more manageable linear model (see figure 6.1).

The BENDS algorithm projects the data points into n-dimensional hyperspace and calculates a piecewise cubic spline through the points in increasing concentration. The cubic spline is calculated using the following minimization function:

$$P\sum_{i=1}^{n} w_i|y_i - S(x_i)|^2 + (1-m)\int \lambda(t)\left|S''(t)\right|^2 dt \qquad (6.1)$$

Here, $|z|^2$ stands for the sum of the squares of all the entries of z and the integral is over the smallest interval containing all the entries of x. The weights are inserted as w and the smoothing parameter is m. λ is the piecewise constant weight function.

Similarly to principal component analysis, the data points are projected orthogonally onto the manifold. The distance is calculated from point zero to each point along the curve, and these distances are regressed against their corresponding concentrations. A least-squares regression is used and the resulting model is linear.

The bootstrap technique is used during the splining calculation in order to find the best possible curve for the data set. The bootstrap technique was devised in order to calculate the error in high order statistical measurements [28-29]. Two hundred data sets of the same size as the original data set are calculated by randomly selecting values from the original data set. The bootstrap data sets that are created are meant to represent a pooled set of values from the "population." The method is treating the original data set as the statistical population like traditional statistics have an acquired data set representing the true population. A well-executed experimental method and large data set are very important when using a bootstrap method since the acquired data set is being considered a true population.

A cubic spline is calculated through each of the two hundred data sets with the weight function holding all data points fully represented (w equal to 1). The standard deviation is calculated for each point and is normalized to 1. The standard deviation is called the bootstrap estimate of the error and can be used to find the areas of the data that comprise the most error. The inverse of the bootstrap estimates of the error are then used as the weights for each of the points. A final cubic spline is calculated through the original data set using the same minimization as in equation 1, but the weights from the bootstrap sampling are inserted. By incorporating these weights, the spline is calculated with the areas with the most error having the least effect on the trend.

Materials and Methods

*Tablets*Three identical batches of experimental 500mg tablets were pressed in-house. Resveratrol tablets (Source Naturals, Inc., 40mg) were mixed with varying amounts of D-

tag (Spherix, Inc.) and D-glucose (Fisher Scientific) and ground in a mortar and pestle into a fine powder. D-glucose was added to make consistent 2000mg samples. 500mg aliquots of the powder samples were used to make three batches of 28 tablets comprising 10mg resveratrol and D-tag from 0 to 50% by weight.

ARS Data Collection Acoustic resonance spectra were collected from an in-house built quartz rod spectrometer. The simple spectrometer comprises sending and receiving piezo electric transducers connected to the top ends of the V-shaped quartz rod. Constant contact between the quartz rod and the sample was maintained at 100g using a scale (Model 3120, Health O Meter, Bridgeview, IL, USA). Random noise was generated at the sensing piezo using a function generator (Stanford Research Systems, Sunnyvale, CA, USA). Data was acquired by connecting the receiving piezo to an external sound card (Model No. SB0490, Creative Labs) which had a frequency response of 20 to 22 kHz. An anti-aliasing filter prevented excitation from frequencies outside the range of the function generator. Scans were recorded for 10 seconds at a sample rate of 44.1 kHz. Each tablet was scanned in triplicate and in random order. The raw data consisted of 252 acoustic spectra.

NIR Data Collection

Near-infrared diffuse reflectance spectra of the tablets were collected from an InfraAlyzer 500 (Bran & Luebbe) spectrometer. The response range of the spectrometer was 1100-2500 nm. Tablets were randomly placed in a conical reflective cup, which was used to maximize diffuse reflectance [86]. The tablets were scanned in random order and randomly rotated in the cup between replicates to average variations in positioning. Each scan took approximately 2 minutes to record with a resolution of 2 nm.

Data Processing All data processing was done in Matlab 7.0.1 (The Mathworks Company, Natick, MA, USA). ARS data was transformed from the time domain to the frequency domain using the Fourier Transform. Each recorded ARS spectrum was the average of 9 scans. The averaged spectra were smoothed with a cubic spline operation. NIR data was averaged and multiplicative scatter correction was used to eliminate baseline differences [87]. Principal components were calculated from both

the AR and NIR spectra [61]. Multivariate analysis was used to determine the PCs that correlated most highly with D-tag concentrations. The BENDS algorithm was performed on the FTARS data using an interval method to find the 10 frequency band that correlated the best. The interval method set a window size of 500 data points and moved across the data in steps of 500. The best correlating section was kept and the window size was reduced to 200, then 100, and so on until a region of 10 frequencies was found. Leave-one-out cross validation was used to determine the effectiveness of each method (NIRS, PCR-ARS and BENDS-ARS) in determining D-tag concentrations.

Results and Discussion

NIR cross validation The NIR spectra were preprocessed by multiplicative scatter correction of the background subtracted spectra. The average corrected NIR spectra of the tablets are shown in figure 6.2. Because the tablets comprised many ingredients, it would not be expected to find a single wavelength calibration useful. Therefore, a multivariate regression and LOO cross validation was performed on PCs and D-tag concentrations. Two PCs comprising 92% of the sample variation were used in the calibration. The correlation ($r^2 = 0.996$) can be found in figure 6.3. The standard error of estimate (SEE) was 0.92% by weight and the SEP was 0.99% by weight.

FTARS PCA vs. BENDS The FTAR spectra were smoothed with a cubic spline operation. The mean smoothed spectra are shown in figure 6.4. Acoustic resonance data contains a high level of information; therefore principal components were calculated to reduce the spectral data to a more usable amount. 15 PCs comprising 60% of the sample variation were needed to validate the acoustic data, suggesting a nonlinear trend for acoustic velocities of the tablets. PCR and LOO cross validation to D-tag concentrations produced $r^2 = 0.974$ with SEE = 3.47% by weight and SEP = 4.78% by weight. The regression and cross validation statistics are represented in figure 6.5. The BENDS algorithm and LOO cross validation calculated on the region between 9.021 to 9.043 kHz (10 data point region) produced $r^2 = 0.994$ with SEE = 1.26% by weight and SEP = 1.34% by weight. Figure 6.6 represents the BENDS regression (left) and the BENDS LOO cross validation predictions (right).

Quantification of D-tag. Two spectroscopic methods, NIR and FTARS, have been studied for the quantification of D-tag in RSV tablets. A non linear and linear multivariate calibration method was used with FTARS. While all three methods (NIRS, PCA-FTARS, BENDS-FTARS) were able to successfully quantify D-tag, it is clear from the results that NIR spectroscopy is the superior method. However, the methods discussed here rely on different interactions. NIR spectroscopy is based on molecular vibrations, and although diffuse reflection does not follow Beer's law explicitly, the relationship between absorbed light and sample can be empirically linear. ARS is dependent on physical properties such as acoustic velocity, density, and sample mass. AR spectra are the result of acoustic interactions with the intrinsic sample properties, and thus may be extremely non linear in nature. ARS may be an alternative spectroscopic choice because of its ability to deeply penetrate many types of material; the idea being that pressure waves can propagate through many types of material when there is contact between them. ARS is limited by large changes in the acoustic spectrum caused by small environmental changes. Pressure to the sample must be held constant and monitored constantly. Tablets must be kept in air tight containers to avoid changes in humidity which can contribute to hydration.

The two data analysis techniques used with ARS (PCA and BENDS) were able to quantify D-tag to a reasonable quantity. The advantage clearly sides with BENDS due to better results and only needing a 10 variable dataset from the original spectra. There is clear evidence that the ARS data have a strong non-linear correlation due to the BENDS method performing superiorly to PCA method and very closely with NIRS.

Speed of method. NIR spectrometry has a much longer integration time than ARS (120-seconds compared to 250-milliseconds). While the NIR results are slightly better than that of ARS, the low cost and rapid scanning abilities of the ARS may make it a better choice where an NIR level of accuracy is not necessary. The incorporation of BENDS with ARS brings the results very close to NIRS but BENDS requires a longer analysis time to mine the data. Once the region was found however, the chemometric methods were similar in their processing time.

Conclusion

This work demonstrated that BENDS has comparable or superior results to that of standard PCR. The BENDS algorithm applied to ARS compares closely with NIR spectroscopy, and the speed and cost of ARS make it an alternative spectroscopic choice. BENDS only used a 10-frequency range, which attributes to the method's advantage over both the PCR and NIRS methods. The logical next step to these experiments is finding further applications of the ARS that pose non-linear problems and then formulating an ISP-BENDS algorithm. The increased throughput (Jacquinot advantage), shorter acquisition time (Fellgett advantage), and decrease in computational demand (ISP advantage) of ISP-ARS make it an even better choice. The prototype AR sensor used in this study was constructed from readily available parts and is inexpensive, suggesting utility as a PAT. The piezoelectric transducers used in the ARS are currently used to monitor various online processes and could be easily networked on the assembly line to monitor 100% of the tablets passing down the line.

Chapter six Figures

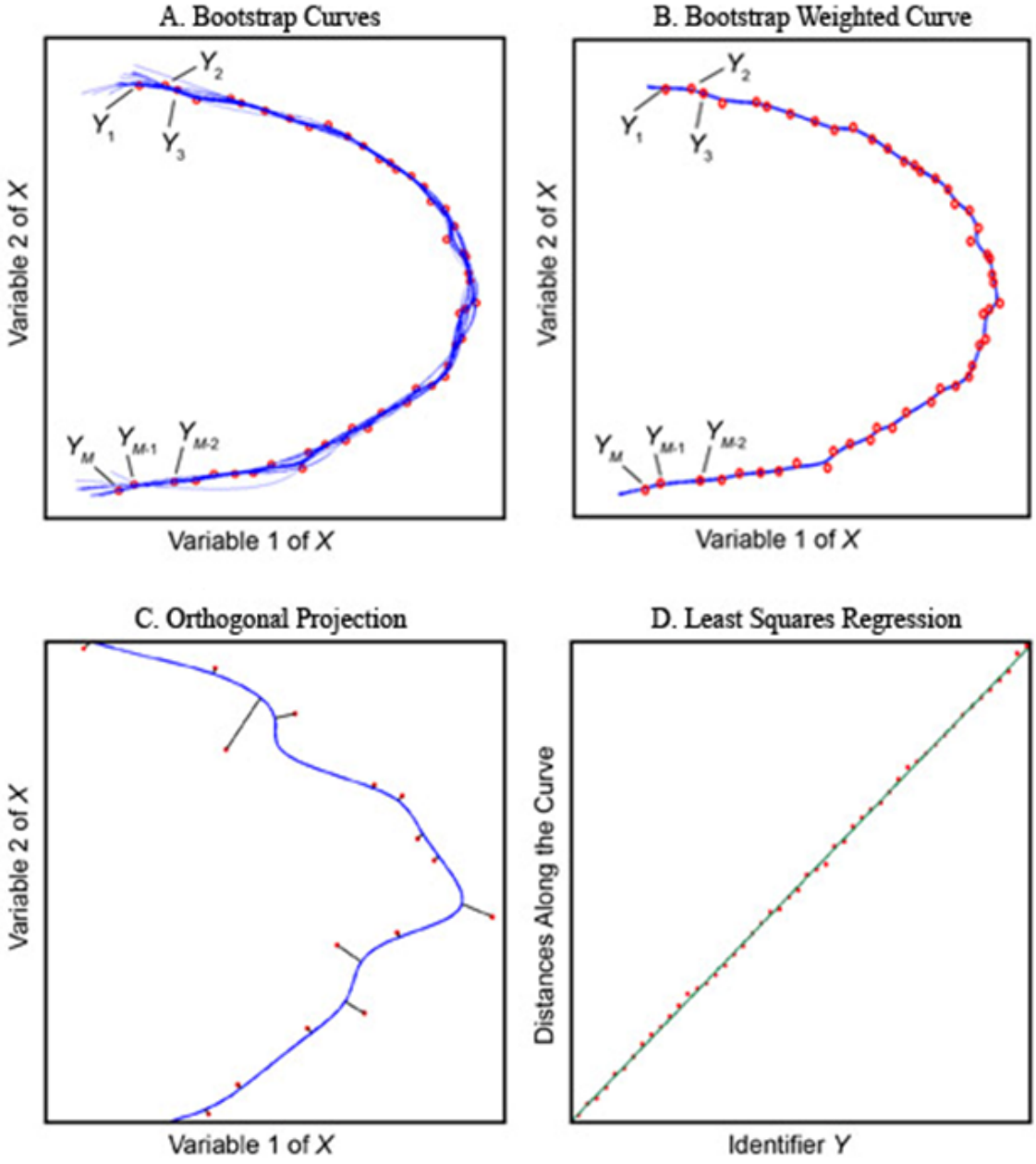


Figure 6.1. Visual representation of the BENDS algorithm.

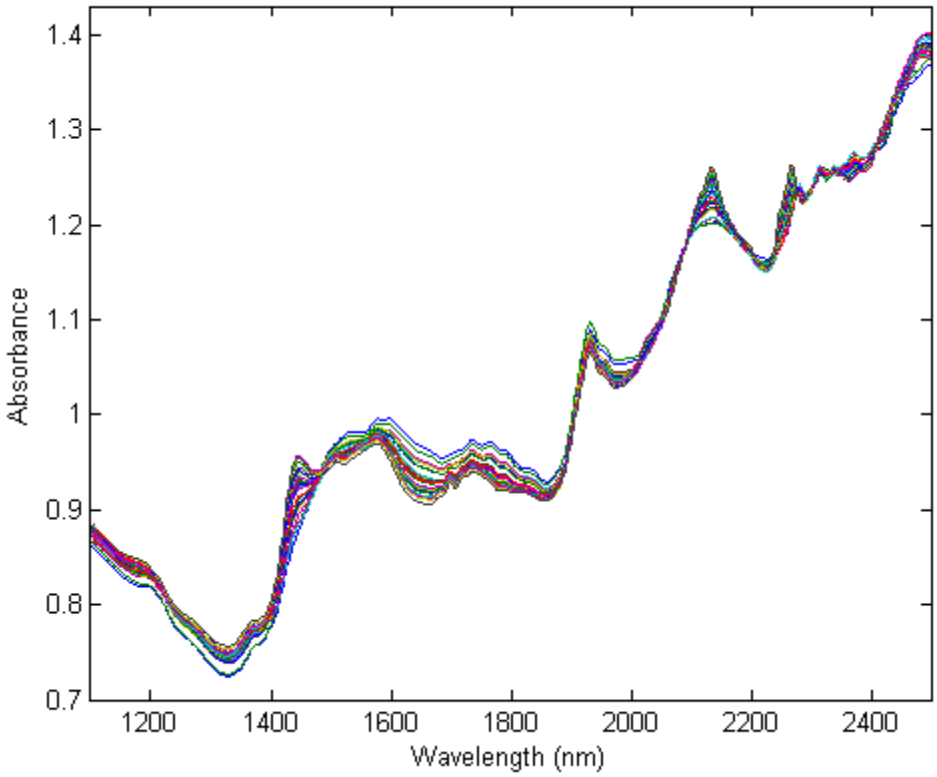


Figure 6.2. Mean corrected NIR spectra over the range of 1100 to 2500 nm

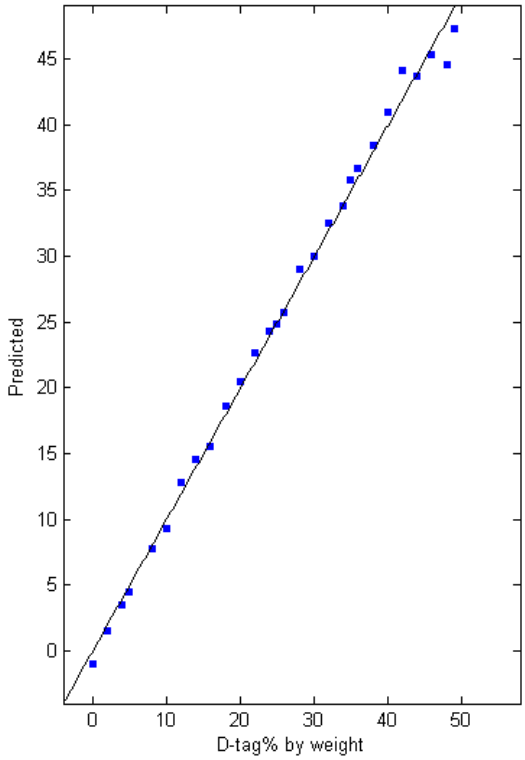


Figure 6.3. Plot of predicted D-tag concentrations versus the actual D-tag concentrations of all 28 samples with NIRS, r^2= 99.6, SEE = 0.92% by weight, and SEP = 0.99% by weight.

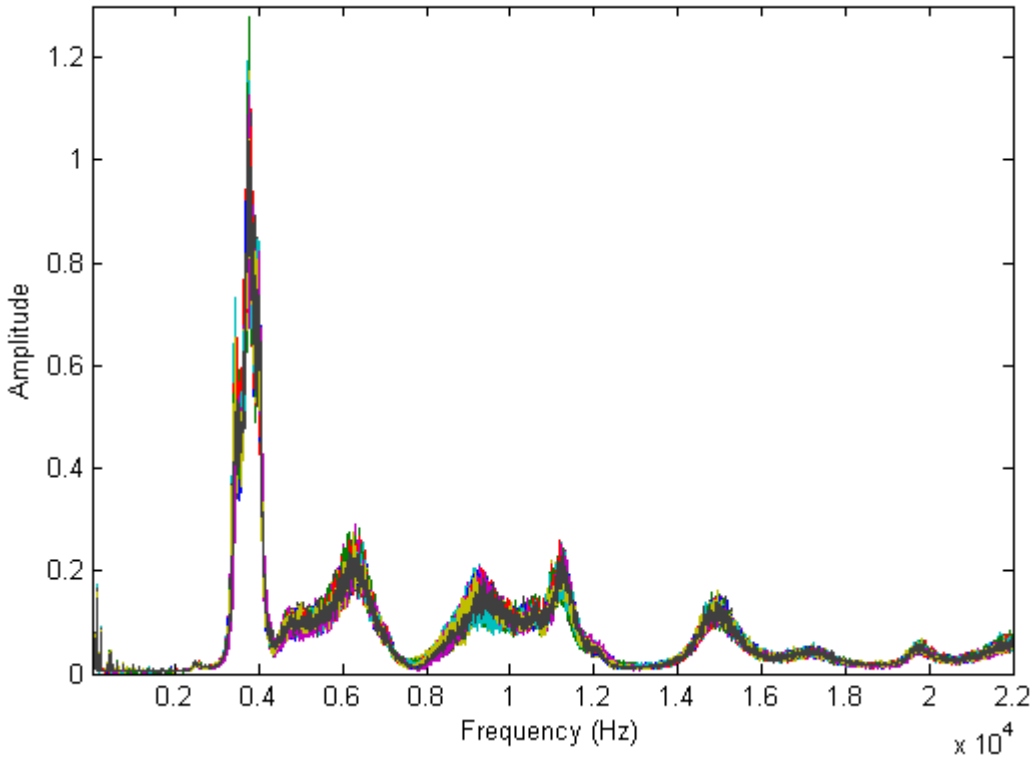


Figure 6.4. Mean smoothed FTARS spectra over the concentration range of 20 to 22000 Hz.

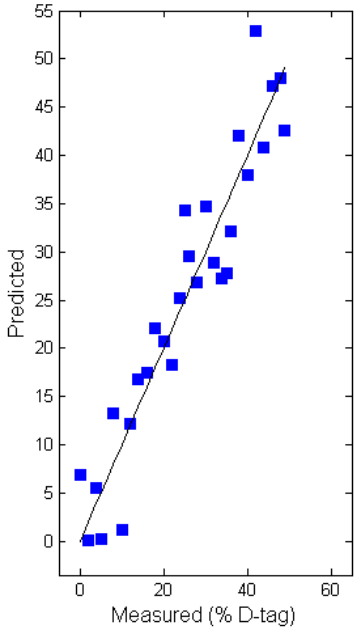


Figure 6.5. Plot of predicted D-tag concentrations versus the actual D-tag concentrations of all 28 samples with PCR-ARS, $r^2 = 0.974$, SEE = 3.47% by weight, and SEP = 4.78% by weight.

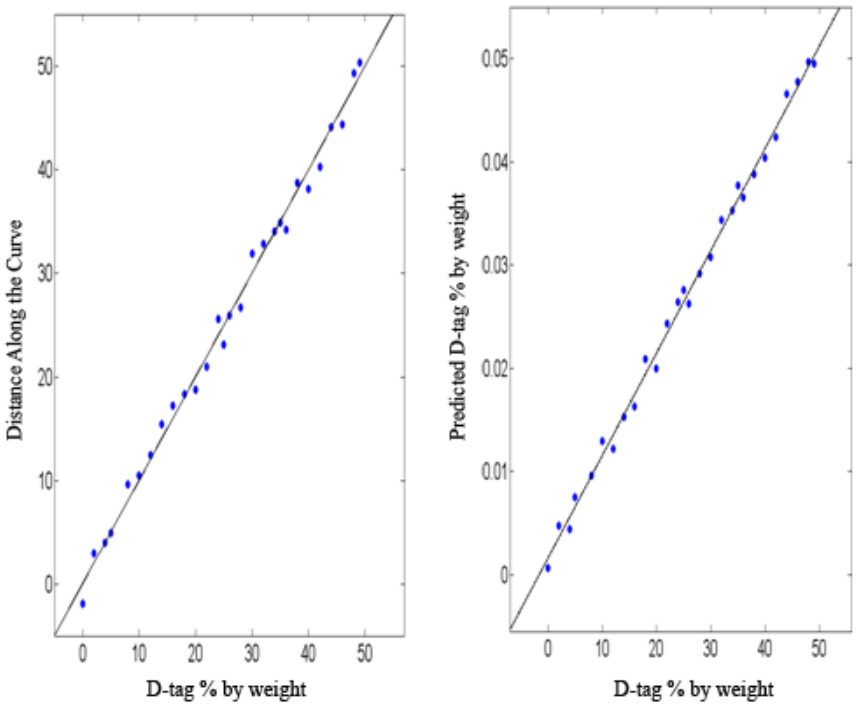


Figure 6.6. Regression of the distances along the curve versus D-tag concentrations (left) and a plot of predicted D-tag concentrations versus the actual D-tag concentrations of all 28 samples with BENDS-ARS, $r^2 = 0.994$, SEE = 1.26% by weight, and SEP = 1.34% by weight.

Copyright Statement

Copyright © AAPS PharmSciTech

Link, D.J.; Hannel, T.A.; Lodder, R.A. *AAPS PharmSciTech* (submitted 23 June 2009)

Chapter seven – ARS with BENDS to Quantify a Contaminant of D-Tagatose

Introduction

D-tag is a drug being tested in phase 3 clinical trials and is being tested as a novel treatment for type 2 diabetes. Type 2 diabetes is responsible for over 220,000 deaths in the U.S. each year [34]. Type 2 diabetes accounts for over 90% of all diabetes diagnosed in adults which amounts to over 19 million Americans [35]. The cost of type 2 diabetes has nearly doubled in the U.S. from 2001 ($6.7 billion) to 2007 ($12.5 billion) [17]. Type 2 diabetes starts as an insulin resistance, and gradually the cells no longer use insulin correctly. As the need for insulin increases, the pancreas progressively loses its ability to produce it. Although D-tag does not improve insulin production, it has been shown to lower glycemic response as well as induce weight loss in clinical trials through a mechanism of action based on several enzymes in the liver [87].

The pharmaceutical industry is built on standards that are enforced by the FDA and other governmental regulatory offices. The current system in detecting possible impurities is batch sample testing in which samples are taken from the total production and tested using time intensive, destructive and costly methods. The portion that is tested statistically represents the total production and therefore can be accepted as the purity for the entire batch. There are multiple recalls each year where a contaminant or other issue has caused a batch of impure product. This provides evidence that the current methods do not find the problems in time. In 2008, Palo Alto Labs issued a recall of Aspire 36 and Aspire Lite due to a contamination of Aildenafil in trace amounts and Dimethyl sildenafil thione (sulfoaildenafil) a purported analog of Sildenafil, an FDA-approved drug used as treatment for male Erectile Dysfunction (ED) [88].

The FDA openly reports that the cGMP are at their limits and better methods need to be explored in order to ensure product safety [37]. PATs are methods designed to monitor the purity of drugs at each step throughout the process [78]. PATs are designed to detect problems before they occur in order to prevent large recalls. NIRS and ARS are two methods under investigation to be used as PATs. NIRS and ARS are rapid, inexpensive and nondestructive quantitative methods.

ARS is a relatively new instrument but has been studied on several different mediums. ARS studies have included analysis of pharmaceutical tablets [9-11], solid fuel mixtures [12], semi-solids and colloidal dispersions [13-14], liquids [49-50][89], and powders[46-48]. ARS has been studied with BENDS to determine hydration levels of aspirin tablets [85] and quantification of D-tag levels in D-tag/Resveratrol tablets [90].

During the production of D-tag there is a possibility that the byproduct sugar, D-galactose (D-gal) will cause impurity issues. Levels of D-gal below 0.5% are acceptable, and therefore a rapid and nondestructive method could be very beneficial. NIRS and ARS are both possible outlets to study. NIRS has been pushed to the forefront of new PATs, and a comparison with ARS could give light to the advantages or disadvantages of each. A major issue is the nonlinear trends that occur in quantitative measurements with ARS. BENDS is a nonlinear MVC that will transform the nonlinear correlations of ARS into a reduced univariate linear calibration. PCR [81] can be used with NIRS but is not as effective with the highly nonlinear behavior with ARS.

The purpose of this study is to demonstrate the advantages and disadvantages of both NIRS with PCR and ARS with BENDS in the determination of D-gal contamination levels of D-tag powder mixtures. The two methods will be compared with the SEP and R-square statistics.

Theory

ARS is based on the principles of acoustic wave propagation, interferences and resonance. Sound waves are passed through a V shaped quartz rod that is in direct contact with the sample. The sound splits into two paths: one through the quartz rod and to the detector and one through the quartz rod, into the sample, back into the quartz rod and then to the detector. The two paths recombine after the point where the rod makes contact with the sample. The sample alters the sound due to changes in the acoustic velocity and amplitude. The recombination creates constructive and destructive interferences that yield a highly unique signal at the detector. The major drawback to ARS is the highly nonlinear correlations that occur in the spectra. The two paths can pass

in and out of phase multiple times throughout a calibration range. The phase changes cause the resulting amplitude to increase and decrease in a nonlinear manner.

BENDS is an algorithm designed to work with AR spectra to transform the highly nonlinear correlations into a reduced linear calibration. Figure 7.1 is an illustration of how the BENDS algorithm is applied to data. First, N bootstrap data sets are randomly selected from the original data where each bootstrap data set is the same size as the original data set. Utilizing the bootstrap method, an error is calculated in order to weight the spline to find the best possible manifold for the data. Bootstrapping is a statistical method for finding estimates of error for statistical measures that are easily defined [28]. The error estimates are found by calculating the standard deviation of the bootstrap curves which are calculated using an n-dimensional piecewise cubic spline for each bootstrap data set.

The standard deviations are normalized so that the greatest standard deviation is equal to one. The inverses of the normalized values are used as the weight values for the optimal nonlinear manifold in order to place the most value on those points with the smallest bootstrap error. Once the optimal nonlinear manifold is found, the data points or orthogonally projected onto the manifold. The distance from point zero (the minimum contaminant concentration) to each data point is calculated, and these distances along the curve are regressed against the known contaminant concentrations.

The resulting calibration is a reduced form of the original data similar to PCR and other MVC methods [15]. The orthogonal relationship between the original data and the reduced form is also similar to the PCR method; however, BENDS does not require a linear correlation [15].

Materials and Methods

Sample Preparation. D-tag and D-gal powders were donated by Spherix Inc. for all studies reported in this paper. D-tag/D-gal mixtures were prepared with D-gal concentrations ranging from 0 to 2% mass by mass. The powder mixtures were ground with mortar and pestle to create a homogenous mixture and batches were split into six samples of 300 mg each. The samples were scanned with both the NIRS and ARS.

Near Infrared Spectroscopy. Powders of both "clean" D-tag and D-gal were obtained via Spherix Incorporated and scanned with a dispersive NIR spectrometer (n = 15). The data was imported into Matlab and scatter corrected with a background. Figure 7.2 is a plot of the absorbance spectra for both D-tag and D-gal which gives evidence of clear differences in their absorbance. The mixtures were scanned with a dispersive NIR spectrometer in triplicate, resetting the sample and tray each time. The samples were scanned according to a randomized list which included six blank measurements which are the empty reflective base of the sample holder. Figure 7.3 is a plot of wavelength versus the mean NIR spectra of the different concentration groups after scatter correction.

The BENDS algorithm was applied to the smoothed AR spectra in an interval manner. The interval method is a common method in data mining for high order multivariate statistical measures [92]. A window size of five frequencies, moved in steps of five frequencies was selected. BENDS was performed on the first 10 frequencies recording the R-square statistic of the distance along the curve versus contamination and the SEP from a LOO cross validation. The window was moved to the right by five frequencies and BENDS was applied to the next five frequencies in line. The region of 10 frequencies that gave the highest R-square with the lowest SEP is represented as the predictability of the method.

Acoustic Resonance Spectroscopy. The quartz rod ARS was used to scan the powders in random order. Each sample was scanned in triplicate resetting the pressure scale before each scan. White noise was used for the excitation waveform and the signal at the receiving piezoelectric discs was recorded through a stereo jack. A sound card with the ability to handle 20 to 22050 Hz was used to record the signal onto a computer for both storage and processing.

Data Processing. The sound files stored from scanning were imported into Matlab 2008a (Mathworks Inc.) and were converted from a time based signal to a frequency based signal using a fast Fourier transform. BENDS was performed on the ARS data and PCR was performed on the NIRS data. The NIRS was first scatter corrected using a multiplicative method. PCA was performed on the non averaged scatter corrected data to make sure that the background measurements were separating and grouping in a different

region in hyperspace. Figure 7.4 is a plot of PC score 1 versus PC score 2 which contribute to the highest amount variance in the data. The blanks are in red while the D-tag/D-gal mixtures are in blue which clearly separate and background subtraction can be used to compensate for any drift.

Results and Discussion

NIRS with PCR. PCR was performed by first calculating new PC scores from the averaged group data. All PC scores contributing to 99 percent of the variance were regressed versus concentration of D-gal in milligrams. Correlating PC scores were selected from the t-stats of the regression. The predictability of the model was found by a LOO cross validation from the PCs yielding an SEP of 0.055% D-gal contamination (0.165 mg of D-gal in a 300 mg sample). Figure 7.5 is a plot of the actual percent D-gal versus the predicted percent D-gal in each sample. The diagonal is a line with slope one representing a perfectly predictive model.

The PCR model provided evidence supporting the hypothesis that NIRS could be used to determine the level of D-gal contamination of D-tag powder samples. The model however needed eleven principal components in order to validate. The drawback is how close the method is approaching an over fitting scenario. Interferences such as water and overtones may be a major factor.

ARS-BENDS. The spectra were slightly smoothed to increase signal to noise and are represented in figure 7.6. Interval BENDS was performed on the data yielding a region consisting of 10 frequencies between 10.13 to 10.24 kHz as the best fit area. The R-square from the least squares regression of the distances along the curve versus contamination level was 0.998 and SEP from the LOO cross validation was 0.047% w/w D-gal contamination (0.141 mg of D-gal in a 300 milligram sample). The linear transformation of BENDS is represented in figure 7.7 along with the LOO cross validation plot.

ARS vs NIRS. Both the NIRS and the ARS results detected the contamination levels of D-gal powders to an acceptable limit. ARS with BENDS predicted slightly better in the LOO cross validation with an SEP of 0.141 mg versus the NIRS SEP of 0.165 mg. Due

to how close the results were however, other criteria should be used to compare the two methods. The speed of data acquisition is terms of milliseconds for the ARS but is in seconds for the NIRS. The time required to mine the data with BENDS is substantially longer than the whole-spectrum approach with PCR for NIRS. Once the area of correlation is found in the AR spectra the processing time is comparable if not faster than the PCR with NIRS. Once a calibration model is set, the two methods take similar processing time to test new samples.

The major risk of using either ARS or NIRS is that the two methods are not stand alone technologies. The accuracy of their results is dependent upon the accuracy and precision of the reference method. NIRS and ARS may be more reproducible than its reference method but it still requires a separate calibration for each new constituent. The reference for the experiments reported in this paper was gravimetric analysis during the formulation of the powders. If there is an issue with the reference method, then the results of the NIRS and ARS methods are useless. In order to ensure the reliability of the calibrations, periodic verifications are needed. This is a major drawback for both methods explored in this report.

Versatility of ARS and NIRS. NIRS has been studied extensively for many different samples though ARS is starting to build a repertoire as well including the study of pharmaceutical tablets [9-11], solid fuel mixtures [12], semi-solids and colloidal dispersions [13-14], liquids [49-50][89], and powders [46-48]. In many ways, calibration transfer issues are simpler to solve in ARS than in other methods like NIRS. ARS does not exhibit baseline variations like NIR spectra, and normalization techniques like multiplicative scatter correction have no effect of the appearance of ARS spectra. The mechanical signature of the apparatus changes with instrument design, of course, but in the acoustic region covered by an MP3 player, minor variations in mechanical construction of nominally identical instruments are very much smaller than the wavelength of the sound. Nevertheless, the true extent of calibration transferability remains to be determined in actual industrial practice.

The ideal analytical method and instrument would not disturb the sample under analysis. Little or no sample preparation would be required to use the technique and very small

sample sizes could be employed. The sample could be used for its originally intended purpose following the analysis, or it can be examined using another analytical technique. ARS and NIRS both fit these criteria and the latter argument regarding reusing samples with other instrumentation was used in these experiments. The same samples were scanned with both NIRS and ARS multiple times with no noticeable change the sample.

One of the most important features of the ARS that could make a preferred choice over NIRS is money. The components involved in the ARS are extremely cheap with the both the source and detectors costing only pennies. The same cannot be said about NIRS instrumentation though it is not as expensive as other alternatives. ARS sensors could be mass produced for a fraction of a penny a piece to make a system of sensors. Money is always a serious deciding factor no matter what field the instruments and sensors may be used for.

Speed of Method. The data collection process itself is very rapid for both ARS and NIRS. Both methods rely on intensive "back-end" processing dealing with chemometrics and data processing. No sample preparation is required but the time for the NIRS to scan each sample was over a minute while ARS scans were acquired for three seconds. It should be noted that FT-NIRS instruments are also able to perform very rapid scans so both NIRS and ARS are similar in data acquisition. NIRS requires an extra data processing step besides the MVC which is multiplicative scatter correction. Though this calculation is minor it should be noted that it does add to the complexity of the calibration model.

BENDS and PCR do vary quite significantly in the time required to calculate the calibration model. BENDS has many more steps for both data mining and can take hours to create a valid predictive model. PCR can be completed in minutes. Once the model is created and an unknown is to be quantified, BENDS is more rapid. For PCR the entire new spectral information must be transformed with PCA with the calibration data and then fit to a least squares regression. BENDS only requires the few frequencies selected during data mining and calculates its distance along the curve resulting in the quantification. BENDS may require more time in the development of the calibration but for each scan after it requires less computation and time for quantification.

Conclusion

NIRS and ARS have the abilities to be viable PATs due to their speed in data acquisition and ease of tailoring excitations. Further studies with NIRS using a filter system (molecular factor computing) and with ARS using tailored excitation (integrated sensing and processing) should be explored to create more robust, rapid and application specific devices. As the need to better the cGMP methods in the pharmaceutical industry grows, more specialized devices will be needed. Scanning nearly 100% of the sample at nearly 100% of the process is the only way to guarantee no recalls, and both NIRS and ARS have shown potential for this application.

Chapter seven Figures

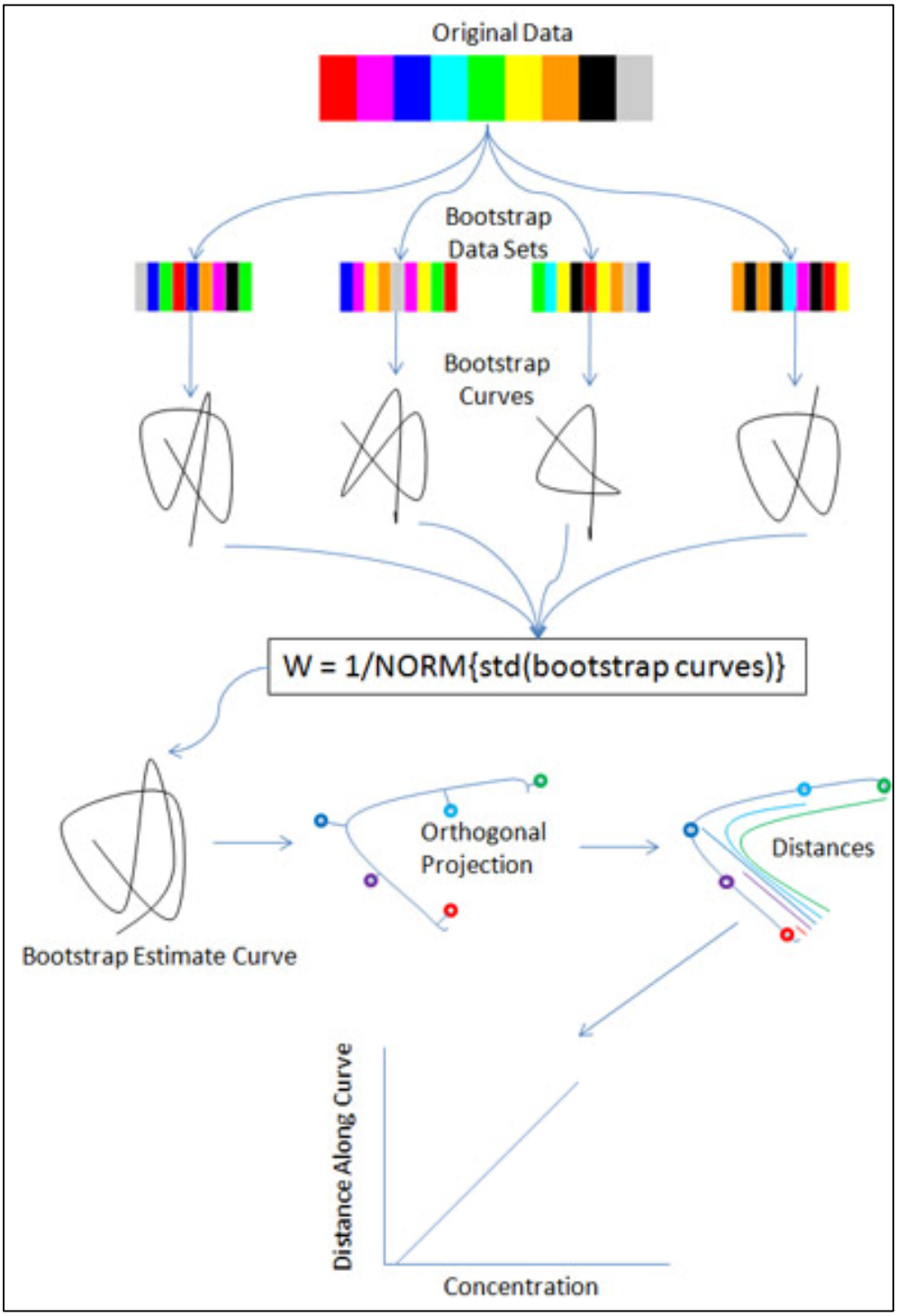


Figure 7.1 Flow-chart describing the BENDS algorithm.

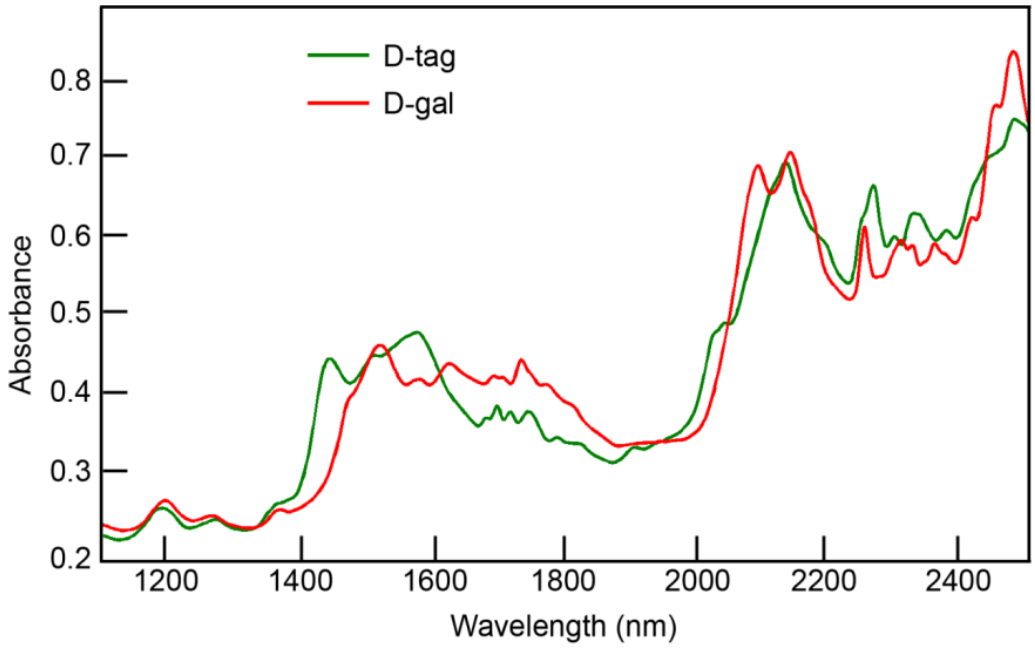


Figure 7.2 NIRS Spectra. NIRS spectra of D-tag and D-gal after scatter correction.

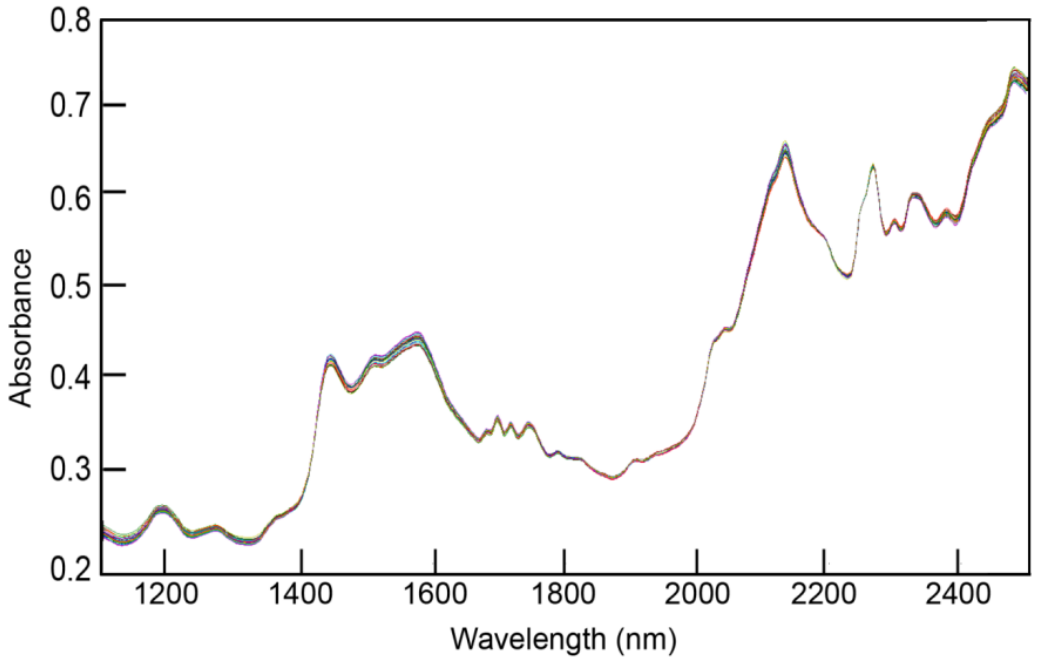


Figure 7.3 NIR Spectra. Mean NIR spectra of the different sample groups of varying D-tag contamination levels.

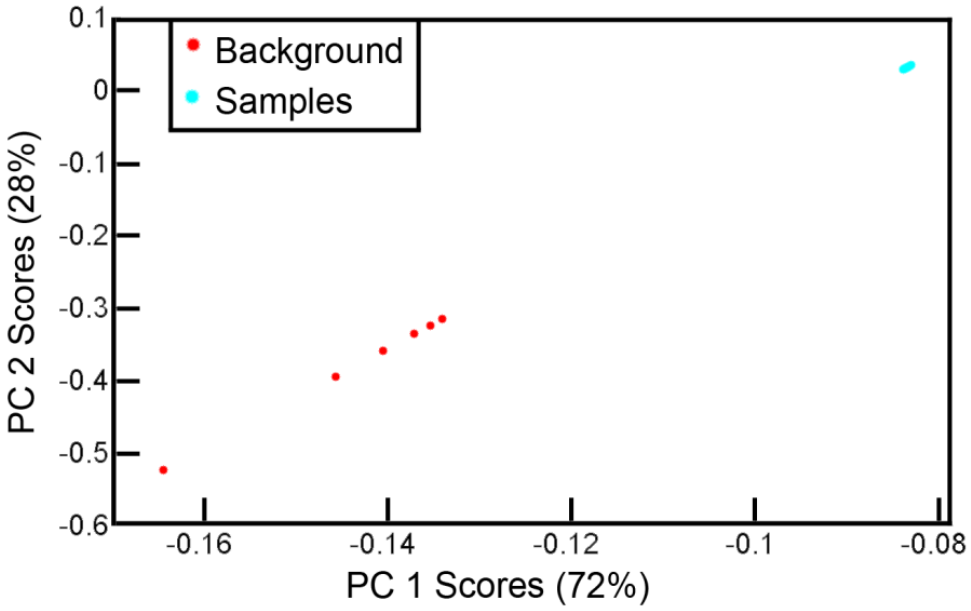


Figure 7.4 NIRS PCA. PC score plot illustrating the separation of the background sample holder from the samples.

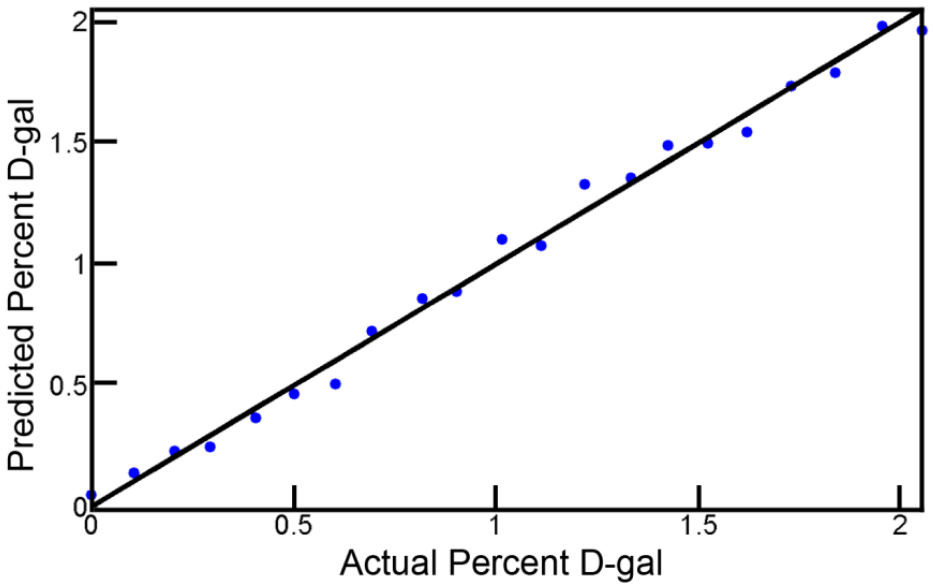


Figure 7.5 NIRS Cross Validation. Actual percent D-gal levels versus the predicted percent D-gal levels by a LOO cross validation of the PC scores.

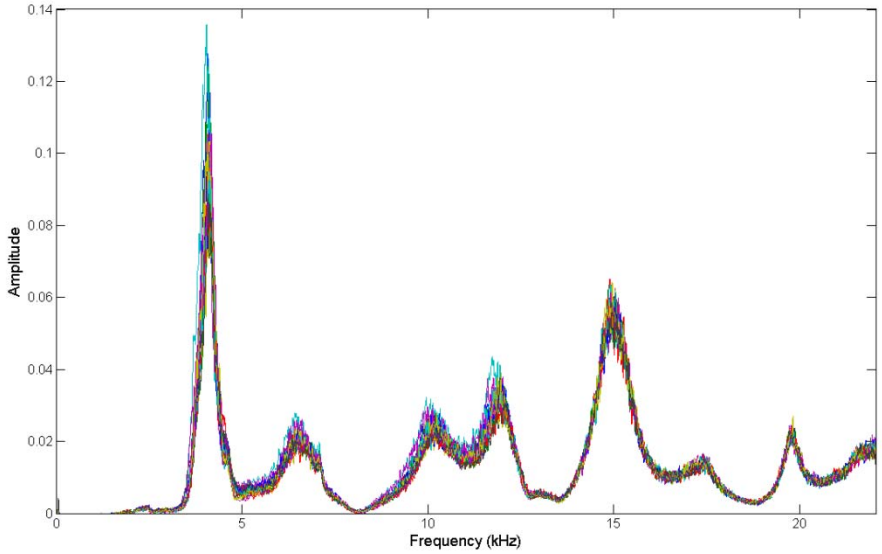


Figure 7.6 AR Spectra. Mean spectra from the ARS experiments of varying D-gal contamination in D-tag powders.

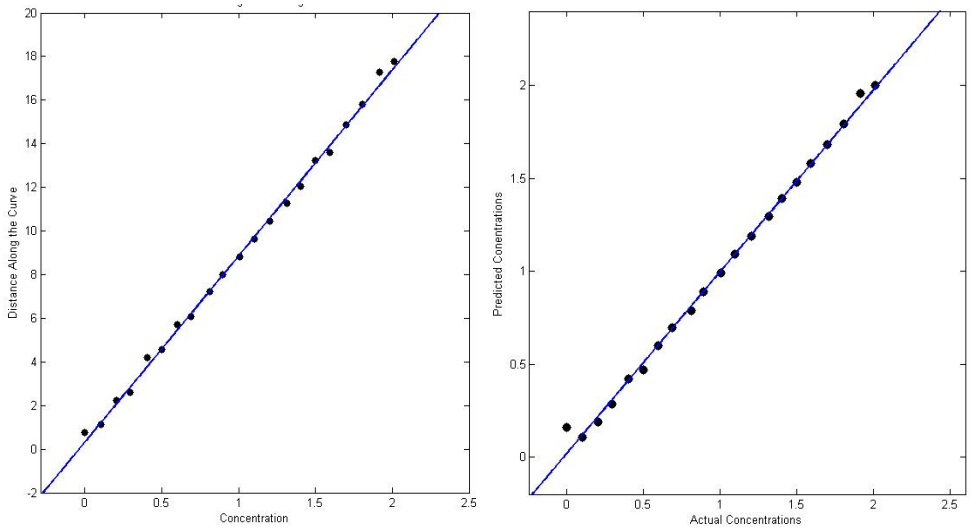


Figure 7.7 BENDS Results. ARS with BENDS results with a LOO cross validation. Least squares regression of the distances along the curve versus D-gal contamination in percent mass by mass (left) and LOO cross validation predicted values of D-gal contamination in percent mass by mass and the actual D-gal contamination in percent mass by mass (right).

Copyright Statement

Copyright © David Link 2009

SECTION FOUR - CONCLUSION OF DISSERTATION

The utility of an analytical method can be determined by examining many different factors. ARS with BENDS was the main instrumentation and chemometric method used in the studies presented in this dissertation, and ARS/BENDS should be evaluated according to all the different factors whether they are positive or negative. A wide dynamic range with the ability to determine concentrations from 10^{-18} to 10^{2} M (though not likely) would be a valuable asset on the resume of any method. Low detection limits and high sensitivity are important, but ideally a method should be able to count single atoms or molecules of a substance as well as determine concentrations of the substance ranging up to 100%. Analyses would be greatly simplified if every analytical signal were linear over this entire range. ARS does have a wide dynamic range as presented with previous work discussed in the introduction and also the experiments reported in this document. The linearity of the method has been a major limitation of the instrument; however, the work done with BENDS shows an ability to eliminate the nonlinear functional relationship between spectroscopic signal and sample property. ARS with BENDS may not be able to detect a single molecule or atom, but for the applications presented, the method was able to perform to current relevant standards.

Versatility and flexibility are two the most important factors because new applications and chemicals are being created every day. Virtually every conceivable analyte and property could be determined using the ideal analytical instrument. Furthermore, these analytes and properties could be determined simultaneously and in any combination from the same sample of a substance. Any sort of sample, solid, liquid, or gas, could be directly analyzed with the instrument. The ideal analytical instrument would be the only instrument in every laboratory. ARS has been shown to work in many different media and phases, and the incorporation of BENDS has given the method the ability to perform qualitatively as well as quantitatively. The ideal instrument is simple, rugged, and easy to maintain, which all apply to the ARS instrument consisting of basic electronics and a quartz rod. Anyone could potentially operate or repair the ARS with little or no training. The results of the ARS are reproducible and have been validated over time, with different operators, different equipment instruments, and different chemical lots. The device could act as a virtual "black box" that always produces the correct result regardless of the skill of the operator.

The highest selectivity or free freedom from interferences means that no characteristic of the sample or the environment will interfere with the measurement of any analyte or sample property. Furthermore, as an instrument approaches the "ideal analytical black box" it will need more and more "false-sample" detection capability; that is, the instrument will need to be able to recognize that it is examining a sample unlike any it has ever examined before, and will need to be able to respond appropriately. Every instrument including the ARS has the problem of interferences, though some instruments have more than others. The ARS has a similar interference profile to NIRS, including water. False-sample analysis could remove this problem by providing an automatic response to unknown samples. This response could take the form of a request for operator assistance, more samples of the same type, and a "second opinion" analysis by another technique, or a library search for the best step to take.

The ARS with BENDs method has been shown to be noninvasive and nondestructive. The method and instrument do not disturb the sample under analysis and little or no sample preparation is required to use the technique. Small sample sizes can be used and those samples are not harmed during analysis. The sample can be used for its originally intended purpose following the analysis or it can be examined using another analytical technique. The nondestructive nature of ARS/BENDS was demonstrated in analysis of D-tagatose and resveratrol tablets and D-galactose contamination of D-tagatose. Because the method is rapid and nondestructive, environmental effects on the sample such as hydration can be avoided keeping the sample the same as when it entered the instrument.

Probably the clearest advantage that ARS with BENDS has over other techniques is the extremely low cost of the method. The sensor and detector are very cheap, costing only pennies. Acoustic electronics are a very large industry because of the popularity of music. With large companies devising ways to make acoustic technology like MP3 players as cheap as possible, the ideal ARS would be built from an MP3 player with all the chemometrics (like BENDS) incorporated in the processing of the unit. BENDS has been shown to create a highly simplified and reduced linear calibration model that most modern cell phones and MP3 players could easily interpret. Ear-bud headphones are

made with piezo electric devices like the current ARS, and therefore the portability of the ARS is easily realized.

The research presented in this dissertation has solved a major issue for the ARS; the issue of highly nonlinear signals. Now that the calibration models are highly simplified and easy to work with, the ARS has the potential of being implemented in extremely small multi-sensor systems or highly portable devices that could be attached to most modern cell phones or MP3 players. All known calibration models spectra could be stored in online database ready for download to anyone's cell phone, or the scanned information could be sent to the server for analysis. In the pharmaceutical industry, the entire production line could be networked with AR sensors connecting wirelessly to a control center. Any detection of something incorrect could easily be detected and fixed, or the process controlled before the problem spread to the rest of the batch or even the intended recipient of the drug.

Future work in the area of AR instrumentation should investigate even simpler analysis methods such as integrated sensing and processing (ISP). BENDS might be coupled with ISP so that voltage at the detector is directly proportional to the analyte concentration. An ISP waveform could be created from the nonlinear manifold found during calibration where the waveform acoustically traces the nonlinear curve. In this scenario, no data analysis is required accept for comparison of the voltage to a table of values. Data acquisition and analysis would then be possible in the time frame of milliseconds, allowing for a rapid, nondestructive whole-system scanning method for the pharmaceutical industry.

Copyright Statement

Copyright © David Link 2009

APPENDIX A – LIST OF ABBREVIATIONS

Abbreviation	**Meaning**
APC	All Possible Combinations
API	Active Pharmaceutical Ingredients
ARS	Acoustic Resonance Spectroscopy
BENDS	Bootstrap Enhanced N-dimensional Deformation of Space
BEST	Bootstrap Error-adjusted Single-sample Technique
cGMP	Current Good Manufacturing Processes
CPS	Cyber Physical System
CV	Canonical Variables
DDDAS	Dynamic Data-Driven Application System
D-gal	D-galactose
D-tag	D-tagatose
ED	Erectile Dysfunction
FFT	Fast Fourier Transform
FTARS	Fourier Transform Acoustic Resonance Spectroscopy
iBENDS	Interval Bootstrap Enhanced N-dimensional Deformation of Space
ISP	Integrated Sensing and Processing
ISP-ARS	Integrated Sensing and Processing Acoustic Resonance Spectroscopy
LOO	Leave One Out
LS	Least Squares
MFC	Molecular Factor Computing
MSD	Multidimensional Standard Deviation
MVC	Multivariate Calibration
NIRS	Near Infrared Spectroscopy
OED	Oxford English Dictionary
PAT	Process Analytical Technology
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares
RSV	3,5,4'-trihydroxystilbene (Resveratrol)
SEE	Standard Error of Estimate
SEP	Standard Error of Prediction
SVD	Single Value Decomposition

REFERENCES

[1] "sound, n.3" The Oxford English Dictionary. 2nd ed. 1989. OED Online. Oxford University Press. 4 April 2007 <http://dictionary.oed.com/cgi/entry/50231531>

[2] Qiong Liua, et. al. "Acoustic velocity measurements on Na2OTiO2SiO2 liquids: Evidence for a highly compressible TiO2 component related to five-coordinated Ti. Geochimica et Cosmochimica Acta 2007, 71, 4314–4326.

[3] Baldwin, Steven, et. al. "Measurements of the anisotropy of ultrasonic attenuation in freshly excised myocardium." The Journal of the Acoustical Society of America 2006, 119, 3130–3139.

[4] Umnova, Olga, et. al. "Deduction of tortuosity and porosity from acoustic reflection and transmission measurements on thick samples of rigid-porous materials." Applied Acoustics 2005, 66, 607–624.

[5] Masudaa, Koji, et. al. "Detailed analysis of acoustic emission activity during catastrophic fracture of faults in rock. Journal of Structural Geology 2004, 26, 247–258.

[6] Kunkler-Peck, A. et. al. "Hearing shape." Journal of Experimental Psychology: Human Perception and Performance 2000, 279–294.

[7] Gordon, M. S.; Rosenblum, L. D. "Perception of sound-obstructing surfaces using body-scaled judgments." Ecological Psychology 2004, 16, 87–113.

[8] Lai, B. L. et. al. "Ultrasonic resonance spectroscopic analysis of liquids." Applied Spectroscopy 1988, 42, 381–529.

[9] Buice, R.; Pinkston, P.; Lodder, R. A. J Appl Spectrosc 1994, 48, 517–524.

[10] Hannel, T.; Link, D. J.; Lodder, R. A. "Integrated Sensing and Processing Acoustic Resonance Spectrometry (ISP-ARS) in Differentiating D-Tagatose and Other Toll Manufactured Drugs." J Pharm Innov 2008, 3, 152–160.

[11] Medendorp, J.; Fackler, J.; Douglas, C.; Lodder, R. J "Integrated Sensing and Processing Acoustic Resonance Spectrometry (ISP-ARS) for Sample Classification." J Pharm Innov 2007, 2, 125–134.

[12] Medendorp, J.; Lodder, R. A. "Acoustic-Resonance Spectrometry as a Process Analytical Technology for Rapid and Accurate Tablet Identification." PharmSciTech 2006, 7, 125–134.

[13] Mills, T.; Jones, A.; Lodder, R. A. "Identification of Wood Species by Acoustic-Resonance Spectrometry Using Multivariate Subpopulation Analysis" Applied Spectroscopy 1993, 47, 1880–1886.

[14] DiGregorio, B. E. "All you need is sound." Analytical Chemistry 2007, 79, 7236.

[15] Johnson, R. A.; Wichern, D. W., Applied Multivariate Statistical Analysis. Prentice Hall: Upper Saddle River, New Jersey, 1998.

[16] Dadhe, K., Nonlinear Calibration for Near-Infrared Spectroscopy. Chemical Engineering & Technology 2004, 27, (9), 946-950.

[17] Thissen, U.; Ustun, B.; Melssen, W.; Buydens, L., Multivariate Calibration with Least-Squares Support Vector Machines. Analyical Chemistry 2004, 76, (11), 3099-3105.

[18] Scholkopf, B.; Smola, A., Learning with Kernels. MIT Press: Cambridge, MA, 2002.

[19] Suykens, J.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J., Least Squares Support Vector Machines. World Scientific: Singapore, 2002.

[20] Hastie, T. a. S., W., Principal Curves. Journal of the American Statistical Association 1989, 84, (406), 502-516.

[21] Banfield, J. D. a. R., A. E., Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. Journal of the American Statistical Association 1992, 87, (417), 7-16.

[22] Raftery, D. S. a. A. E., Principal curve clustering with noise. IEEE Transactions on Pattern Analysis and Machine Intelligence 2001.

[23] De'ath, G., Principal curves: a new technique for indirect and direct gradient analysis. Ecology 1999, 80, (7), 2237-2253.

[24] T. Hermann, P. M., and H. Ritter, Principal curve sonification. International Conference on Auditory Display 2000, 81-86.

[25] Kégl, B., Principal Curves. In University of Montreal.

[26] de Boor, C., A Practical Guide to Splines. Springer-Verlag New York Inc.: New york, 1978.

[27] Asuncion, A. N., D.J. Pima Indians Diabetes Data Set Periodical, 2007. <ttp://www.ics.uci.edu/~mlearn/MLRepository.html> (05 June 07).

[28] Efron, B., Am. Stat. 1986, 40, 1.

[29] Efron, B.; Tibshirani, R., Statistical Data Analysis in the Computer Age. Science 1991, 253, 390-395.

[30] "Type 2 Diabetes" American Heart Association, http://www.americanheart.org/presenter.jhtml?identifier=3044759, September 3, 2007

[31] National Institute of Health, http://www.nih.gov/about/researchresultsforthepublic/Type2Diabetes.pdf, November 4, 2007

[32] "Diabetes" WHO, http://www.who.int/diabetes/facts/en/, September 3, 2007

[33] Hogan P, Dall T, Nikolov P. American Diabetes Association. Economic costs of diabetes in the U.S. in 2002. Diabetes Care. 2003; 26, 917–932.

[34] Moore MC. Curr Opin Invest Drugs. 2006; 7(10):924–5.

[35] Donner TW, Wilber JF, Ostrowski D. Diabet, Obes Metab. 1999; 1:285–21.

[36] Harris, A. Associated Content, 11 Nov 2006. <http://www.associatedcontent.com/article/86060/generic_acetaminophen_recall_by_perrigo.html> (September 3, 2007)

[37] Woodcock, Janet. US Food and Drug Administration. http://www.fda.gov/ohrms/dockets/ac/02/briefing/3869B1_08_woodcock/sld001.htm. 9/12/2007

[38] M. Parashar et al., Towards Dynamic Data-Driven Management of the Ruby Gulch Waste Repository, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2006, 3993, 384 - 392.

[39] NSF, January 2006 DDDAS Workshop Report, http://www.nsf.gov/cise/cns/dddas/2006_Workshop/index.jsp, November 4, 2007

[40] 403rd Wing Public Affairs, http://www.403wg.afrc.af.mil/news/story.asp?id=123066690, Novermber 4, 2007

[41] Medendorp, J.P.; Lodder, R.A. AAPS PharmSciTech. 2006, 7(1), Article 25

[42] Buice R, Pinkston P, Lodder R. Optimization of acoustic-resonance spectrometry for analysis of intact tablets and prediction of dissolution rate. Appl Spectrosc. 1994;48:517-524. DOI: 10.1366/000370294775268929

[43] Medendorp J, Lodder RA. Integrated Sensing and Processing and a Novel Acoustic-Resonance Spectrometer. Baltimore, MD: American Association of Pharmaceutical Sciences; 2004.

[44] Medendorp J, Lodder RA. Acoustic Resonance Spectrometry and Analysis of Powder Drying. Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy; March 7-12, 2004; Chicago, IL.

[45] Medendorp J, Lodder RA. Applications of Integrated Sensing and Processing (ISP) in Acoustic and Optical Sceptroscopy. Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy; February 27-March 4, 2005; Orlando, FL.

[46] Serris E, Camby-Perier L, Thomas G, Desfontaines M, Fantozzi G. Acoustic emission of pharmaceutical powders during compaction. Powder Technol. 2002;128:296-299. DOI: 10.1016/S0032-5910(02)00174-2

[47] Reynaud P, Dubois J, Rouby D, Fantozzi G. Acoustic emission monitoring of uniaxial pressing of ceramic powders. Ceramics Int. 1992;18:391-397. DOI: 10.1016/0272-8842(92) 90071-K

[48] Martin L, Poret J, Danon A, Rosen M. Effect of adsorbed water on the ultrasonic velocity in alumina powder compacts. Mater Sci Eng. 1998;252:27-35. DOI: 10.1016/S0921-5093(98)00669-8

[49] Kaatze U, Wehrmann B, Pottel R. Acoustical absorption spectroscopy of liquids between 0.15 and 3000 MHz, I: high resolution ultrasonic resonator method. J Phys E Sci Instrum. 1987;20:1025-1030. DOI: 10.1088/0022-3735/20/8/014

[50] Bolotnikov M, Neruchev Y. Speed of sound of hexane + 1-chlorohexane, hexane + 1- iodohexane, and 1-chlorohexane + 1-iodohexane at saturation condition. J Chem Eng Data. 2003;48:411-415. DOI: 10.1021/je0256129

[51] L. E. Rodriguez-Saona, F. M. Khambaty, F. S. Fry, and E. M. Calvey, Rapid Detection and Identification of Bacterial Strains By Fourier Transform Near-Infrared Spectroscopy, J. Agric. Food Chem. 2001, 49, 574-579

[52] Iola F. Duarte, Antonio Barros, Claudia Almeida, Manfred Spraul, and Ana M. Gil, Multivariate Analysis of NMR and FTIR Data as a Potential Tool for the Quality Control of Beer, J. Agric. Food Chem. 2004, 52, 1031-1038

[53] Miller, James N., Miller, Jane C., Statistics and Chemometrics for Analytical Chemistry 4th ed., Person Education Limited. 2000, 217-221

[54] Lodder, R. Hieftje, G. Quantile BEAST Attacks the False-Sample Problem in Near Infrared Reflectance Analysis. Applied Spectroscopy 42: 8, p1500-1512, 1988.

[55] Medendorp, J.P.; Fackler, J.A; Douglas, C.C; Lodder, R.A. J Pharm Innov 2007; 2:125-134.

[56] Linda T. Kohn, Janet M. Corrigan, and Molla S. Donaldson (eds.), "To Err is Human", Journal, National Academies Press, Washington, 2000.

[57] M. Parashar, et Al. *Computational Science - ICCS 2006: 6th International Conference*, Reading, UK, May 28-31, 2006, Proceedings, Part III, Vassil N. Alexandrov, Geert Dick van Albada, Peter M.A. Sloot, Jack J. Dongarra (eds.), Lecture Notes in *Computer Science* series, Springer-Verlag Heidelberg, **2006**; 3993:384-392.

[58] "Recall – Firm Press Release." FDA.org. 02 May 2006. <http://www.fda.gov/oc/po/firmrecalls/ivax05_06.html>

[59] Cutnell, J., Johnson, K. Physics. Wiley, New York, 1997.

[60] Jolliffe, I. T. Principal Component Analysis. Springer, New York, 2002.

[61] Leardi, R.; Norgaard, L. J Chemometrics 2004, 18, 486–497.

[62] Noord, O. E. Analytical Chemistry 1996, 68, 3851–3858.

[63] Blanco, M.; Coello, J.; Iturriaga, H.; Maspoch, S.; Pages, J. Chemometrics and Intelligent Laboratory Systems 2000, 50, 75–82.

[64] Seasholtz, M.; Wang, Z.; Kowalski, B.; Lee, S.; Hold, B. Analytical Chemistry 1993, 65, 835–845.

[65] "Principal Curves." University of Montreal. June 12, 2008. <http://www.iro.umontreal.ca/~kegl/research/pcurves/>

[67] "About NIH Obesity Research." National Institute of Health. <http://www.obesityresearch.nih.gov/About/about.htm> (14 September 2008)

[68] "Metabolic Syndrome." American Heart Association. <http://www.americanheart.org/presenter.jhtml?identifier=4756> (15 June 2009)

[69] "Leading Causes of Death." Centers for Disease Control and Prevention. <http://www.cdc.gov/nchs/fastats/lcod.htm> (April 11, 2008).

[70] Joseph A. Baur, et.al. Nature, 2006; 444.

[71] Jang M, Cai L, Udeani GO, Slowing KV, Thomas CF, Beecher CW, Fong HH, Farnsworth NR, Kinghorn AD, Methta RG, Moon RC, and Pezzuto JM. Science. 1997; 275: 218–220.

[72] Jang DS, Kang BS, Ryu SY, Chang IM, Min KR, and Kim Y. Biochem Pharmacol. 1999; 57: 705–712, 1999.

[73] Rotondo S, Rajtar G, Manarinis S, Celardo A, Rotillio D, de Gaetano G, Evangelista V, and Cerletti C. Br J Pharmacol 1998; 123: 1691–1699.

[74] Bertelli AA, Giovannini L, Giannessi D, Migliori M, Bernini W, and Fregoni M. Int J Tissue React 1995; 17: 1–3.

[75] Hui-Chen Su, Li-Man Hung and Jan-Kan Chen. Am J Physiol Endocrinol Metab 2006; 290:1339-1346.

[76] Satheesh, M.A.; Pari, L. Journal of Applied Biomedicine. 2008; 6(1):1-14.

[77] Police, S. B.; Harris, J. C.; Lodder, R. A.; Cassis, L. A. Obesity. 2008. doi:10.1038/oby.2008.508.

[78] "Process Analytical Technology." FDA.gov 06 February 2008. <http://www.fda.gov/AboutFDA/CentersOffices/CDER/ucm088828.htm> (27 August 2008)

[79] Robert P. Cogdill, Carl A. Anderson, and James K. Drennen, III. Spectroscopy. 2004; 19(12):104-109.

[80] Corti P, Ceramelli G, Dreassi E, Mattii S. Analyst. 1999; 124:755–758.

[81] Wang Q, DeJesus S. J Near Infrared Spectroscopy. 1998; 6:A223–A226.

[82] Trafford AD, Jee RD, Moffat AC, Graham P. Analyst. 1999; 124:163–167.

[83] Lodder, R. A.; Hieftje, G. M. Appl. Spectroscopy. 1988; 42(4), 556-558.

[84] Medendorp J; Lodder R.A. J Pharm Innov. 2006; 54-61.

[85] Link, D; Hannel, T; Lodder, R.A. Algorithms. 2009 (submitted 27 May 2009).

[86] Lodder R, Selby M, Nieftje G. Anal. Chem. 1987; 59:1921-1930.

[87] Wulfert, F.; Kok, W. T.; Smilde, A. K., Anal. Chem. 1998, 70, 1761-1767.

[88] "Recall -- Firm Press Release." U.S. Food and Drug Administration. 28 February 2008. <http://www.fda.gov/Safety/Recalls/ArchiveRecalls/2008/ucm112386.htm> (01 June 2009)

[89] Leveque, G.; Ferrandis, J.; Van Est, J.; Cros, B. Rev. Sci Instrum. 2000, 71(3), 1433-1440.

[90] Link, D.J.; Hannel, T.S.; Lodder, R.A. AAPS PharmSciTech. 2009 (submitted 23 June 2009)

[91] Maronna, R. Technometrics. 2005, 47(3), 264-273.

[92] Pei, Lei; et. Al. Energy Fuels. 2008, 22(2), 1059-1072.

VITA

DAVID JOHN LINK

Born: 08 April 1984

Butler, Pennsylvania

EDUCATION

- Berea College, Bachelor of Arts in Chemistry, May 2006

PROFESSIONAL

- Student Employee, University of Kentucky, 2008 – 2009
- Research Assistant, University of Kentucky, 2007 – 2008
- Teaching Assistant, University of Kentucky, 2006 – 2007
- Teaching Assistant, Berea College, 2004 – 2006
- Web Designer, Berea College, 2002 – 2004

HONORS

- 2006 – 2009 Chemistry Add-on Fellowship

PUBLICATIONS AND PRESENTATIONS

- Link, D; Lodder, R.A. *AAPS PharmaSciTech*. **2009**. (submitted 23 Jun 2009).
- Link, D; Hannel, T; Lodder, R.A. *Algorithms*. **2009** (submitted 27 May 2009).
- Hannel, T. S.; Link, D. J.; Lodder, R. A. *Journal of Pharmaceutical Innovation*. **2008**, 3(3), 152-160.
- Douglas, C. C.; Hannel, T. S.; Link, D. J.; Lodder, R. A. *Cyber Physical Systems*. **2008**. Preprinted at http://drake.contactincontext.org/david/pdf/incorrectdefectivepilldetection.pdf
- Link, D. J. *Contact in Context* 2(2), **2007**.

- Link, D. J.; Hannel, T. S.; Lodder, R.A. *2008 AAPS Annual Meeting and Exposition*. Atlanta, GA. November, 19 **2008**; abstract 5012.
- Hannel, T. S.; Link, D. J.; Lodder, R.A. *2008 AAPS Annual Meeting and Exposition*. Atlanta, GA. November, 19 **2008**; abstract 4052.
- Link, D. J.; Hannel, T. S.; Lodder, R.A. *FACSS 2008 Annual Meeting*, Reno, NV. Sept, 30 **2008**; abstract 150.
- Hannel, T. S.; Link, D. J.; Lodder, R.A. *FACSS 2008 Annual Meeting*, Reno, NV. Sept, 30 **2008**; abstract 151.
- Hannel, T. S.; Link, D. J.; Lodder, R.A. *4th Kentucky Innovation and Enterprise Conference (KIEC),* Lexington, KY. April, 17 **2008**.
- Link, D. J.; Hannel, T. S.; Lodder, R.A. *Ohio Valley Affiliates for Life Sciences (OVALS) 6th Annual Conference: Transformational Research: A Bridge to Building Economies*. Louisville, KY. April, 14 **2008**.
- Link, D. J.; Hannel, T. S.; Lodder, R.A. *34th Annual Naff Symposium on Chemistry and Molecular Biology*. Lexington, KY. April, **2008**.
- Link, D. J.; Hannel, T. S.; Lodder, R.A. *First Annual UK Cognitive Sciences Day.* Lexington, KY. February, 16 **2008**.