AN EXPLORATION OF CHEMOMETRIC REGRESSION TECHNIQUES TO ANALYZE INFRARED SPECTRA OF AQUEOUS SUGAR MIXTURES

A thesis presented to the faculty of the Graduate School of Western Carolina University in partial fulfillment of the requirements for the degree of Masters of Science in Chemistry.

By

Morgan Elise Cheek

Advisor: Dr. Scott Huffman
Associate Professor of Analytical Chemistry
Department of Chemistry & Physics

Committee Members: Dr. Carmen Huffman, Department of Chemistry & Physics
Dr. Jamie Wallen, Department of Chemistry & Physics

April 2019

ACKNOWLEDGEMENTS

I would like to thank the Department of Chemistry and Physics at Western Carolina University for providing me the opportunity to complete my M.S. degree thesis work. I would also like to thank the Honors College at WCU for their financial support via the Academic Project Grant (APG) and the Graduate School at WCU for support via scholarships. I am very grateful to my thesis research advisory committee members Associate Professor Dr. Scott Huffman, Associate Professor Dr. Carmen Huffman, and Assistant Professor Dr. Jamie Wallen, for their feedback and assistance with this project. In particular, I would like to thank my research advisor, Dr. Scott Huffman, for his years of assistance and his patience in teaching me programming with Python. I would also like to extend my gratitude to all faculty members in the Department of Chemistry and Physics at WCU for furthering my education and bringing me to this point. Lastly, I'd like to thank my family and friends for their continued support and for listening while I talk about science.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ATR	attenuated total reflectance
BSTFA	N,O-bis(trimethylsilyl)trifluoroacetamide
CLS	classical least squares
CLSR	classical least squares regression
DP	data preprocessing
FID	flame ionization detector
FT	Fourier transform
GC	gas chromatography
HPLC	high performance liquid chromatography
IR	infrared
LVs	loading vectors
MS	mass spectrometry
NIR	near infrared
PCA	principal component analysis
PCAR	principal component analysis regression
PCs	principal components
PLS	partial least squares
PLRS	partial least squares regression
$RSEP$	relative standard error of prediction
$RSEP_{cv}$	relative standard error of prediction by cross validation
$RSEP_{sv}$	relative standard error of prediction by separate validation
SEP	standard error of prediction
SEP_{cv}	standard error of prediction by cross validation
SEP_{sv}	standard error of prediction by separate validation
TCMS	trimethylchlorosilane
TMSI	trimethylsilyl imidazole
UV-VIS	ultraviolet and visible

ABSTRACT

AN EXPLORATION OF CHEMOMETRIC REGRESSION TECHNIQUES TO ANALYZE INFRARED SPECTRA OF AQUEOUS SUGAR MIXTURES

Morgan Elise Cheek, Masters of Science in Chemistry

Western Carolina University (April 2019)

Advisor: Dr. Scott Huffman

Infrared spectroscopy (IR) is a valuable tool for both qualitative and quantitative studies in chemistry. This is due to its high sensitivity, robustness, short measurement time, and ease of use. However, IR has several disadvantages when it comes to quantitation of mixtures. The most notable is that due to its high sensitivity, spectra of mixture samples become highly convoluted. Additionally, intermolecular forces in a mixture can shift the frequency of a vibration, which complicates analyses that rely on only one wavelength. Aqueous samples, and particularly aqueous sugar samples, are mixtures that exemplify these problems. Sugars form hydrates in water, which have different IR spectra than pure sugars. These mixtures violate the assumptions of Beer's Law, the basis for quantitative spectroscopy. Therefore, quantitative analysis of aqueous sugar samples by IR does not give accurate results when using normal regression techniques.

The goal of this project was to improve the accuracy of this type of analysis by using advanced multivariate regression techniques and data preprocessing. Simple linear models like classical least squares regression (CLSR) were expected to give less accurate results than models like principal component analysis (PCAR) and partial least squares (PLSR), which can use more variables to explain unknown complexes in a mixture, like sugar hydrates. These regression techniques were used to predict one-component aqueous fructose mixtures as well as three-component aqueous sugar mixtures. Data preprocessing was used to optimize the parameters of these techniques. The accuracy of these analyses were validated by standard error of prediction

(*SEP*) and relative standard error of prediction (*RSEP*).

It was found that for both one-component and three-component mixtures, PLSR was the most accurate regression model. CLSR gave the highest errors, which was expected due to its reliance on Beer's Law assumptions. PCAR performed better than CLSR, but worse than both forms of PLSR. PLSR1 and PLSR2 gave similar error values, and performed better with different components.

This work helped show that it is possible to somewhat accurately model data from IR spectra of aqueous sugar mixtures. This is beneficial in that current analysis of these samples is time-consuming and expensive. With more development, these techniques could be applied to more complex samples for use in industry.

CHAPTER ONE: INTRODUCTION

Current Techniques and Applications for Analysis of Carbohydrates

The identity and quantification of carbohydrates in mixtures is of importance in several fields including food science,[1–12] environmental research,[13–16] biochemistry,[17–20] and the medical field.[21–23] There are several techniques that can be used for this analysis, each with a unique set of advantages and disadvantages. These analyses are complicated by the similar moieties of complex carbohydrates as well as high interference from the sample matrix, especially water.

Gas chromatography coupled with mass spectrometry (GC-MS) is a widely used technique for complicated mixtures due to its excellent separation of components. This method has been used for analysis of flour,[1,2] honey,[3] sports drinks,[4] and aerosols.[13–15] Due to the nature of GC-MS, all components must be volatile in order to pass through the column, as well as thermally stable. Carbohydrates are neither. Therefore, a silylation step is required using N,O-bis(trimethylsilyl)-trifluoroacetamide (BSTFA) and trimethylchlorosilane (TMCS)[4,13,15,24] or trimethylsilyl imidazole (TMSI)[1,2] to silylate the sugars and increase their volatility and stability. These silylation reactions require expensive, air- and water- sensitive reagents, which is a problem especially when discussing aqueous mixtures. The samples must be dried completely before the silylation, usually with an inert gas like argon. This sample preparation therefore takes a long time to perform, even before the time-consuming GC-MS analysis. Additionally, this preparation requires added expense and a skilled chemist to perform.

An alternative method to GC-MS is high performance liquid chromatography (HPLC), which does not require volatile components. HPLC has been used for samples such as molasses,[5] honey,[6] beverages,[7,8] fruit,[9] and berries.[10–12] HPLC is used in conjunction with some sort of detector. This detector is often an ultraviolet and visible (UV-VIS) detector but can also be a refractive index detector,[7,8,11,12] evaporative light scattering detector,[5,10] or pulsed amperometric detector.[9] The main drawback of HPLC is its use of high volumes of solvents like water, acetonitrile, and

methanol.[5,6] While GC uses a relatively cheap and inert carrier gas, HPLC requires two solvents in a gradient to gradually move the sample through the column.

These solvents generate large volumes of chemical waste, which can be harmful to the environment, and disposal can be expensive. This method does not require the silylation step, which does save some expense and time. However, the use of the mobile phase solvents, any needed sample prep, calibration standards, and internal standards all add time and expense to this technique.

Another less common technique is enzymatic determination.[25] One subset of this technique includes the enzyme glucose oxidase, which catalyzes the oxidation of glucose to gluconolactone and hydrogen peroxide.[26] The hydrogen peroxide then oxidizes another reagent to form chromophores which are detected by UV-VIS spectroscopy.[26] There are several variations of this technique for various sugars and detectors.[27–29] These enzymatic techniques are relatively simple to perform compared to the other methods, and they require simpler instrumentation. However, this still requires sample preparation in a laboratory, which the IR method could potentially avoid.

The Beer's Law Problem

In contrast to the above methods, infrared (IR) spectroscopy is faster, cheaper, easier, safer, and more environmentally friendly. This is due to the lack of sample preparation (and therefore solvents, reagents, and chemical waste) and need for physical separation of components.[17,30,31] IR spectroscopy also has the capacity of field-deployment and automation.[32] From a spectroscopic standpoint, IR spectroscopy is beneficial because it is highly sensitive and robust.[30] Almost every single molecule has IR active vibrational modes, which can be a double-edged sword when it comes to analyzing complicated mixtures.

Because of the sensitivity and the lack of physical separation of components, IR spectra can easily become highly convoluted and difficult to interpret with mixture samples. Because of this, IR spectroscopy is often used alongside chemometric methods such as data preprocessing (DP)

and multivariate statistical analyses.

Quantitative absorbance spectroscopy is based on the Beer-Lambert-Bouguer Law (Beer's Law),[33]

$$A = \epsilon b C \tag{1}$$

where A is the absorbance at a certain wavelength, ϵ is the absorptivity constant for a particular substance at a particular wavelength, b is the pathlength, and C is the concentration of the substance. When dealing with a mixture of various compounds, the law can be adapted

$$A = \epsilon_1 b_1 C_1 + \epsilon_2 b_2 C_2 + ... + \epsilon_n b_n C_n \tag{2}$$

where n is the number of components. The signal from each component is additive to get the total absorbance. This law relies on the assumption that the compounds are not chemically or spectroscopically interfering, or that all interfering chemical components are known. It is rare to have a prior knowledge of how compounds interfere to affect Beer's Law, so one typically relies on mixtures with non-interfering components.[34]

The biggest disadvantage of using IR spectroscopy with aqueous samples is the interference from liquid water. Water displays large, characteristic bands at 3200-3600 cm^{-1} (O-H stretching modes) and 1600 cm^{-1} (bending mode). The high concentration of water in aqueous samples means high-absorbance bands in the IR region, as concentration and absorbance are directly related. These bands can obscure the absorption bands of other compounds in those regions.[35]

Additionally, the problems involved in this analysis stem from the fact that water and sugars form hydrates when mixed, and these mixtures exhibit a strong hydrogen bonding network.[35–40] These interacting species and intermolecular forces can alter the vibrational modes, which alters the position and intensity of bands. In other words, Beer's Law, which assumes an additive signal from each component at one particular wavelength, is violated.[34,41] This is demonstrated in Figure 1 where it is shown that the shape of the experimental spectrum is different from the theo-

retical spectrum. The wavenumber position of absorbance bands has changed due to the bonding effects.

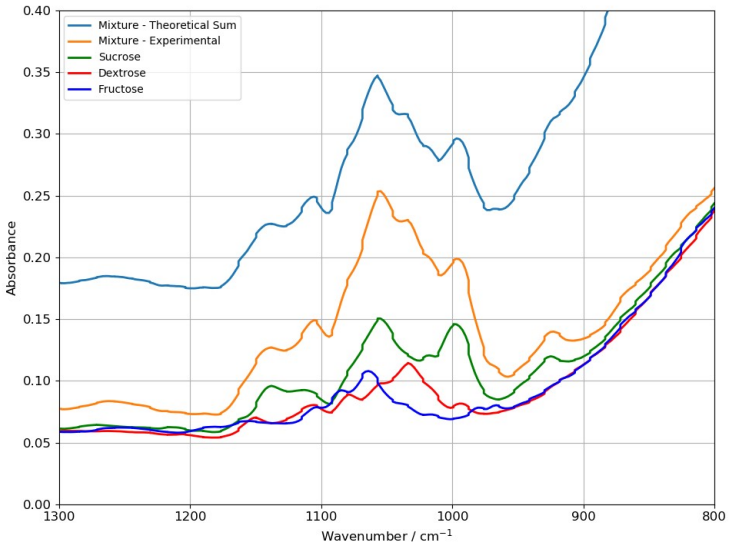

Figure 1. Exemplary IR spectra of aqueous sugar mixtures to demonstrate that the signals are not additive. Concentrations of individual samples of fructose, dextrose, and sucrose were 0.6004, 0.6008, and 0.6005 M, respectively. The mixtures had concentrations of 0.5995, 0.6014, and 0.6005 M. (The maximum difference between individual and mixture concentration was 0.15% for fructose.)

To further illustrate these concepts, observe the spectra in Figure 2 of solid sugars and their aqueous counterparts. The spectra of the solids all display bands that are not visible in the spectra of the aqueous counterparts, notably in the region 1500-500 cm^{-1}. Many of these signals are completely obscured by signals from water due to its high concentration in the mixture. Additionally, there are noticeable wavenumber shifts of the most intense bands between solid and aqueous samples. The aqueous samples also cause baseline problems in the region 1500-900 cm^{-1}. In the aqueous samples, the baseline in this region is about 0.8 when it should be closer

to 0. This baseline issue can cause problems with regression.

Another way to observe the limitations of Beer's Law is by simply creating a calibration curve and observing that the data are not linear. An aqueous fructose mixture was created by combining 0.053731 moles (9.6801 g) of fructose and 0.17860 moles of water (3.2183 g) to achieve a mixture of 23 mol % fructose. This was then serially diluted to achieve concentrations of 21, 17, 13, 11, 9, 6, 4, 2, and 0.5 mol % fructose. These were measured with attenuated total reflectance (ATR) IR spectroscopy and several bands attributed to fructose were chosen to create calibration curves. The spectra of these samples are shown in Figure 3.

The wavenumbers indicated in Figure 3 were chosen to make calibration curves. These are shown in Figure 4. A line interpolating the first and last data point show where the data would lie if it were linear. The points corresponding to the band at 1035 cm^{-1} (yellow) show this non-linearity the best, and is one of the most intense bands in the fructose spectrum. The chosen vibrational bands are identified in Table 1 derived from Max and Chapados' work.[36] Their work indicates that the band at 1035 cm^{-1} is a shoulder for fructose pentahydrate and fructose monohydrate. It is possible that this sample has so many different hydrates absorbing here that the band is more intense (not a shoulder), which would also explain why the calibration curve for this band deviates from the data so much.

Table 1. Band positions of aqueous fructose. Wavenumbers of fructose pentahydrate are from Max and Chapados.[36]

Wavenumber / cm^{-1}	**Assignment**	**Wavenumber / cm^{-1} of fructose pentahydrate**
1061	C-O stretch	1062 *strong*
1035	C-O stretch	1033 *shoulder*
995	C-O stretch *exocyclic*	993
967	C-O stretch *exocyclic*	966 *medium*

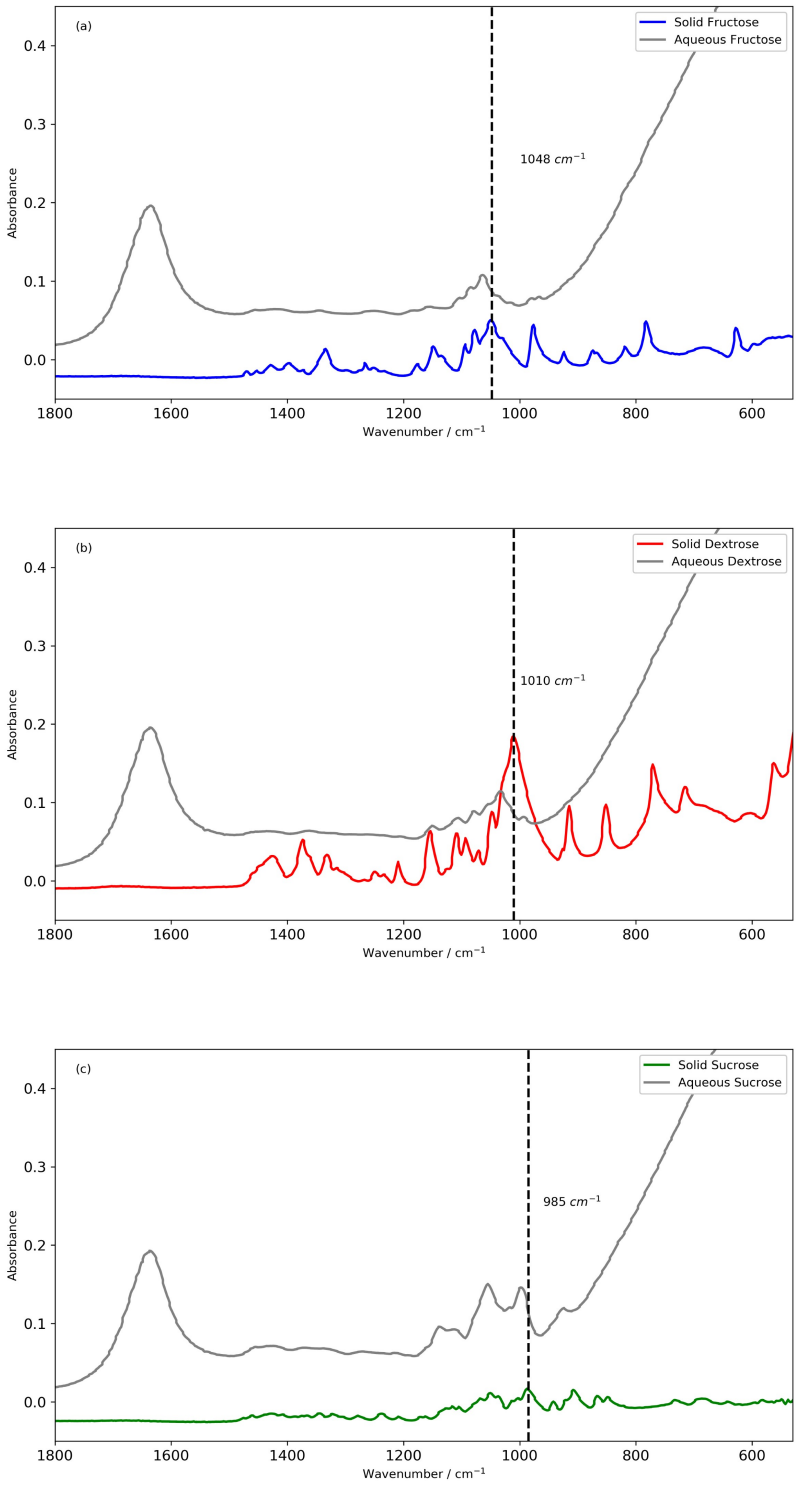

Figure 2. IR spectra of (a) fructose, (b) dextrose, and (c) sucrose in solid and aqueous form. Vertical lines are used to show the difference in band position.

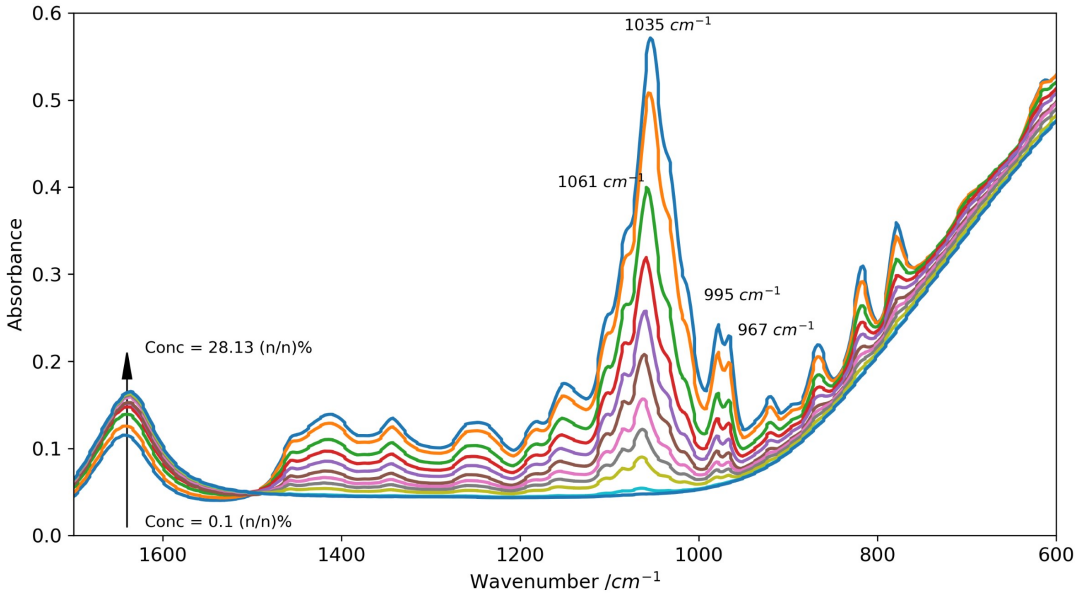

Figure 3. IR Spectra of various concentrations of fructose in water. Bands used in the calibration curves are indicated.

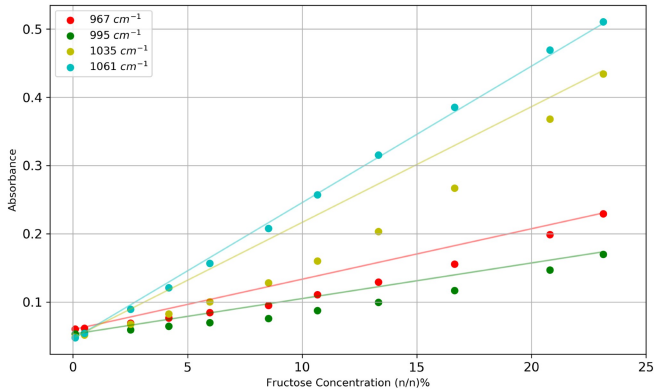

Figure 4. Scatter plot to show nonlinear trend between concentration of fructose and absorbance in IR spectra. Not all bands are included. Lines connect the first and last datapoint for each series.

CHAPTER TWO: THEORY

Notation

Throughout this paper, all matrices are denoted in bold with a capital letter, all vectors are denoted in bold with a lowercase letter, and all scalars are normal typeface. Matrices that describe parameters for particular sets of data are denoted with a subscript.

Classical Least Squares

The simplest and most commonly used regression technique is known as classical least squares (CLS). This method has been described in detail elsewhere,[34,41] but it is briefly described here in order to discuss the differences among regression techniques. CLS uses the same type of math as a calibration curve, but can be applied to multivariate, or full-spectrum, data. A typical calibration curve for spectroscopy relies on Beer's Law, Equation /refkbeer,

$$A = \epsilon b C \tag{1}$$

If one assumes a constant pathlength, this equation can be expressed as a linear relationship

$$A = CK \tag{3}$$

where K is ϵb, henceforth referred to as the model, or the relationship between concentration and absorbance.

To apply this technique to multivariate data, one can convert these variables to matrices that hold more information. A becomes $\mathbf{A}_k$ to represent the absorbance of known standards, or the training set. The spectral data is contained in $\mathbf{A}_k$, spanning several different wavelengths and containing the spectra from each standard in the training set. The dimensions of $\mathbf{A}_k$ are (*nspec*, *npts*), or number of spectra by number of datapoints on each spectrum. C becomes $\mathbf{C}_k$, and has the di-

mensions (*nspec*, *ncomp*), or number of samples by number of known components. The model, K, has the dimensions (*ncomp*, *npts*), leaving

$$\mathbf{A}_k = \mathbf{C}_k\mathbf{K} \tag{4}$$

In CLS, $\mathbf{C}_k$ and $\mathbf{K}$ represent physical information about the samples. $\mathbf{C}_k$ holds the concentration information, while each row of $\mathbf{K}$ represents the pure spectrum for the corresponding component.[34] For a univariate model, these matrices are absorptivity scalars, or single values.

To fit the data or solve for $\mathbf{K}$, one must multiply both sides of the equation by the inverse of $\mathbf{C}_k$.[41] For a true inverse to be calculated, a matrix must be square, or have the same number of rows as columns. Since $\mathbf{C}_k$ is not square, the Moore-Penrose pseudo-inverse of $\mathbf{C}_k$, $\mathbf{C}_k^+$, is calculated:[42,43]

$$\mathbf{C}_k^+ = (\mathbf{C}_k^T\ \mathbf{C}_k)^{-1}\mathbf{C}^T \tag{5}$$

This can then be used to solve for $\mathbf{K}$

$$\begin{aligned} \mathbf{A}_k\mathbf{C}_k &= \mathbf{K}\mathbf{C}_k\mathbf{C}_k^+ \\ \mathbf{K} &= \mathbf{A}_k\mathbf{C}_k^+ \end{aligned} \tag{6}$$

In a univariate model, $\mathbf{K}$ represents the slope of the best fit line from a calibration curve.

Once $\mathbf{K}$ is known, it can be used to predict the unknown concentrations, $\mathbf{C}_u$, by using the spectra from a validation or testing set, $\mathbf{A}_u$. These matrices are very similar to $\mathbf{A}_k$ and $\mathbf{C}_k$, but hold information from the validation set instead of the training set. The number of spectra, *nspec*, may be different here than that of the training set. To predict the concentrations, one must solve Equation 3 for $\mathbf{C}$:

$$\mathbf{A}_u = \mathbf{C}_u\,\mathbf{K}$$
$$\mathbf{C}_u = \mathbf{A}_u\mathbf{K}^{+} \tag{7}$$

For a univariate model, one can simply take the equation of the best fit line

$$A = Cm + A_0 \tag{8}$$

where m is the slope and A_0 is the intercept and solve for C.

Principal Component Analysis

While CLS is easy to perform, it does not give accurate predictions when the samples violate Beer's Law and are nonlinear.[34,41] Principal component analysis (PCA) is more suited for this type of analysis, as it can account for unknown species in a mixture, like the sugar hydrates. While this has been detailed elsewhere,[34] a brief explanation follows. For the purposes of this project, the independent variables are contained in the spectra. The spectral data can be expressed as an array of absorbance values each corresponding to a particular wavenumber. The dependent variables are the concentrations of the different components in the sample.

To demonstrate how PCA is used to represent data, a 2-dimensional example will be used. While this data was taken from IR spectra of aqueous sugar mixtures, it is for demonstrative purposes only. This example will show how the model is built by observing only the absorbance values at two wavenumbers. In Figure 5 is shown the absorbance values of 5 different spectra at these two wavenumbers.

There is some positive correlation between these two sets of absorbance values. One way of representing the variance of this data is by using a vector. In PCA, this vector is an eigenvector (a scaled vector with a magnitude of one) that spans one dimension of variance, or a principal component (PC). These can also be called latent variables or loading vectors (LVs). Figure 6 shows

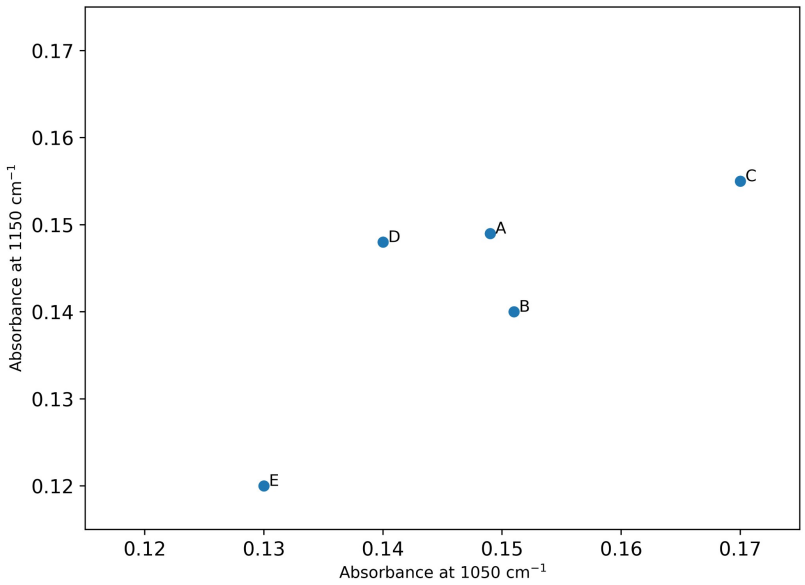

Figure 5. Plot of example spectral data.

the first eigenvector, or principal component 1 (**PC1**), which represents the most variance in this data.[34] Since this is a 2-dimensional example, there are at most 2 principal components. The second principal component (**PC2**) is orthogonal to **PC1** and spans the variance in this direction. This simple example has two variables contained in 2 PCs. However, many analyses have thousands of variables (absorbance values). PCA can represent those variables into only a few PCs, which greatly reduces the dimensionality of the data, which is why PCA is so beneficial for multivariate analysis.

These eigenvectors are determined through an iterative process to minimize the distance from all data points to the vector. To do this, the data is centered around an origin, the center of the data, and a random vector is drawn through this origin. The data points can then be projected onto the vector, and the distance from that point to the PC, d_i, is found. Equation 9 shows how

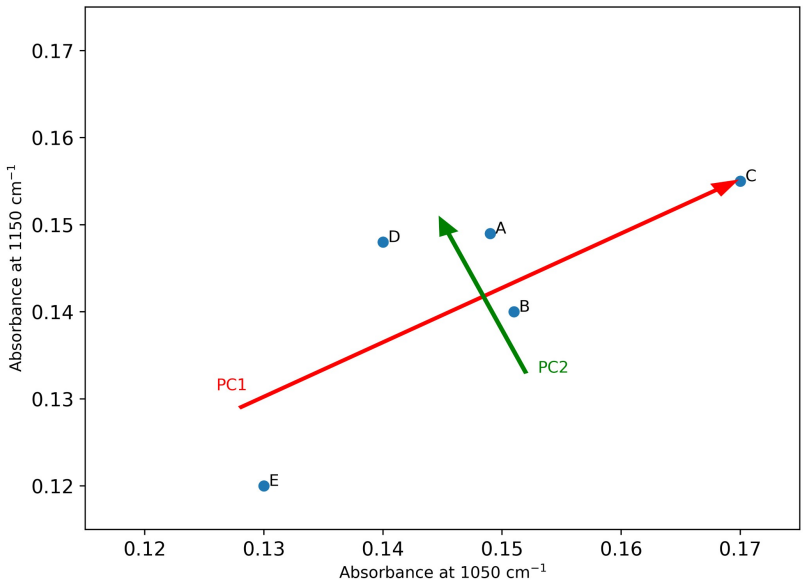

Figure 6. Example data plotted with principal components **PC1** and **PC2**.

these distances can be used to calculate the sum of the squared distances, or SSD:

$$SSD = \sum_{i=1}^{npts} d_i^2 \tag{9}$$

Since the distance from each data point to a PC should be as small as possible, minimizing the SSD by altering the vector's coordinates works to find the best vector to fit the data. Minimizing SSD is equivalent to maximizing the variance contained in the PC. This SSD is known as the eigenvalue for **PC1**. To find **PC2**, the second eigenvector, the same process is used with the constraint that **PC2** must be orthogonal to **PC1**. Therefore, **PC2** spans the second most amount of spectral variance. For data sets that have more than 2 dimensions, this process can be continued until the principal components represent all the variance of the data. When building a model with PCA, one must input the minimum number of principal components that correctly fits the

data without fitting the noise or overfitting the data.

Each PC is associated with a vector of weights, or scores. These scores are calculated by projecting each datapoint onto the PC and finding the distance to the origin of the data. Scores can be positive or negative depending on the point's relation to the origin. A visual representation of centering the data on an origin, the distance from the i^{th} datapoint to **PC1**, d_i, and the score for the i^{th} datapoint on **PC1**, $t_{i,1}$, is shown in Figure 7.

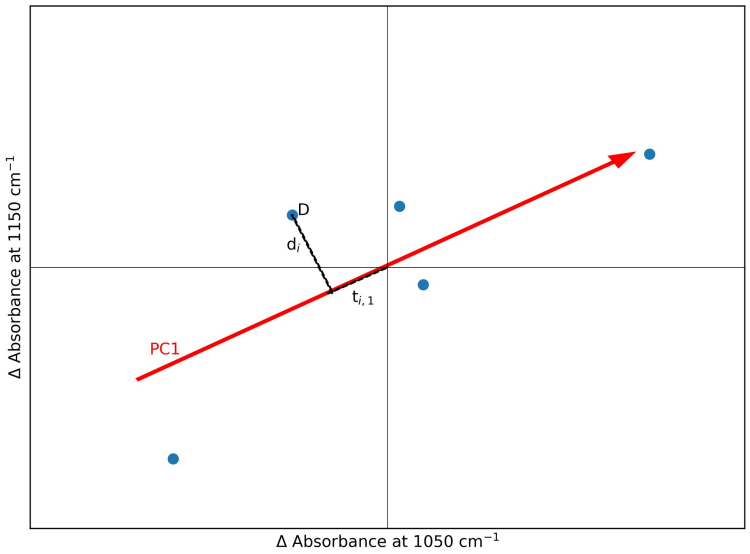

Figure 7. Example data centered on the origin to show d_i and $t_{i,1}$.

To perform regression, we can return to our adapted Beer's Law in Equation 4. However, for PCA, $\mathbf{A}_k$ is broken down quite differently. Instead of $\mathbf{C}_k$ and $\mathbf{K}$ which represent physical information, $\mathbf{A}_k$ is broken down into matrices that hold the PCs and scores. This looks like

$$\mathbf{A}_k = \mathbf{SL} \tag{10}$$

where **S** is the scores matrix and **L** is the matrix of PCs. These matrices are similar to $\mathbf{C}_k$ and **K**, but represent different information. First, the number of PCs is often larger than the number of components in the mixture, which is how this technique incorporates unknown complexes or components in a mixture. **L** holds what can be thought of as composite components, meaning these vectors do not represent pure component spectra.[34] Rather, they contain composite information of various spectral features.

To predict the validation set based on their spectra, one can use the same PCs and calculate the scores for that sample. This involves the pseudoinverse of **L**.

$$\begin{aligned} \mathbf{A}_u &= \mathbf{S}_u\mathbf{L} \\ \mathbf{S}_u &= \mathbf{A}_u\mathbf{L}^+ \end{aligned} \tag{11}$$

In this way, the scores are related to the loading vectors and absorbance values. However, the scores need to be related to the concentrations of the samples. This correlation exists, just in a different dimension. To determine the concentrations, $\mathbf{C}_u$, we have to rotate the vectors in $\mathbf{S}_u$ using a rotation matrix, **P**.

$$\mathbf{S}_u\mathbf{P} = \mathbf{C}_u \tag{12}$$

This rotation matrix can be optimized iteratively to provide the most accurate results when rotating the scores vectors to concentration.

Partial Least Squares

Partial least squares analysis (PLS) is very similar to PCA in that it also relies on ”PCs” and scores. In PLS, the PCs are known as loading vectors (LVs) and are calculated differently. This is described in detail elsewhere,[34,44] but is briefly explained here for comparative purposes. While PCs are found by maximizing the spectral variance, LVs are found by (1) best explaining the dependent variables (2) best explaining the independent variables, and (3) maximizing the correlation between those two variables. PLS therefore looks at both $\mathbf{A}_k$ and $\mathbf{C}_k$ and extracts a single set

of scores from each simultaneously.[34] This can be written

$$\begin{aligned} \mathbf{A}_k &= \mathbf{SL} \\ \mathbf{C}_k &= \mathbf{UV} \end{aligned} \tag{13}$$

where U is the scores matrix for $\mathbf{C}_k$ and V is the loading vector matrix for $\mathbf{C}_k$.

The objective then is to maximize the covariance between the scores matrices **S** and **U**. Covariance (Cov) is found by

$$Cov(\mathbf{S}, \mathbf{U}) = \sum (t_{i,j} - \overline{t_j})(u_{i,j} - \overline{u_j}) \tag{14}$$

where $t_{i,j}$ is the score for the i^{th} sample on the j^{th} LV from $\mathbf{A}_k$, $\overline{t_j}$ is the mean of those scores, $u_{i,j}$ is the score for the i^{th} sample on the j^{th} LV from $\mathbf{C}_k$, and $\overline{u_j}$ is the mean of those scores. The greater the covariance, the greater the correlation between **S** and **U**. This is the same as maximizing the correlation between $\mathbf{A}_k$ and $\mathbf{C}_k$ and best explaining $\mathbf{A}_k$ and $\mathbf{C}_k$.

Other than this differing optimization of loading vectors, PLS and PCA work in the same way. For regression, the LVs can be used to find the scores $\mathbf{S}_u$, which can then be rotated with **P** to find the predicted concentrations, $\mathbf{C}_u$.

There are two different ways to perform PLS, known simply as PLS1 and PLS2. PLS2 is what was described above, where $\mathbf{C}_k$ holds the concentrations of each component for each sample. PLS1 differs in that $\mathbf{C}_k$ is sliced so that it is one-dimensional and only contains information about one component. This means that for PLS1, the analysis is performed as many times as there are number of components. This is sometimes helpful in cases where the different components are not correlated, or when they require different numbers of loading vectors to achieve the best fit.[34]

CHAPTER THREE: EXPERIMENTAL

Materials and Instrumentation

Materials

Fructose was obtained from MP Biomedicals, LLC and had a purity of over 99%. Lab grade dextrose was obtained from Fisher Scientific. Sucrose was reagent grade and was obtained from Fisher Scientific. The structures of these sugars are shown in Figure 8. Fructose and dextrose are the two monosaccharides that form sucrose. These three sugars were chosen to see how the regression models perform on monosaccharides vs disaccharides as well as how they perform when the components have similar structure.

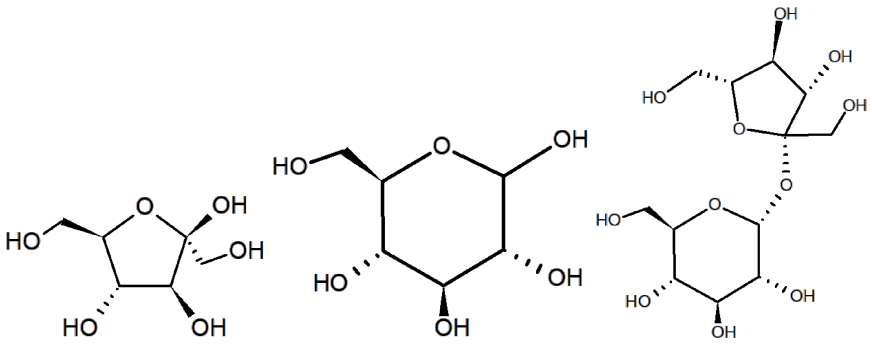

Figure 8. Structures of chosen sugars.

Attenuated Total Reflectance Fourier-Transform Infrared Spectroscopy

All aqueous sugar spectra described in this work were acquired on a Thermo iS10 Fourier transform infrared (FTIR) spectrometer with a diamond attenuated total reflectance (ATR) attachment. Each spectrum is the result of 32 scans with a resolution of 4 cm^{-1} with a new background acquired every 10 min to account for changes in atmosphere and instrumental drift. Atmospheric temperature, humidity, and pressure were recorded during each measurement session to ensure

conditions would not affect spectral data. Between each sample, the ATR was wiped clean with a KimWipe, then washed with water followed by acetone.

Sample Preparation

For an ideal set of standards to serve as a training set, a range of concentrations from each component must be achieved. Additionally, these concentrations must vary in relation to each other. For example, if one standard has the ratio of 3:2:1 for each component, the training set should also include standards of 1:2:3 and 2:3:1. This helps build a more robust model in which to describe the data, as there is a greater variety of spectral information in the training set. In other words, we want to represent as many types of mixtures as possible in the training set. For this experiment, the chosen components were fructose, dextrose, and sucrose.

Additionally, to test the model that is built from these standards, a separate set of "unknowns," or a validation set, were created in the same manner. To some of these unknowns, various levels of interferents were added to see if the model could overcome the chemical and spectral interferent effects. We chose to use sodium acetate, ethyl acetate, and sorbitol to act as interferents to the model. These were chosen as they all contain C=O bonds and some C-O-H bonds like the sugar molecules which were of interest. The structures of these interferents are shown in Figure 9. Sorbitol is a sugar found in some fruits like pears,[45] ethyl acetate is present during wine fermentation,,[46] and sodium acetate is a common food additive.[47]

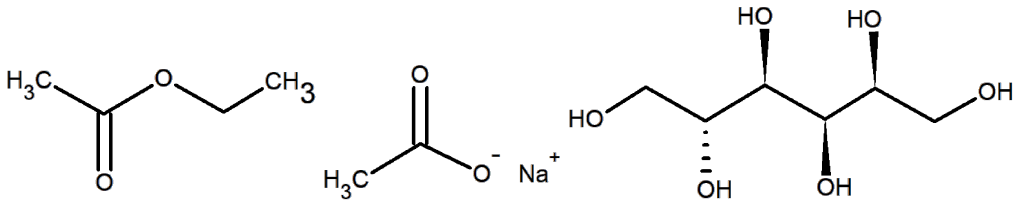

Figure 9. Structures of chosen interferents.

To make the training and validation sets, the chosen quantities of sugars and/or interferents were added to a 25 mL volumetric flask, then nanopure water was added, and the solution was sonicated to dissolve all solids. The solution was then diluted to 25 mL before samples were stored in 50 mL conical tubes in a refrigerator. Table 2 shows the concentrations of all components from both sets of samples. Table 3 shows the ratios of these concentrations in relation to each other for each sample.

Table 2. Concentrations of aqueous sugar samples. Standards from the training set are denoted by "S" and have no interferents, while validation standards are denoted by "U."

Sample ID	Fructose / (M	Dextrose / M	Sucrose / M	Sodium Acetate / M	Ethyl Acetate / M	Sorbitol / M
S1	0.0203	0.0367	0.0078			
S2	0.1085	0.0724	0.0093			
S3	0.1832	0.1311	0.0199			
S4	0.2096	0.1821	0.0283			
S5	0.2624	0.2102	0.0042			
S6	0.2778	0.2370	0.0444			
S7	0.3398	0.2798	0.0497			
S8	0.3688	0.3001	0.0587			
S9	0.4254	0.3354	0.0727			
S10	0.4555	0.3697	0.0792			
S11	0.6524	0.3034	0.1333			
S12	0.2697	0.5639	0.2392			
S13	0.2216	0.1988	0.3538			
S14	0.2392	0.1886	0.2340			
S15	0.2635	0.1806	0.6397			
S16	0.2822	0.1908	0.4868			
S17	0.2521	0.7596	0.1130			
S18	0.9408	0.1970	0.1481			
S19	0.2593	0.5736	0.3841			
S20	0.6944	0.6288	0.1232			
U1	0.2004	0.1600	0.0246			
U2	0.2501	0.2990	0.0398			
U3	0.3010	0.2624	0.0505			
U4	0.1998	0.1615	0.0596			
U5	0.2489	0.3026	0.0797			
U6	0.2027	0.1621	0.0251	0.3273		
U7	0.1843	0.1220	0.0202	0.0406		
U8	0.1809	0.1199	0.0205		0.0396	
U9	0.1800	0.1203	0.0201		0.0407	
U10	0.1804	0.0800	0.1244	0.0634		
U11	0.1814	0.0798	0.1205	0.0528		
U12	0.1810	0.0807	0.1196		0.0493	
U13	0.1801	0.1206	0.0200	0.4116		0.0406
U14	0.1813	0.1206	0.0205	0.0420	0.0784	

Table 3. Ratios of concentrations of aqueous sugar samples. Standards from the training set are denoted by "S" and have no interferents, while validation standards are denoted by "U."

Sample ID	Fructose	Dextrose	Sucrose	Sodium Acetate	Ethyl Acetate	Sorbitol
S1	0.55	1	0.21			
S2	1	0.67	0.09			
S3	1	0.72	0.11			
S4	1	0.87	0.14			
S5	1	0.80	0.02			
S6	1	0.85	0.16			
S7	1	0.82	0.15			
S8	1	0.81	0.16			
S9	1	0.79	0.17			
S10	1	0.81	0.17			
S11	1	0.47	0.20			
S12	0.48	1	0.42			
S13	0.63	0.56	1			
S14	1	0.79	0.98			
S15	0.41	0.28	1			
S16	0.58	0.39	1			
S17	0.33	1	0.15			
S18	1	0.21	0.16			
S19	0.45	1	0.67			
S20	1	0.91	0.18			
U1	1	0.80	0.12			
U2	0.84	1	0.13			
U3	1	0.87	0.17			
U4	1	0.81	0.30			
U5	0.82	1	0.26			
U6	1	0.80	0.12	1.61		
U7	1	0.66	0.11	0.22		
U8	1	0.66	0.11		0.22	
U9	1	0.67	0.11		0.23	
U10	1	0.44	0.69	0.35		
U11	1	0.44	0.66	0.29		
U12	1	0.45	0.66		0.27	
U13	1	0.67	0.11	2.29		0.23
U14	1	0.67	0.11	0.23	0.43	

Data Preprocessing

Before the spectral data was used for regression analysis, it required some data preprocessing (DP) steps to maximize important information and minimize noise and interference from the sample matrix. While this section describes each possible step individually, several variations and

combinations of these steps were used in final data analysis. All processing was performed using Python (version 3.7.2) and Jupyter Notebook (version 5.7.4), an open-source web application that allows for writing and sharing code. For all steps below, data was loaded and stored in a Python dataframe, which organized the following information: an index for each sample, the name of the sample, the spectrum file associated with that sample, and some information about the composition of that sample (metadata). Spectral data was pulled from the dataframe and stored in matrix **A**, which contained the absorbance values of each spectrum as rows. Wavenumbers were stored in a separate matrix, **x**.

Spectral Window Selection

It is common in spectroscopy that the components of interest in a sample and the sample matrix absorb at different wavelengths in a spectrum. Before performing any regression, it is therefore often helpful to cut out regions that correspond only to the sample matrix, as there is no relevant information contained in those regions.

Second Derivative

One step that helps emphasize changes in the data is to use the second derivative of the spectra. The benefit of this technique is that low intensity bands are emphasized more from the baseline, and there is a baseline correction where no important bands are located. Using the second derivative rather than the first preserves band position, making it slightly easier to interpret the derivative spectra after preprocessing. Figure 10 shows how a small band in the original spectrum is greatly emphasized by the second derivative. This band points down in the derivative spectrum, as it is where there first derivative crosses zero. To perform this preprocessing step, a polynomial is convoluted with the original spectrum, which both smooths and calculates the second derivative. This process is performed with a Savitsky-Golay filter as described in Savitsky and Golay.[48]

Mean Center

Another way to emphasize the difference in the spectra and correct the baseline is known as a Z mean center. This step is done by taking the mean of all spectra in **A** to get an average spectrum,

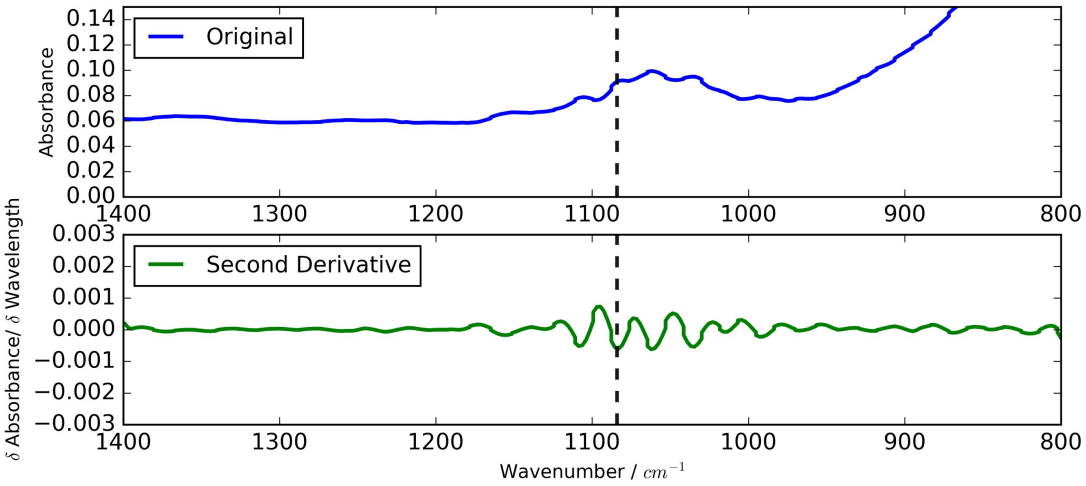

Figure 10. Unprocessed (top) and second derivative (bottom) spectra. Note that band position is preserved, bands are emphasized, and the baseline is set to zero.

$\mathbf{a}_{mean}$, and subtracting this from each individual spectrum in the dataset.

$$\mathbf{a}_{i,mc} = \mathbf{a}_i - \mathbf{a}_{mean} \tag{15}$$

where $\mathbf{a}_{i,mc}$ is the i^{th} spectrum in $\mathbf{A}$ that has been mean centered and $\mathbf{a}_i$ is the i^{th} spectrum in $\mathbf{A}$. Mean centering is especially useful in cases where all samples have the same or similar sample matrix, as this information is removed from the data when subtracting the average spectrum, leaving spectral information that is more likely to be just from the components of interest in the sample.

Autoscale / Z Transform

A z-transformation can be used to emphasize data that is changing over the dataset. A z-transformation is similar to a mean center, with the additional step of dividing by the standard deviation at each wavenumber, $\mathbf{a}_{std}$. This calculation, shown in Equation 16, helps put all the data on the same

scale, which normalizes the areas under each band.

$$\mathbf{a}_{i,z} = \frac{\mathbf{a}_i - \mathbf{a}_{mean}}{\mathbf{a}_{std}} \tag{16}$$

Normalization

There are several ways to get all data on the same scale. One of those is to normalize by the entire spectral area by dividing the spectrum by the magnitude of the total area for each spectrum, m_i. In this way, the total spectral area of each spectrum is equal to 1, as shown in Equation 17. This method is known simply as normalization by total area.

$$\mathbf{a}_{i,nt} = \frac{\mathbf{a}_i}{m_i} \tag{17}$$

It is sometimes more beneficial to normalize by the height of a specific band rather than the area, or normalize by band height. Typically, the most intense band, at least in the region of interest, should be used to normalize the spectrum as it will be set to equal 1. However, it is sometimes likely that different samples/spectra have different bands that act as the most intense. For consistency, the same band should be used for each one, so several different bands were tested to see which improved the analysis more. This method is done in the same manner as normalization by total area, but we divide by the absorbance at a particular wavelength, a_b. This is shown in Equation 18.

$$\mathbf{a}_{i,nb} = \frac{\mathbf{a}_i}{a_b} \tag{18}$$

Remove Background

Another way to avoid effects from the sample matrix is to simply subtract out the spectrum of the solvent or, equivalently, use the solvent as a background during measurement. By doing this, the bands corresponding to the components of interest are emphasized. This background removal is easily done by subtracting the absorbance values of the solvent spectrum from the absorbance

values of the sample spectra.

Baseline Correction

While previously mentioned preprocessing steps do have some degree of baseline correction, it is sometimes useful to manually improve the baseline without the added effects of those steps. This correction is done by fitting a polynomial to the existing baseline, then subtracting that polynomial from the original data, setting the new baseline to zero. To fit the baseline, points along the baseline that do not contain any relevant data and only the baseline intensity values are identified. These points are then used to create a polynomial with a specified degree, which is subtracted from the original spectrum.

Validation

In order to compare the various regression techniques to each other, a metric is needed to measure the error of the fit. The error was calculated in two different ways: cross validation and separate validation. This section describes how these errors were calculated.

Cross validation is a method that observes only one set of data: the training set. This is what was used to build the model. For this project, the standard error of prediction by cross validation, or *SEPcv* was chosen. To do this, one spectrum was removed from the training set. The training set was used to build the model and predict the concentrations of each sugar component in the removed spectrum. Since the true concentrations of that sample are known, the square of the difference between true and predicted concentrations, R, can be calculated. The spectrum is then returned to the training set. This process is done in turn with each spectrum being removed one at a time and the residual, R, stored. Equation 19 shows how *SEP* is calculated based on R and the number of spectra in the training set, n.

$$SEP = \sqrt{\frac{\sum R^2}{n}} \tag{19}$$

This process is done for each component of the mixture, so there will be as many *SEPcv* values

as components.

Notice that Equation 19 is for *SEP* and not *SEPcv*, because this value is calculated the same way for both cross validation and separate validation. For separate validation, R is calculated by comparing the predicted values of the validation set to their actual values. The validation set is separate from the training set that was used to build the model. It also contained some samples that have interferents, or extra components that are not predicted. The standard error of prediction for separate validation is referred to as *SEPsv*.

SEP values have the same units as the predicted and true concentrations, so the magnitude of *SEP* is dependent on the samples in question. Additionally, a large *SEP* value can indicate low error of the model if the concentration of the samples are suitably large. Therefore, it is helpful to scale the *SEP* values to a relative standard error of prediction, *RSEP*. This value is a percentage, so it is easy to compare these values across different models and samples. Equation 20 shows this calculation, where C_{avg} is the average concentration of the samples.

$$RSEP = \frac{SEP}{C_{avg}} \times 100 \tag{20}$$

Regression Parameters

All analyses with the exception of univariate CLSR involved preprocessing the data with a spectral window selection of 900-1300 cm^{-1}. For PCAR and PLSR2, three loading vectors were used. PLSR1 used five loading vectors for fructose, three for dextrose, and two for sucrose. This section describes how these parameters were chosen.

Both PLS and PCA require a selection of the number of loading vectors, or principal components, as described in Chapter 2. To determine the number of loading vectors to use for PLS2, a plot was generated as shown in Figure 11. Plot (a) shows these calculations performed on unprocessed data. Plot (b) shows how error improved with the use of DP methods as described above. The goal of this analysis was to minimize the *SEPcv* value for all of the components in the mix-

ture. Based on plot (a) we chose five loading vectors, as the sum line had the lowest *SEPcv* value at five loading vectors. However, based on plot (b), three loading vectors would be ideal. The difference between these plots demonstrate how DP and regression techniques are dependent on each other. Ultimately, determining the number of loading vectors to use with which preprocessing steps takes some back and forth.

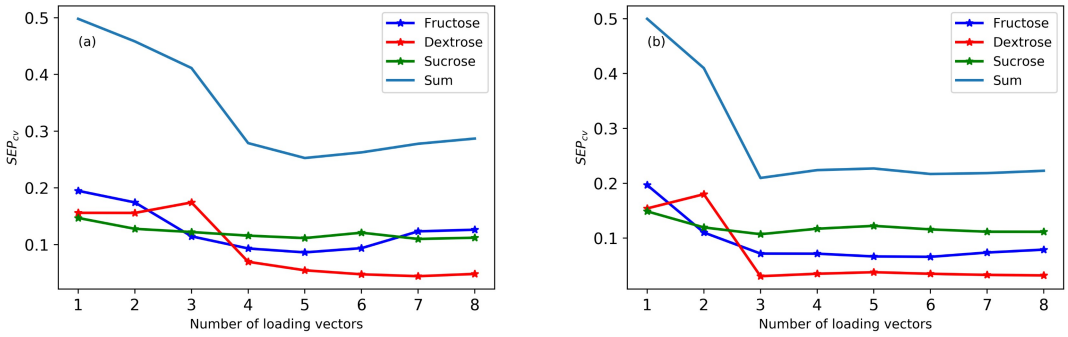

Figure 11. Error plots to show the best number of loading vectors for PLSR2. Plot (a) shows unprocessed data, while plot (b) includes preprocessing steps.

A similar analysis was done for PLS1, which observes each component separately. The plot in Figure 12 shows the results of this analysis. It was found that the ideal number of loading vectors were five for fructose, three for dextrose, and two for sucrose.

This analysis was also done for PCA. The results of this are shown in Figure 13. In plot (a), the unprocessed data, the error tends to decrease by adding more loading vectors until leveling out at around six. However, the data in plot (b) shows the opposite trend. There is a barely noticeable minimum at three loading vectors for sucrose in this plot, so three loading vectors were chosen to try to improve the sucrose predictions.

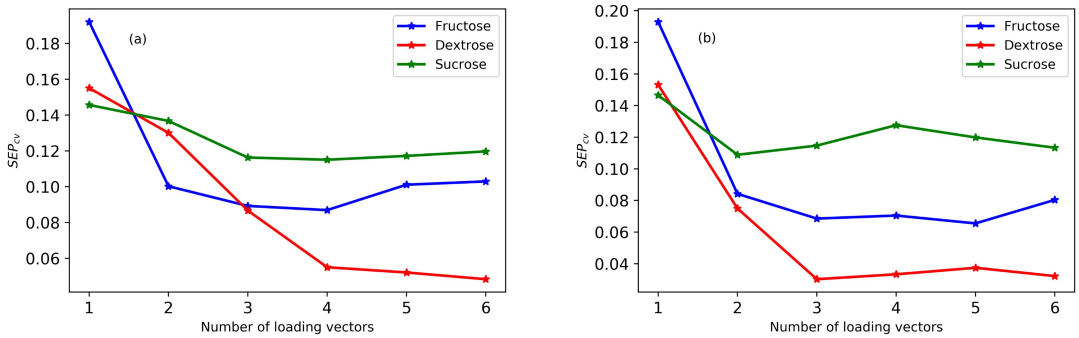

Figure 12. Error plots to show the best number of loading vectors for PLSR1. Plot (a) shows unprocessed data, while plot (b) includes preprocessing steps.

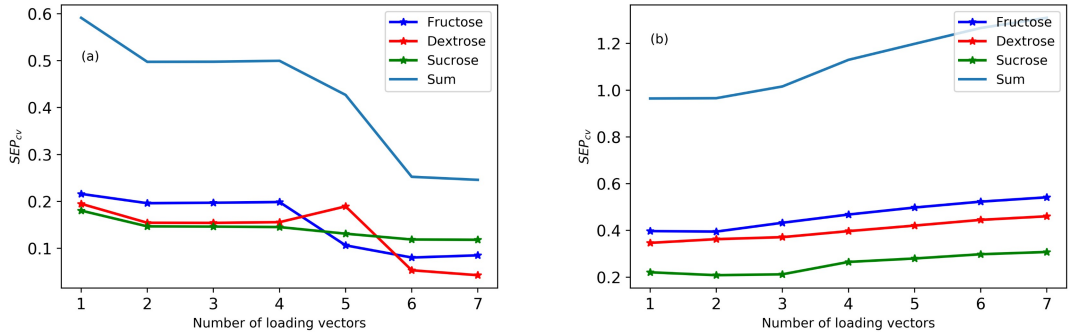

Figure 13. Error plots to show the best number of loading vectors for PCAR. Plot (a) shows unprocessed data, while plot (b) includes preprocessing steps.

Data Preprocessing Parameters

A variety of different preprocessing techniques were performed on both training sets, but most of these did not improve error and were therefore not utilized for regression. The one preprocessing step that did improve error was a spectral window selection as described above. The large bands corresponding to water did not contain much spectral information about the sugar components, while bands in the region 900-1600 cm^{-1} did reflect the sugar components. Figure 14 shows the full spectra including the large O-H stretching and bending band from water, and a zoomed ver-

sion showing only the bands from the components of interest.

In order to choose which specific wavelengths should be used to fit the data, several different wavelength ranges were chosen. These ranges were all used with PLS2 regression to calculate the *SEPcv* value for each range. Shown in Figure 15 is a graphical representation of this analysis. On the x-axis is the *SEPcv* value from PLS2 associated with each spectral window selection. PLS2 was arbitrarily chosen as it is likely that the regression models would all benefit from the same spectral window selection. However, this is an assumption and was not proven. Each *x*-value (standard error value) is associated with two *y*-values. These two *y*-values represent the wavenumbers at which to select the spectral window. From this plot, one can see how different components were predicted more accurately by using different wavelength ranges.

This figure also demonstrates how the different sugars benefit from using different wavelength ranges. In order to choose the best spectral window for fitting all sugars, the sum as well as the average of the standard errors were considered. By analyzing the minimum *SEPcv* for each subplot, the range 900-1300 cm^{-1} was chosen.

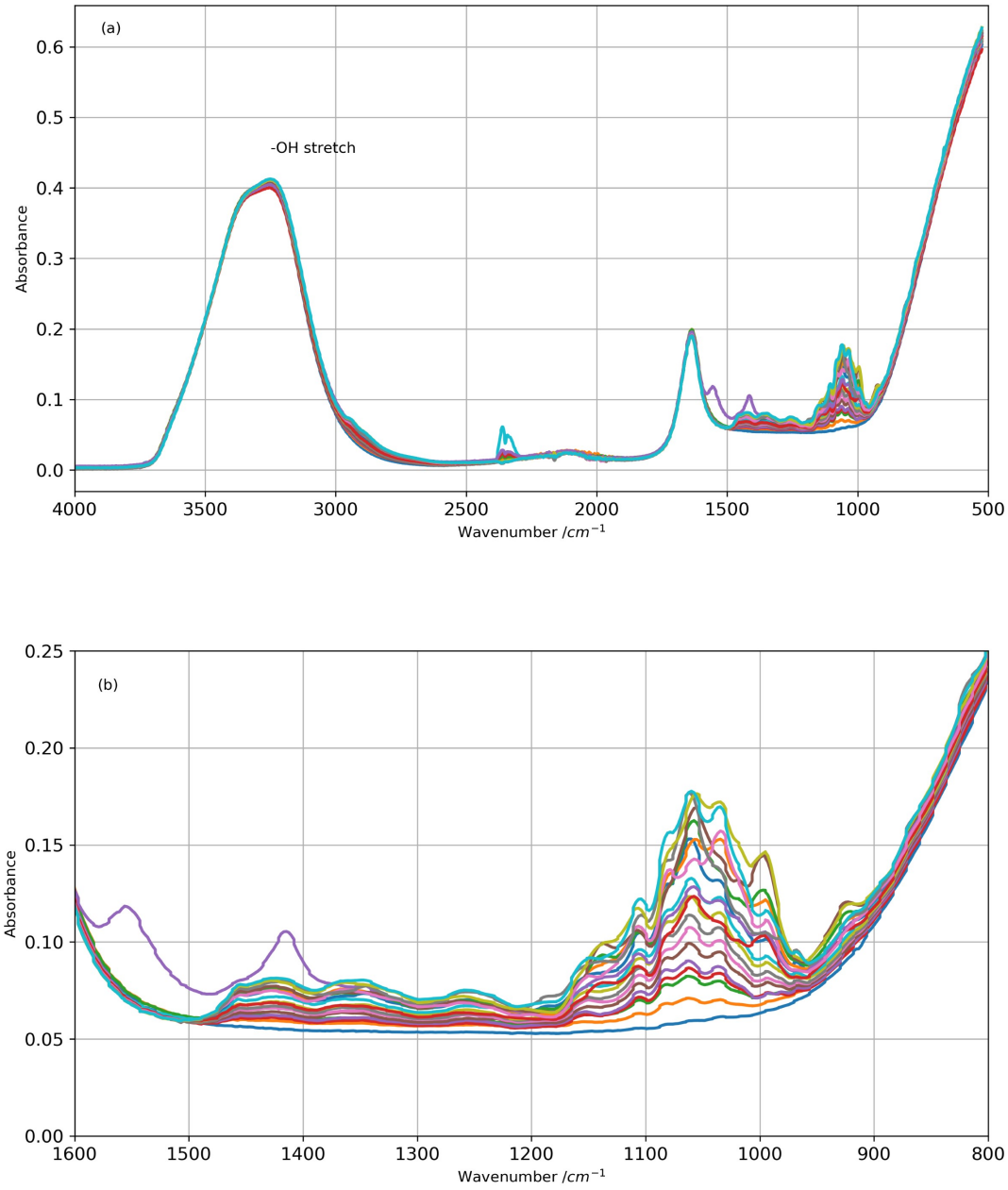

Figure 14. Infrared spectra of three-component training set. Plot (a) shows the full spectra, while plot (b) is zoomed in to only include bands from the sugar components. Regression used bands 900-1300 cm^{-1}.

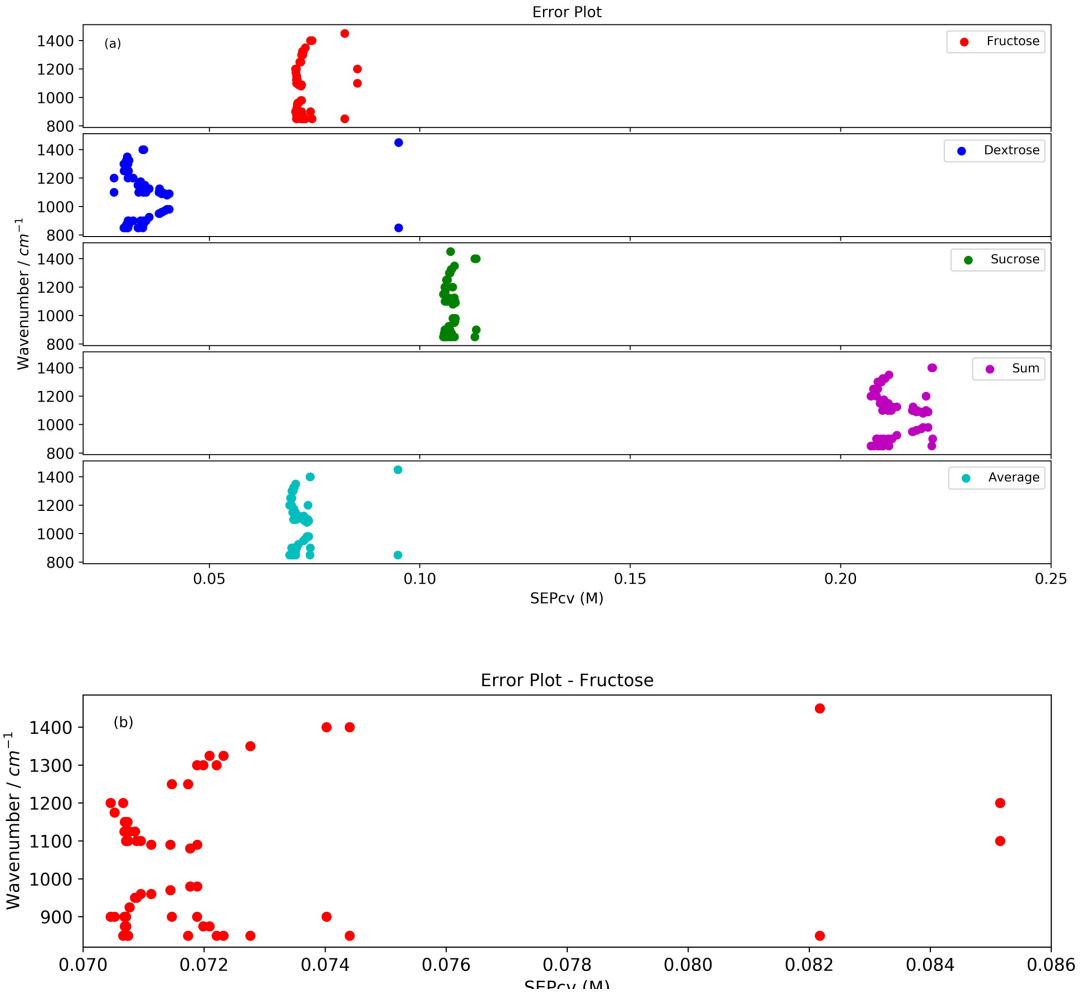

Figure 15. Error plot relating spectral window and *SEPcv* from PLS2. Plot (a) shows data for all components. Plot (b) shows the fructose plot in more detail.

CHAPTER FOUR: RESULTS AND DISCUSSION

The following section shows the results of each analysis performed. This is organized first by regression technique, then by training set (one-component mixtures and three-component mixtures). A comparison of the modeling results is presented as a summary at the end.

Univariate Classical Least Squares

One-Component Mixture

The univariate model of regression was an analysis at one wavelength. This data was plotted in Figure 3. The wavenumber 1055 cm^{-1} was chosen to make the calibration curve as it was the most intense band. The absorbance for each sample at 1055 cm^{-1} was plotted against fructose concentration, as shown in Figure 16. Then, a line of best fit was determined for use in regression. The equation of this line was

$$A = 0.02281(\text{mol }\%)^{-1}C - 0.01584\text{AU} \tag{21}$$

and this was used to calculate the *SEPcv* for this model. The *SEPcv* was 1.1826 and the *RSEPcv* was 12.23%. The plot shown in Figure 17 shows how the predicted values compare to the actual values. This is henceforth referred to as a linearity plot.

The errors from this analysis are not outrageously high. With only one component in the mixture, the chemical and spectral interference are minimized, and a traditional calibration curve works okay. However, the error is still a bit high for what is considered a high accuracy. The definition of high accuracy is dependent on what this modeling is used for, but for the purposes of this paper, we consider high accuracy to be greater 90 % or error to be less than 10 %. This error shows that the hydrogen bonding effects are already affecting the linearity of the data.

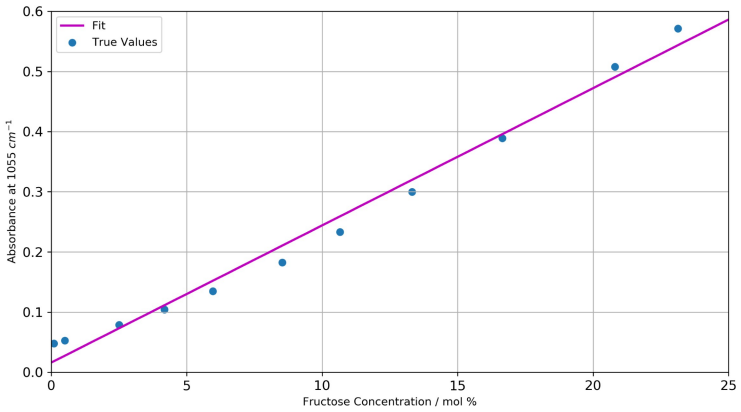

Figure 16. Calibration curve of fructose samples at 1055 cm^{-1}.

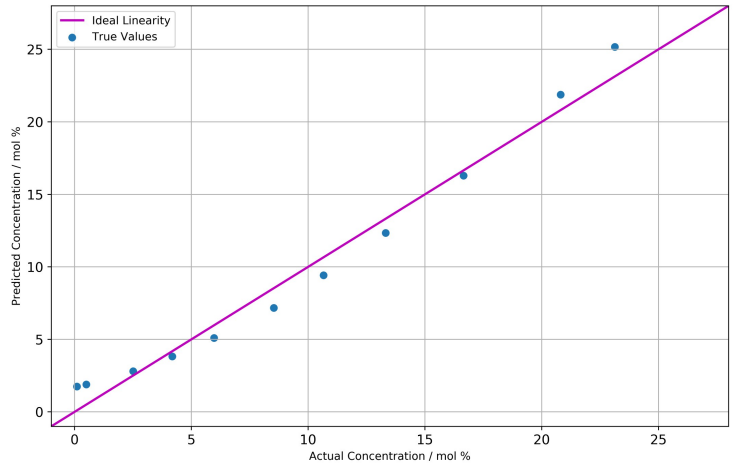

Figure 17. Linearity plot comparing true and predicted concentration for fructose from univariate CLSR. The perfect correlation line indicates where data would fall if the prediction were perfect.

Three-Component Mixtures

To expand on the previous example, univariate CLSR was used to predict three-component mixtures. To evaluate these mixtures, a separate calibration curve was made for each component, again using 1055 cm^{-1} as the wavenumber. These calibration curves are shown in Figure 18 and show a significant change from the one-component example, in that the data does not fall on the trendline whatsoever. Adding just two more components to the mixture removed almost all linearity as compared to the previous example. This decrease in accuracy is due to the increased amount of hydrogen bonding that occurs when more sugars are available to participate. This causes the wavenumber shifts and absorbance changes discussed above.

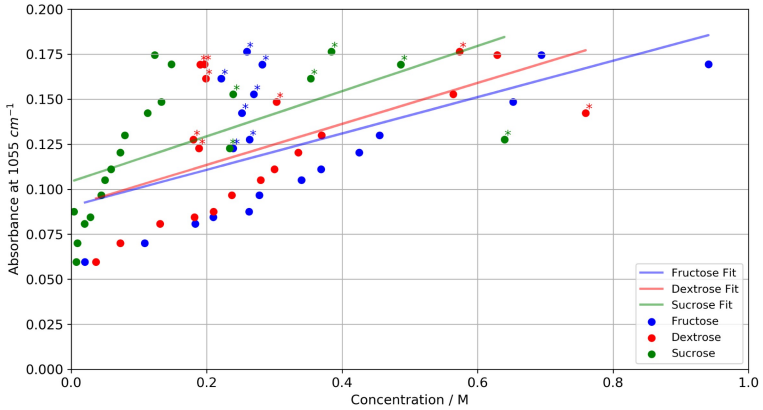

Figure 18. Calibration curves of 3-component samples at 1055 cm^{-1}. Starred datapoints indicate data that does not follow a linear fit.

The starred data points in Figure 18 indicate data that does not follow a linear fit, and is therefore negatively affecting the calibration curves. For fructose, these samples were S12-S17 and S19. For dextrose, those samples were S11 and S13-S19. The samples S12-S16 and S19 did not fit with sucrose. It is interesting to note that many of these samples were not linear for any sugar:

S13-16 and S19. All of these samples for the most part have similar concentrations for each component. When a mixture is mostly made up of one component, the absorbance is mostly a result of that component. That would allow the univariate CLSR (which assumes only one component) to model the data accurately. However, when the samples have almost equal contribution to absorbance from each component, univariate CLSR could not accurately model the data.

The equations for the trendlines are as follows for fructose, dextrose, and sucrose, respectively.

$$\text{Fructose:} A = 0.1009\text{M}C + 0.09058 \tag{22}$$

$$\text{Dextrose:} A = 0.1138\text{M}C + 0.09072 \tag{23}$$

$$\text{Sucrose:} A = 0.1255\text{M}C + 0.1043 \tag{24}$$

The *SEPcv* values were calculated using cross validation and were 0.2880 M, 0.2890 M, and 0.3351 M for fructose, dextrose, and sucrose, respectively. The *RSEPcv* values were 85.62%, 97.28%, and 207.6% (sum = 390.5%, mean = 130.2%). The linearity plot is shown in Figure 19. As shown, this model did not give accurate predictions. The *RSEPcv* values also dramatically increased as compared to the one-component mixture.

Three-Component Mixture with Interferents

To complicate the data even further, the model built from the three-component mixtures was used to predict a separate validation set. Some of the samples in this validation set contained various concentrations of interferents. The *SEPsv* values were calculated using separate validation and were 0.2389 M, 0.1916 M, and 0.1916 M for fructose, dextrose, and sucrose, respectively. The *RSEPsv* values were 117.2%, 122.4%, and 360.0% (sum = 599.7%, mean = 199.9%). The results of this analysis are shown in Table 4.

Based on the high error values, this model does not properly handle the presence of interferents that are not present in the training set. The blue cells in Table 4 indicate which samples and sugars were predicted with an R value $>$120%. Sucrose was part of this group for every single

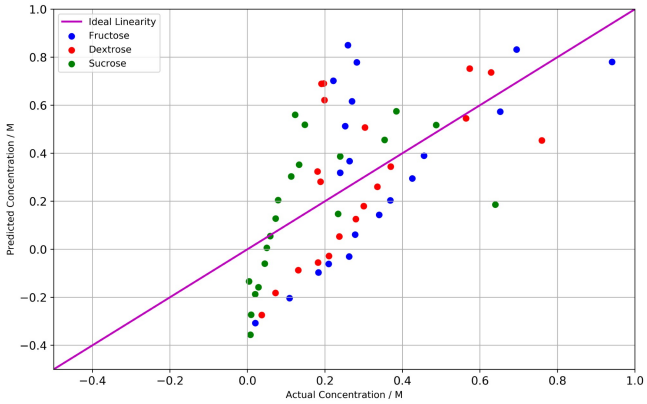

Figure 19. Linearity plot comparing actual and predicted concentrations from univariate CLSR.

Table 4. Predicted and actual concentrations from validation by univariate CLSR. All concentrations are molarity. Blue cells indicate a prediction with an R value $>120\%$.

Sample ID	**Fructose - Prediction**	**Fructose - Actual**	**Dextrose - Prediction**	**Dextrose - Actual**	**Sucrose - Prediction**	**Sucrose - Actual**
U1	-0.079697	0.2004	-0.071948	0.1600	-0.173191	0.0246
U2	0.074942	0.2501	0.065186	0.2990	-0.048795	0.0398
U3	0.110693	0.3010	0.096890	0.2624	-0.020035	0.0505
U4	-0.023543	0.1998	-0.022151	0.1615	-0.128019	0.0596
U5	0.149868	0.2489	0.131630	0.3026	0.011478	0.0797
U6	-0.057013	0.2027	-0.051831	0.1621	-0.154943	0.0251
U7	-0.124714	0.1843	-0.111869	0.1220	-0.209404	0.0202
U8	-0.127359	0.1809	-0.114214	0.1199	-0.211531	0.0205
U9	-0.110239	0.1800	-0.099032	0.1203	-0.197760	0.0201
U10	0.015929	0.1804	0.012854	0.0800	-0.096266	0.1244
U11	0.013043	0.1814	0.010294	0.0798	-0.098588	0.1205
U12	0.032348	0.1810	0.027414	0.0807	-0.083058	0.1196
U13	-0.109426	0.1801	-0.098311	0.1206	-0.197105	0.0200
U14	-0.119135	0.1813	-0.106921	0.1206	-0.204916	0.0205

sample, showing that this model struggled to predict sucrose the most. The samples with relatively lower errors were U3-U5, which had no interferents, and U10-U12, which had interferent

concentrations smaller than any of the sugar concentrations. Samples U10-U12 also had much higher sucrose concentrations than the others

Multivariate Classical Least Squares

One-Component Mixture

The multivariate CLSR model is an extension of univariate CLSR which utilizes multiple wavelengths as part of the model. It was expected that this model would predict the data poorly, as the univariate model did, due to its reliance on Beer's Law assumptions. However, this data was preprocessed as described and multivariate CLS has shown some promising results in studies of other types of samples.[49] The *SEPcv* value was 2.294 mol % with an *RSEPcv* value of 23.72 %. The linearity plot is shown in Figure 20.

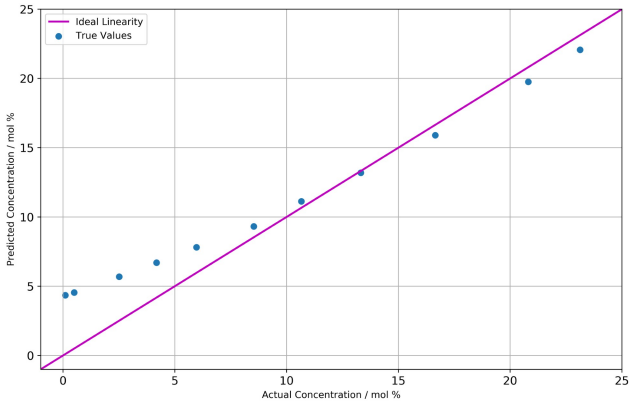

Figure 20. Linearity plot comparing actual and predicted concentrations from multivariate CLSR

Three-Component Mixture

The *SEPcv* values for each component were calculated, and are 0.3367 M, 0.0980 M, 0.1927 M for fructose, dextrose, and sucrose, respectively with *RSEPcv* values of 100.1%, 33.00%, and 119.4%(sum = 252.5%, mean = 84.16%). The linearity plot shown in Figure 21 shows graphi-

cally how inaccurate the predictions are. Clearly, the use of more wavelengths did not improve CLS enough for it to be useful. The presence of sugar hydrates just aren't accounted for with CLSR, so the errors are very high. When comparing the predictions of the one-component and three-component mixtures, it is clear that adding more sugars to the mixture causes the errors to be much higher. This is due to the increase in sugar hydrates.

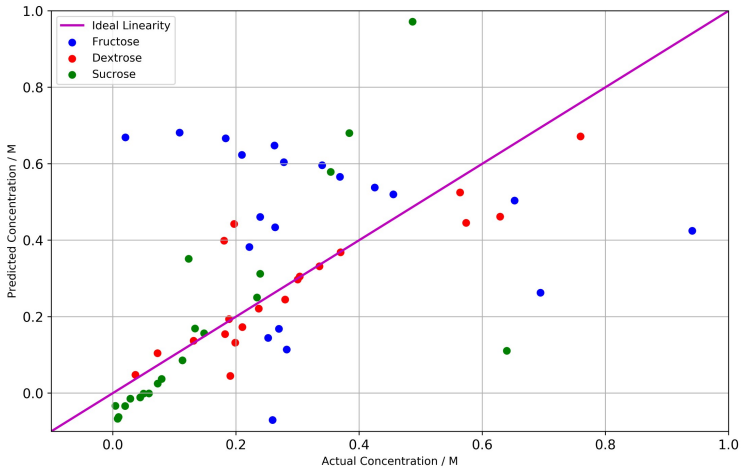

Figure 21. Linearity plot comparing actual and predicted concentrations from CLSR.

Three-Component Mixture with Interferents

This model was also used to predict the separate validation set. Shown in Table 5 are the predicted results for each component and the actual values. The *SEPsv* values were calculated using separate validation and were 0.4106 M, 0.0349 M, and 0.0872 M for fructose, dextrose, and sucrose, respectively. The *RSEPsv* values were 201.5%, 22.31%, and 164.0% (sum = 387.9%, mean = 129.3%). The high errors show how the presence of interferents in the training set and using a separate validation set can affect the analysis. This regression did not perform well, again because CLSR relies on the assumptions of Beer's Law which are violated with these types of

mixtures.

Table 5. Predicted and actual concentrations from separate validation by CLSR. All concentrations are molarity. Blue cells indicate a prediction with an R value $>120\%$.

Sample ID	Fructose - Prediction	Fructose - Actual	Dextrose - Prediction	Dextrose - Actual	Sucrose - Prediction	Sucrose - Actual
U1	0.631780	0.2004	0.187430	0.1600	-0.075851	0.0246
U2	0.555129	0.2501	0.311211	0.2990	-0.058460	0.0398
U3	0.578284	0.3010	0.266814	0.2624	-0.027641	0.0505
U4	0.604555	0.1998	0.176812	0.1615	-0.018034	0.0596
U5	0.515805	0.2489	0.301714	0.3026	0.014371	0.0797
U6	0.625727	0.2027	0.216806	0.1621	-0.081913	0.0251
U7	0.647515	0.1843	0.161364	0.1220	-0.081680	0.0202
U8	0.648740	0.1809	0.161912	0.1199	-0.085103	0.0205
U9	0.643401	0.1800	0.174731	0.1203	-0.085003	0.0201
U10	0.591521	0.1804	0.093034	0.0800	0.090908	0.1244
U11	0.594716	0.1814	0.095176	0.0798	0.083927	0.1205
U12	0.591528	0.1810	0.109855	0.0807	0.083430	0.1196
U13	0.644962	0.1801	0.176892	0.1206	-0.086311	0.0200
U14	0.647718	0.1813	0.167788	0.1206	-0.086360	0.0205

The blue cells in Table 5 indicate a prediction with an R value greater than 120%. None of the dextrose predictions fit this criteria, showing that dextrose was predicted the most accurately. All of the fructose predictions were predicted poorly, as well as most of the sucrose predictions. Samples U10-U12 had sucrose predicted accurately, which were some of the same samples that were predicted slightly better with univariate CLSR. These were the unknowns which had interferent concentrations lower than sugar concentrations, as well has higher concentrations of sucrose. It is possible that these high sucrose concentrations improved the prediction for sucrose.

Principal Component Regression

One-Component Mixture

This model worked quite well in predicting the one-component fructose mixtures. The *SEPcv* from cross validation was 0.2536 mol %, with a *RSEPcv* value of only 2.623%. Shown in Figure 22 is the linearity plot, which demonstrates how well this model worked. The values all fall extremely close to or on the perfect correlation line. This error is a marked improvement over CLSR, as PCAR can compensate for the hydrogen bonding network and presence of hydrates.

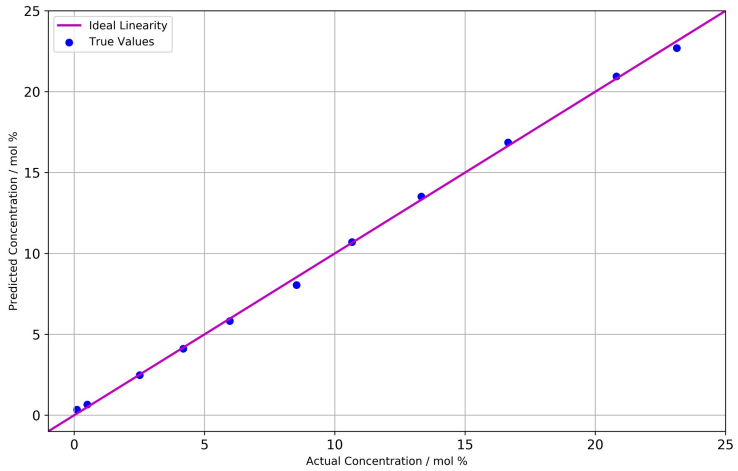

Figure 22. Linearity plot comparing actual and predicted concentrations of one-component mixtures from PCAR.

Three-Component Mixture

After performing this analysis on the three-component mixtures, the *SEPcv* values were found to be 0.1046 M, 0.1719 M, and 0.1237 M for fructose, dextrose, and sucrose. The *RSEPcv* values were 31.10%, 57.87%, and 76.61% (sum = 165.6%, mean = 55.19%). While these errors are still quite high, they are a notable improvement over the CLSR technique. The data in Figure 23

shows how predicted values compare to actual values.

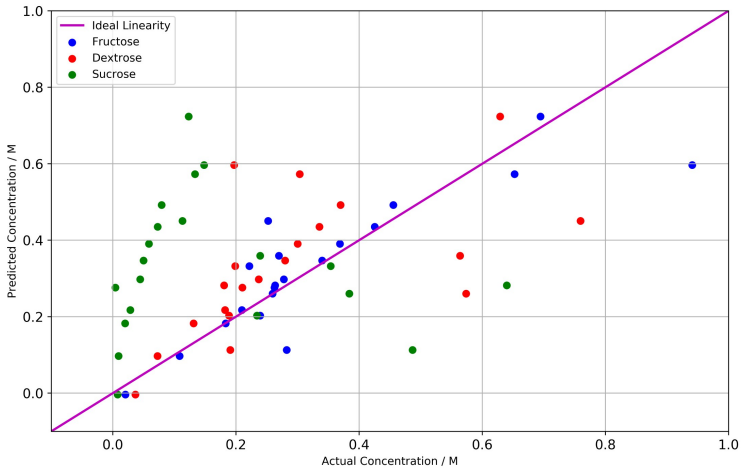

Figure 23. Linearity plot comparing actual and predicted concentrations from PCAR.

Three-Component Mixture with Interferents

When predicting the separate validation set, the *SEPsv* values were 0.0245 M, 0.0479 M and 0.0284 M for fructose, dextrose, and sucrose, respectively. The *RSEPsv* values were 12.01%, 30.61%, and 53.41% (sum = 96.03%, mean = 32.01%). These values are lower than the cross-validation, which is interesting as one would expect separate validation to give a higher error due to the interferents. This improvement shows that PCA handles interferent effects very well. The results of this regression are shown in Table 6. Once again, this model performed better than did CLSR, but is still not very accurate.

The blue cells in Table 6 indicate where R values are greater than 40%. Notice that this threshold is much lower than for CLSR, as the prediction has improved considerably. The most noticeable trend here is that PCAR predicted sucrose poorly except for samples U10-U12, which are the samples with the highest sucrose concentrations. This helps to show that these models do not

Table 6. Predicted and actual concentrations from separate validation by PCAR. All concentrations are molarity. Blue cells indicated an R value >40%.

Sample ID	Fructose - Prediction	Fructose - Actual	Dextrose - Prediction	Dextrose - Actual	Sucrose - Prediction	Sucrose - Actual
U1	0.200448	0.2004	0.135709	0.1600	0.055954	0.0246
U2	0.296342	0.2501	0.209071	0.2990	0.080566	0.0398
U3	0.317311	0.3010	0.217297	0.2624	0.075788	0.0505
U4	0.192761	0.1998	0.151273	0.1615	0.082213	0.0596
U5	0.283634	0.2489	0.228418	0.3026	0.115517	0.0797
U6	0.206475	0.2027	0.143451	0.1621	0.062069	0.0251
U7	0.170110	0.1843	0.115691	0.1220	0.052404	0.0202
U8	0.170353	0.1809	0.114568	0.1199	0.050732	0.0205
U9	0.183783	0.1800	0.122941	0.1203	0.051706	0.0201
U10	0.135198	0.1804	0.148470	0.0800	0.121829	0.1244
U11	0.140511	0.1814	0.148398	0.0798	0.117726	0.1205
U12	0.156266	0.1810	0.158047	0.0807	0.118740	0.1196
U13	0.185304	0.1801	0.123682	0.1206	0.051623	0.0200
U14	0.176358	0.1813	0.117816	0.1206	0.050550	0.0205

perform well in predicting mixtures with low sucrose concentrations. Fructose was all predicted with an R under the threshold, but dextrose had a few samples that were not: U2, U5, and U10-U12. In samples U10-U12, the dextrose concentrations are the lowest. These two trends may indicate that there is a problem with limit of quantitation. Samples U2 and U5, however, have the highest dextrose concentrations and were also predicted poorly. This could indicate a problem with the training set data not being robust enough.

Partial Least Squares Regression 2

One-Component Mixture

Similarly to PCAR, PLSR2 modeled the one-component mixture very well. The *SEPcv* was 0.2201 mol % and the *RSEPcv* value was 2.276%. These errors are slightly better than PCA, but these two models have comparable error values for this sample. Shown in Figure 24 is the linearity plot for this analysis. The data points are all very close to the ideal linearity line.

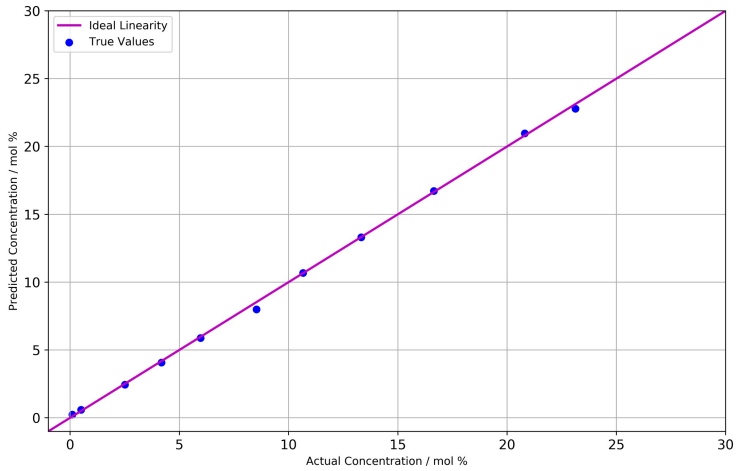

Figure 24. Linearity plot comparing actual and predicted concentrations of one-component mixtures from PLSR2.

Three-Component Mixture

The *SEPcv* values were 0.0719 M, 0.0306 M, and 0.1072 M for fructose, dextrose, and sucrose. The *RSEPcv* values were 21.37%, 10.31%, and 66.38% (sum = 98.07%, mean = 32.69%). This is an improvement over PCAR, but the errors are still larger than what would be ideal in a typical regression model. We expected both types of PLSR to perform better than PCAR, as PLSR optimizes the model by best describing the correlation between the absorbance and concentration, rather than best describing the variance. The results of PLSR2 can be seen in the linearity plot shown in Figure 25, where the data points now fall much closer to the ideal linearity line than they did with any previous model.

Three-Component Mixture with Interferents

When predicting the validation set, PLSR2 worked better than expected based on the cross validation errors. The *SEPsv* values were 0.01212 M, 0.01408 M, and 0.01474 M for fructose, dextrose, and sucrose. The *RSEPsv* values were 5.95%, 8.99%, and 27.7% (sum = 42.65%, mean

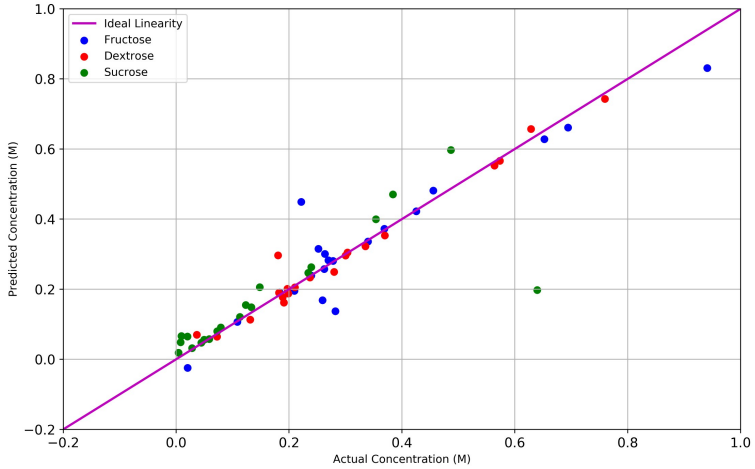

Figure 25. Linearity plot comparing actual and predicted concentrations from PLSR2.

= 14.22%). The errors for fructose and dextrose are $<$10%, which means a $>$90% accuracy for those components. The error for sucrose is a bit high at 27.7 %, but still much lower than in all the other models so far. This improvement from cross validation shows that PLSR2 can handle the interferent effects quite well.

The blue cells in Table 7 indicate where R values are greater than 15%. Again, this threshold has been lowered with the improvement of PLSR2 over PCAR. As shown in the table, both fructose and dextrose were all predicted with errors under the threshold. The problem is largely with sucrose, with only U2 and U3 being predicted accurately. These two samples do not appear to be much different from samples U1 and U4-U5, so it is unclear what is causing this improvement.

Partial Least Squares Regression 1

One-Component Mixture

The difference between PLSR1 and PLSR2 is that PLSR1 evaluates one component at a time, while PLSR2 evaluates all components at once. Therefore, for a one-component mixture, these techniques are equivalent.

Table 7. Predicted and actual concentrations from separate validation by PLSR2. All concentrations are molarity. Blue cells indicate where R values were $>15\%$.

Sample ID	Fructose - Prediction	Fructose - Actual	Dextrose - Prediction	Dextrose - Actual	Sucrose - Prediction	Sucrose - Actual
U1	0.189983	0.2004	0.158922	0.1600	0.037073	0.0246
U2	0.264502	0.2501	0.282154	0.2990	0.046110	0.0398
U3	0.305000	0.3010	0.244975	0.2624	0.058307	0.0505
U4	0.190528	0.1998	0.155273	0.1615	0.071553	0.0596
U5	0.259620	0.2489	0.282563	0.3026	0.088716	0.0797
U6	0.196462	0.2027	0.170690	0.1621	0.051012	0.0251
U7	0.163657	0.1843	0.130247	0.1220	0.037098	0.0202
U8	0.162832	0.1809	0.130894	0.1199	0.034630	0.0205
U9	0.174946	0.1800	0.142855	0.1203	0.034876	0.0201
U10	0.162671	0.1804	0.083562	0.0800	0.139280	0.1244
U11	0.166639	0.1814	0.086078	0.0798	0.133629	0.1205
U12	0.182983	0.1810	0.096379	0.0807	0.136875	0.1196
U13	0.177612	0.1801	0.142075	0.1206	0.036863	0.0200
U14	0.168412	0.1813	0.135260	0.1206	0.035051	0.0205

Three-Component Mixture

The errors for PLSR1 were expected to be less than, but still comparable to the errors from PLSR2. The reason for this is the ability to choose an ideal number of loading vectors for each individual component with PLSR1. The *SEPcv* values for PLSR1 were 0.06551 M, 0.03021 M, and 0.1089 M for fructose, dextrose, and sucrose. The *RSEPcv* values were 19.48%, 10.17%, and 67.45% (sum = 97.09%, mean = 32.36%). The predictions can be seen in the linearity plot shown in Figure 26. These values are almost identical to those of PLSR2. For fructose and dextrose, the errors were just slightly better. However, the error for sucrose is a bit higher with PLSR1.

Three-Component Mixture with Interferents

For separate validation, the *SEPsv* values were 0.01054 M, 0.01416 M, and 0.01521 M for fructose, dextrose, and sucrose. The *RSEPsv* values were 5.18%, 9.04%, and 28.6% (sum = 42.79%, mean = 14.26%). These values are all comparable to the PLSR2 values. Shown in Table 8 are the

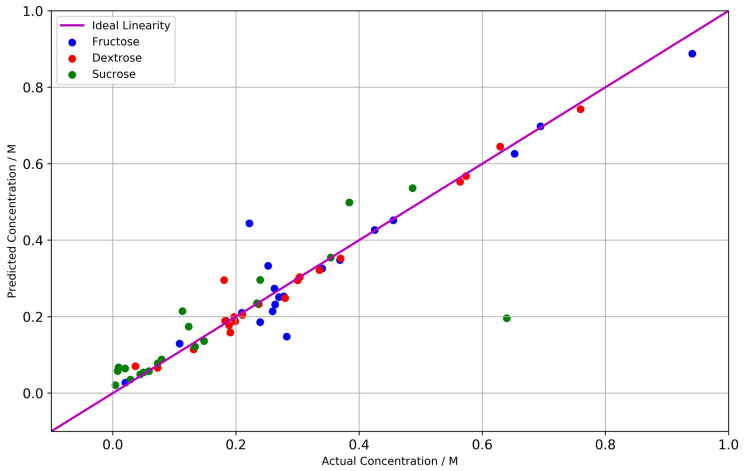

Figure 26. Linearity plot comparing actual and predicted concentrations from PLSR1.

results from this analysis. The blue cells indicate where R values were greater than 15%. Like with PLSR2, fructose and dextrose were all predicted with low errors. Sucrose was mainly predicted with an error over the threshold value, with the exception of samples U10-U12. These are the same samples that have been discussed previously with the highest sucrose concentrations.

Summary

To compare these techniques, two summary tables listing the *SEPcv* and *RSEPcv* values were generated: one for the one-component mixtures and one for the three-component mixtures. Another table was generated with the *SEPsv* and *RSEPsv* values for the mixtures with interferents.

In Table 9 is shown the results of the one-component mixture. In this case, PLSR1 and PLSR2 are equivalent, so PLSR1 is excluded from the table. Based on the *RSEPcv* values, it is clear that the multivariate CLSR performed the worst, with univariate CLSR performing the second worst. This is different from expected, as multivariate CLSR looked at more spectral data. It is possible that since the preprocessing steps were optimized to improve the three-component mixtures, that the chosen spectral window selection is what increased the error in the multivariate calculation.

Table 8. Results of separate validation by PLSR1. All concentrations are molarity. Blue cells indicate where R values are $>15\%$.

Sample ID	Fructose - Prediction	Fructose - Actual	Dextrose - Prediction	Dextrose - Actual	Sucrose - Prediction	Sucrose - Actual
U1	0.214375	0.2004	0.158504	0.1600	0.038138	0.0246
U2	0.265787	0.2501	0.281826	0.2990	0.055302	0.0398
U3	0.312740	0.3010	0.244470	0.2624	0.058728	0.0505
U4	0.207586	0.1998	0.154964	0.1615	0.070553	0.0596
U5	0.257503	0.2489	0.282329	0.3026	0.096270	0.0797
U6	0.195238	0.2027	0.171359	0.1621	0.055486	0.0251
U7	0.183216	0.1843	0.129948	0.1220	0.037300	0.0202
U8	0.174076	0.1809	0.130642	0.1199	0.035270	0.0205
U9	0.177705	0.1800	0.142575	0.1203	0.035861	0.0201
U10	0.157655	0.1804	0.083655	0.0800	0.129614	0.1244
U11	0.169085	0.1814	0.086136	0.0798	0.124126	0.1205
U12	0.173859	0.1810	0.096573	0.0807	0.127495	0.1196
U13	0.183009	0.1801	0.141948	0.1206	0.037875	0.0200
U14	0.183698	0.1813	0.135159	0.1206	0.036208	0.0205

PCAR and PLSR had comparable *RSEPcv* values, with PLSR performing slightly better.

Table 9. Summary of results from all regression models on one-component mixtures.

Technique	Fructose *SEPcv* / mol %	Fructose *RSEPcv* / mol %
Univariate CLSR	1.1826	12.23
Multivariate CLSR	2.2936	23.72
PCAR	0.2536	2.623
PLSR2	0.2201	2.276

The results from cross-validation of the three-component mixtures are shown in Table 10. For these mixtures, the PCAR did not perform very well with errors all greater than 31 %, but PCAR was still better than CLSR. The two PLSR methods performed rather well, though both had problems with predicting sucrose concentrations. The results from separate validation in Table 11

show the same results: both PLSR methods worked well for this analysis.

Table 10. Summary of results from all regression models on three-component mixtures. *SEPcv* values have units of molarity, while *RSEPcv* values are percentages.

Technique	**Fructose *SEPcv***	**Fructose *RSEPcv***	**Dextrose *SEPcv***	**Dextrose *RSEPcv***	**Sucrose *SEPcv***	**Sucrose *RSEPcv***
Univariate CLSR	0.2880	85.62	0.2890	97.23	0.3351	207.6
Multivariate CLSR	0.3367	100.1	00980	33.00	0.1927	119.4
PCAR	0.1046	31.10	0.1719	57.87	0.1237	76.61
PLSR2	0.0718	21.37	0.0306	10.31	0.1072	66.38
PLSR1	0.0656	19.48	0.0302	10.17	0.1089	67.45

The consistent problem with predicting sucrose likely stems from the fact that in the training set, about half of the sucrose concentrations are quite low. With an over-abundance of low sucrose concentrations, the models did not have as diverse a range of information to work with. As stated earlier, it is important to have varying ratios of concentrations from all components to explain a greater variety of data in the training set. This problem with predicting sucrose could potentially be fixed by adding more samples to the training set that have more diverse ratios.

Alternatively, the sucrose problem could be due to the fact that sucrose is a disaccharide while fructose and dextrose are monosaccharides. The larger structure of sucrose could be affecting the spectral interference more so than with the monosaccharides. It was found that using 2 loading vectors gave the lowest *SEPcv* value for sucrose. It is possible that this is because sucrose is a combination of the other two components (fructose and dextrose) and the model struggled to separate sucrose from its two monosaccharides.

Table 11. Summary of results from all regression models on 3-component mixtures with interferents. All *SEPsv* values are in units of molarity and all *RSEPsv* values are percentages.

Technique	**Fructose *SEPcv***	**Fructose *RSEPcv***	**Dextrose *SEPcv***	**Dextrose *RSEPcv***	**Sucrose *SEPcv***	**Sucrose *RSEPcv***
Univariate CLSR	0.2389	117.2	0.1916	122.4	0.1916	360.0
Multivariate CLSR	0.4106	201.5	0.0349	22.31	0.0872	164.01
PCAR	0.0245	12.01	0.0479	30.61	0.0284	53.41
PLSR2	0.0121	5.95	0.0141	8.99	0.0147	27.7
PLSR1	0.0105	5.18	0.0142	9.04	0.0152	28.6

CHAPTER FIVE: CONCLUSIONS AND FUTURE DIRECTIONS

It was confirmed that CLSR is not a good choice for analysis of aqueous sugar mixtures. The reason is due to its reliance on the assumptions of Beer's Law, which are violated in these mixtures. The univariate CLSR, which is what chemists traditionally use in quantitative spectroscopy, provided the highest errors of all the techniques. Multivariate CLSR performed slightly better, but still gave extremely high errors. CLSR serves to demonstrate the problem that this project sought to address: the assumptions of Beer's Law are violated in aqueous sugar mixtures.

PCAR performed significantly better than CLSR, especially for the one-component mixture. For the three-component mixtures, the errors were a bit higher than is acceptable (10%), but the one-component mixture was predicted very accurately (>5%). This difference is likely due to the lower concentrations of sugar in the one-component sample forming less hydrates and altering the spectra less. Adding more components to predict further complicated the mixture, and the violations of Beer's Law were more pronounced.

The concentrations of sugars in all samples were varied in relation to each other, so an increase in concentration of one component does not correlate with concentrations of the other components. Because of this, it was expected that PLSR1 would perform better than PLSR2, as PLSR2 looks at all components at once. Additionally, PLSR1 has the benefit of varying the number of loading vectors for each component to model each one separately. However, both of these models performed at about the same level. PLSR1 had slightly lower errors for some components, but PLSR2 was had lower errors for the others. These two models predicted with the lowest error compared to the other techniques, so they should be chosen for future studies. This result was anticipated, since PLS works to calculate the loading vectors to best explain the relationship between the two variables (concentration and absorbance), as opposed to PCA which calculates loading vectors to explain the greatest variance in the data.

The issue with predicting sucrose could possibly be fixed by expanding the training set to in-

clude more samples with higher concentrations of sucrose as compared to fructose and dextrose. It would be interesting to see if this decreases error and if the ideal number of loading vectors for sucrose by PLSR1 would increase from two.

With the success of this project proving that this type of analysis is possible with decent accuracy for fructose and dextrose ($RSEP_{cv}$ $<22\%$ and $RSEP_{sv}$ $<10\%$), future projects can be done which analyze more complex mixtures. These could include food and beverages, human bodily fluids, plant products, insects, and more. The ultimate goal of this research is to make these types of analyses easier to perform and more accessible to scientists of all disciplines. With more comprehensive sugar standards (i.e. more types of sugars), a more widely applicable model could be developed. For example, the addition of lactose could help with analysis of non-lactose food products. Analysis of the monosaccharides galactose and mannose is important for biochemical research as they are synthesized by the body for a variety of biochemical processes. The analysis of maltose can help with production of fermented beverages like beer.

Additionally, this analysis could potentially be automated for greater ease of use and accessibility. IR spectroscopy is capable of remote operation, so the need for taking aliquots of samples and sending them to a lab would not be necessary. With a working model already in place in the software, spectral analysis and prediction of components could be performed almost immediately and software could output the predictions to a technician. This advantage would greatly speed up and simplify these analyses.

REFERENCES

[1] Ačanski, M. M.; Vujić, D. N. Comparing sugar components of cereal and pseudocereal flour by GC-MS analysis. *Food Chemistry* **2014**, *145*, 743–748.

[2] Psodorov, D.; Acanski, M.; Vujic, D.; Brkljaca, J.; Psodorov, D. Homogenity of oil and sugar components of flour amaranth investigated by GC-MS. *Chemical Industry and Chemical Engineering Quarterly* **2015**, *21*, 71–76.

[3] Terrab, A.; Vega-Pérez, J. M.; Díez, M. J.; Heredia, F. J. Characterisation of northwest Moroccan honeys by gas chromatographic-mass spectrometric analysis of their sugar components. *Journal of the Science of Food and Agriculture* **2002**, *82*, 179–185.

[4] Craig, A. P.; Fields, C. C.; Simpson, J. V. Development of a gas chromatography-mass spectrometry method for the quantification of glucaric acid derivatives in beverage substrates. *International Journal of Analytical Chemistry* **2014**, *2014*, 402938.

[5] Xu, W.; Liang, L.; Zhu, M. Determination of sugars in molasses by HPLC following solid-phase extraction. *International Journal of Food Properties* **2015**, *18*, 547–557.

[6] Özkök, A.; Sorkun, K.; D'Arcy, B. Sugar analysis of pine honey from Mugla region using HPLC. *Mellifera* **2014**, *32*, 27–32.

[7] Young, J. E.; Josephy, J.; Matyska, M. T. Robust HPLC-refractive index analysis of simple sugars in beverages using silica hydride columns. *Current Nutrition and Food Science* **2016**, *12*, 125–131.

[8] Di Egidio, V.; Sinelli, N.; Giovanelli, G.; Moles, A.; Casiraghi, E. NIR and MIR spectroscopy as rapid methods to monitor red wine fermentation. *European Food Research and Technology* **2010**, *230*, 947–955.

[9] Detrain, C.; Verheggen, F. J.; Diez, L.; Wathelet, B.; Haubruge, E. Aphid-ant mutualism: How honeydew sugars influence the behaviour of ant scouts. *Physiological Entomology* **2010**, *35*, 168–174.

[10] Montesano, D.; Cossignani, L.; Giua, L.; Urbani, E.; Simonetti, M. S.; Blasi, F. A simple HPLC-ELSD method for sugar analysis in goji berry. *Journal of Chemistry* **2016**, *2016*, 6271808.

[11] Varandas, S.; Teixeira, M. J.; Marques, J. C.; Aguiar, A.; Alves, A.; Bastos, M. M. Glucose and fructose levels on grape skin: Interference in *Lobesia botrana* behaviour. *Analytica Chimica Acta* **2004**, *513*, 351–355.

[12] Kurt, A.; Torun, H.; Colak, N.; Seiler, G.; Hayirlioglu-Ayaz, S.; Ayaz, F. A. Nutrient profiles of the hybrid grape cultivar 'Isabel' during berry maturation and ripening. *Journal of the Science of Food and Agriculture* **2017**, *97*, 2468–2479.

[13] Pietrogrande, M. C.; Bacco, D. GC-MS analysis of water-soluble organics in atmospheric aerosol: Response surface methodology for optimizing silyl-derivatization for simultaneous analysis of carboxylic acids and sugars. *Analytica Chimica Acta* **2011**, *689*, 257–264.

[14] Pietrogrande, M. C.; Bacco, D.; Chiereghin, S. GC/MS analysis of water-soluble organics in atmospheric aerosol: Optimization of a solvent extraction procedure for simultaneous analysis of carboxylic acids and sugars. *Analytical and Bioanalytical Chemistry* **2013**, *405*, 1095–1104.

[15] Pietrogrande, M. C.; Bacco, D.; Mercuriali, M. GC-MS analysis of low-molecular-weight dicarboxylic acids in atmospheric aerosol: Comparison between silylation and esterification derivatization procedures. *Analytical and Bioanalytical Chemistry* **2010**, *396*, 877–885.

[16] Kačuráková, M.; Capek, P.; Sasinkova, V.; Wellner, N.; Ebringerova, A. FT-IR study of

plant cell wall model compounds: pectic polysaccharides and hemicelluloses. *Carbohydrate Polymers* **2000**, *43*, 195–203.

[17] Grube, M.; Bekers, M.; Upite, D.; Kaminska, E. IR-spectroscopic studies of *Zymomonas mobilis* and levan precipitate. *Vibrational Spectroscopy* **2002**, *28*, 277–285.

[18] Marcotte, L.; Kegelaer, G.; Sandt, C.; Barbeau, J.; Lafleur, M. An alternative infrared spectroscopy assay for the quantification of polysaccharides in bacterial samples. *Analytical Biochemistry* **2007**, *361*, 7–14.

[19] Trollope, K. M.; Volschenk, H.; Görgens, J. F.; Bro, R.; Nieuwoudt, H. H. Direct, simultaneous quantification of fructooligosaccharides by FT-MIR ATR spectroscopy and chemometrics for rapid identification of superior, engineered β-fructofuranosidases. *Analytical and Bioanalytical Chemistry* **2014**, *407*, 1661–1671.

[20] Coimbra, M. A.; Barros, A.; Barros, M.; Rutledge, D. N.; Delgadillo, I. Multivariate analysis of uronic acid and neutral sugars in whole pectic samples by FT-IR spectroscopy. *Carbohydrate Polymers* **1998**, *37*, 241–248.

[21] Wang, Y.; Rong, R.; Chen, H.; Zhu, M.; Wang, B.; Li, X. Triazole-linked fluorescent bisboronic acid capable of selective recognition of the Lewis Y antigen. *Bioorganic & Medicinal Chemistry Letters* **2017**, *27*, 1983–1988.

[22] Nait Chabane, Y.; Marti, S.; Rihouey, C.; Alexandre, S.; Hardouin, J. Characterisation of pellicles formed by *Acinetobacter baumannii* at the air-liquid interface. *PLoS ONE* **2014**, *9*.

[23] Blazey, T.; Snyder, A. Z.; Goyal, M. S.; Vlassenko, A. G.; Raichle, M. E. A systematic meta-analysis of oxygen-to-glucose and oxygen-to-carbohydrate ratios in the resting human brain. *PLoS ONE* **2018**, *13*.

[24] Ferrer, I.; Thurman, M.; Churley, M.; Prest, H.; Stremple, P. Analysis of phytoestrogens in soy milk by GC / MS / MS with the Agilent 7000 Series Triple Quadrupole GC / MS. *Agilent Technologies* **2009**, 1–8.

[25] Serrano, S.; Villarejo, M.; Espejo, R.; Jodral, M. Chemical and physical parameters of Andalusian honey: Classification of citrus and eucalyptus honeys by discriminant analysis. *Food Chemistry* **2004**, *87*, 619–625.

[26] Park, H.-S.; Jun, S.-C.; Han, K.-H.; Hong, S.-B.; Yu, J.-H. *Advances in Applied Microbiology*; Academic Press, 2017; Vol. 100; pp 161–202.

[27] Torres-Gamez, J.; Rodriguez, J. A.; Elena Paez-Hernandez, M.; Galan-Vidal, C. A. Application of multivariate statistical analysis to simultaneous spectrophotometric enzymatic determination of glucose and cholesterol in serum samples. *International Journal of Analytical Chemistry* **2019**, 7532687.

[28] Pacifique Mutuyimana, F.; Liu, J.; Na, M.; Nsanzamahoro, S.; Rao, Z.; Chen, H.; Chen, X. Synthesis of orange-red emissive carbon dots for fluorometric enzymatic determination of glucose. *Microchimica Acta* **2018**, *185*, 518–528.

[29] Velterop, J. S.; Vos, F. A rapid and inexpensive microplate assay for the enzymatic determination of glucose, fructose, sucrose, L-malate and citrate in tomato (*Lycopersicon esculentum*) extracts and in orange juice. *Phytochemical Analysis* **2001**, *12*, 299–304.

[30] Gordon, R.; Chapman, J.; Power, A.; Chandra, S.; Roberts, J.; Cozzolino, D. Unfrazzled by fizziness: Identification of beers using attenuated total reflectance mid-infrared spectroscopy and multivariate analysis. *Food Analytical Methods* **2018**, 2360–2367.

[31] Grassi, S.; Amigo, J. M.; Lyndgaard, C. B.; Foschino, R.; Casiraghi, E. Assessment of the sugars and ethanol development in beer fermentation with FT-IR and multivariate curve resolution models. *Food Research International* **2014**, *62*, 602–608.

[32] García-González, D. L.; Sedman, J.; Van De Voort, F. R. Principles, performance, and applications of spectral reconstitution (SR) in quantitative analysis of oils by Fourier transform infrared spectroscopy (FT-IR). *Applied Spectroscopy* **2013**, *67*, 448–456.

[33] Beer, J. Determination of the absorption of red light in colored liquids. *Annalen der Physik und Chemie* **1852**, *86*, 78–88.

[34] Haaland, D. M.; Thomas, E. V. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry* **1988**, *60*, 1193–1202.

[35] Shiraga, K.; Adachi, A.; Ogawa, Y. Characterization of the hydrogen-bond network of water around sucrose and trehalose: H-O-H bending analysis. *Chemical Physics Letters* **2017**, *678*, 59–64.

[36] Max, J. J.; Chapados, C. Glucose and fructose hydrates in aqueous solution by IR spectroscopy. *Journal of Physical Chemistry A* **2007**, *111*, 2679–2689.

[37] Chen, C.; Li, W. Z.; Song, Y. C.; Weng, L. D.; Zhang, N. Formation of water and glucose clusters by hydrogen bonds in glucose aqueous solutions. *Computational and Theoretical Chemistry* **2012**, *984*, 85–92.

[38] Brizuela, A. B.; Bichara, L. C.; Romano, E.; Yurquina, A.; Locatelli, S.; Brandán, S. A. A complete characterization of the vibrational spectra of sucrose. *Carbohydrate Research* **2012**, *361*, 212–218.

[39] Brizuela, A. B.; Castillo, M. V.; Raschi, A. B.; Davies, L.; Romano, E.; Brandán, S. A. A complete assignment of the vibrational spectra of sucrose in aqueous medium based on the SQM methodology and SCRF calculations. *Carbohydrate Research* **2014**, *388*, 112–124.

[40] Márquez, M. J.; Brizuela, A. B.; Davies, L.; Brandán, S. A. Spectroscopic and structural studies on lactose species in aqueous solution combining the HATR and Raman spectra with SCRF calculations. *Carbohydrate Research* **2015**, *407*, 34–41.

[41] Fearn, T. Classical least squares. *Chemometric Space* **2010**, *21*, 16–17.

[42] Moore, E. H. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society* **1920**, *26*, 394–395.

[43] Penrose, R. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society* **1955**, *51*, 406–413.

[44] Rosipal, R.; Krämer, N. In *Lecture Notes in Computer Science - Subspace, Latent Structure and Feature Selection*, 3940th ed.; Saunders, C., Gunn, S., Shawe-Taylor, J., Grobelnik, M., Eds.; Springer-Verlag Berlin Heidelberg, 2006; pp 34–51.

[45] Tanaka, M.; Kojima, T. Near-Infrared Monitoring of the Growth Period of Japanese Pear Fruit Based on Constituent Sugar Concentrations. *Journal of Agricultural and Food Chemistry* **1996**, *44*, 2272–2277.

[46] Miranda, A.; Pereira, V.; Pontes, M.; Albuquerque, F.; Marques, J. C. Acetic acid and ethyl acetate in Madeira wines: Evolution with ageing and assessment of the odour rejection threshold. *Ciência e Técnica Vitivinícola* **2017**, *32*, 1–11.

[47] Mohammadzadeh-Aghdash, H.; Sohrabi, Y.; Mohammadi, A.; Shanehbandi, D.; Dehghan, P.; Ezzati, J.; Dolatabadi, N. Safety assessment of sodium acetate, sodium diacetate and potassium sorbate food additives. *Food Chemistry* **2018**, *257*, 211–215.

[48] Savitzky, A.; Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* **1964**, *36*, 1627–1639.

[49] Mahamuni, N. N.; Adewuyi, Y. G. Fourier transform infrared spectroscopy (FTIR) method to monitor soy biodiesel and soybean oil in transesterification reactions, petrodiesel-biodiesel blends, and blend adulteration with soy oil. *Energy and Fuels* **2009**, *23*, 3773–3782.