

FT-IR SPECTROMETRY AND CHEMOMETRICS FOR SPECTRAL DISCRIMINATION OF COTTON CONTAMINANTS

by

JAMES BRIAN LOUDERMILK

(Under the Direction of James A. de Haseth)

ABSTRACT

During the cotton harvest, other parts of the cotton plants in addition to the cotton fibers are collected. This plant debris causes a number of problems in cotton processing, including decreased yarn quality and increased number of yarn breakages during spinning. Because these debris break down in size physically after harvest, human visual identification of and discrimination among these contaminants are difficult or impossible, but identification of the contaminant type is desirable for investigations of process breakdowns and efficiency monitoring of cleaning processes. To achieve contaminant identification and discrimination, a number of projects that explore spectral comparison in general and develop solutions to the specific problem of cotton contaminant discrimination is reported. Chemometric methods for contaminant class discrimination have been developed for and applied to the FT-IR attenuated total reflection (ATR) spectra of cotton contaminants. Novel voting scheme algorithms were developed to improve spectral identification by library searching for a USDA spectral library of cotton contaminants. Improvements in contaminant identification were also

realized via partial least squares discriminant analysis (PLS-DA). Quantitative analysis of cotton contaminant mixtures was achieved with the use of partial least squares (PLS) regression and a novel error correction algorithm that was developed. This work also reports the development of a mixture generator algorithm to generate sets of mixtures representative of mixture spaces of arbitrary dimensions. Finally, the spectral differences caused by the use of different FT-IR spectrometers and ATR accessories were investigated by measuring spectra of a polyethylene terephthalate film with the use of several different spectrometers and accessories. The spectra were compared before and after corrections for depth of penetration and anomalous dispersion effects. The results show that these correction methods do not always achieve the goal of increased spectral similarity.

INDEX WORDS: Chemometrics, FT-IR, ATR, cotton, contaminants, spectral library, spectral search system, PLS, PLS-DA, voting scheme algorithms, quantitative analysis, calibration, mixtures, experimental design, depth of penetration, anomalous dispersion

FT-IR SPECTROMETRY AND CHEMOMETRICS FOR SPECTRAL DISCRIMINATION
OF COTTON CONTAMINANTS

by

JAMES BRIAN LOUDERMILK

B.S., North Georgia College and State University, 2004

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in
Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008

© 2008

James Brian Loudermilk

All Rights Reserved

FT-IR SPECTROMETRY AND CHEMOMETRICS FOR SPECTRAL DISCRIMINATION
OF COTTON CONTAMINANTS

by

JAMES BRIAN LOUDERMILK

Major Professor: James A. de Haseth

Committee: James L. Anderson
Lionel A. Carreira
David S. Himmelsbach

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2008

DEDICATION

This work is dedicated to my wife. Beth, you have been a constant source of inspiration and support throughout these last four years. Thank you for your love and for making all of this possible.

ACKNOWLEDGEMENTS

I begin by thanking my mentor and advisor, Prof. James de Haseth, for his assistance and encouragement. Under his guidance, I have gone from student to scientist, and I am honored to have studied under him these last four years. I also thank Drs. David S. Himmelsbach and Franklin E. Barton, II, for giving me opportunities that few graduate students experience. I thank Dr. Brad Herbert and Mr. Garry McGlaun for their friendship. Without their support and encouragement, I would not have begun this journey. I also thank all of my advisory committee members including Profs. James Anderson and Lionel Carreira for their work. Of course, I have only named a few of those who have helped me along the way, but I thank everyone who has been there to guide me throughout the many years of education in both science and life.

I especially thank my wife Beth. Without her, this work would have been meaningless. She deserves an enormous prize for the love and support she has shown me throughout these four years. I also thank my parents, Henry and Joyce. They often receive little credit for the work I do, but they helped to lay the foundation upon which all has been built. Family and friends too numerous to list deserve to be mentioned here. (Jasper, this includes you.) Let me at least say thank you for your love and support throughout everything.

Above everyone else, I thank God for His constant provision. If I have done anything worthwhile or have gained anything of value, it is only because He has blessed me, and knowing Christ has been the greatest blessing.

TABLE OF CONTENTS

| | Page |
|---|------|
| ACKNOWLEDGEMENTS | v |
| LIST OF TABLES | x |
| LIST OF FIGURES | xii |
| CHAPTER | |
| 1 INTRODUCTION AND LITERATURE REVIEW | 1 |
| COTTON CONTAMINATION..... | 2 |
| SPECTRAL DISCRIMINATION | 9 |
| MULTIVARIATE CALIBRATION..... | 18 |
| ATR SPECTROMETRY..... | 25 |
| REFERENCES..... | 31 |
| 2 NOVEL SEARCH ALGORITHMS FOR A MID-IR SPECTRAL LIBRARY | |
| OF COTTON CONTAMINANTS | 39 |
| ABSTRACT..... | 40 |
| INTRODUCTION | 41 |
| BACKGROUND..... | 43 |
| MATERIALS AND METHODS | 44 |
| RESULTS AND DISCUSSION | 51 |
| CONCLUSION | 79 |

| | | |
|----------|---|------------|
| | ACKNOWLEDGEMENT | 80 |
| | REFERENCES..... | 80 |
| 3 | QUALITATIVE IDENTIFICATION OF COTTON CONTAMINANTS BY | |
| | PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS | 82 |
| | ABSTRACT..... | 83 |
| | INTRODUCTION | 84 |
| | MATERIALS AND METHODS | 85 |
| | RESULTS AND DISCUSSION | 89 |
| | CONCLUSION | 102 |
| | REFERENCES..... | 103 |
| 4 | QUANTITATIVE PREDICTION OF THE COMPOSITION OF COTTON | |
| | DEBRIS POWDER AND AN ITERATIVE PREDICTION ERROR | |
| | CORRECTION ALGORITHM | 105 |
| | ABSTRACT..... | 106 |
| | INTRODUCTION | 107 |
| | MATERIALS AND METHODS | 113 |
| | RESULTS AND DISCUSSION | 117 |
| | CONCLUSION | 132 |
| | ACKNOWLEDGEMENT | 135 |
| | REFERENCES..... | 135 |

| | | |
|----------|--|------------|
| 5 | A MIXTURE GENERATOR ALGORITHM FOR GENERATION OF CALIBRATION MIXTURE STANDARDS FOR CLOSED MIXTURE SPACES | 138 |
| | ABSTRACT..... | 139 |
| | INTRODUCTION | 140 |
| | MATERIALS AND METHODS | 149 |
| | RESULTS AND DISCUSSION | 157 |
| | CONCLUSION | 173 |
| | REFERENCES..... | 174 |
| 6 | THE EFFECTS OF DIFFERENT ATR ACCESSORIES ON SPECTRAL COMPARISON IN FT-IR SPECTROMETRY | 175 |
| | ABSTRACT..... | 176 |
| | INTRODUCTION | 177 |
| | MATERIALS AND METHODS | 182 |
| | RESULTS AND DISCUSSION | 185 |
| | CONCLUSION | 213 |
| | REFERENCES..... | 214 |
| 7 | CONCLUSION AND FUTURE STUDIES..... | 216 |

LIST OF TABLES

| | Page |
|--|-------|
| Table 2.1: Search results for a leaf powder sample..... | 65 |
| Table 2.2: Top 10 ranked results for all standard algorithms for a cotton seed coat powder spectrum..... | 67-68 |
| Table 2.3: Results from voting scheme algorithms for the same seed coat powder sample for which standard algorithm results are shown in Table 2.2..... | 69 |
| Table 3.1: Summary of spectra sets used in experiments. | 86 |
| Table 4.1: Describes the preprocessing conditions and regression type used for each of the eleven conditions sets used in the experiments. | 118 |
| Table 4.2: The model conditions numbers ranked from lowest to highest RMSECV values for the three calibrations sets A, B, and C..... | 120 |
| Table 4.3: The model conditions numbers ranked from lowest to highest RMSEP values for the three calibrations sets A, B, and C | 121 |
| Table 4.4: RMSECV values and their variances for the three test sets..... | 129 |
| Table 4.5: Example of RMSEP values for test set B before and after application of the iterative error redistribution algorithm..... | 131 |
| Table 5.1: Example calculations from algorithm for three component system with $z = 3$ | 158 |
| Table 6.1: Experimental design..... | 183 |

| | |
|---|-----|
| Table 6.2: Correlation coefficients for PETE spectra. | 194 |
| Table 6.3: Correlation coefficients for PETE spectra after depth of penetration correction.. | 197 |
| Table 6.4: Correlation coefficients of PETE spectra after “Advanced ATR Correction” | 201 |
| Table 6.5: Correlation coefficients for baseline corrected PETE spectra..... | 207 |
| Table 6.6: Correlation coefficients for baseline corrected PETE spectra after “Advanced ATR Correction” | 210 |

LIST OF FIGURES

| | Page |
|---|------|
| Figure 1.1: Structure of cellulose | 5 |
| Figure 1.2: Spectra of leaf, stem, seed coat, and hull from the cotton plant. The spectra have been offset for clarity..... | 7 |
| Figure 1.3: Spectra of four different cotton leaves | 10 |
| Figure 1.4: Snell’s Law diagram..... | 27 |
| Figure 2.1: Spectra of leaf, stem, seed coat, and hull from the cotton plant. The spectra have been offset for clarity..... | 53 |
| Figure 2.2: Spectra of four different cotton leaves | 55 |
| Figure 2.3: Number of samples in the 20 member test set correctly identified by the first result returned by each standard algorithm. The results are broken down by sample type as indicated in the legend..... | 58 |
| Figure 2.4: Total number of results in the top 10 results returned by each of the standard algorithms for each of the 20 test set samples that correctly identified a test spectrum..... | 60 |
| Figure 2.5: Total number of results in the top 10 results returned by each of the standard algorithms for each of the 20 test set samples that correctly identified a test spectrum. The results are broken down by sample type as indicated in the legend..... | 62 |

| | |
|---|-----|
| Figure 2.6: Number of samples in the 20 member test set correctly identified by the first result returned by each of the voting scheme algorithms..... | 71 |
| Figure 2.7: Number of samples in the 24 member test set searched against the augmented library that was correctly identified by the first result returned..... | 75 |
| Figure 2.8: Number of samples in the 24 member test set searched against the augmented library that was correctly identified by the first result returned. The results are broken down by sample type as indicated in the legend..... | 77 |
| Figure 3.1: Spectra of leaf, stem, seed coat, and hull from the cotton plant. The spectra have been offset for clarity..... | 91 |
| Figure 3.2: Spectra of four different cotton leaves | 93 |
| Figure 3.3: The columns of this figure represent the model A test set samples. The rows correspond to the group the PLS-DA algorithm indicated the test sample belonged to. The patterns representing true positive, false positive, and false negative identifications are shown in the legend. White boxes represent true negative identifications. | 96 |
| Figure 3.4: This figure representing the results of the second experiment should be read the same way as Figure 3.1. Samples leaf 5, stem 1, seed coat 1, and hull 2 were samples from the model A test set that were not identified correctly by model A. Samples leaf 4, stem 3, seed coat 4, and hull 5 were samples from the model A test set that were identified correctly by model A. The remaining samples were taken from the model A calibration set. | 100 |

| | |
|---|-----|
| Figure 4.1: Spectra of leaf, stem, seed coat, and hull from the cotton plant. The spectra have been offset for clarity..... | 109 |
| Figure 4.2: Spectra of four different cotton leaves | 111 |
| Figure 4.3: Number of predictions within ± 10 percentage points for test sets A, B, and C. A total of 40 predictions were made for each test set | 123 |
| Figure 4.4: Comparison of the decrease in RMSEP seen for each of three test sets A, B, and C as a result of the three error redistribution methods used for the iterative error redistribution algorithm | 127 |
| Figure 4.5: Comparison of the number of predictions within ± 10 percentage points for test sets A, B, and C before and after application of the iterative error redistribution algorithm with method 1 used to redistribute the error..... | 133 |
| Figure 5.1: Ten point simplex-lattice experimental design for three components..... | 143 |
| Figure 5.2: Shows the location of the component axes for a three component system .. | 146 |
| Figure 5.3: Algorithm flow chart | 150 |
| Figure 5.4: Pattern followed by algorithm | 153 |
| Figure 5.5: Visual depiction of algorithm's output for 15 mixtures of 3 components | 159 |
| Figure 5.6: Pattern followed by algorithm to produce 10 mixtures in 4 component space..... | 162 |
| Figure 5.7: Visual depiction of algorithm's output for 10 mixtures of 4 components | 164 |
| Figure 5.8: Demonstrates how the pattern shown in Fig. 5.4 is an unfolded representation of the mixture sets | 167 |
| Figure 5.9: Vector plot demonstrating linear independence of data set..... | 170 |

| | |
|--|-----|
| Figure 6.1: PETE spectra IC1 and IVC2 | 186 |
| Figure 6.2: PETE spectra IC1 and IA4..... | 189 |
| Figure 6.3: Scale normalized PETE spectra from 1800 to 650 cm^{-1} : a) IA4, b) IC1, c) IVC2, d) IIIA5, and e) IIB3. The vertical dotted lines draw attention to the spectral shifts. | 192 |
| Figure 6.4: PETE spectra after depth of penetration correction: a) IA4, b) IC1, c) IVC2, d) IIIA5, and e) IIB3 | 199 |
| Figure 6.5: PETE spectra after “Advanced ATR Correction”: a) IA4, b) IC1, c) IVC2, d) IIIA5, and e) IIB3 | 202 |
| Figure 6.6: PETE spectra from 4000 to 650 cm^{-1} : a) IA4, b) IC1, c) IVC2, d) IIIA5, and e) IIB3 | 204 |
| Figure 6.7: Non-scale normalized PETE spectra after baseline correction: a) IA4, b) IC1, c) IVC2, d) IIIA5, and e) IIB3 | 208 |
| Figure 6.8: Baseline corrected PETE spectra after “Advanced ATR Correction”: a) IA4, b) IC1, c) IVC2, d) IIIA5, and e) IIB3..... | 211 |

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

The research projects presented in this dissertation relate to two important areas of analytical chemistry: chemometrics and spectrometry. Specifically, Fourier transform infrared spectrometry (FT-IR) with the use of attenuated total reflection (ATR) sampling accessories has been the spectrometric technique used and explored in these projects. Because of its inherent ability to produce unique spectra of numerous organic and inorganic substances, FT-IR proves to be an ideal candidate with which to combine the power of chemometrics to achieve discrimination and quantitative analysis of complex samples. Chemometrics has been defined as “the chemical discipline that uses mathematics and statistical methods, (a) to design or select optimal measurement procedures and experiments; and (b) to provide maximum chemical information by analyzing chemical data”¹. This non-specific definition allows the study of chemometrics to encompass many different techniques and approaches. The studies that make up this work are generally divided into three sub-disciplines of chemometrics: (1) spectral analysis and discrimination, (2) quantitative analysis of mixtures, and (3) experimental design.

The projects presented here were pursued with a two-fold purpose in mind. The first purpose was to deal with immediate challenges that presented themselves in the areas of discrimination of cotton contaminants and quantitation of cotton contaminant

mixtures, but the second and broader purpose was to develop and investigate analysis methods applicable to situations where complex samples or feedstocks occur. These projects provide both novel methods of analysis for cotton contamination, and fundamental advances in the understanding and application of the chemometric sub-disciplines listed above. In this introductory chapter, the problem of cotton contamination and the complex nature of cotton contaminants will be discussed. In addition, the fundamentals of ATR spectrometry and of the chemometric sub-disciplines that relate to the projects will be reviewed. Beyond this introductory chapter, five research projects presented in the manuscript style are followed by a final conclusion chapter that briefly discusses the prospects for future studies.

COTTON CONTAMINATION

Harvested cotton fibers can be contaminated with a variety of natural and synthetic substances²⁻⁶. Most of the contaminants are debris from other parts of the cotton plants themselves. The hulls, stems, leaves, seeds, and other parts of the plants are unavoidably harvested along with the cotton fibers. Other contaminants such as soil, greases, oils, and plastics can at times be found in harvested cotton fibers as well, but the primary contaminants present are those from the cotton plants. Cotton fibers can also be contaminated by sugary residues left behind when white flies and aphids feed on cotton fibers late in the growing season. Although sticky cotton is a serious problem for the cotton industry, it will not be discussed further in this introduction because none of the following research projects directly relate to the problem of sticky

cotton. The interested reader should consult the large volume of work that has been and continues to be published in this area⁶⁻⁵⁵.

Cotton debris causes a number of problems during cotton processing, including an increased number of yarn breakages during spinning and an increased number of yarn imperfections^{2, 56-58}. Because of these issues, cotton contaminated with debris undergoes cleaning processes at the cotton gin location to lessen the impact of the debris on production, but the cleaning processes shorten the length of the cotton fibers⁵⁹. All of these factors decrease the profitability of producing products composed of cotton and decrease the value of raw cotton contaminated with large amounts of debris. For increased profitability, it would be beneficial to identify the types of debris present at different stages of the production process. This information can allow process engineers to make process and machinery adjustments to limit the effects of particular types of debris. For instance in the work by Foulk et al.², it was discovered that the dust accumulating in the rotor groove during rotor spinning of cotton fibers into yarn was composed primarily of hull and shale. The authors speculate that a more profitable cleaning process that focused specifically on these types of debris instead of indiscriminately cleaning out every type of debris could be implemented to prevent the buildup of dust in the rotor groove. By limiting the buildup of dust in the rotor groove, the number of yarn imperfections and breakages could potentially be limited. A more efficient cleaning process would save time and limit the detrimental effects of cleaning on the fiber quality mentioned above. This case study provides only one example of the

potential advantages of being able to discriminate among the different types of plant debris present in cotton.

Although the potential advantages of debris identification are great, discrimination of different types of debris is not a trivial task. Because debris breaks down in size physically from the time it is harvested along with the cotton fibers until spinning is complete, visual identification of the type of debris can be difficult or impossible. In some cases, debris particles found in the system may be as small as the size of particles found in ground pepper². In the past, work has been reported on a method of debris identification relying on color in the visible region of the spectrum and/or geometric analysis of debris shape for larger pieces of debris^{4, 5}. These systems were primarily intended for large scale classing of raw cotton with regards to its trash content. They have not been developed for investigative work of the type described in the case study above. In addition, they fail to make use of the more precise chemical information contained in spectra from the mid-IR region of the spectrum.

Recently, USDA scientists have been developing an ATR FT-IR spectral library of cotton contaminants that provides a searchable repository of debris spectra to be used for the identification of unknown cotton contaminants^{2, 3}. Currently, this library contains over 900 spectra, and although it contains spectra of both cotton plant debris and the other non-plant related contaminants mentioned above, the majority of spectra in the library are of cotton plant parts. All of the plant parts are cellulose based. The structure of cellulose is given in Fig. 1.1. Figure 1.2 shows representative spectra for four of the most common parts of plant debris found in cotton: hull, leaf, stem, and seed

Figure 1.1: Structure of cellulose

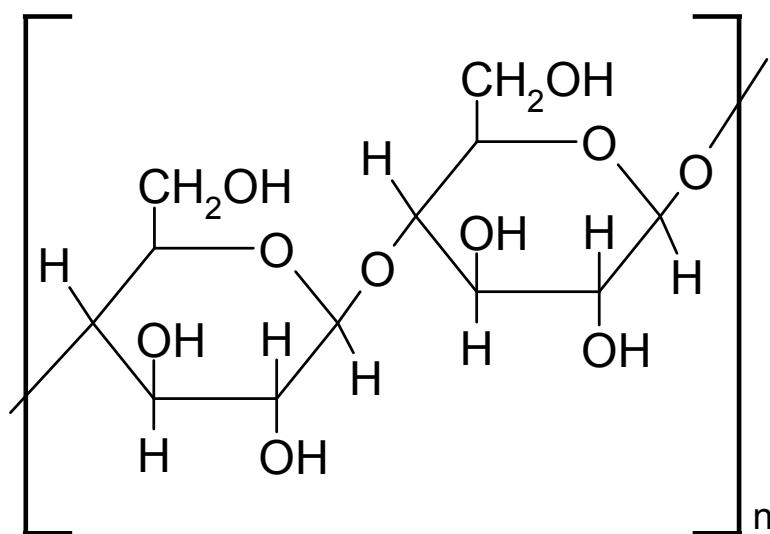
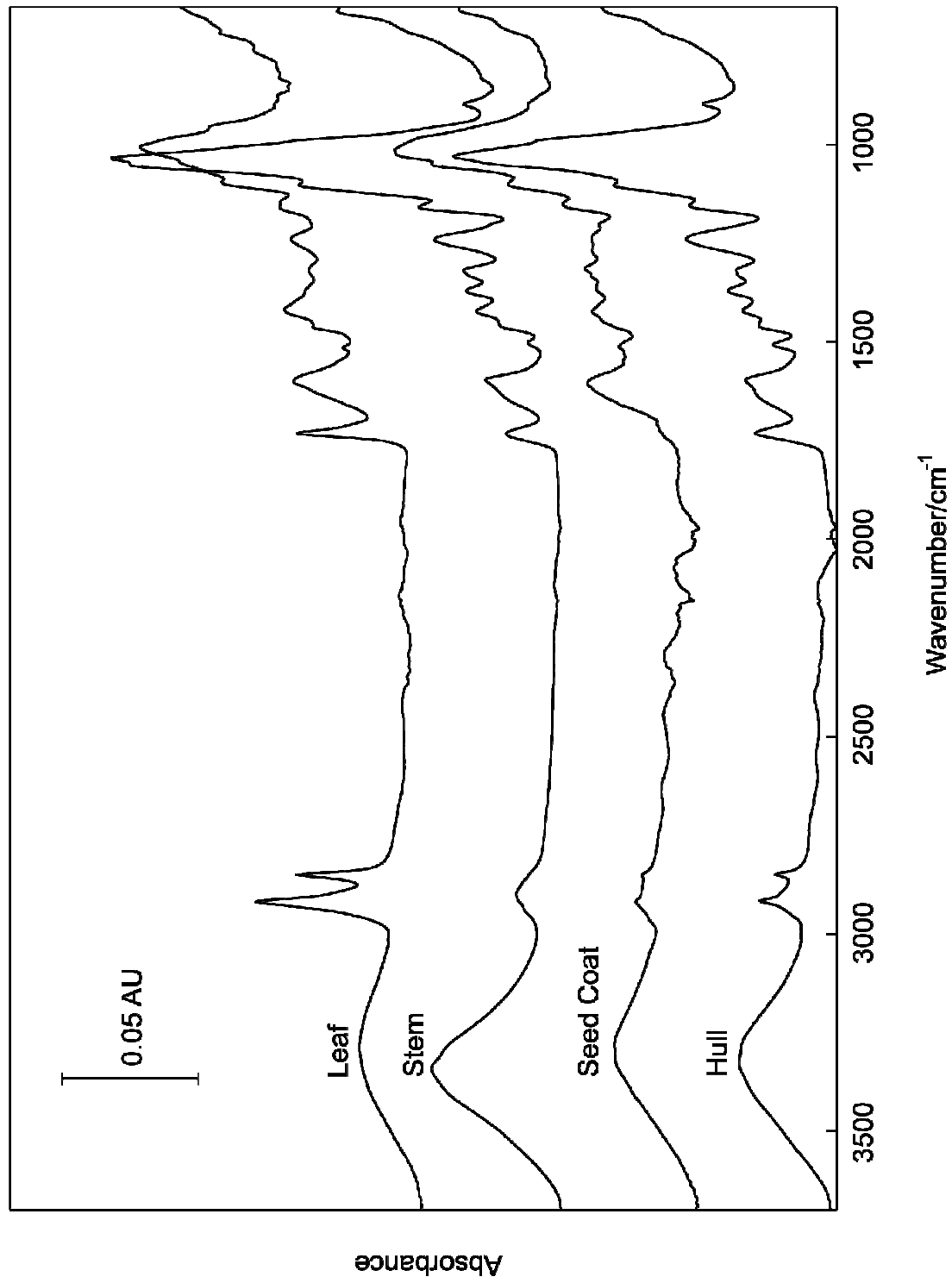


Figure 1.2: Spectra of leaf, stem, seed coat, and hull from the cotton plant. The spectra have been offset for clarity.

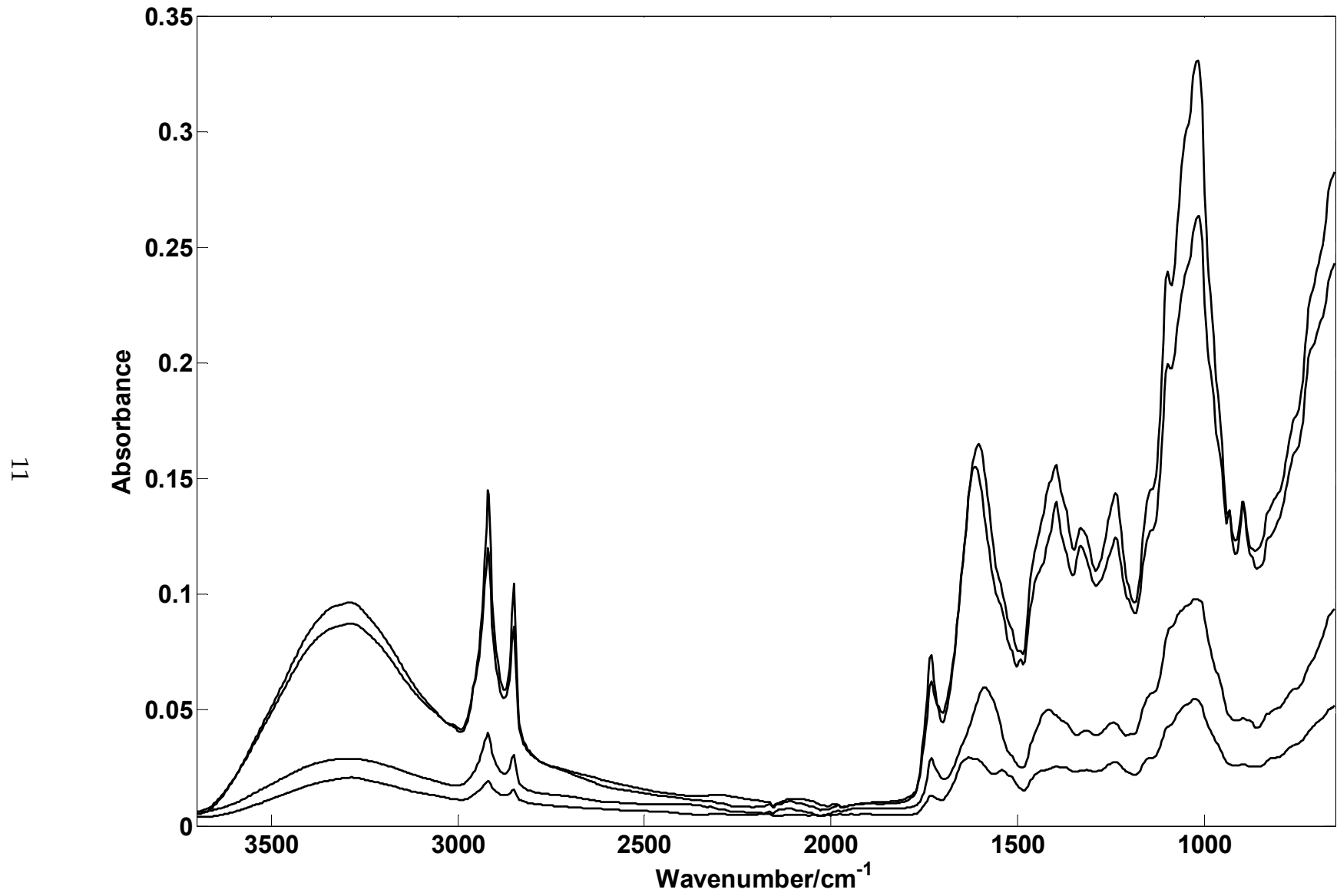


coat. One will notice that the spectra of different types of plant parts are very similar because of the common major chemical component; however, the spectra differ because of changes in the other chemical components of the plant parts. Figure 1.3 demonstrates the variability that can occur in a single category of plant debris present in the library. The extreme similarity of the spectra from different debris categories combined with the wide variability that exists within the spectra from individual categories of debris make debris discrimination a challenge. Research has shown, though, a greater than 90% identification rate for a test set where the test set spectra were obtained on the same spectrometer and ATR accessory as the library spectra and the library spectra were generally representative of the plant variety and plant growing conditions for the test set³. The work presented in this dissertation continues the exploration and development of cotton contaminant discrimination and identification methods for cases where the unknown spectra are not well represented by the library.

SPECTRAL DISCRIMINATION

Spectral identification is a key concern of the projects presented in this work, and the research reported augments the line of spectral discrimination tools that have been reported previously. Lowery et al.⁶⁰ have given a succinct history of infrared spectral search systems. In the 1950s, the first infrared spectral search systems emerged^{61, 62}. For these systems, a library of spectra would be encoded into punch cards. The cards denoted whether or not a peak existed at each spectral resolution element. When an unknown spectrum was obtained, an electric sorter was employed to determine the cards in the library that had peaks at the same resolution elements as the unknown

Figure 1.3: Spectra of four different cotton leaves.



spectrum. These systems allowed one to search for the absence of absorption bands at particular resolution elements as well. The computerized ASTM binary spectral library was the major development in spectral search systems in the 1960s⁶³⁻⁶⁵. Like the punch card system, the binary spectral library only recorded whether a band was present or absent at a particular resolution element in the spectrum. The advantage of this system was that the time required to search through the library was much faster than with punch cards and an electric sorter. This innovation made it much easier to perform multiple spectral searches in a reasonable time frame. The EPA vapor phase library, the first major digitized infrared spectral library, came into use in the 1970s⁶⁶. With truly digitized spectra, the actual intensity values at each resolution element were stored in the library. For the first time, both band location and intensity could easily be used to identify unknown spectra. The introduction of a digitized spectral library required new types of search algorithms that went beyond logic comparisons of binary data. At this time, scientists began to use Euclidean distance metrics to measure the similarity of an unknown spectrum to the library spectra. In the late 1970s and into the 1980s, correlation metrics^{67, 68} and metrics based on differences in derivatives of spectra came into use⁶⁰. Since the 1980s, there have been no fundamental advances in the way FT-IR spectral search systems operate. They still rely on the same comparison metrics that have been listed.

A brief description of the standard search metrics in use follows. A spectrum can be thought of as a vector where the intensity of each resolution element forms an entry in the vector, i.e. each resolution element can be thought of as a direction in space. The

descriptions of spectral similarity metrics that follow assume that each spectrum in the library and being searched against the library is represented by a vector. Let the vector $\bar{\mathbf{l}}_j$ represent the j^{th} spectrum in the library, and the vector $\bar{\mathbf{u}}$ represent a spectrum searched against the library. The subscript i represents the i^{th} element of a vector, and n equals the number of elements in each vector or the number of resolution elements in each spectrum. The score describing how closely related the spectrum being searched is to the j^{th} spectrum in the library is represented by S_j . Before calculating the scores for these search metrics, all spectra are usually normalized by some method to eliminate the spectral variation due only to overall intensity variations.

Equation 1.1 is the formula for the Euclidean distance metric:

$$S_j = \left[\sum_{i=1}^n (\bar{\mathbf{l}}_{ji} - \bar{\mathbf{u}}_{ji})^2 \right]^{\frac{1}{2}} \quad (1.1)$$

The smaller the value of S_j the more similar two spectra are. A value of 0 indicates identical spectra. Equation 1.2 is the formula for the dot product metric:

$$S_j = \bar{\mathbf{l}}_j \bullet \bar{\mathbf{u}}_j \quad (1.2)$$

When ordering a set of library spectra from most similar to least similar, the dot product metric will yield the same order as the Euclidean distance metric, but calculation of the dot product metric requires fewer calculations than the Euclidean distance metric. For spectra vector normalized to unit magnitude, the score from the dot product metric equals the cosine of the angle between the two spectral vectors being compared. A perfect score for the dot product metric is 1, and scores decrease as spectra

become more dissimilar. Equation 1.3 gives the formula for a variation of the Euclidean distance metric called the absolute value metric:

$$S_j = \sum_{i=1}^n |\bar{\mathbf{I}}_{ji} - \bar{\mathbf{u}}_{ji}| \quad (1.3)$$

Equations 1.4 and 1.5 show the formulas for two metrics that are based on the differences in the derivatives between two spectra:

$$S_j = \sum_{i=1}^n |(\bar{\mathbf{I}}_{j,i+1} - \bar{\mathbf{I}}_{ji}) - (\bar{\mathbf{u}}_{j,i+1} - \bar{\mathbf{u}}_{ji})| \quad (1.4)$$

$$S_j = \sum_{i=1}^n ((\bar{\mathbf{I}}_{j,i+1} - \bar{\mathbf{I}}_{ji}) - (\bar{\mathbf{u}}_{j,i+1} - \bar{\mathbf{u}}_{ji}))^2 \quad (1.5)$$

Difference of derivative metrics work well in situations where there are varying baseline offsets in the library, the unknowns, or both. For metrics in Eqs. 1.3-1.5, a smaller score means more similar spectra versus a larger score. A score of 0 indicates identical spectra. Finally, Eq. 6 gives the formula for a correlation similarity metric:

$$S_j = \frac{\sum_{i=1}^n \bar{\mathbf{I}}_{ji} \bar{\mathbf{u}}_{ji} - \frac{\left(\sum_{i=1}^n \bar{\mathbf{I}}_{ji} \sum_{i=1}^n \bar{\mathbf{u}}_{ji} \right)}{n}}{\left[\left(\sum_{i=1}^n \bar{\mathbf{I}}_{ji}^2 - \frac{\left(\sum_{i=1}^n \bar{\mathbf{I}}_{ji} \right)^2}{n} \right) \left(\sum_{i=1}^n \bar{\mathbf{u}}_{ji}^2 - \frac{\left(\sum_{i=1}^n \bar{\mathbf{u}}_{ji} \right)^2}{n} \right) \right]^{\frac{1}{2}}} \quad (1.6)$$

Like the dot product metric, identical spectra have a score of 1, and the score decreases as similarity decreases. The scores from different metrics cannot be compared directly, but the rankings from different metrics can be compared. The ranking of the library

spectra from most similar to least similar to an unknown spectrum will not necessarily be identical when using the different metrics.

Chemometric classification methods present a different approach to spectral discrimination than library searching. The disadvantage of replacing standard library searching with a classification technique is the time involved in creating classification models versus simply searching a spectrum against a spectral library, but the potential gain in discrimination power warrants exploration of classification techniques for cotton debris discrimination. SIMCA and partial least squares discriminant analysis (PLS-DA) are the two major classification methods that could be used for spectral discrimination. Wise et al.^{69, 70} have given a detailed overview of these two techniques. PLS-DA is arguably the most promising classification technique available for this purpose, but one needs to understand the basics of the two methods to understand why this is the case.

There is disagreement on what the acronym SIMCA actually stands for, but one choice is soft independent method of class analogy⁷⁰. SIMCA requires a separate principal component analysis (PCA) model to be constructed for the calibration spectra in each sample class that is to be distinguished. PCA is a data decomposition and compression method that works by decomposing a matrix of data into scores and loadings vectors. Equation 1.7 describes the PCA decomposition:

$$\mathbf{A} = \mathbf{T}_k \mathbf{V}_k^T + \mathbf{E} \quad (1.7)$$

The square matrix of data is \mathbf{A} , \mathbf{T}_k is the matrix of scores, \mathbf{V}_k is the matrix of loadings, \mathbf{E} is the residual matrix, and the subscript k represents the number of principal components (PCs) that have been retained in the model. If the original data matrix, \mathbf{A} , is not square, $\mathbf{A}^T\mathbf{A}$ must be used in place of \mathbf{A} in the decomposition. The loadings in PCA are the orthonormal eigenvectors of the matrix \mathbf{A} , and the scores are given by Eq. 1.8:

$$\mathbf{T} = \mathbf{A}\mathbf{V} \quad (1.8)$$

A single PC is composed of the i^{th} column of \mathbf{T} and the i^{th} row of \mathbf{V}^T . The first PC captures the maximum variance that can possibly be described in a single linear direction in the data space. The next PC captures the maximum variance possible in a single linear direction that is orthogonal to the direction of the first principal component, and the pattern continues for higher numbers of PCs. There will be as many PCs as there are rows or columns in the data matrix \mathbf{A} . Because the first PCs explain the largest directions of variance in the data, the last PCs usually capture mostly random noise in the data. By discarding the later PCs, one can actually create a data set where most of the noise has been extracted from the original data. Because the PCs are orthogonal, PCA is also a method of removing repetitious or collinear information within the data. The removal of noise and repetitious information explain how PCA is used to compress data.

For each class considered in SIMCA, the data matrix has the spectra representing that class as its row vectors. This fact means that the columns of the data matrix are the resolution elements of the spectra. The use of PCA models to represent spectral data is

advantageous because many of the resolution elements in spectra are describing the same information. Applying PCA to the spectra both eliminates random noise and removes collinear information recorded in the spectra, making comparisons of unknown spectra to the compressed spectra quicker than to uncompressed spectra after the initial models have been created. After a PCA model has been created for each class, one can determine the class that best represents an unknown spectrum by projecting that spectrum vector onto the lower dimensional PCA models that have been created, and then finding the class that the projection of the unknown spectrum vector lies closest to.

PLS-DA is based on PLS multivariate regression. See the next section of the introduction on multivariate calibration for a discussion of multivariate regression methods and an explanation of PLS regression. In PLS-DA, categorical variables are regressed onto the spectral data by means of a PLS regression. A matrix of categorical variables or so called dummy variables take the place of the matrix of analyte concentrations that would normally be used to build a spectral regression model to predict concentrations. The goal of PLS-DA is of course not to predict analyte concentrations, but to predict class membership. There will be a separate categorical variable for each class that one is interested in predicting membership for. In the calibration set, all of the spectra that belong to a given class are assigned a value of 1, and all other spectra that do not belong to that class are assigned a value of 0. After building the PLS model, a probabilistic threshold value between 0 and 1 is chosen for each class variable. When the value of the class variable for an unknown spectrum is

predicted by the PLS model, the unknown will be considered part of the class if the predicted value is greater than the threshold value and not part of the class if the predicted value is less than the threshold value. This threshold value is chosen so that the misclassification rate for the calibration set is as small as possible.

The advantages of the PLS-DA method versus PCA methods like SIMCA have been discussed by Barker and Rayens⁷¹. As discussed above, the data decomposition and compression in PCA is guided by total variance in the spectral data. This type of decomposition could be helpful when the largest sources of variance in the spectral data are related to class differences; however, if the among class variation is smaller than the within class variation, PCA based methods will do a very poor job of class discrimination. The data decomposition in PLS is guided by among class variation instead of total variation, so the PLS-DA method is ideally suited to the class discrimination application. As was demonstrated by Figs. 2.2 and 2.3, the within class variation for the classes of cotton debris is substantial compared to the between class variation. For this reason, PLS-DA was one of the classification methods explored in the work presented in this dissertation.

MULTIVARIATE CALIBRATION

Several excellent texts discuss the topic of multivariate calibration⁷²⁻⁷⁴. To understand multivariate calibration, it is helpful to start with a review of univariate calibration. In univariate least squares regression for spectral quantitative analysis, the absorbance values of a set of calibration solutions at a particular frequency related only to the species of interest are regressed onto the concentrations of that species in the

corresponding solutions. The regression yields a proportionality constant that can be used to predict the concentration of the species of interest in unknown solutions from the absorbance values of those solutions. This process is based on Beer's Law which is shown in Eq. 1.9:

$$A = \epsilon bc \quad (1.9)$$

In this equation, A is absorbance at a particular frequency, ϵ is the absorption coefficient at that frequency, b is the pathlength or sample thickness, and c is the concentration of the analyte. To simplify the equation, the absorption coefficient and the pathlength can be combined to form a single proportionality constant. In short, Eq. 1.9 shows that absorbance is linearly proportional to concentration. Of course, there are conditions where this relationship can break down, but assuming linearity holds, the relationship given in Eq. 1.9 forms the basis for univariate calibration. In simple terms, the least squares regression works by finding the line of best fit through the data points. In this process, the line of best fit is found by choosing the line that minimizes the error between the actual absorbance values for all the calibration samples and the values predicted for those samples by the line. The data points do not fall exactly on the line because in any experiment random errors from a variety of sources exist.

Equation 1.9 also leads us to multivariate least squares calibration if one incorporates the fact that the total absorbance at a certain frequency is equal to the absorbance of all the analytes in a mixture that absorb radiation of that energy. This fact is important because in mixtures absorption bands are many times attributed to more than one component of the mixture⁷². In fact, a univariate calibration to predict

the concentration of an individual analyte in a mixture is impossible if the concentrations of different components in a mixture are changing and a band attributable only to the single analyte of interest does not exist. If one wants to develop a multivariate calibration model, spectral measurements must be made for at least as many frequencies as components in the mixture. The information at these frequencies must correspond to the chemical structure of the analytes in such a way that information about all of the analytes is obtained. In other words, if one of the analytes does not absorb at any of the frequencies measured, one will not be able to make predictions about the concentration of this analyte in mixtures of unknown concentration. As the number of frequencies measured is increased above the number of components in the system, one achieves the same effect as is achieved by adding more calibration points in a univariate calibration, i.e. the precision of the calibration model will increase as random errors are averaged out.

Equation 1.10 shows the multivariate version of Eq.1.9:

$$\mathbf{A} = \mathbf{K}\mathbf{C} + \mathbf{E} \quad (1.10)$$

This relationship is the basis for a multivariate calibration method called classical least squares (CLS). In this equation, \mathbf{A} is a matrix of absorbance values, \mathbf{K} is a matrix of proportionality constants, \mathbf{C} is a matrix of concentrations, and \mathbf{E} is a matrix of random errors. The rows of \mathbf{A} are sample spectra, and each column of \mathbf{C} records the concentrations for all the components of a particular calibration sample. The proportionality constants are found by solving Eq. 1.10 for \mathbf{K} . The proportionality constants can then be used to predict analytes' concentrations from spectra of mixtures of

unknown concentrations. The major disadvantage to this method lies in the calibration step. In order to calculate the proportionality matrix, one must have access to the pure spectra of every component contributing to the spectra of the mixtures because one must account for the total absorbance, which is the sum of the absorbance values for all components that absorb at a given frequency. In many cases, it may be difficult or impossible to obtain the pure component spectra for all components in a mixture system.

Inverse least square (ILS) regression methods do not require access to the pure spectra of each component in a mixture or the knowledge of all of the components in a mixture. Equation 1.11 describes the ILS model:

$$\mathbf{C} = \mathbf{A}\mathbf{P} + \mathbf{E} \quad (1.11)$$

In this equation, \mathbf{C} is a matrix of concentrations, \mathbf{A} is a matrix of absorbance values, \mathbf{P} is a matrix of proportionality constants, and \mathbf{E} is a matrix of random errors. This method looks like Beer's Law in an inverse fashion: Instead of treating total absorbance as being proportional to the concentrations of all of a mixture's components, ILS works by treating the concentration of a particular component as being proportional to the absorbance values at all the resolution elements of a spectrum. During the calibration step, ILS only requires one to know the concentrations of the analytes of interest. Because in practice one many times is only interested in a single component or a few components in a complex system, ILS is a much more practical method than CLS in many circumstances. The problem with ILS is its sensitivity to collinearity. Equation 1.12 shows the least squares solution for \mathbf{P} :

$$\mathbf{P} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{C} \quad (1.12)$$

In order for the inverse of $\mathbf{A}^T \mathbf{A}$ to exist, the columns of \mathbf{A} must be linearly independent. The columns of \mathbf{A} are the absorbance values of the calibration spectra at the frequencies measured, and all of the frequencies measured in the spectra are usually not linearly independent. In addition to the independence requirement, the number of calibration spectra in \mathbf{A} must be at least as large as the number of resolution elements in each spectrum for the inverse of $\mathbf{A}^T \mathbf{A}$ to exist. A typical infrared spectrum may have several thousand resolution elements, but the experimentalist will seldom have that many linearly independent samples to work with. A solution to this problem is to choose a subset of the frequencies measured instead of the full spectra. Methods of frequency selection have been discussed in the literature⁷⁵. This process can be time consuming and tricky because algorithms designed to choose the best subset of spectral frequencies can often reach what might be termed local minima. In other words, the algorithm might select a good subset, but not the best subset.

Multivariate data reduction methods and regression techniques provide a more popular solution to the disadvantages of ILS. One popular method is principal component regression (PCR). In this method, the scores matrix, \mathbf{T}_k , from Eq. 1.7 replaces \mathbf{A} in Eq. 1.11. PCR is ILS where the data has been dimensionally reduced via PCA before regression takes place. One can think of the rows of \mathbf{T}_k as being dimensionally reduced spectra that represent the original calibration spectra. Instead of frequencies, the columns of \mathbf{T}_k represent linearly independent combinations of spectral frequencies, i.e. the loadings. PRC allows most of the original spectral information to be represented by a small number of PCs, which means that the restriction that the number

of sample spectra must be at least the number of columns is easily met. Because of this reason, PCR is many times a very effective regression method; however, the disadvantage to PCR is that the data reduction and compression that takes place only considers the variation in the spectral data. No attempt is made to maximize the covariance between the spectral data and the concentration data for the calibration set, and because factors other than variation in concentration affect the spectral variation, the PCs calculated in PCR may not be well correlated to the changes in concentration that the model is trying to predict. In other words, if there are large variations in the measurement system that do not directly relate to changes in analyte concentration, PCR may produce a poor prediction model

PLS regression, which has already been mentioned in the section of the introduction on spectral discrimination, does consider the covariance between the spectral data and the concentration data. In PLS, the abstract factors that are created are called latent variables (LVs) instead of principal components, and the LVs calculated by PLS will be different for a given set of data than the PCs that would be calculated by PCA of the spectral data. PLS requires the LVs to be orthogonal like the PCs must be in PCA. PLS begins by decomposing the spectral data similarly to PCA. The loading capturing the largest amount of variation is then rotated so that it achieves the best possible correlation to the concentration data. This iterative process continues until the desired number of latent variables has been calculated. One can essentially view the LVs in PLS as the PCs from PCR that have been rotated to achieve the best correlation between the PCs and the changes in concentration. Because PLS chooses factors that

correlate best to concentration variance, PLS models many times outperform PCR models with higher numbers of factors. Because of all its potential advantages, PLS is the regression method used in the projects in this dissertation.

Two major types of PLS regression exist. In PLS-1, a separate regression model is developed for each analyte for which concentration is to be predicted. This method only considers the concentration of one analyte when the LVs for a particular model are calculated. When the concentration of more than one analyte is of interest, the matrix of concentration values can also be regressed simultaneously onto the spectral data with a method called PLS-2. In PLS-2, the concentration matrix is decomposed like the spectral data, and the loadings of the spectral data are rotated to achieve the best correlation possible with the loadings of the concentration matrix. In general, a priori prediction of whether a PLS-1 or a PLS-2 will be the best model for a certain dataset cannot be made⁷³.

Experimental design is an important part of any calibration experiment. Many authors have addressed this topic in detail⁷⁶⁻⁷⁸, and there are many different options and viewpoints to consider when the samples for a calibration set are chosen. In dealing with multi-component systems, some general guidelines have been discussed by Kramer⁷³. First, even for linear systems, regression models should not be used to extrapolate predictions beyond the mixture space covered by the calibration set. Second, the mixtures in the calibration set must cover the mixture space of interest and be linearly independent. Third, one needs to be able to visualize the mixture space in order to clearly understand whether or not the mixture space is being representatively

covered by the calibration set. This last consideration can be difficult or impossible to achieve for some higher dimensional mixture spaces. One of the projects presented here addresses this point through the use of experimental design and novel algorithm development.

ATR SPECTROMETRY

In the traditional FT-IR sampling method of transmission, the sample must be prepared in such away that light can be absorbed by the analyte as radiation passes through the sample, but the sample must be thin enough or dilute enough that total absorption of the radiation does not occur. For liquid samples or solutions, transmission sampling requires special plates or sample cells that are transparent to the light frequencies of interest. For polymers, transmission requires the preparation of thin films. For other solids, the sample must be pressed with KBr into a pellet or mixed with mineral oil into a mull. These sample preparation techniques can be expensive and difficult or impossible to achieve for some samples. Some transmission measurements also require the sample compartment to be exposed to the atmosphere each time the sample is changed. This increases the sampling time for good measurements because one must usually wait for the IR active atmospheric gases to be purged from the sample compartment before measurements can be made. ATR accessories provide a quick and easy alternative to the transmission mode of measurement. A number of texts have discussed the ATR technique^{72, 79, 80}.

When light strikes the interface between a higher refractive index medium and a lower refractive index medium from the higher refractive index side, total internal

reflection of the light can occur. Total internal reflection will occur when the angle of the radiation to the normal is at or above the critical angle for the higher refractive index medium. Snell's Law is given in Eq. 1.13:

$$\eta_1 \sin \theta_1 = \eta_2 \sin \theta_2 \quad (1.13)$$

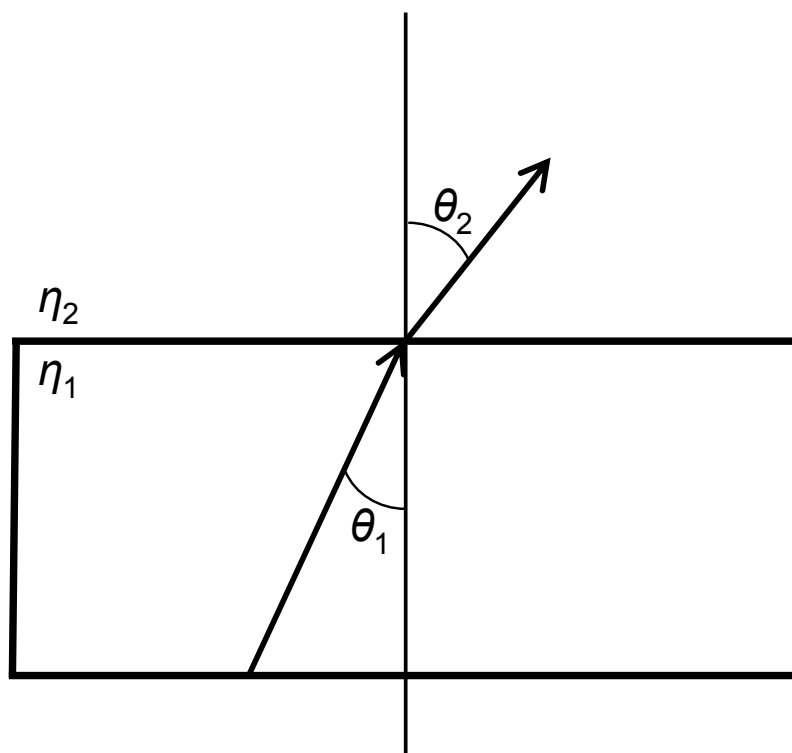
In this well known equation, η_1 is the refractive index of medium 1, η_2 is the refractive index of the medium 2, θ_1 is the angle of incidence for the radiation, and θ_2 is the angle of refraction. Figure 1.4 demonstrates Snell's Law for a beam of light traveling from material 1 to material 2 where η_1 is greater than η_2 . In this case, the angle of refraction is greater than the angle of incidence. As θ_1 is increased, θ_2 increases until at some critical angle, $\theta_1 = \theta_C$, θ_2 equals 90° and total internal reflection begins to occur. The formula for the critical angle is given in Eq. 1.14:

$$\theta_C = \arcsin \frac{\eta_2}{\eta_1} \quad (1.14)$$

The critical angle is simply determined by the refractive indices for the two media being considered.

Although the photons in the light wave are completely reflected inside the higher refractive index medium during total internal reflection, the electric field of the light wave penetrates a short distance beyond the media interface. This electric field is known as the evanescent wave, and it decays exponentially as the distance into the lower refractive index medium increases. Eq. 1.15 gives the effective depth of penetration of the evanescent wave into the lower refractive index medium:

Figure 1.4: Snell's Law diagram.



$$d_p = \frac{\lambda}{2\pi\eta_1\sqrt{\sin^2\theta - (\eta_2/\eta_1)}} \quad (1.15)$$

In Eq. 1.15, E_θ is the electric field strength of the evanescent wave at the media interface, θ is the angle of incidence, and d_p is the effective depth of penetration of the evanescent wave. This equation shows that the depth of penetration at a given angle of incidence will vary depending on the refractive indices of the media.

In the ATR sampling mode, the lower refractive index medium is the sample. Most organic materials have a refractive index of approximately 1.5 in the mid-IR region of the spectrum. Various materials such as Zinc selenide ($\eta=2.40$), Silicon ($\eta=3.41$), and Germanium ($\eta=4.00$) can be used for the higher refractive index medium. In an ATR accessory, the higher refractive index medium is called the internal reflection element (IRE). Because all but the surface of the IRE can be enclosed in a purged atmosphere that allows for seamless connection of the accessory to the spectrometer, ATR is a very convenient technique. Liquid, solid, and film samples can simply be pressed onto the IRE and the spectrum can be obtained without other sample preparation. One potential disadvantage of this technique is that it is essentially a surface technique. As Eq. 1.15 shows, the depth of penetration into the sample is governed by the wavelength of the light and the refractive index of the IRE. In most cases, the depth of penetration will be no more than a few micrometers. In many cases though, the relatively low depth of penetration does not present a problem, and if one is interested in studying layers, ATR can be an ideal technique because the depth of penetration can be varied by the angle of incidence.

Differences exist between transmission and ATR spectra that must be considered when comparing ATR and transmission spectra. Equation 1.9 describes the factors that contribute to absorbance in transmission measurements. The equation for absorbance in ATR is more complicated and is given in Eq. 1.16:

$$A = (\log e) \frac{\eta_2 E_0^2 d_p a}{\eta_1 (\cos \theta)^2} \quad (1.16)$$

In this equation, all of the variables have already been defined except a , which is the linear absorption coefficient per unit thickness of sample. In ATR, depth of penetration is the analog of pathlength in transmission. In transmission the pathlength is constant, but as Eq. 1.15 shows, depth of penetration increases with wavelength. In ATR, the intensity of longer wavelength absorption bands will be greater than that of smaller wavelength bands simply because of the greater depth of penetration. This effect does not occur in transmission. The second major difference between ATR and transmission spectra is due to anomalous dispersion. It is well known that the refractive index of a material changes sharply around the absorption bands for that material. On the smaller wavelength side of an absorption band, the refractive index of the material falls lower than the average refractive index of the material in spectral regions where there are no absorption bands. On the longer wavelength side of the band, the refractive index rises higher than the average refractive index of the material. Because of the dependence of Eqs. 1.15 and 1.16 on η_2 , anomalous dispersion causes the absorption bands in an ATR spectrum to be shifted slightly to longer wavelengths compared to a transmission

spectrum of the same sample. In one of the projects presented here, the implications of these effects on spectral comparison are investigated.

REFERENCES

1. J. Workman, *Chemom. Intell. Lab. Syst.* **60**, 1-2, 13 (2002).
2. J. Foulk, D. McAlister, D. Himmelsbach, and E. Hughs, *J. Cotton Sci.* **8**, 243 (2004).
3. D. S. Himmelsbach, J. W. Hellgeth, and D. D. McAlister, *J. Agric. Food Chem.* **54**, 20, 7405 (2006).
4. B. Xu and C. Fang, *Text. Res. J.* **69**, 9, 656 (1999).
5. B. Xu, C. Fang, and R. Huang, *Text. Res. J.* **67**, 12, 881 (1997).
6. H. H. Perkins, Jr., "Cotton Stickiness--A Major Problem in Modern Textile Processing", in *Proc. Beltwide Cotton Conferences* (1991), p. 523.
7. N. Abidi and E. Hequet, *Text. Res. J.* **75**, 4, 362 (2005).
8. W. S. Anthony and R. K. Byler, Patent Application Country: Application: US; Patent Country: US Patent 5700961 (1997).
9. J. Bourely, J. Gutknecht, and J. Fournier, *Coton et Fibres Tropicales* (French Edition) **39**, 3, 47 (1984).
10. D. B. Brushwood, "Relationship of individual honeydew sugar concentrations on cotton lint stickiness potential and measured sugar content", in *Proc. Beltwide Cotton Conferences* (2000), p. 1518.

11. D. E. Brushwood, "Measurement of sugar on raw cotton by HPLC, individual carbohydrate concentrations and their relationship to stickiness potential", in *Proc. Beltwide Cotton Conferences* (1997), p. 1656.
12. D. E. Brushwood, *Textile Chemist and Colorist* **30**, 2, 33 (1998).
13. D. E. Brushwood, *Journal of Cotton Science* [online computer file] **4**, 3, 202 (2000).
14. D. E. Brushwood and Y. J. Han, *J. of Cotton Science* [online computer file] **4**, 2, 137 (2000).
15. D. E. Brushwood and Y. J. Han, "Characteristics of entomological sugars applied to the surface of raw cotton", in *Proc. Beltwide Cotton Conferences* (2000), p. 1522.
16. D. E. Brushwood and H. H. Perkins, Jr., "Characterization of sugars from honeydew contaminated and normal cottons", in *Proc. Beltwide Cotton Conferences* (1994), p. 1408.
17. D. E. Brushwood and H. H. Perkins, "Cotton plant sugars and insect honeydew characterized by high performance liquid chromatography", in *Proc. Beltwide Cotton Conferences* (1996), p. 1310.
18. P. S. R. Cheung, C. W. Roberts, and H. H. Perkins, Jr., *Text. Res. J.* **50**, 1, 55 (1980).
19. D. T. W. Chun and D. Brushwood, *Text. Res. J.* **68**, 9, 642 (1998).
20. O. Elsner, *Text. Res. J.* **52**, 8, 538 (1982).
21. O. Fonteneau-Tamime, R. Frydrych, and J. Y. Drean, *Text. Res. J.* **71**, 11, 1023 (2001).
22. O. Fonteneau-Tamime, J. P. Gurlot, and E. Goze, *Text. Res. J.* **71**, 12, 1046 (2001).

23. J. Fournier, J. Gutknecht, E. Jallas, and J. Bourely, *Coton et Fibres Tropicales* (French Edition) **40**, 2, 113 (1985).
24. G. R. Gamble, *J. of Cotton Science* [online computer file] **5**, 3, 169 (2001).
25. G. R. Gamble, *J. Cotton Sci.* **6**, 4, 143 (2002).
26. G. R. Gamble, *Text. Res. J.* **72**, 2, 174 (2002).
27. G. R. Gamble, "Detection of sticky cotton via release of volatile compounds", in *Proc. Beltwide Cotton Conferences* (2003), p. 1953.
28. G. R. Gamble, *J. Cotton Sci.* **7**, 2, 45 (2003).
29. A. V. Ghule, R. K. Chen, S. H. Tzing, J. Lo, and Y. C. Ling, *Anal. Chim. Acta* **502**, 2, 251 (2004).
30. L. D. Godfrey, K. E. Keillor, P. B. Goodell, M. R. McGuire, J. Bancroft, and R. B. Hutmacher, "Management of late-season insect pests for protection of cotton quality in the San Joaquin Valley", in *Proc. Beltwide Cotton Conferences* (2003), p. 1089.
31. A. Gray, N. C. North, and A. N. Wright, *Coton et Fibres Tropicales* (French Edition) **40**, 2, 105 (1985).
32. C. Heinrichs, H. Dobbelsstein, S. Dugal, and E. Schollmeyer, *Melliand Textilberichte* **65**, 10, 710 (1984).
33. D. L. Hendrix and T. J. Henneberry, "Differences in polyol accumulation and honeydew excretion in sweetpotato whitefly and cotton aphid", in *Proc. Beltwide Cotton Conferences* (2000), p. 1296.

34. T. J. Henneberry, B. Blackledge, T. Steele, D. L. Hendrix, H. H. Perkins, and R. L. Nichols, "Preliminary evaluations of an enzyme approach to reduce cotton lint stickiness", in *Proc. Beltwide Cotton Conferences* (1997), p. 430.
35. T. J. Henneberry, L. F. Jech, and D. L. Hendrix, *Southwestern Entomologist* **23**, 2, 105 (1998).
36. T. J. Henneberry, L. F. Jech, and D. L. Hendrix, "Sweet potato whiteflies, cotton aphids, and sticky cotton", in *Proc. Beltwide Cotton Conferences* (2000), p. 1160.
37. T. J. Henneberry, L. F. Jech, D. L. Hendrix, and T. Steele, *Southwestern Entomologist* **25**, 1, 1 (2000).
38. E. Hequet and N. Abidi, *Journal of Cotton Science* [online computer file] **6**, 1, 77 (2002).
39. E. Hequet, R. Frydrych, and M. Watson, *World Textile Congress on Natural and Natural-Polymer Fibres*, University of Huddersfield, Queensgate, UK, July 9-11, 1997, 200 (1997).
40. E. F. Hequet and N. Abidi, Patent Application Country: Application: US; Patent Country: US Patent 2002083764 (2002).
41. E. F. Hequet, N. Abidi, and D. Ethridge, *Text. Res. J.* **75**, 5, 402 (2005).
42. B. Heuer and Z. Plaut, *Text. Res. J.* **55**, 5, 263 (1985).
43. O. J. Lantero and J. K. Shetty, Reg.Pat.Tr.Des.States: Designated States R: BE, CH, DE, DK, ES, FR, GB, GR, IT, LI, NL.; Patent Application Country: Application: EP; Patent Country: EP; Priority Application Country: US Patent 622487 (1994).

44. A. Luchian, G. Mihala, and C. Popa, *Industria Usoara: Textile, Tricotaje, Confectii Textile* **-9**, 12, 524 (1978).
45. C. Marquie, J. Bourelly, and A. Bonvalet, *Coton et Fibres Tropicales* (French Edition) **38**, 4, 323 (1983).
46. W. B. Miller, E. Peralta, D. R. Ellis, and H. H. Perkins, Jr., *Text. Res. J.* **64**, 6, 344 (1994).
47. U. Mor, PCT Designated States: Designated States W: AM, AT, AU, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, JP, KE, KG, KP, KR, KZ, LK, LR, LT, LU, LV, MD, MG, MN, MW, NL, NO, NZ, PL, PT, RO, RU, SD, SE, SI, SK, TJ, TT, UA, US, UZ.; (TRUNCATED) Patent 9522762 (1995).
48. R. L. Nichols, W. B. Miller, K. Mysore, and H. H. Perkins, Jr., "Honeydew sugar estimates differ among reducing-sugar test methods", in *Proc. Beltwide Cotton Conferences* (1998), p. 1547.
49. H. H. Perkins, Jr., *Textile Bulletin* **97**, 8, 21 (1971).
50. H. H. Perkins, Jr., *Textile Industries* (Atlanta) **135**, 3, 49 (1971).
51. H. H. Perkins, Jr., *Text. Res. J.* **53**, 8, 508 (1983).
52. C. W. Roberts, P. S. R. Cheung, and H. H. Perkins, Jr., *Text. Res. J.* **48**, 2, 91 (1978).
53. C. W. Roberts, H. S. Koenig, R. G. Merrill, P. S. R. Cheung, and H. H. Perkins, Jr., *Text. Res. J.* **46**, 5, 374 (1976).
54. J. E. Slosser, M. N. Parajulee, D. L. Hendrix, T. J. Henneberry, and D. R. Rummel, *Journal of economic entomology* **95**, 2, 299 (2002).
55. B. M. Talpay, *Melliand Textilberichte* (1923-1969) **49**, 6, 641 (1968).

56. A. D. Brashears, R. V. Baker, and C. K. Bragg, "Effect of Bark on Spinning Efficiency of Cotton", in *Proc. Beltwide Cotton Conference* (1992), p. 1218.
57. J. D. Barger and T. H. Garner, "The role of seed-coat and mote-fragment neps in yarn and fabric imperfections: a survey", in *Proc. Beltwide Cotton Production Research Conferences* (1988), p. 586.
58. C. K. Bragg, C. L. Simpson, A. D. Brashears, and R. V. Baker, *Trans. ASAE* **38**, 1, 57 (1995).
59. R. V. Baker, "Influence of Lint Cleaning on Fiber Quality", in *Proc. Beltwide Cotton Production Research Conferences* (1987), p. 535.
60. S. R. Lowry, D. A. Huppler, and C. R. Anderson, *J. Chem. Inf. Comp. Sci.* **25**, 3, 235 (1985).
61. A. W. Baker, N. Wright, and A. Opler, *Anal. Chem.* **25**, 10, 1457 (1953).
62. L. E. Kuentzel, *Anal. Chem.* **23**, 10, 1413 (1952).
63. D. H. Anderson and G. L. Covert, *Anal. Chem.* **39**, 11 (1967).
64. D. S. Erley, *Anal. Chem.* **40**, 6, 894 (1968).
65. R. A. Sparks, "Storage and Retrieval of Wyandotte-ASTM Infrared Spectral Data Using an IBM 1401 Computer", (ASTM, Philadelphia, PA, 1964).
66. A. Hanna, J. C. Marshall, and T. L. Isenhour, *J. Chromatogr. Sci.* **17**, 434 (1979).
67. K. Tanabe and S. Saeki, *Anal. Chem.* **47**, 1, 118 (1975).
68. L. A. Powell and G. M. Hieftje, *Anal. Chim. Acta* **100**, 313 (1978).

- 69. B. M. Wise, J. M. Shaver, N. B. Gallagher, W. Windig, R. Bro, and R. S. Koch, "PLS-DA", in *PLS_Toolbox 3.5* (Eigenvector Research, Inc., Manson, WA, 2005), p. 185.
- 70. B. M. Wise, J. M. Shaver, N. B. Gallagher, W. Windig, R. Bro, and R. S. Koch, "SIMCA", in *PLS_Toolbox 3.5* (Eigenvector Research, Inc., Manson, WA, 2005), p. 180.
- 71. M. Barker and W. Rayens, *J. Chemom.* **17**, 166 (2003).
- 72. P. R. Griffiths and J. A. de Haseth, *Fourier Transform Infrared Spectrometry* (John Wiley & Sons, Hoboken, New Jersey, 2007), 2nd ed.
- 73. R. Kramer, *Chemometric Techniques for Quantitative Analysis* (Marcel Dekker, New York, 1998).
- 74. P. J. Gemperline, Ed., *Practical Guide to Chemometrics* (Taylor & Francis, Boca Raton, 2006), 2nd ed.
- 75. J. H. Kalivas and P. J. Gemperline, "Variable Selection", in *Practical Guide to Chemometrics*, P. Gemperline, Ed. (Taylor & Francis, Boca Raton, 2006), p. 135.
- 76. J. Cornell, *Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data* (John Wiley & Sons, New York, 2002), 3rd ed.
- 77. K. Stoyanov and A. D. Walmsley, "Response-Surface Modeling and Experimental Design", in *Practical Guide to Chemometrics*, P. Gemperline, Ed. (Taylor & Francis, Boca Raton, 2006), p. 263.

78. J. H. Kalivas and P. J. Gemperline, "Selection of Calibration and Validation Samples", in *Practical Guide to Chemometrics*, P. Gemperline, Ed. (Taylor & Francis, Boca Raton, 2006), p. 113.
79. N. J. Harrick, *Internal Reflection Spectroscopy* (Harrick Scientific Corporation, Ossining, New York, 1967).
80. F. M. Mirabella, Jr. and N. J. Harrick, *Internal Reflection Spectroscopy: Review and Supplement* (Harrick Scientific Corporation, Ossining, New York, 1985).

CHAPTER 2

NOVEL SEARCH ALGORITHMS FOR A MID-IR SPECTRAL LIBRARY OF COTTON CONTAMINANTS¹

¹ J.B. Loudermilk, D.S. Himmelsbach, F.E. Barton II, J.A. de Haseth, Appl. Spectrosc. In press. (2008).
Reprinted here with permission of publisher.

ABSTRACT

During harvest, a variety of plant based contaminants are collected along with cotton lint. The USDA previously created a mid-IR, ATR, FT-IR spectral library of cotton contaminants for contaminant identification as the contaminants have negative impacts on yarn quality. This library has shown impressive identification rates for extremely similar cellulose based contaminants in cases where the library was representative of the samples searched. When spectra of contaminant samples from crops grown in different geographic locations, seasons, and conditions and measured with a different spectrometer and accessories were searched, identification rates for standard search algorithms decreased significantly. Six standard algorithms were examined: dot product, correlation, sum of absolute values of differences, sum of the square root of the absolute values of differences, sum of absolute values of differences of derivatives, and sum of squared differences of derivatives. Four categories of contaminants derived from cotton plants were considered: leaf, stem, seed coat, and hull. Experiments revealed that the performance of the standard search algorithms depended upon the category of sample being searched and that different algorithms provided complementary information about sample identity. These results indicated that choosing a single standard algorithm to search the library was not possible. Three voting scheme algorithms based on result frequency, result rank, category frequency, or a combination of these factors for the results returned by the standard algorithms were developed and tested for their capability to overcome the unpredictability of the standard algorithms' performances. The group voting scheme search was based on the

number of spectra from each category of samples represented in the library returned in the top 10 results of the standard algorithms. This group algorithm was able to identify correctly as many test spectra as the best standard algorithm without relying on human choice to select a standard algorithm to perform the searches.

Index Headings: Spectral search system; Search algorithm; Voting scheme algorithm; Spectral library; Spectral library; Infrared; FT-IR; ATR; Chemometrics; Cotton; Contaminants.

INTRODUCTION

When cotton is mechanically harvested from the field, it can be contaminated by a variety of foreign matter. In addition to the desired cotton fiber, parts of the plants such as leaves, stems, seeds, and hulls are also collected. Parts of other plants growing in the field, inorganic matter such as clay or sand, plastic waste, and greases and oils from machinery can also contaminate the cotton, but the majority of contamination comes from the cotton plants themselves¹. These contaminants have major impacts on the quality of cotton yarn produced and, ultimately, on the profitability of processing. Contaminants are responsible for an increased number of yarn breakages during spinning and lower the quality and price of the final yarn product^{2,3}.

Because of the negative impact of these contaminants on quality and profit, their detection and removal is important. Knowledge of which contaminants are most abundant in the cotton and which contaminants create the most difficulties during processing allow the entire process from harvesting through production of the final

product to be streamlined to limit those contaminants, but the identification of contaminants is not straightforward. From the time organic contaminants are first harvested along with the cotton, they begin to break down in size physically making human visual identification difficult or impossible.³

Some attempts to identify contaminant particles by color⁴ or geometric features⁵ have been made in the past. These methods did not make use of the chemical information available through spectroscopic analysis of trash components. Over several years, the United States Department of Agriculture—Agricultural Research Service (USDA-ARS) has developed a mid-IR, FT-IR, attenuated total reflection spectral library of cotton contaminants^{1, 3}. This library allows one to employ the power of molecular fingerprinting, inherent to the mid-IR region of the spectrum, to be applied to the problem of cotton contaminant identification. Work by Himmelsbach et al.^{1, 3} has shown the utility of the library and the high percentage of correct matches returned by searching when the library spectra are representative of the unknown samples. It should be noted that identification of cotton plant parts by library searching is significantly more difficult than standard searches for a diverse library, such as most commercial libraries. In the current study, all the samples have both spectra and chemical compositions that are extremely similar. Standard algorithms have been designed to distinguish between relatively different library entries.

The current work demonstrates that when spectra of plant parts grown in different seasons and geographic locations and measured with a different spectrometer and ATR accessories from those represented in the library are searched, the

identification rate drops significantly compared to searching spectra that have representatives from the same growing regions and seasons and measured under the same conditions. The performance of a variety of standard searching algorithms was investigated, and it was found that identification depends on the type of plant part spectrum being searched. Additionally, different algorithms were found to give complementary information about sample identity instead of completely repetitive information. In other words, each algorithm may return correct answers that are not returned by every other algorithm. These results meant that it was impossible to pick a single standard algorithm that would yield the best matches from the library for every category of contaminant considered. This work focuses on the creation of voting scheme algorithms that overcome the disadvantages of the standard algorithms and improve identification.

BACKGROUND

A thorough history of the development of infrared spectral library search systems through the mid 1980s has been reported by Lowry et al.⁶ The first automated spectral library search systems based on punch cards and electric sorters were reported in the 1950s^{7, 8}. These systems relied on encoded cards to record the location of spectral absorption bands. Electric sorters were then used to locate cards that represented spectra with the same bands as the unknown spectrum searched. These early systems even made use of the absence as well as the presence of bands to match an unknown spectrum to the most similar spectra in a library. In the 1960s, the computerized ASTM binary encoded spectral library came into use⁹⁻¹¹. Because this system allowed for

computerized searching of electronically stored information, the time required to search through a given number of spectra decreased, but only band location, and not band intensity, information was still used to match spectra. In the 1970s, Azaraga¹² introduced the EPA vapor phase library, which was the first major digitized spectral library. With the introduction of this library, scientists began to use Euclidean distance as a metric to determine how closely related spectra in a library were to an unknown spectrum. Into the 1980s, work continued with correlation algorithms^{13, 14} and variations on the Euclidean distance metric, such as metrics based on the differences between the derivatives of spectra⁶, but in practice there has been little change in infrared spectral library searching since that time.

MATERIALS AND METHODS

USDA ATR FT-IR Spectral Library. This library contained 929 FT-IR spectra measured with a Nicolet Magna 850 FT-IR spectrometer (Thermo Fisher Scientific, Waltham, MA) and a DuraScope attenuated total reflection (ATR) sampling accessory (Smiths Detection, Danbury, CT). The spectrometer contained a ceramic source, a KBr beamsplitter, and a deuterated triglycine sulfate (DGTS) detector. The ATR accessory had a diamond-coated ZnSe internal reflection element (IRE). Spectra were measured over the range of 4000 to 650 cm^{-1} at 8 cm^{-1} resolution, with 128 interferograms co-added, and interferograms were processed with Happ-Genzel apodization prior to Fourier transformation. Spectra were collected with the use of Omnic E.S.P. 5.2 software (Thermo Fisher Scientific, Waltham, MA). All spectra were first converted to GRAMS format (Thermo Fisher Scientific, Waltham, MA) and then imported into

MATLAB (The Math Works, Natick, MA). The library was comprised of the spectra of foreign materials found in harvested cotton and included the following types of samples: cotton plant parts including bloom, bract, hull, leaf, seed coat, shale, and stem; other organic matter such as poultry feathers, cow leather, and weed parts; inorganic materials such as sand and clay; greases and oils; and plastics. Organic samples were representative of several geographic locations and growing conditions.

Test Set Spectra. A set of 75 test spectra of samples from several geographic locations and seasons different from the samples in the USDA library were measured with the use of a Varian Excalibur Series FTS-4000 FT-IR spectrometer (Varian, Palo Alto, CA) and three different ATR sampling accessories. The samples were obtained from the USDA-ARS Cotton Quality Research Laboratory (Clemson, SC). The spectrometer contained a ceramic source, a KBr beamsplitter, and a DGTS detector. The ATR accessories were the Specac Golden Gate (Specac, Woodstock, GA) with a diamond-coated ZnSe IRE and the Harrick SplitPea and Seagull (Harrick Scientific, Pleasantville, NY) with Si and ZnSe IREs, respectively. Spectra were measured over the range of 4000 to 400 cm^{-1} at 4 cm^{-1} resolution, with 256 interferograms co-added. Interferograms were processed with Happ-Genzel apodization to be consistent with the USDA library. Each spectrum in the test set was the average of three replicate spectra. Spectra were collected with the use of Varian Resolutions Pro 4.0.5.009 and WinIR Pro 3.2 software (Varian, Palo Alto, CA). All spectra were first converted to GRAMS format and then imported into MATLAB. The test set samples contained hull, leaf, seed coat, and stem samples, both intact and powdered, from nine different growing locations.

Standard Search Algorithms. Six standard search algorithms were used in the experiments: dot product, correlation, sum of the absolute values of differences, square root of the sum of the absolute values of differences, sum of the absolute values of differences of the derivatives, and sum of the squared differences of the derivatives. Each equation below can be thought of as a metric for the comparison of two spectra represented as vectors. The vector $\vec{\mathbf{I}}_j$ represents the j^{th} spectrum in the library, and the vector $\vec{\mathbf{u}}$ represents the spectrum searched against the library. The subscript i represents the i^{th} element of a vector, and n equals the number of elements in each vector or the number of resolution elements in each spectrum. The score describing how closely related the spectrum being searched is to the j^{th} spectrum in the library is represented by S_j . The dot product metric is given by Eq. 2.1:

$$S_j = \vec{\mathbf{I}}_j \bullet \vec{\mathbf{u}}_j \quad (2.1)$$

It should be noted that the dot product metric gives the same comparative information as an algorithm based on the sum of the squares of the differences between spectral resolution elements (often called a least squares algorithm in the literature), but the dot product metric requires a much smaller number of computations. For spectra vectors normalized to unit magnitude, the dot product metric is equivalent to the cosine of the angle between the vectors. A perfect match has a score of 1 and orthogonal spectra have a score of 0. Equation 2.2 is the sum of the absolute values of the differences algorithm, and Eq. 2.3 is the sum of the square root of the absolute values of the differences.

$$S_j = \sum_{i=1}^n |\tilde{\mathbf{I}}_{ji} - \bar{\mathbf{u}}_{ji}| \quad (2.2)$$

$$S_j = \sum_{i=1}^n |\tilde{\mathbf{I}}_{ji} - \bar{\mathbf{u}}_{ji}|^{\frac{1}{2}} \quad (2.3)$$

These first three metrics differ in the emphasis placed on small versus large differences between the spectra compared: from Eq. 2.1 to Eq. 2.2 to Eq. 2.3 more emphasis is placed on small differences between the spectra versus large differences. Equations 2.4 and 2.5 represent the sum of the absolute values of the differences of the derivatives and sum of the squared differences of the derivatives metrics.

$$S_j = \sum_{i=1}^n |(\tilde{\mathbf{I}}_{j,i+1} - \tilde{\mathbf{I}}_{ji}) - (\bar{\mathbf{u}}_{j,i+1} - \bar{\mathbf{u}}_{ji})| \quad (2.4)$$

$$S_j = \sum_{i=1}^n ((\tilde{\mathbf{I}}_{j,i+1} - \tilde{\mathbf{I}}_{ji}) - (\bar{\mathbf{u}}_{j,i+1} - \bar{\mathbf{u}}_{ji}))^2 \quad (2.5)$$

These derivative metrics are especially useful for the comparison of spectra with varying baseline shifts. For the metrics represented by Eqs. 2.2-2.5, the smaller the score between unknown and library spectra the more spectrally related the unknown spectrum is to the library spectrum, and a perfect match has a score of 0. Equation 2.6 is for the correlation metric:

$$S_j = \frac{\sum_{i=1}^n \bar{\mathbf{I}}_{ji} \bar{\mathbf{u}}_{ji} - \frac{\left(\sum_{i=1}^n \bar{\mathbf{I}}_{ji} \sum_{i=1}^n \bar{\mathbf{u}}_{ji} \right)}{n}}{\left[\left(\sum_{i=1}^n \bar{\mathbf{I}}_{ji}^2 - \frac{\left(\sum_{i=1}^n \bar{\mathbf{I}}_{ji} \right)^2}{n} \right) \left(\sum_{i=1}^n \bar{\mathbf{u}}_{ji}^2 - \frac{\left(\sum_{i=1}^n \bar{\mathbf{u}}_{ji} \right)^2}{n} \right) \right]^{\frac{1}{2}}} \quad (2.6)$$

For the correlation metric, a perfect match has a value of 1, and values of the score decrease as the similarity of the spectra decrease. In order to find the spectrum in a library that most closely matches the spectrum that is searched for a given algorithm, the score, S_j , for each spectrum in the library is calculated. The scores for a given standard algorithm are then ordered so that the spectra in the library are ranked by their similarity to the spectrum being searched. The scores returned by different standard search algorithms are not directly comparable, but the ranks of the results returned by each algorithm are directly comparable. The voting scheme algorithms described in the next section are a method of incorporating information from all of the standard search algorithms described for a given spectrum searched against the library.

Voting Scheme Algorithms. Three voting scheme algorithms were developed. Each algorithm combined the top 10 ranked results from the library spectra returned by each standard algorithm and then ranked these 60 matches according to three different criteria. The weighted frequency algorithm assigned a weight to each of the 60 matches. Any match that was ranked first by a standard algorithm received a weight of 10. Any match ranked second by a standard algorithm received a weight of 9. This pattern

continued until the matches given a rank of 10 by each of the standard algorithms were given a weight of 1. All of the weights associated with a given spectrum from the library were summed, and all of the spectra were then ranked by their associated sum of weights from highest to lowest. The highest sums were considered to be the best matches. This algorithm considered both the ranks of a given spectrum in the results of the standard algorithms and the number of standard algorithms that matched the spectrum.

The frequency algorithm ranked spectra solely on the number of standard algorithms that returned the given spectrum. This search would have been equivalent to a weighted frequency search where all 60 spectra were assigned a weight of 1.

The group algorithm counted the number of matches out of 60 that represented sample categories found in the library. The spectra in the library were assigned to belong to 1 of 21 possible groups. Group divisions included leaf, stem, seed coat, and hull. The number of matches that belonged to a particular group counted as the number of votes for a particular group. The group with the highest number of votes was considered to be the best match for the test spectrum that was searched.

Library Searching. Before searching, all spectra were truncated so that they ranged from 3700 to 2700 cm^{-1} and 1800 to 650 cm^{-1} . The region between 2700 to 1800 cm^{-1} was removed because none of the library or test spectra contained significant absorption bands in this region. The spectra in the test set were also de-resolved to 8 cm^{-1} resolution to match the library spectra. After these preprocessing steps, all spectra contained 558 resolution elements. All spectra were also given a common minimum by

subtraction of the lowest absorbance value in each spectrum from the absorbance values for every resolution element in that spectrum. All spectra were vector normalized to unit magnitude.

A 20 member subset of the test spectra set was chosen to test the performance of standard spectral library search algorithms. This subset was comprised of five spectra each of hull, leaf, seed coat, and stem spectra. Spectra of powdered samples (80 mesh) accounted for two or three spectra out of each group of five spectra. These spectra were searched against the library with the use of the six standard search algorithms, and the top 10 results returned by each algorithm for each spectrum in the subset were recorded. The combined top 10 results returned for each test spectrum by the standard algorithms were then searched with the use of the three voting scheme algorithms.

A second subset of 12 spectra from the 75 test spectra set and 12 spectra from the library were chosen to test the performance of the standard algorithms when test spectra were searched against a library that had been augmented by combining the 75 spectra test set into the USDA library. This 24 spectra test set contained six spectra each of hull, leaf, seed coat, and stem. The spectra for each category were comprised of three spectra from the original 929 spectra library and three spectra from the 75 member test set. When each spectrum was searched against the augmented library, all spectra in the library that were replicates of the spectrum to be searched were removed from the library before the search was executed. All algorithms were programmed in the MATLAB programming language and executed with MATLAB 7 software.

The performances of the standard algorithms were judged by the number of rank one results returned that correctly identified the test samples and by the total number of correct matches in the top 10 results returned for the test samples (e.g. If a leaf is being searched against the library, any leaf spectrum that is returned will be considered a correct result). The performances of the voting scheme algorithms were compared to the standard algorithms by the number of correct rank one results returned.

RESULTS AND DISCUSSION

To distinguish among the infrared spectra of plant based contaminants found in cotton is not a trivial task. Figure 1.2 shows the spectra of four cotton contaminant samples: a leaf, stem, seed coat, and hull. Each of these materials has cellulose as its main component, and thus, they all have very similar spectra. The absorbance band locations in these spectra are nearly identical from one spectrum to the next; however, the band intensities do vary. If all stems from all cotton plants had the same spectrum and if all hulls from all cotton plants had the same spectrum but the hull spectrum was different from the stem spectrum, then one would be able to distinguish the spectrum of a stem contaminant from the spectrum of a hull contaminant by the differences in band intensities between the two spectra. Unfortunately, this is not the case. Because of the differing concentrations of components that make up parts from different plants and multiple parts from the same plant, relatively large variations exist among the spectra of a given type of plant part when compared to the similarity of the spectra of different types of plant parts seen in Fig. 2.1. In other words, the within group variation is significant compared to the between group variation. Figure 2.2 demonstrates this

point for a group of leaf spectra. This figure shows the spectra of four different cotton leaves.

Despite these challenges, previous research has shown that more than 90% of organic based contaminant samples can be identified with the use of the USDA cotton contaminant library and standard spectral search algorithms when the spectra in the library are representative of the spectra being searched and the unknown spectra are acquired with the same instrument used to measure the library spectra¹. Representative means that spectra of cultivars from the same geographic region and time period and grown under similar conditions as the sample represented by the unknown spectrum being searched must be contained in the library. In this work, spectra of samples grown in different geographic regions and different time periods than the library spectra were used to test the performance of standard algorithms. Additionally, the test spectra were measured with a different instrument and different ATR accessories than the library spectra. It is well known that transferring spectral calibrations from one spectrometer to another can be difficult because of instrumental differences, and these instrumental differences also affect the performance of library searches. For true versatility, a spectral library needs to provide consistent results despite these sources of variation, and the experiments conducted in this study were designed to check the performance of the library under these conditions.

When a test set of 20 spectra that represented samples from growing regions, years, and environmental conditions not represented in the library and that were measured with different spectrometers and accessories was searched with the use of

Figure 2.1: Spectra of leaf, stem, seed coat, and hull from the cotton plant. The spectra have been offset for clarity.

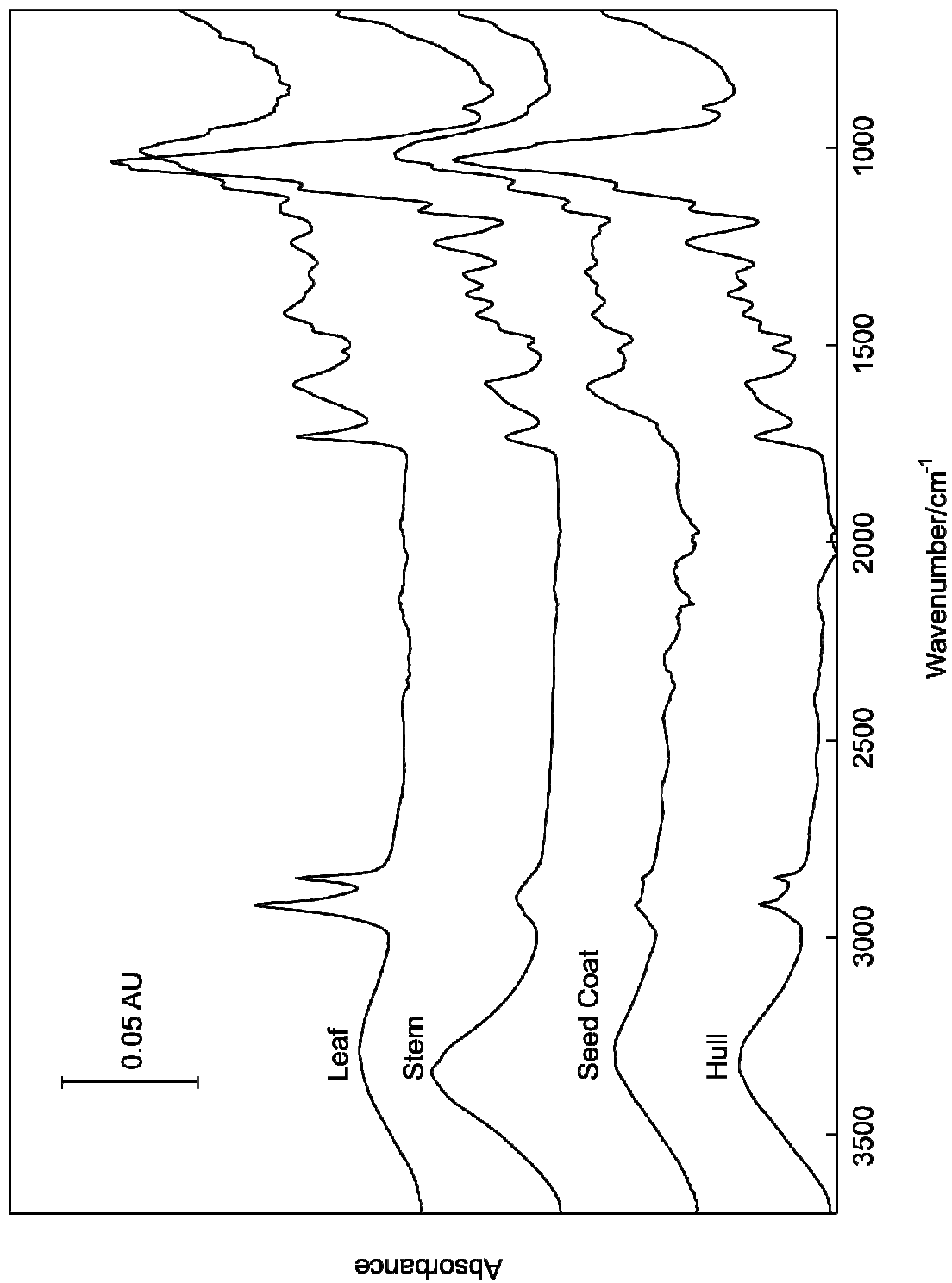
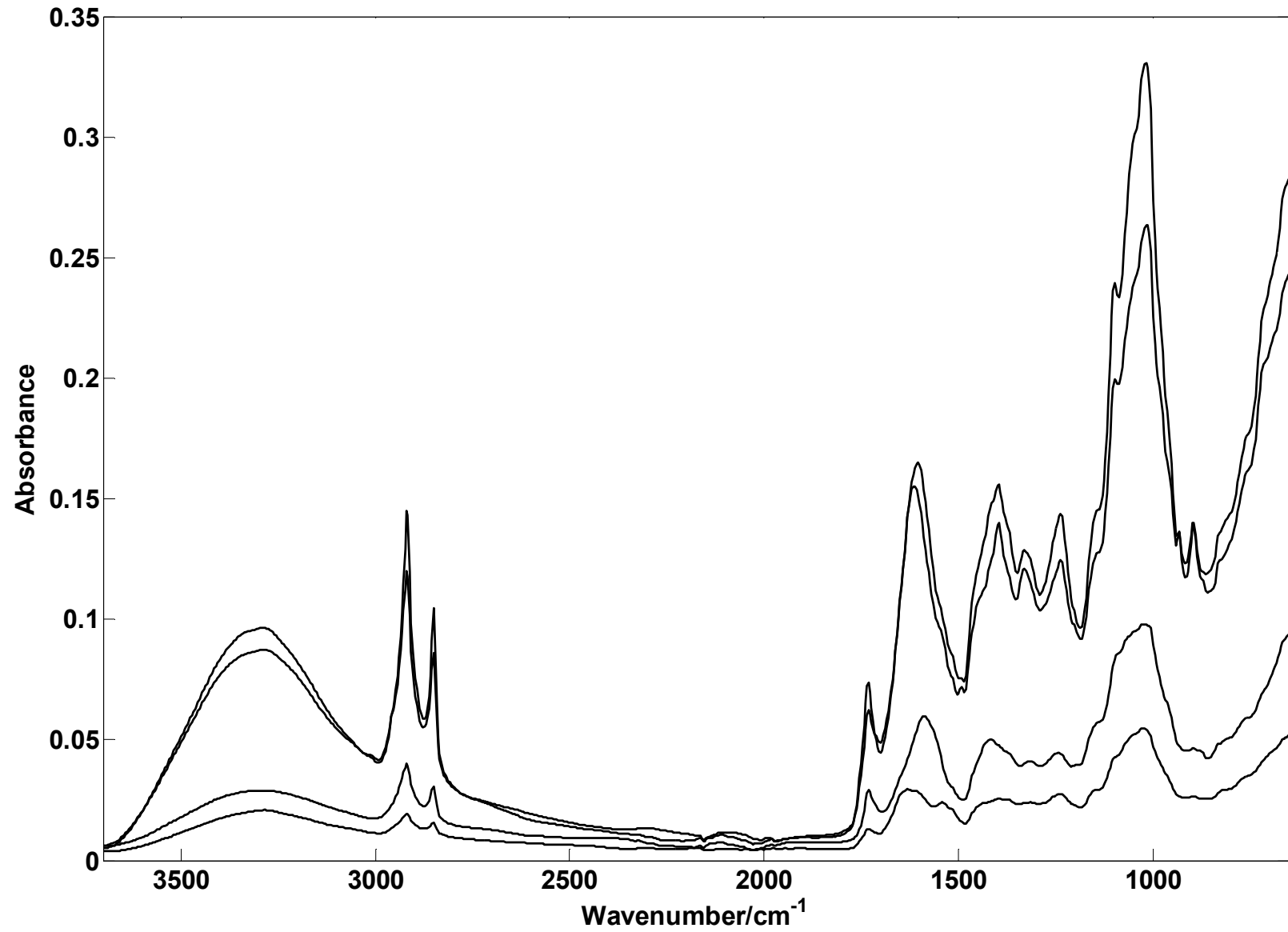


Figure 2.2: Spectra of four different cotton leaves.



standard search algorithms, the standard search algorithm that produced the highest number of rank one identifications was able to correctly identify only 12 out of 20 spectra. More serious was the differing performance of the standard search algorithms. Figure 2.3 shows the number of correct rank one results returned by each of the six standard algorithms, broken down by the four categories of plant parts. One can see that a different algorithm performed best for each of the four categories of plant parts tested against the library. Furthermore, in some cases, the best performing algorithms for a given sample category were the worst performing algorithms for a different category of samples. For example, the correlation and square root algorithms correctly identified the most seed coat samples of any of the algorithms, but the correlation and square root algorithms identified fewer stem samples than any of the other algorithms. These data revealed that one can have significantly different rates of identification that depend upon the algorithm chosen to search this library. Since there is no single standard algorithm capable of identifying all sample types, one's chances of successfully identifying an unknown spectrum depend on the choice of algorithm, but for a true unknown sample, there is no way to know if the best algorithm has been chosen. These facts pointed to the need for a voting scheme algorithm.

Figure 2.4 shows the total number of correct results returned in the top 10 results by each of the six standard algorithms. From looking at these results, one might predict that it would be best to use the top 10 matches returned by one of the derivative algorithms as a basis for a voting algorithm, but Fig. 2.5 shows why this approach is not

Figure 2.3: Number of samples in the 20 member test set correctly identified by the first result returned by each standard algorithm. The results are broken down by sample type as indicated in the legend.

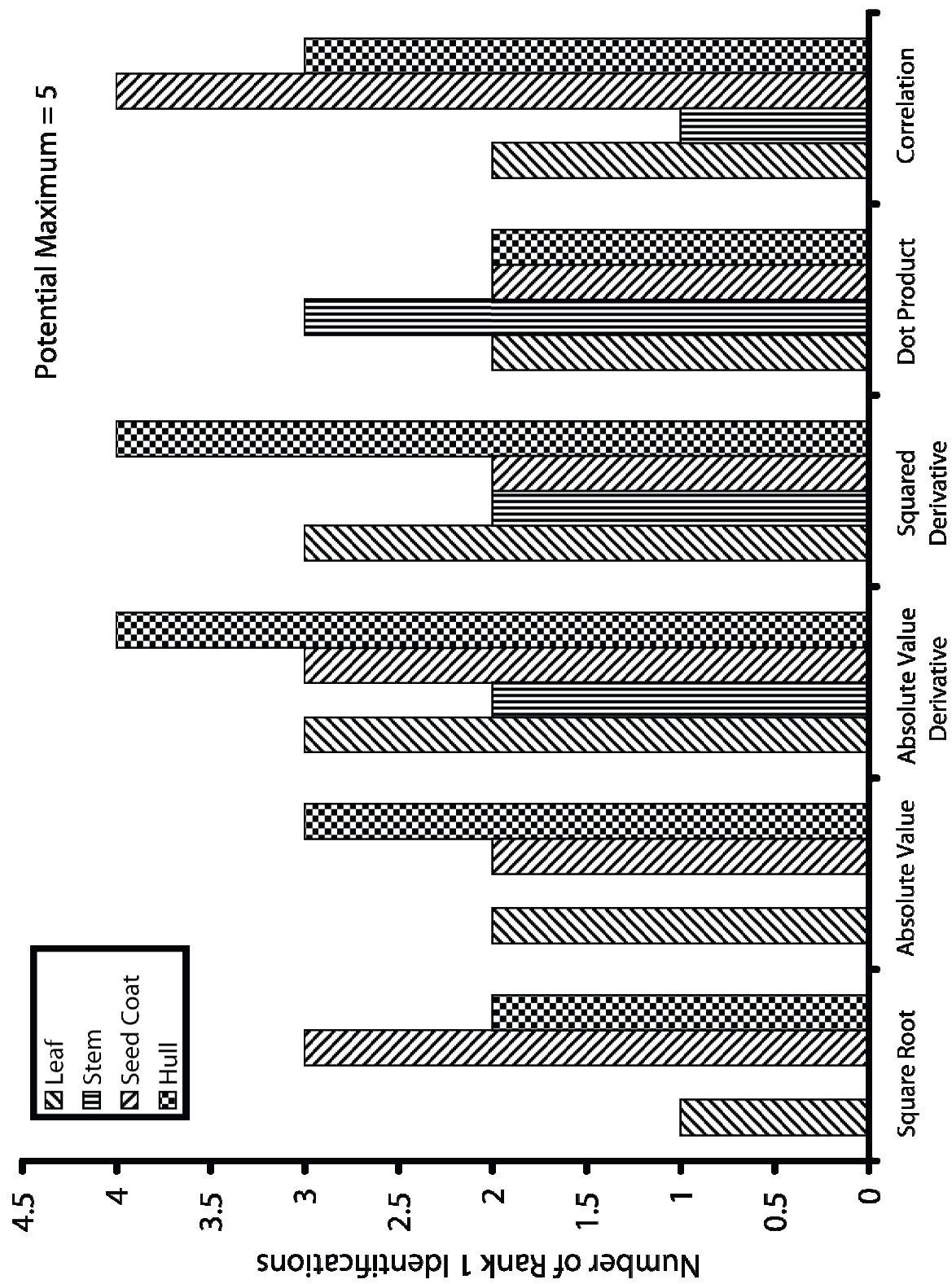


Figure 2.4: Total number of results in the top 10 results returned by each of the standard algorithms for each of the 20 test set samples that correctly identified a test spectrum.

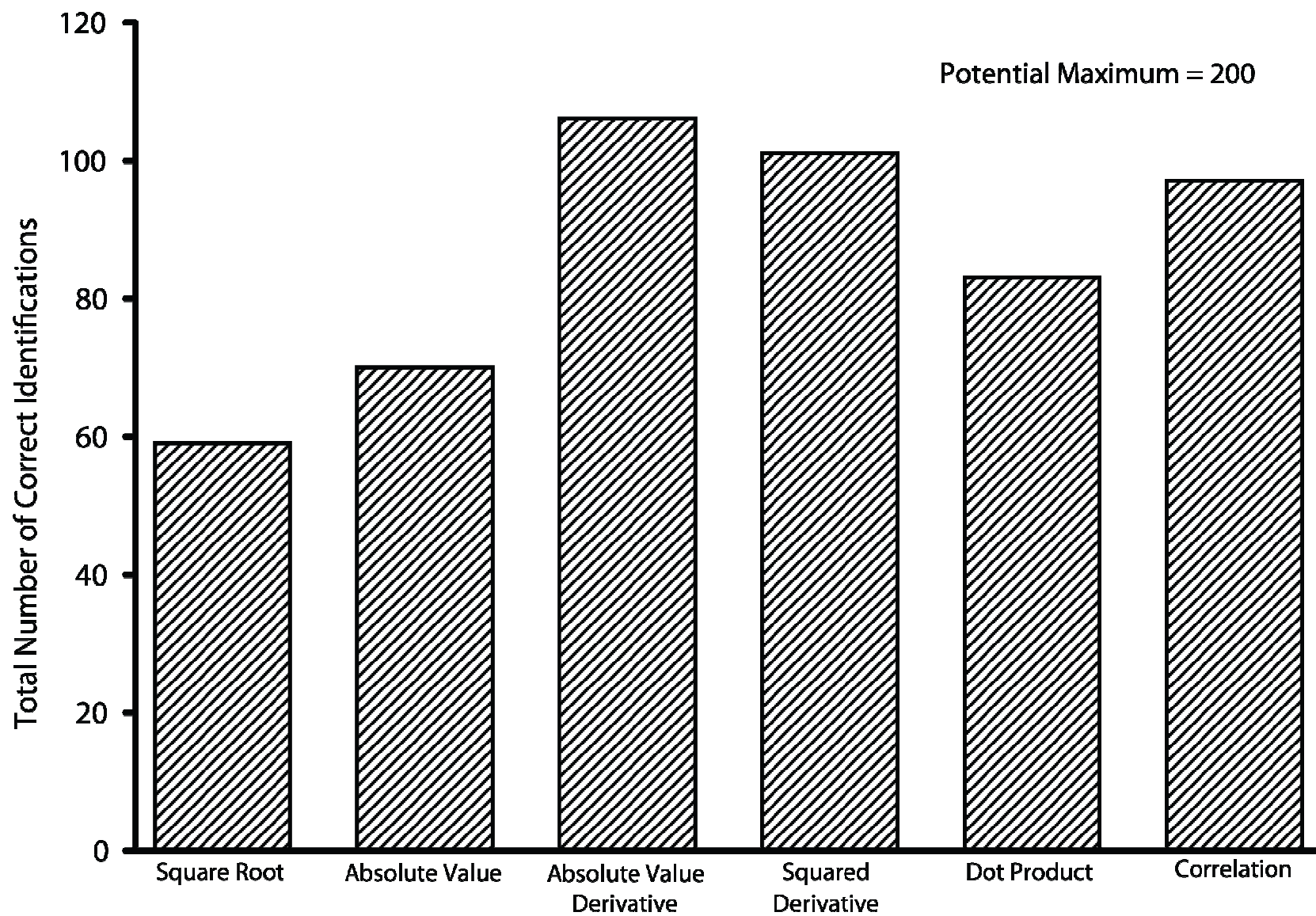
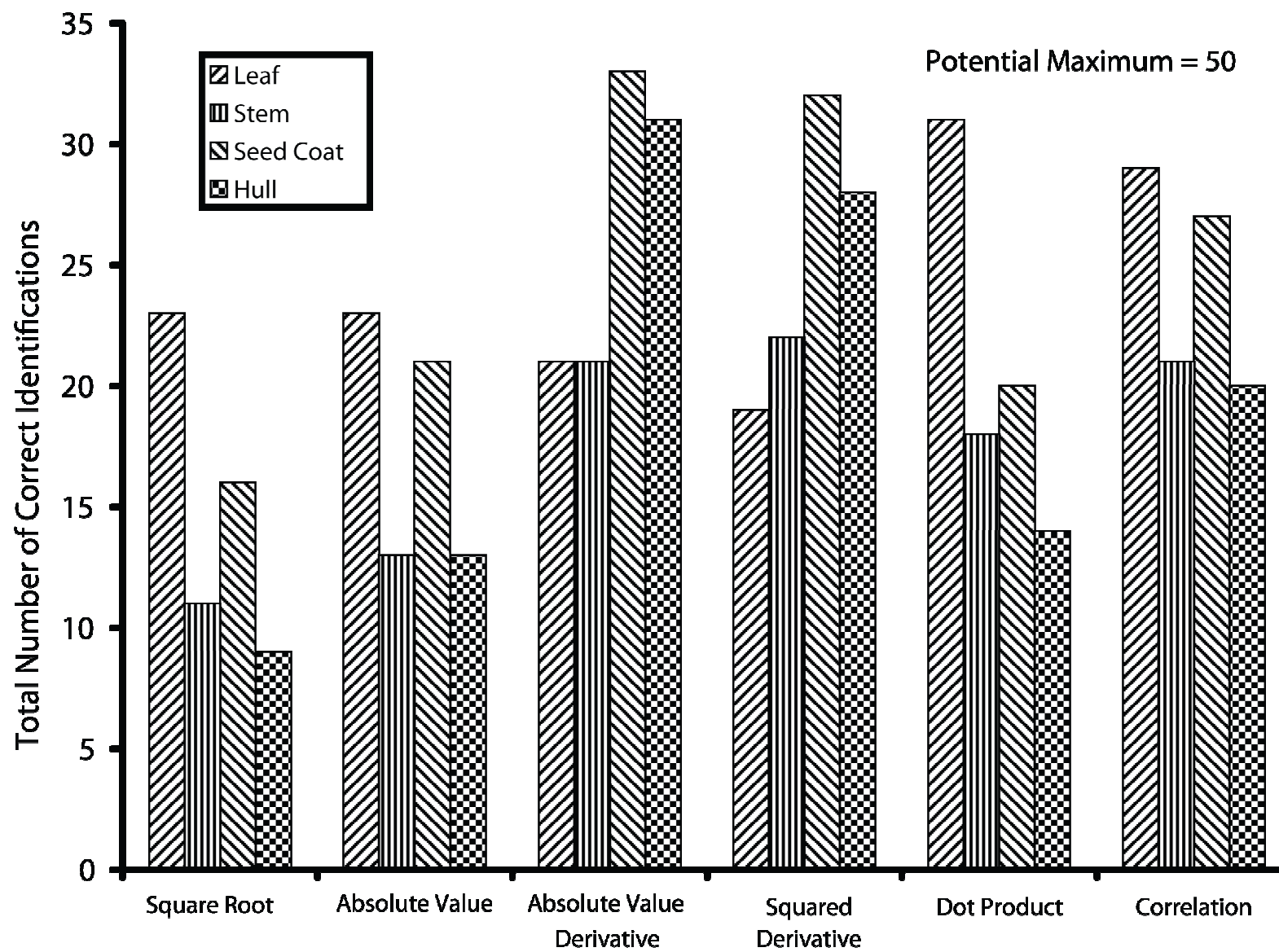


Figure 2.5: Total number of results in the top 10 results returned by each of the standard algorithms for each of the 20 test set samples that correctly identified a test spectrum.

The results are broken down by sample type as indicated in the legend.



ideal. As with the number of correct rank one results (see Fig. 2.5), the algorithm that produces the greatest number of correct matches in the top ten results returned varies by sample category. These results are indicative of the fact that for a given unknown sample the algorithm that produces the greatest number of correct answers cannot be predicted. For instance, Table 2.1 shows the actual results returned for a leaf sample from the test set when searching with the absolute value and the absolute value derivatives algorithms. From the results shown in Fig. 2.4, one would have predicted that the derivative algorithm would have given the greatest number of correct answers, but in this case (see Table 2.1) the absolute value algorithm yielded five correct matches to the derivative algorithm's one correct match. This example demonstrates why any successful voting scheme algorithm must incorporate information from all six standard algorithms.

The three voting scheme algorithms developed were designed to take advantage of both complementary and repetitive information among the top 10 results returned by the six algorithms. The results returned by the standard algorithms varied because each algorithm measures the similarity of spectra by a different metric, and since these spectra are so similar to begin with, different orderings of the most likely candidate spectra occur. Despite this fact, one still expects that some of the algorithms will return some of the exact same spectra. This is the principle behind the frequency algorithm. The number of times a particular result shows up in the hit lists of the six algorithms the more likely that result correctly identifies the unknown contaminant.

Table 2.1: Search results for a leaf powder sample.

| Rank | Absolute Value | | | Absolute Value Derivative | | |
|------|----------------|----------|------------------------------|---------------------------|----------|------------------------------|
| | Score | Category | Index Number ^a | Score | Category | Index Number ^a |
| 1 | 1.318 | Hull | 703 | 0.2118 | Leaf | 363 |
| 2 | 1.636 | Hull | 697 | 0.2129 | Hull | 703 |
| 3 | 1.651 | Bract | 65 | 0.2130 | Hull | 478 |
| 4 | 1.702 | Stem | 518 | 0.2157 | Hull | 481 |
| 5 | 1.785 | Leaf | 363 | 0.2178 | Hull | 697 |
| 6 | 1.788 | Leaf | 364 | 0.2214 | Stem | 21 |
| 7 | 1.815 | Stem | 805 | 0.2219 | Hull | 696 |
| 8 | 1.843 | Leaf | 525 | 0.2223 | Bloom | 35 |
| 9 | 1.874 | Leaf | 523 | 0.2256 | Hull | 699 |
| 10 | 1.877 | Leaf | 526 | 0.2261 | Hull | 695 |

^aUnique index number assigned to the spectra from the library.

The weighted frequency algorithm modified this approach by considering both the number of algorithms that returned a particular result and the algorithm specific ranks of the results returned. This algorithm considers the facts that both frequency of a particular result among the 60 hits and the rank of those results among the individual 10 member hit lists are indicators of the probability of a particular result correctly matching the unknown. The group algorithm uses the frequency of particular results returned in a different manner. This algorithm is only considering the number of results returned in a particular category.

Table 2.2 shows an example of how these algorithms work. The results returned by searching a particular seed coat spectrum from the test set with the six standard algorithms are shown. The group search algorithm would reveal that the group returned most often was seed coat with 36 out of 60 results returned. The frequency search would focus on the fact that seed coat samples 167 and 624, appearing in the results five times each, tied for the first ranked result with hull samples 481 and 703, which also appeared five times each. Finally, the weighted voting search revealed that when frequency and rank are considered seed coat sample 624 is ranked number 1 with a score of 42.

Table 2.3 summarizes the top matches returned by each of the voting scheme algorithms for the same seed coat test spectrum for which standard algorithm results are shown in Table 2.2. The results of the weighted frequency voting and group voting searches definitely identify the test spectrum as seed coat, while only four of the six standard algorithms had a correct rank one match for this test spectrum (see Table 2.2).

Table 2.2: Top 10 ranked results for all standard algorithms for a cotton seed coat powder spectrum.

| Rank | Square Root | | | Squared Derivative | | | Absolute Value | | |
|------|-------------|--------------|-----------------|--------------------|--------------|-----------------|----------------|--------------|-----------------|
| | Score | Category | Index Number | Score | Category | Index Number | Score | Category | Index Number |
| 1 | 25.15722 | Seed Coat | 165 | 0.0001371 | Stem | 20 | 1.49286 | Seed Coat | 166 |
| 2 | 26.28441 | Seed Coat | 166 | 0.0001542 | Hull | 481 | 1.50666 | Seed Coat | 165 |
| 3 | 27.73352 | Seed Coat | 624 | 0.0001582 | Seed Coat | 624 | 1.61960 | Seed Coat | 624 |
| 4 | 27.75385 | Hull | 703 | 0.0001615 | Seed Coat | 618 | 1.75337 | Hull | 703 |
| 5 | 29.44865 | Hull | 481 | 0.0001620 | Seed Coat | 167 | 1.83423 | Seed Coat | 164 |
| 6 | 29.76650 | Seed Coat | 164 | 0.0001622 | Hull | 697 | 1.83423 | Seed Coat | 168 |
| 7 | 29.76650 | Seed Coat | 168 | 0.0001630 | Seed Coat | 625 | 1.89044 | Hull | 481 |
| 8 | 29.81726 | Hull | 696 | 0.0001632 | Seed Coat | 628 | 1.91286 | Seed Coat | 167 |
| 9 | 29.83374 | Cacyx | 122 | 0.0001641 | Hull | 696 | 1.93826 | Hull | 696 |
| 10 | 29.84427 | Bloom | 35 | 0.0001726 | Hull | 703 | 1.95001 | Seed Coat | 623 |

Table 2.2: (continued)

| Rank | Dot Product | | | Absolute Value Derivative | | | Correlation | | |
|------|-------------|--------------|-----------------|---------------------------|--------------|-----------------|-------------|--------------|-----------------|
| | Score | Category | Index Number | Score | Category | Index Number | Score | Category | Index Number |
| 1 | 0.99224 | Seed Coat | 166 | 0.18702 | Stem | 20 | 0.99080 | Seed Coat | 166 |
| 2 | 0.99150 | Seed Coat | 164 | 0.19129 | Seed Coat | 624 | 0.98959 | Seed Coat | 624 |
| 3 | 0.99150 | Seed Coat | 168 | 0.19322 | Seed Coat | 628 | 0.98957 | Seed Coat | 165 |
| 4 | 0.99100 | Seed Coat | 165 | 0.19403 | Hull | 481 | 0.98754 | Seed Coat | 164 |
| 5 | 0.99080 | Leaf | 244 | 0.19486 | Hull | 697 | 0.98754 | Seed Coat | 168 |
| 6 | 0.99050 | Seed Coat | 623 | 0.19535 | Seed Coat | 167 | 0.98581 | Seed Coat | 167 |
| 7 | 0.99048 | Bract | 76 | 0.19782 | Hull | 489 | 0.98573 | Seed Coat | 627 |
| 8 | 0.99044 | Hull | 699 | 0.19816 | Seed Coat | 625 | 0.98537 | Hull | 703 |
| 9 | 0.99032 | Seed Coat | 167 | 0.19844 | Hull | 703 | 0.98441 | Seed Coat | 623 |
| 10 | 0.99017 | Stem | 26 | 0.19854 | Seed Coat | 618 | 0.98384 | Hull | 481 |

Table 2.3: Results from voting scheme algorithms for the same seed coat powder sample for which standard algorithm results are shown in Table 2.2.

| Rank | Weighted Frequency | | | Score | Frequency | | | Group | |
|------|--------------------|----------------|---------------------------------|-------|----------------|---------------------------------|--------------------|----------------|---------------------------------|
| | Score | Substance Name | Index/Group Number ^a | | Substance Name | Index/Group Number ^a | Score ^b | Substance Name | Index/Group Number ^c |
| 1 | 42 | Seed Coat | 624 | 5 | Seed Coat | 167 | 36 | Seed Coat | 11 |
| 2 | 39 | Seed Coat | 166 | 5 | Hull | 481 | 17 | Hull | 18 |
| 3 | 34 | Seed Coat | 165 | 5 | Seed Coat | 624 | 3 | Stem | 5 |
| 4 | 27 | Hull | 481 | 5 | Hull | 703 | 1 | Bloom | 6 |
| 5 | 27 | Seed Coat | 164 | 4 | Seed Coat | 164 | 1 | Bract | 7 |
| 6 | 23 | Seed Coat | 168 | 4 | Seed Coat | 165 | 1 | Leaf | 8 |
| 7 | 21 | Seed Coat | 167 | 4 | Seed Coat | 166 | 1 | Cacyx | 10 |
| 8 | 20 | Hull | 703 | 4 | Seed Coat | 168 | | | |
| 9 | 20 | Stem | 20 | 3 | Seed Coat | 623 | | | |
| 10 | 11 | Seed Coat | 628 | 3 | Hull | 696 | | | |

^aUnique index number assigned to the spectra from the library.

^bAll 60 possible spectra to be ranked are included in 7 results.

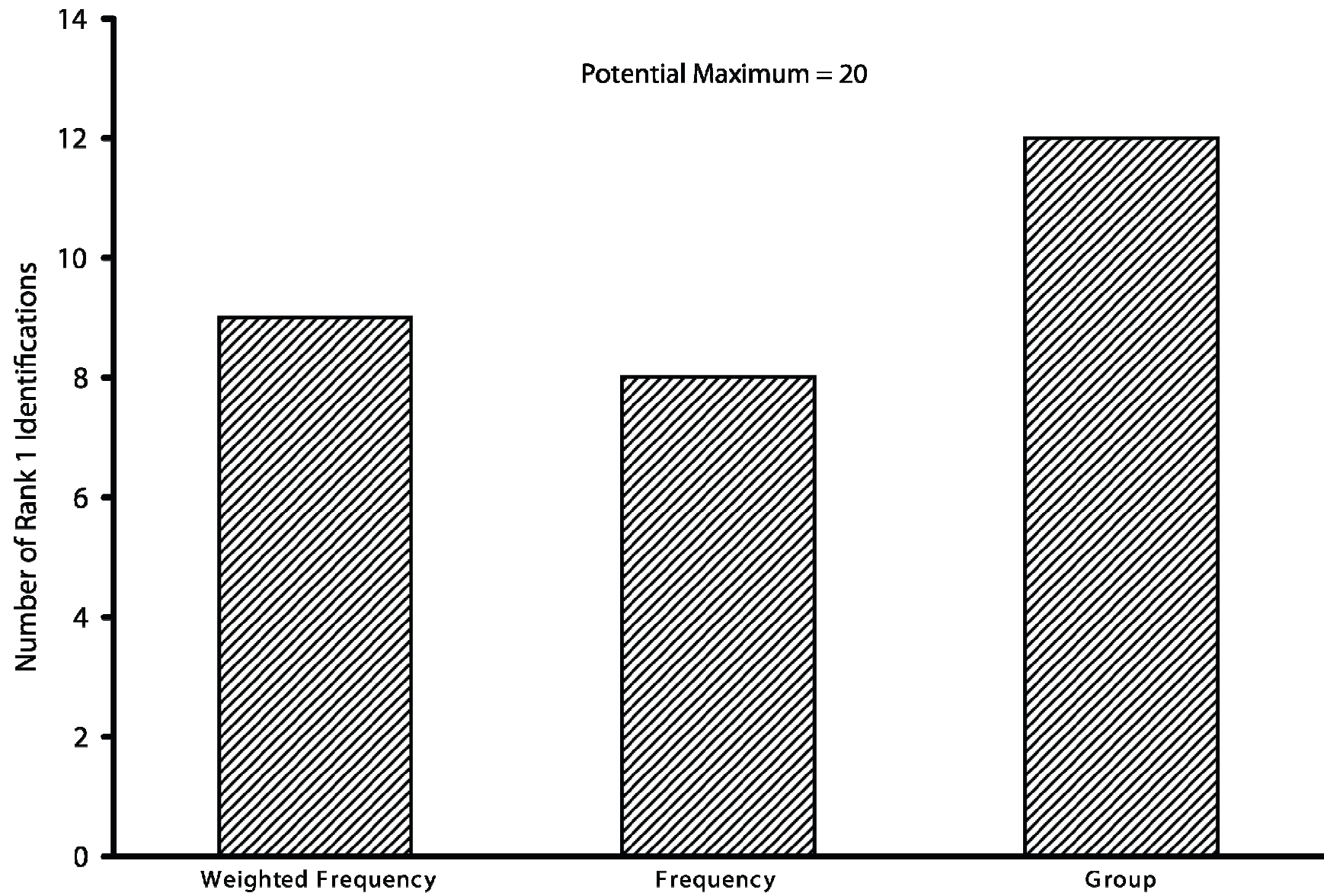
^cUnique numbers assigned to each category of spectra represented in the library.

One should also notice that the derivative algorithms are the algorithms that did not have a correct rank one match for the test spectrum. Recall that in Fig. 2.4 the derivative algorithms were shown to provide the largest total of correct results in the top ten results returned by the six standard algorithm, but in this case the derivative algorithms did not give a correct rank 1 result. This outcome reiterates the fact that the six standard algorithms are not simply giving repetitive information: All of the algorithms seem to be giving some distinct information useful to correctly identify unknown spectra. The ability of these voting scheme algorithms to take advantage of both complementary and repetitive information provided by the standard algorithms' results makes these voting scheme algorithms valuable.

The results for the voting scheme algorithms for the entire test set are shown in Fig. 2.6. The data for the group algorithm show impressive performance: The group search was able to yield as many correct rank one matches as the best standard algorithm. By use of the group search algorithm, one does not need to make a decision as to which standard algorithm should be used and risk choosing a poorly performing algorithm. The group search uses the discriminating information given by all of the standard algorithms to overcome the problem of different standard algorithms performing differently for different sample types (see Figs. 2.3 and 2.4).

Figure 2.6 also reveals that while the frequency and weighted frequency algorithms performed better than some of the standard search algorithms they do not have the same discriminating power that the group search algorithm has. One can gain some insight into this outcome by looking at the results shown in Table 2.3. Regarding

Figure 2.6: Number of samples in the 20 member test set correctly identified by the first result returned by each of the voting scheme algorithms.



the frequency algorithm, the results show that many ties occur among the hits. This behavior is typical of the results obtained by this algorithm for all of the test samples, and this lack of discrimination means that this algorithm is not as successful at distinguishing among spectra as the other algorithms. The rank information considered by the weighted frequency algorithm, in addition to the frequency information, generally gives the weighted frequency algorithm more discrimination power than the frequency algorithm. This trend is also demonstrated in the results contained in Table 2.3. The example in Table 2.3 cannot show us why the group search generally performs better than the weighted voting search, but one can infer the reasons from the search data for the test set on the whole. In the top 10 results returned by the six standard algorithms, correct answers show up often, but are not always highly ranked. Also, these six standard algorithms all calculate spectral similarity so that they often return the correct category of sample in their results, but these algorithms use metrics to calculate spectral similarity that differ enough that particular spectra from the library are not consistently returned by all of the standard algorithms. These factors hurt both the frequency search and the weighted frequency search since these algorithms depend upon a particular spectrum from the library being returned consistently, or consistently and highly ranked, respectively, in the results from the standard algorithms. These same factors are beneficial to the group search algorithm because it does not rely on rank or the consistency of particular results being returned. The group search only considers the category of spectrum being returned.

In order to demonstrate that the dependence of algorithm performance on sample category was related to the test set samples not being represented in the library, the 75 test spectra were added to the library. Twelve test spectra along with 12 spectra from the original library were searched against this augmented library, as described earlier, to show that when spectra representative of the test set were included in the library a high identification rate could be obtained by the standard algorithms. This experiment was a legitimate use of the 75 member test set because replicate spectra of the same plant parts represented by the test set were removed from the augmented library before each search was conducted. The results of this experiment are summarized in Fig. 2.7. The absolute value derivative and squared derivative algorithms yielded correct rank one matches for 22 out of 24 test spectra, and the number of correct rank one results improved for all of the standard algorithms. Figure 2.8 shows how each of the standard algorithms performed by sample category. The absolute value derivative and squared derivative algorithms consistently performed the best for all sample categories tested. The data show that when spectra representative of the test set were found in the library the performances of the standard search algorithms were predictable. These results demonstrated the importance of making the library representative of the spectra that are going to be searched against it, but in cases where the library cannot be made completely representative of the unknown spectra to be searched, the voting scheme algorithms provide a way to overcome the problem of not being able to choose a single standard algorithm for the searches.

Figure 2.7: Number of samples in the 24 member test set searched against the augmented library that was correctly identified by the first result returned.

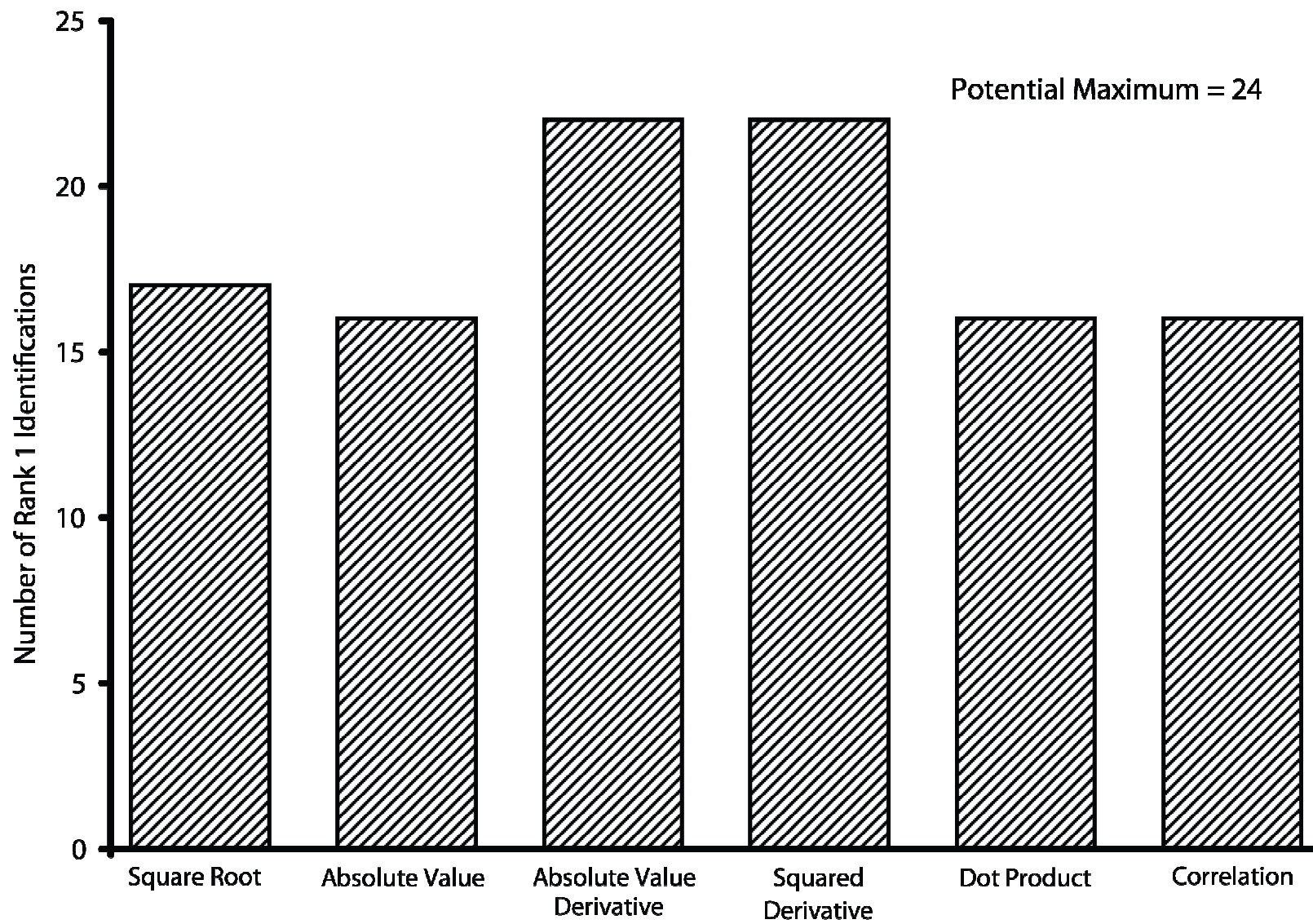
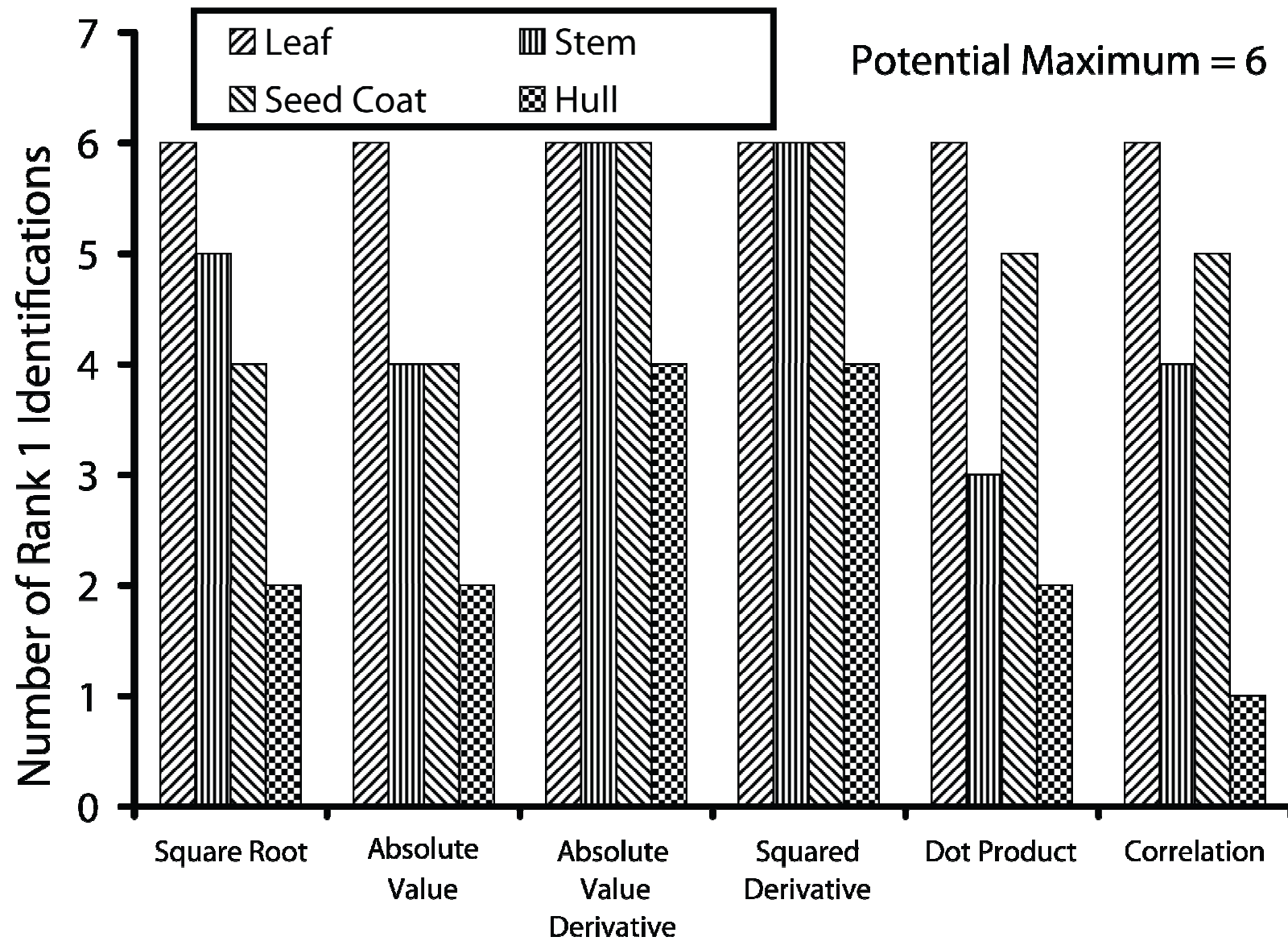


Figure 2.8: Number of samples in the 24 member test set searched against the augmented library that was correctly identified by the first result returned. The results are broken down by sample type as indicated in the legend.



CONCLUSION

A spectral library of cotton contaminants had previously been developed to aid in the identification of foreign matter of extremely similar chemical composition and with closely related spectra found in cotton lint. Our experiments demonstrated that when this library was representative of the types of samples being searched against it, standard library searching algorithms accurately identified test samples, but when spectra of samples grown in different geographic locations, seasons, and environmental conditions and measured with a different spectrometer and ATR accessories were searched against the library, the identification rates for standard spectral search algorithms decreased significantly. Compounding this problem was the fact that under these conditions one could not reliably choose a standard search algorithm to search unknown spectra against the library because the performances of the standard algorithms varied by sample type; consequently, the best performing algorithm could not be predicted. Our experiments showed that by using the group voting scheme algorithm based on the numbers of samples returned from each category of samples represented in the library, a number of rank one identifications equal to the best standard algorithm could be obtained. The success of this voting scheme is due to the fact that the information gained from different standard search algorithms is complementary and repetitive. By using the information gained from multiple standard search algorithms, a more reliable and robust search algorithm was created. In addition to the identification of cotton contaminants in an underrepresented library, these voting scheme techniques could have applications to other spectral libraries with

significant within group variation compared to between group variation, such as biological or forensic spectral libraries.

ACKNOWLEDGEMENT

The authors sincerely thank Drs. John Foulk and Angela Allen from the USDA-ARS Cotton Quality Research Laboratory in Clemson, South Carolina for providing the contaminant samples that were used to measure the spectra test set.

REFERENCES

1. D. S. Himmelsbach, J. W. Hellgeth, and D. D. McAlister, *J. Agric. Food Chem.* **54**, 20, 7405 (2006).
2. A. D. Brashears, R. V. Baker, and C. K. Bragg, "Effect of Bark on Spinning Efficiency of Cotton", in *Proceedings of the Beltwide Cotton Conference* (1992), p. 1218.
3. J. Foulk, D. McAlister, D. Himmelsbach, and E. Hughs, *J. Cotton Sci.* **8**, 243 (2004).
4. B. Xu, C. Fang, and R. Huang, *Text. Res. J.* **67**, 12, 881 (1997).
5. B. Xu and C. Fang, *Text. Res. J.* **69**, 9, 656 (1999).
6. S. R. Lowry, D. A. Huppler, and C. R. Anderson, *J. Chem. Inf. Comp. Sci.* **25**, 3, 235 (1985).
7. A. W. Baker, N. Wright, and A. Opler, *Anal. Chem.* **25**, 10, 1457 (1953).
8. L. E. Kuentzel, *Anal. Chem.* **23**, 10, 1413 (1952).
9. D. H. Anderson and G. L. Covert, *Anal. Chem.* **39**, 11 (1967).
10. D. S. Erley, *Anal. Chem.* **40**, 6, 894 (1968).

11. R. A. Sparks, "Storage and Retrieval of Wyandotte-ASTM Infrared Spectral Data Using an IBM 1401 Computer", (ASTM, Philadelphia, PA, 1964).
12. A. Hanna, J. C. Marshall, and T. L. Isenhour, *J. Chromatogr. Sci.* **17**, 434 (1979).
13. K. Tanabe and S. Saëki, *Anal. Chem.* **47**, 1, 118 (1975).
14. L. A. Powell and G. M. Hieftje, *Anal. Chim. Acta* **100**, 313 (1978).

CHAPTER 3

QUALITATIVE IDENTIFICATION OF COTTON CONTAMINANTS BY PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS¹

¹ J.B. Loudermilk, D.S. Himmelsbach, F.E. Barton II, J.A. de Haseth, To be submitted to Appl. Spectrosc.

ABSTRACT

Identification and removal of cotton contaminants are important goals of cotton research. Contaminants harvested along with the desired cotton fibers include other parts of the cotton plant such as leaves, stems, seeds, and hulls. Identification and removal of these contaminants is important because of the detrimental effects contaminants have on product quality and profitability. During transportation and processing, these contaminants break down in size, which makes visual identification difficult at best. The United States Department of Agriculture (USDA) created a mid-IR spectral library to enable identification of contaminants by the use of infrared spectra. In a previous paper, the authors described the difficulty of searching samples not representative of the geographic regions and growing conditions represented in the library against the library with standard spectral search algorithms. In that previous work, a novel voting scheme algorithm capable of overcoming the unreliability of standard algorithms was introduced. The current work explores the use of partial least squares discriminant analysis (PLS-DA) for contaminant identification of non-represented samples. PLS-DA identified 80% of the test set samples compared to 60% for the previously described voting scheme algorithm. PLS-DA was shown to be a good choice for classification of samples that have extremely similar spectra.

Index Headings: Discriminant Analysis; Partial least squares; PLS-DA; cotton; contaminants; Spectral discrimination; Spectral library.

INTRODUCTION

During harvest, organic contaminants such as cotton leaf, stem, seed coat, and hull are harvested along with fibers from the cotton plant that are to be spun into yarn. In cotton research, there has been a large and sustained emphasis on the investigation of the effects of these contaminants on cotton quality and processing efficiency¹⁻¹³. These contaminants cause an increased number of yarn breakages during the spinning process⁶, and are also a cause of yarn imperfections⁴. Since any factors that lower quality of the final product or decrease production efficiency negatively affect profitability of textile manufacturing, identification and removal of cotton contaminants is important. A method to identify the debris present in cotton would allow for correlation of yarn quality and process efficiency to the types of contaminants present¹⁴. This information would allow production of yarn from plant harvest to completed product to be optimized for contaminant removal.

As contaminants travel from the fields to the processing facilities and as they travel through different stages of cotton processing, they break down in size. While it is often possible to recognize the intact contaminants, the small pieces and contaminant powders that result from breakdown are difficult or impossible to distinguish visually. To solve this problem, the USDA created a mid-IR spectral library of cotton contaminants to identify contaminants from their infrared spectra. Previous work has described the introduction and use of this spectral library¹⁴⁻¹⁶.

In a previous article, the authors described the difficulties associated with searching this library when the library is not sufficiently representative of the

geographic regions and growing seasons of the unknown samples to be identified¹⁵. Despite vector normalization, spectra of unknown samples from instruments different from the one used to obtain the library can also be a challenge for a non-representative library. This previous work examined the performance of several standard library search algorithms and found that when the library is not sufficiently representative the performance of the standard algorithms varied unpredictably by sample type. That paper reported the development of a novel library searching method based on a voting scheme algorithm that solved the problem of not being able to choose a best standard search algorithm *a priori*.

The purpose of the current work is to explore qualitative identification of cotton contaminants with PLS-DA. This work demonstrates that PLS-DA performs better as an identification method for cotton contaminants than the previously described voting scheme algorithm¹⁵.

MATERIALS AND METHODS

Several sets of spectra used in the creation and testing of two PLS-DA models are described below. The models will be referred to as model A and model B. The sets of spectra are summarized in Table 3.1 for the reader's convenience.

Model A Calibration Set. A set of 354 spectra from a USDA cotton contaminant spectral library¹⁶ was selected for a PLS-DA calibration set. The set contained 87 cotton leaf spectra, 73 cotton stem spectra, 48 cotton seed coat spectra, and 146 cotton hull spectra. The spectra represented intact plant parts. The original library spectra were obtained with the use of a Nicolet Magna 850 FT-IR spectrometer (Thermo Fisher

Table 3.1: Summary of spectra sets used in experiments.

| Set Name | Number of Spectra | Description |
|-------------------------|----------------------|--|
| Model A Calibration Set | 354 | Spectra from the USDA Cotton Contaminant Library |
| Augmentation Set | 75 | Spectra of samples from different growing locations and seasons than represented in the USDA Cotton Contaminant Library |
| Model B Calibration Set | 330 | Spectra chosen from the model A calibration set combined with spectra chosen from the augmentation set |
| Model A Test Set | 20 | Spectra chosen from the augmentation set |
| Model B Test Set | 12 | Spectra chosen from the model A test set and the model A calibration set |

Scientific, Waltham, MA) and a DuraScope attenuated total reflection (ATR) sampling accessory (Smiths Detection, Danbury, CT). The spectrometer contained a ceramic source, a KBr beamsplitter, and a L-Alanine doped DGTS detector. The ATR accessory contained a diamond coated ZnSe internal reflection element (IRE). Spectra were measured over the range of 4000 to 650 cm^{-1} at 8 cm^{-1} resolution, with 128 interferograms co-added. Interferograms were processed with Happ-Genzel apodization. Spectra were collected with the use of Omnic E.S.P. 5.2 software (Thermo Fisher Scientific, Waltham, MA). All spectra were first converted to GRAMS format (Thermo Fisher Scientific, Waltham, MA) and then imported into MATLAB (The Math Works, Natick, MA).

Augmentation Set. A set of 75 spectra of samples from several geographic locations and seasons different from the samples in the USDA library were measured with the use of a Varian Excalibur Series FTS-4000 FT-IR spectrometer (Varian, Palo Alto, CA) and three different ATR sampling accessories. The set contained hull, leaf, seed coat, and stem samples, both intact and powdered, from 9 different growing locations. The spectrometer contained a ceramic source, a KBr beamsplitter, and a L-Alanine doped DGTS detector. The ATR accessories were the Specac Golden Gate (Specac, Woodstock, GA) with a diamond coated ZnSe IRE and the Harrick SplitPea and Seagull (Harrick Scientific, Pleasantville, NY) with Si and ZnSe IREs, respectively. Spectra were measured over the range of 4000 to 400 cm^{-1} at 4 cm^{-1} resolution, with 256 interferograms co-added. Interferograms were processed with Happ-Genzel apodization to be consistent. Each spectrum in the test set was the average of three

replicate spectra. Spectra were collected with the use of Varian Resolutions Pro 4.0.5.009 and WinIR Pro 3.2 software (Varian, Palo Alto, CA). All spectra were first converted to GRAMS format and then imported into MATLAB.

Model A Test Set. A set of 20 spectra was randomly chosen from the augmentation set to serve as a test set for the PLS-DA model created from the model A calibration set. This test set included five spectra each of leaf, stem, seed coat, and hull samples.

Model B Calibration Set. A set of 330 spectra was chosen from the model A calibration set and the augmentation set of spectra to create a PLS-DA model representative of both the model A calibration set and the augmentation set of spectra.

Model B Test Set. A set of 12 spectra was drawn from the model A calibration set and the augmentation set to test the PLS-DA model formed from the Model B calibration set. This test set included four spectra from the Model A test set that were not identified by Model A, four spectra from the Model A test set that were identified by Model A, and four spectra from the Model A calibration set.

Spectra Pretreatment. Before spectra were used in any of the experiments, they were truncated to include only the regions from 3700 to 2700 cm^{-1} and 1800 to 650 cm^{-1} for consistency with earlier work¹⁵. All spectra were corrected to give each spectrum a common minimum intensity value. Spectra were vector normalized to unit magnitude, and all spectra were mean centered.

PLS-DA Model A. A PLS-DA model was constructed from the model A calibration set to predict the class membership of the model A test set samples from one

of four classes: leaf, stem, seed coat, and hull. A PLS-2 model was created, so that the PLS regression included all four class membership response variables. Plots of the magnitudes of regression vectors versus root mean squared error of calibration (RMSEC) and plots of the A-values versus RMSEC were used to determine the number of significant latent variables (LVs). These methods of choosing the significant number of LVs and the definitions of RMSEC and A-values have been described by Green and Kalivas¹⁷. A model with 21 LVs was chosen. All models presented in this work were built with the use of MATLAB 7 and PLS_toolbox 3.5 (Eigenvector Research, Wenatchee, WA).

PLS-DA Model B. A PLS-DA model was constructed from the model B calibration set to predict the group membership of the initial test set samples in one of four classes: leaf, stem, seed coat, and hull. A PLS-1 model was created, so that a separate PLS-DA model was made for each of the four group membership variables. The same methods of choosing significant LVs as described for model A was used for model B. The number of LVs chosen for the leaf, stem, seed coat, and hull models were 15, 25, 8, and 22, respectively.

RESULTS AND DISCUSSION

PLS-DA is a method of classification that relies on PLS regression. Wise et al.¹⁸ have given a thorough description of the PLS-DA method. The calibration spectra make up the **X** block of data, and vectors of dummy variables for each class in the data make up the **Y** block. For this work, the **Y** block contained four column vectors: one each for leaf, stem, seed coat, and hull. For example if a calibration sample is a leaf, the

component of the leaf column vector for the corresponding spectrum will contain a 1. If the sample is not a leaf, the leaf column vector will contain a 0. These dummy variables are regressed onto the spectra using PLS. Predicted values are then obtained for the calibration set and a probabilistic threshold value between 0 and 1 is determined that will minimize the rate of misclassifications for the calibration set. Predictions can then be made for unknown samples, and class membership decided by determining if the predicted value of the dummy variable is above or below the threshold value.

Barker and Rayens¹⁹ have reported the advantages of PLS-DA compared to classification methods that rely upon principal component analysis (PCA). The data reduction in PLS-DA is determined by the largest directions of between class variance in the data space, but in PCA the reduction is determined by the largest directions of total variance. In other words, the PLS-DA algorithm uses the directions in the data space that show the greatest separation among classes. The directions showing greatest separation among classes will not always be the directions showing the largest total variance in the data space, and total variance is what guides the PCA method. Since the total variance is the sum of the between class and within class variance, PCA will perform well for classification tasks where the between class variance is much larger than the within class variance, but PCA will not yield good classification for situations where the within class variance is significant compared to the between class variance. Figure 3.1 shows the spectra of samples of the four different classes for which we were interested in classifying our samples into. Figure 3.2 shows the spectra of four different cotton leaves. These figures demonstrate that for these cotton plant parts the within

Figure 3.1: Spectra of leaf, stem, seed coat, and hull from the cotton plant. The spectra have been offset for clarity.

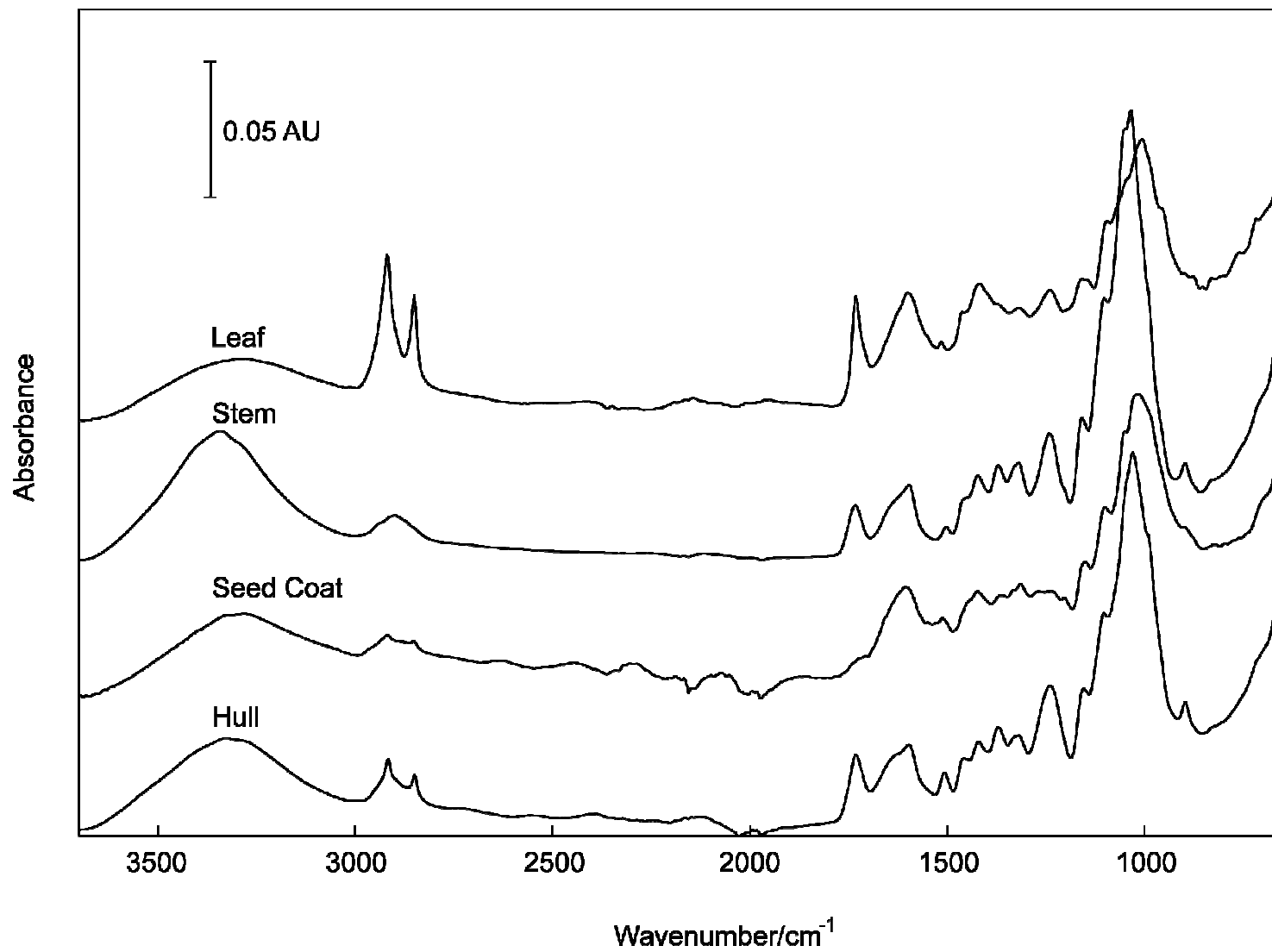
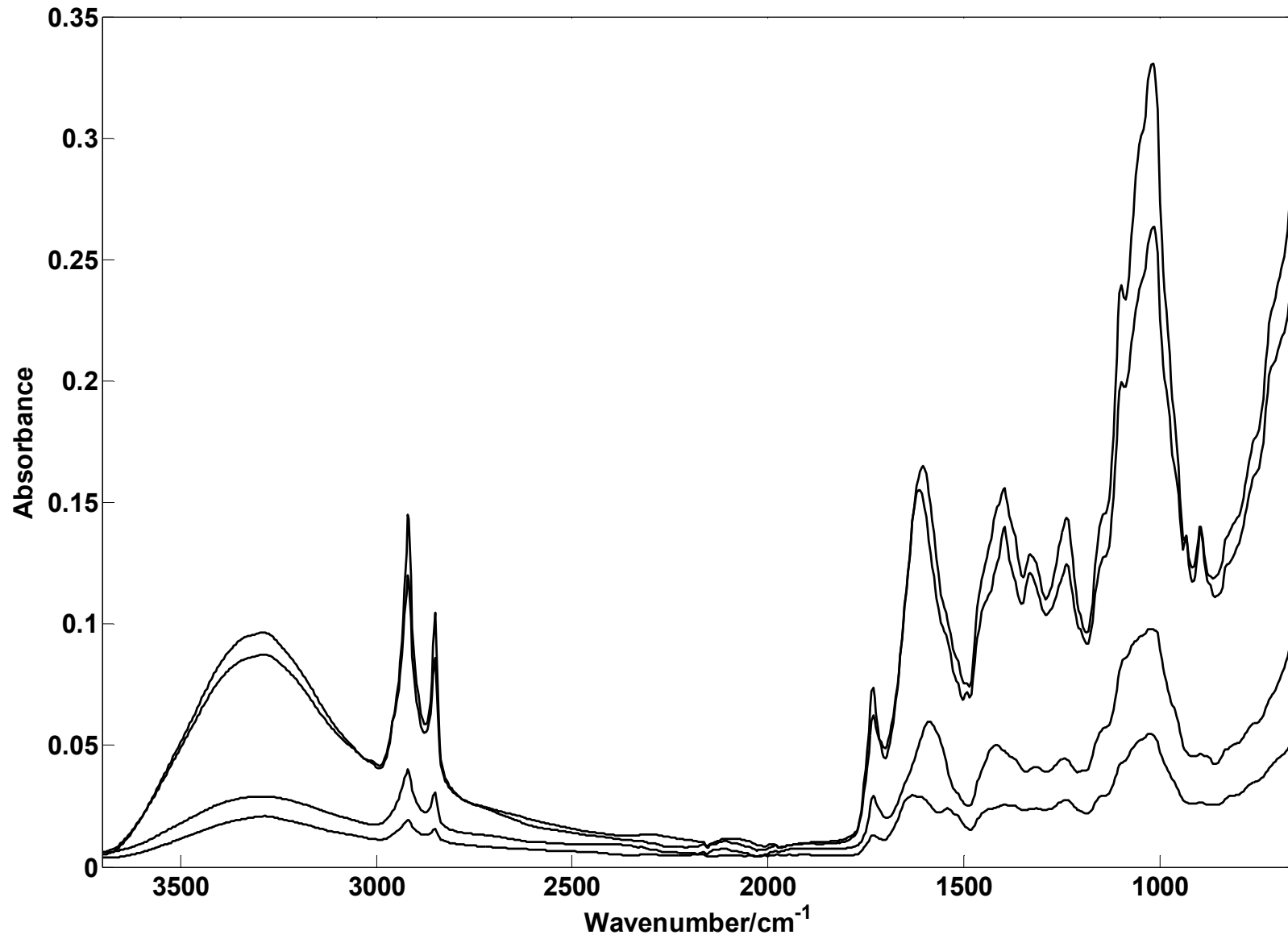


Figure 3.2: Spectra of four different cotton leaves.




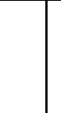
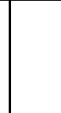
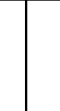
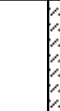
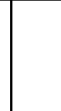





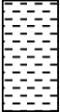

class spectral variance is large compared to the between class variance, so PLS-DA is a wise choice for the types of samples that were considered in this work.

In the first experiment that was performed, a set of calibration spectra (model A calibration set) taken from the USDA library was used to create a single PLS-DA model (model A) that included all of the response variables. The response variables were dummy variables corresponding to class membership in the leaf, stem, seed coat, and hull classes. When the plots of the magnitude of the regression vectors versus RMSEC and the A-Values versus RMSEC (see Methods and Materials) were examined to determine the optimum number of LVs to use, the plots for leaf, stem, seed coat, and hull indicated 21 LVs should be used. The performance of model A was tested by examining the accuracy of class membership predictions made for a set of 20 test spectra (model A test set) representative of samples from different growing locations and seasons than those spectra in the model A calibration set. Model A was created to test the ability of PLS-DA to predict sample class when the test set was not representative of the calibration set. Figure 3.3 shows the results of the analysis. One can see from the figure that model A unambiguously identified 16 of the 20 samples from the model A test set. In the previous work¹⁵, we reported that a group voting scheme algorithm identified 12 of 20 test samples. With PLS-DA instead of the voting scheme algorithm, a 20% increase in the identification rate was seen. The test set in both cases was the model A test set.

The second experiment was designed to test the performance of PLS-DA when the calibration set used to build the model (model B) was representative of the test set

Figure 3.3: The columns of this figure represent the model A test set samples. The rows correspond to the group the PLS-DA algorithm indicated the test sample belonged to. The patterns representing true positive, false positive, and false negative identifications are shown in the legend. White boxes represent true negative identifications.

| | | True Identity | | | | | | | | | | | | | | | | | | | |
|-------------|-----------|---|---|---|--|---|---|---|---|---|---|---|--|--|--|--|---|---|---|---|---|
| | | Leaf | | | | | Stem | | | | | Seed Coat | | | | | Hull | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Predictions | Leaf |  |  |  |  |  | | | | | | | | | | | | | | | |
| | Stem | | | | | |  |  |  |  |  | | | | | | | | | | |
| | Seed Coat | | | | | | | | | | | | | | | | | | | | |
| | Hull | | | | | | | | | | | | | | | | | | | | |

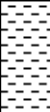


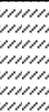









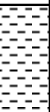
 = False positive
  = False negative
  = True positive

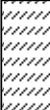


being used (model B test set). This situation differed from the first experiment because the model A test set was not representative of the model A calibration set. To form the model B calibration set, a set of 75 spectra (augmentation set) from which the model A test set was drawn were combined with the model A calibration set. These 75 spectra were representative of samples from different geographic regions and growing seasons from those represented by the model A calibration set. Of the 429 combined spectra, 330 spectra were chosen for the model B calibration set. Of the 99 remaining spectra, 12 spectra were chosen for the model B test set. This combined test set included 4 spectra from the model A test set that were correctly identified by model A, 4 spectra from the model A test set that were not identified by model A, and 4 spectra from the model A calibration. The four spectra from the model A test set that were identified correctly by model A were included to test for the reliability of the PLS-DA method. The four spectra from the model A test set that were not identified correctly by model A were included to test for prediction improvements with adding representative spectra to the model. The four spectra from the model A calibration set were added to determine the ability of model B to identify samples from the original USDA library. Of the 12 members of the model B test set, there were three spectra representing each of the following classes: leaf, stem, seed coat, and hull. The remaining 87 of the 429 spectra combined from the augmentation set and the model A test set were discarded to ensure that the model B test set was independent of the model B calibration set. These 87 spectra were the spectra of plant parts from the same plants that were represented in the model B test set.

When the plots of magnitudes of regression vectors and A-values versus RMSEC were studied for PLS-2 models, it was discovered that the optimum number of LVs suggested for each response category differed greatly. Because of this result, it was decided that model B would contain a PLS model for each response variable. The number of LVs chosen for these models was 15, 25, 8, and 22 for the leaf, stem, seed coat, and hull variables, respectively. The lower number of LVs chosen for the leaf and seed coat classes indicates that these classes are easier to distinguish from the other sample classes than the stem and hull classes are. These results generally agree with the results reported in the previous library search paper¹⁵.

Figure 3.4 shows the favorable results of the predictions made by model B. One should first note that the four spectra identified correctly by model A were also identified correctly by model B. This fact demonstrates that broadening the range of samples represented in the calibration set did not degrade the performance of PLS-DA. Three of the four spectra that were not identified correctly by model A were identified correctly by model B. The sample leaf 5 which was incorrectly identified by both model A and model B was on the threshold of identification by model B, so even though it was not clearly identified in model B, it was closer to being identified as leaf by model B than by model A. The fact that the spectra that were not identified correctly by model A were identified correctly by model B demonstrates the importance of using a calibration set as representative of unknown samples as possible. For the case of cotton contaminants, these results illustrate that it is best to continue updating the library with samples from different geographic locations and growing conditions as the samples to

Figure 3.4: This figure representing the results of the second experiment should be read the same way as Figure 3.1. Samples leaf 5, stem 1, seed coat 1, and hull 2 were samples from the model A test set that were not identified correctly by model A. Samples leaf 4, stem 3, seed coat 4, and hull 5 were samples from the model A test set that were identified correctly by model A. The remaining samples were taken from the model A calibration set.

| | | True Identity | | | | | | | | | | | |
|-------------|-----------|---|---|---|--|---|---|---|---|---|--|--|--|
| | | Leaf | | | Stem | | | Seed Coat | | | Hull | | |
| | | 5 | 4 | C | 1 | 3 | C | 1 | 4 | C | 2 | 5 | C |
| Predictions | Leaf |  |  |  | | |  | | | | | | |
| | Stem | | | |  |  |  | | | | | |  |
| | Seed Coat | | | | | | |  |  |  | | | |
| | Hull | | | | | | | | | |  |  |  |
| | | | | | | | | | | | | | |

 = False positive
  = False negative
  = True positive

be identified change. Only two of the four model B test set spectra obtained from the model A calibration set were identified correctly. This was probably due to the fact that once these two spectra and their closely related spectra were removed from the 429 pool of spectra, the remaining calibration spectra were not representative of these two spectra. If a larger pool of spectra had been available to choose the model B calibration and test sets from, these unidentified test set spectra may have been identified. Overall, the results from model B demonstrate that improvement in PLS-DA results when spectra representative of the same growing regions, conditions, and instruments as those used in the test set are present in the library.

CONCLUSION

The use of PLS-DA as a classification technique for cotton contaminants has been shown to be an effective method for contaminant identification. The results of the predictions from model A revealed that when the library was not representative of the test set samples, PLS-DA achieved an 80% unambiguous identification rate. This rate is 20% greater than the group voting scheme algorithm introduced in a previous paper¹⁵. The results from model B showed that further improvements in the performance of PLS-DA models for cotton contaminant identification can be obtained with representative calibration sets. In short, when a representative calibration set is unavailable, PLS-DA more accurately identifies cotton contaminants than library searching methods, but if the calibration set can be updated to include spectra more representative of the unknown spectra to be identified, the accuracy of PLS-DA

predictions will further increase. In all, PLS-DA was shown to be a good choice for classification of unknown cotton contaminants.

REFERENCES

1. R. V. Baker, "Influence of Lint Cleaning on Fiber Quality", in *Proc. Beltwide Cotton Production Research Conferences* (1987), p. 535.
2. R. V. Baker, J. B. Price, and K. Q. Robert, Jr., "Cleaning cotton for effective rotor spinning", in *Proc. Beltwide Cotton Conferences* (1994), p. 1605.
3. R. V. Baker, J. B. Price, and K. Q. Rovert, *Trans. ASAE* **37**, 4, 1077 (1994).
4. J. D. Barger and T. H. Garner, "The role of seed-coat and mote-fragment neps in yarn and fabric imperfections: a survey", in *Proc. Beltwide Cotton Production Research Conferences* (1988), p. 586.
5. C. K. Bragg and C. L. Simpson, "The effect of mechanical cleaning on processing efficiency and yarn quality in rotor spinning", in *Proc. Beltwide Cotton Production Research Conferences* (1988), p. 584.
6. C. K. Bragg, C. L. Simpson, A. D. Brashears, and R. V. Baker, *Trans. ASAE* **38**, 1, 57 (1995).
7. A. D. Brashears, R. V. Baker, C. K. Bragg, and C. L. Simpson, "Effect of bark on spinning efficiency of cotton", in *Proc. Beltwide Cotton Conferences* (1992), p. 1206.
8. D. E. Brushwood, *Appl. Eng. Agric.* **20**, 407 (2004).
9. G. J. Mangialardi, *Text. Res. J.* **39**, 1, 11 (1969).
10. G. J. Mangialardi, Jr., "Relationship of lint cleaning to seed coat fragments", in *Proc. Beltwide Cotton Production Research Conferences* (1987), p. 535.

11. G. R. Pilsbury, "Eliminating bark and seed cost fragments from cotton card silver", in *Proc. Beltwide Cotton Production Research Conferences* (1992), p. 1258.
12. K. Q. Robert and L. J. Blanchard, "Fiber breakage in cotton processing. I. A model", in *Proc. Beltwide Cotton Conferences* (1991), p. 894.
13. J. Simpson, *Text. Res. J.* **52**, 1, 52 (1982).
14. J. Foulk, D. McAlister, D. Himmelsbach, and E. Hughs, *J. Cotton Sci.* **8**, 243 (2004).
15. J. B. Loudermilk, D. S. Himmelsbach, F. E. Barton, II, and J. A. de Haseth, "Novel Search Algorithms for a Mid-IR Spectral Library", *Appl. Spectrosc.* **in press** (2008).
16. D. S. Himmelsbach, J. W. Hellgeth, and D. D. McAlister, *J. Agric. Food Chem.* **54**, 20, 7405 (2006).
17. R. L. Green and J. H. Kalivas, *Chemom. Intell. Lab. Syst.* **60**, 173 (2002).
18. B. M. Wise, J. M. Shaver, N. B. Gallagher, W. Windig, R. Bro, and R. S. Koch, "PLS-DA", in *PLS_Toolbox 3.5* (Eigenvector Research, Inc., Manson, WA, 2005), p. 185.
19. M. Barker and W. Rayens, *J. Chemom.* **17**, 166 (2003).

CHAPTER 4

QUANTITATIVE PREDICTION OF THE COMPOSITION OF COTTON DEBRIS POWDER AND AN ITERATIVE PREDICTION ERROR CORRECTION ALGORITHM¹

¹ J.B. Loudermilk, D.S. Himmelsbach, F.E. Barton II, J.A. de Haseth, To be submitted to J. Agric. Food Chem.

ABSTRACT

During the harvest and ginning processes, cotton fibers are contaminated with a variety of plant debris from the cotton plants themselves. Cotton leaves, stems, seeds, and hulls are the major constituents of this debris. Because debris leads to increased numbers of yarn breakages during spinning and imperfections in the finished yarn, debris has serious negative effects on the profit margin of the cotton industry. These problems point to the importance of cleaning debris from the cotton before the cotton is spun into yarn. Much of this cleaning takes place in different processes at the gin location. The operators manually adjust the cleaning machinery responsible for removing different types of debris to optimize debris removal for different batches of cotton containing different debris compositions. Real time feedback on the effects machinery adjustments have on the composition of the debris being removed at a particular point in the process would enable faster and better optimization of the cleaning processes. This work focused on creating partial least squares (PLS) regression models for quantitative analysis of simulated debris mixtures and on creating an iterative error correction algorithm to improve the prediction accuracy of the models' predictions.

Keywords: Quantitative analysis; FT-IR; ATR; Chemometrics; Cotton; Contaminants, Prediction error.

INTRODUCTION

Historically, contamination of cotton by plant debris has been an important and difficult problem for the cotton industry, and this problem is one that continues today(1-17). Contamination takes place when materials in the growing fields other than the desired cotton fibers are harvested along with the fibers. The greatest source of these contaminants is the cotton plants themselves. The major parts of the plants that are harvested along with the fibers are the leaves, stems, hulls, and seeds. Although steps are taken to reduce the amounts of these unwanted plant parts during harvest, extraction of debris from the cotton is still an important part of processing that occurs at the cotton gin.

Numerous papers have described the problems caused by debris remaining in cotton during spinning of the cotton into yarn(2-6, 13). One important impact of debris is increases in the number of yarn breakages during spinning(6). As the number of breakages increases, the time and, thus, the cost required to produce yarn increase. Debris also causes imperfections in yarn(4). As the number of imperfections in yarn increases, the value of the yarn decreases. Because of these problems, cotton contamination decreases the profit margin of the cotton industry. The detriments of cotton contamination reveal the need for cotton to undergo cleaning during processing at the gin location.

During processing, different types of debris are removed at different times(2-4). Currently, the gin operators must manually make adjustments to much of the cleaning machinery to ensure efficient operation and cleaning. Adjustments are necessary

because as the composition of the plant debris in the cotton changes between different batches the machinery must be optimized to operate efficiently under different conditions. As with any manufacturing process, faster processing times mean more profit if a faster production rate can be achieved without sacrificing the quality of the product being produced, but the time required to adjust the machinery for optimum operation is limited by the difficulty of identifying the types and amounts of debris present in the cotton. When debris is intact and sufficiently large, it may be identified visually, but from the time the debris leaves the field with the cotton, the debris begins to break down in size. This breakdown continues as the debris travels to the gin and the cotton begins to undergo processing(4, 8). Much of the debris eventually breaks down into powder sized particles making visual identification impossible.

Recent work has shown that mid-IR spectrometry can be used to successfully distinguish among different types of plant debris found in cotton(8, 9, 18). Figure 4.1 shows the spectra of cotton leaf, stem, seed coat, and hull. One can see that the spectra of these cellulose based plant parts are very similar. Figure 4.2 shows an example of the variation that can occur within a single class of these natural products. These figures demonstrate that the within group variation is significant compared to the between group variation. One can see that to visually discriminate among spectra from these different classes is not a trivial task; however, these previous reports have shown that mid-IR spectrometry combined with the appropriate chemometric methods has the ability to distinguish among these different types of debris. One of the important advantages of using spectroscopy for this application is its speed and ease of sampling.

Figure 4.1: Spectra of leaf, stem, seed coat, and hull from the cotton plant. The spectra have been offset for clarity.

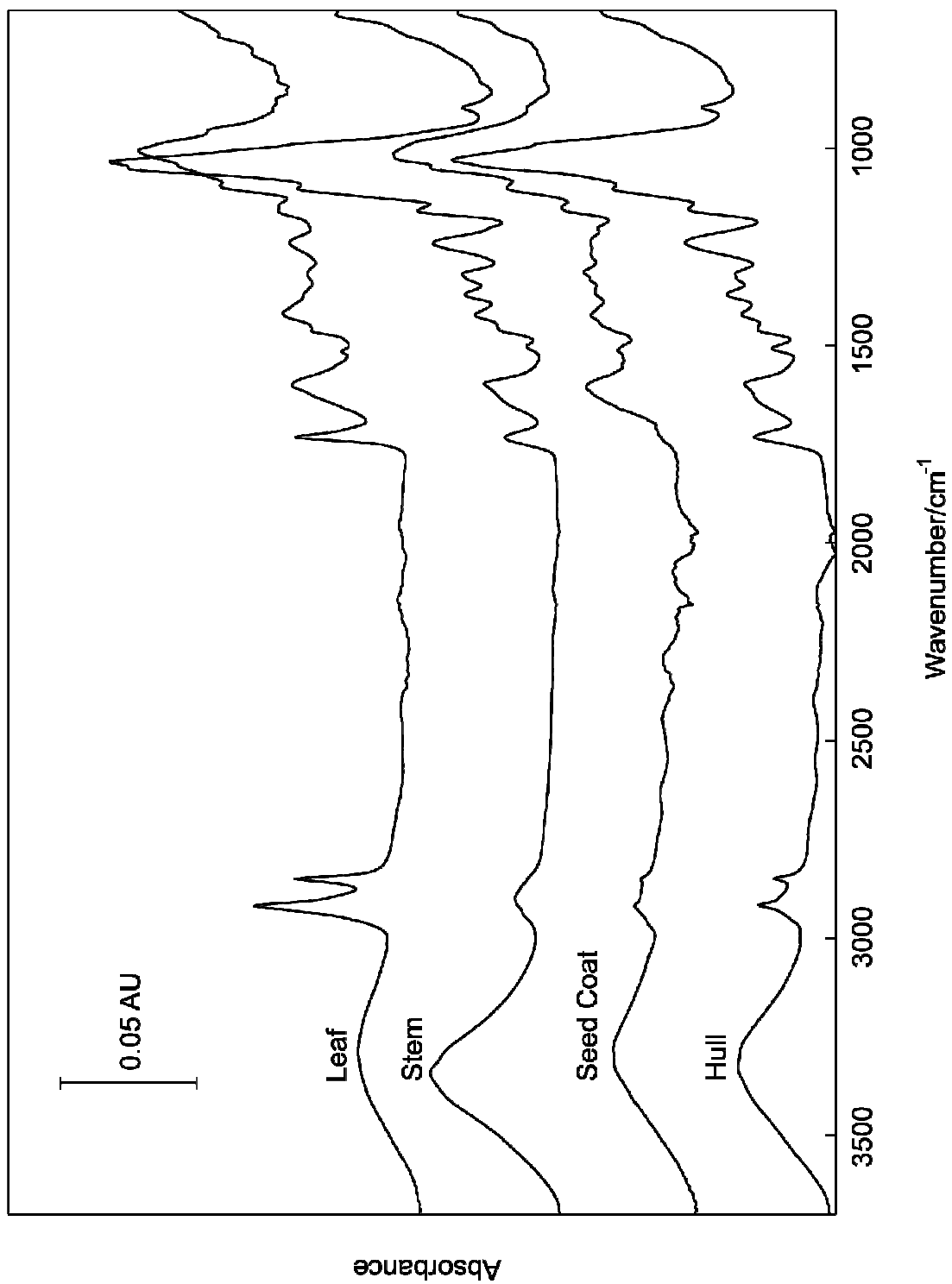
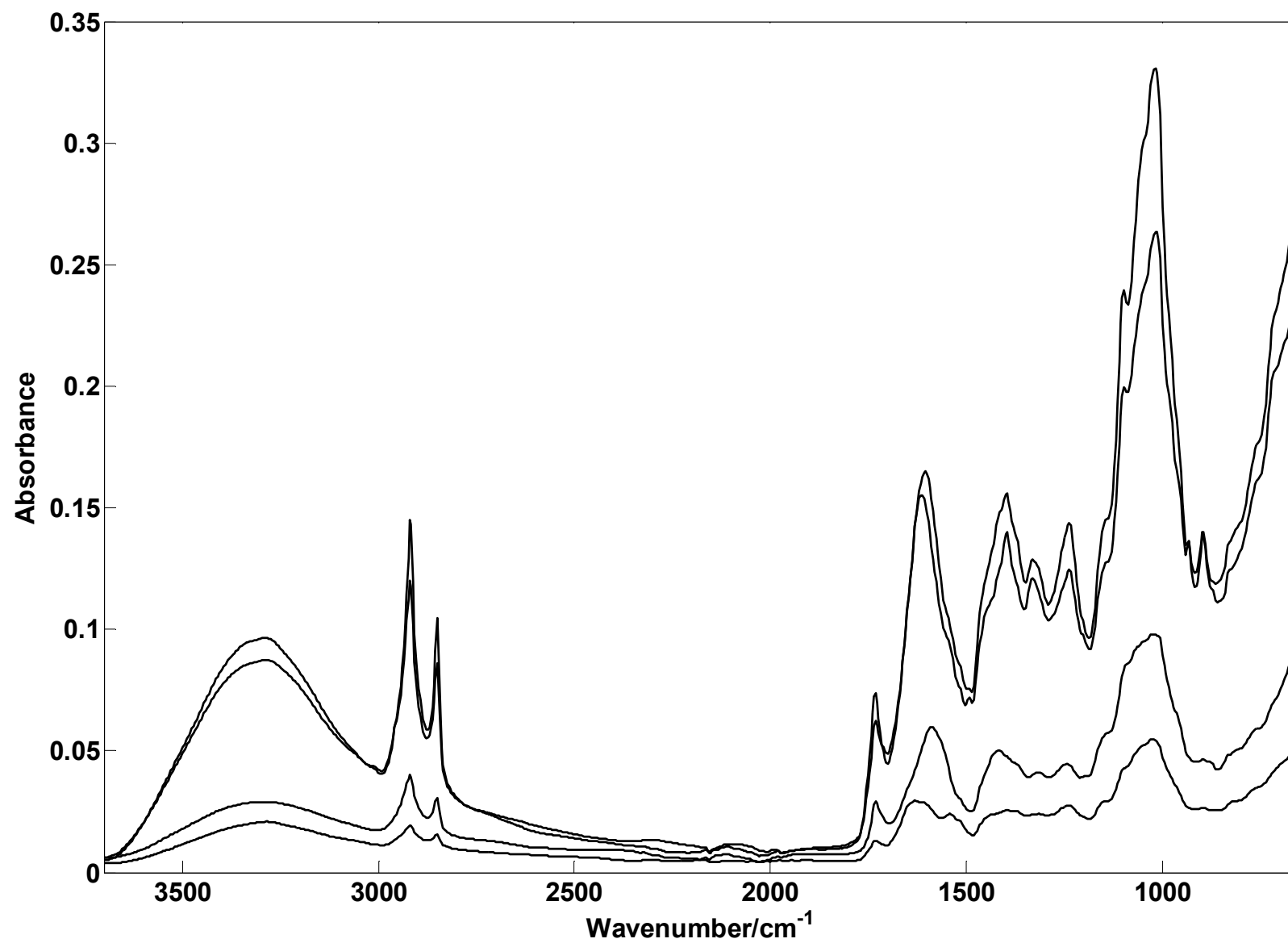


Figure 4.2: Spectra of four different cotton leaves.



These advantages mean that if quantitative regression models could be developed, process efficiency monitoring could be incorporated on-line or at-line during the cleaning processes, providing real or near real-time feedback to operators.

The successes of previous studies to distinguish qualitatively among the spectra of the different types of plant debris and the potential gain to the industry if quantitative analysis of debris could be achieved, led us to investigate the use of mid-IR spectra to develop chemometric regression models for quantitative analysis. Our work focused on the creation of PLS regression models to predict the percent of leaf, stem, seed coat, and hull present in simulated debris powders. Additionally, an iterative error redistribution algorithm that has potential applications to other quantitative analysis problems was developed to improve the accuracy of the predictions obtained from the models.

MATERIALS AND METHODS

Debris Mixture Samples. Samples of powdered debris from three different growing locations were obtained from the USDA-ARS Cotton Quality Research Laboratory (Clemson, SC). The locations will be referred to as A, B, and C. For each location, we received several grams of powder of cotton leaves, stems, seed coats, and hulls that had been harvested from cotton plants and then ground to a size of 80 mesh. Thirty mixtures of these four types of debris powders were prepared for each growing location by accurate weighing of aliquots of each type of powder. The total mass for each mixture was 0.5000 g. The samples' masses were measured with a Fisher A-200DS

analytical balance (Fisher Scientific, Hampton, NH). The powders were measured onto clean watch glasses, and the samples were mixed thoroughly with the use of a spatula.

Calibration and Test Spectra Sets. Spectra of the 30 samples from each growing location were measured with a Varian Excalibur Series FTS-4000 FT-IR spectrometer (Varian, Palo Alto, CA) and a Specac Golden Gate attenuated total reflection (ATR) sampling accessory (Specac, Woodstock, GA) with a diamond coated ZeSe Internal Reflection Element (IRE). The spectrometer contained a ceramic source, a KBr beamsplitter, and an L-Alanine-doped DGTS detector. Spectra were measured over the range of 3700 to 650 cm^{-1} at 4 cm^{-1} resolution, with 256 interferograms co-added. Interferograms were processed with Norton-Beer Medium apodization before Fourier transformation. Each spectrum used in the experiments was the average of three sample replicate spectra. Spectra were collected with the use of Varian Resolutions Pro 4.0.5.009 software (Varian, Palo Alto, CA). All spectra were first imported into GRAMS (Thermo Fisher Scientific, Waltham, MA) and then imported into MATLAB (The MathWorks, Natick, MA). Twenty spectra that represented each growing location were used for calibration sets, and the remaining 10 spectra from each growing location were used for test sets.

Regression Models. The calibration sets were used to create PLS models to predict the percent cotton leaf, stem, seed coat, and hull in each test sample. Both PLS-1 and PLS-2 models were investigated. In PLS-1, a separate model is built for each variable to be predicted. In PLS-2, a single model is developed to predict all of the variables of interest. All of the various permutations of mean centering and autoscaling the spectra

and composition data of the calibration sets were investigated. In this paper, autoscaling is defined as transforming the values of a variable to have mean equal to 0 and variance equal to 1. The PLS_Toolbox 3.5 software (Eigenvector Research, Wenatchee, WA) and MATLAB 7 software (The MathWorks, Natick, MA) were used to build and test the calibration models. The appropriate number of latent variables (LVs) for the models was chosen based on the root mean squared error of cross validation (RMSECV) obtained from random subset cross validation (4 subsets, 20 iterations). Equation 1 shows the formula for RMSECV.

$$RMSECV = \left(\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n} \right)^{\frac{1}{2}} \quad (1)$$

In this equation, \hat{y}_i represents the predicted value for the i^{th} sample left out of the calibration set for the cross validation, y_i represents the true value of the i^{th} sample, and n is the number of samples in the calibration set. The number of LVs used in the models ranged from 2 to 11.

Iterative Error Correction Algorithm. An algorithm was developed to redistribute the error in the prediction values due to negative predictions and sum of percent compositions over 100% for individual samples. The algorithm was written in MATLAB. First, negative predictions for each sample were set to 0. A value of 100 was then subtracted from the new sum of the predictions for the percent leaf, stem, seed coat, and hull predicted for each sample. Let this difference be denoted by d . If d was

negative, the absolute value of d was divided proportionally into four parts, and the parts were added to the four predictions for each sample. If d was positive, d was divided proportionally into four parts, and the parts were subtracted from the four predictions for each sample. This procedure was repeated until the sum of the predictions for a given sample was within ± 0.005 of 100% percent. The number of iterations required for convergence ranged from 6 to 27. Only samples with negative predictions were corrected.

Three methods for proportionally dividing the error, d , were examined. The first method used the ratios of the individual squared RMSECV values for each response variable to the sum of the squared RMSECV values for all four response variables to estimate the prediction error. The second method simply assumed that the proportion of prediction error for each variable was equal, i.e. 25% per variable. The third method determined the mean positive and negative differences between the predicted values and the true values for each response variable for all of the samples in a calibration set. One set of proportions was obtained by dividing the mean positive difference by the sum of the mean positive differences for all four variables. The other set of proportions was obtained by dividing the mean negative difference by the sum of the mean negative differences for all four variables. If d was positive, the proportions from the positive differences were used to divide the error, and if d was negative, the proportions from the negative differences were used.

RESULTS AND DISCUSSION

Since preprocessing data can have significant effects on the accuracy of predictions obtained from a regression model, we compared different methods of preprocessing the **X** and **Y** block data. The **X** block refers to the matrix of spectra, and the **Y** block refers to the matrix of percent leaf, stem, seed coat, and hull response variables for PLS-2 models and vectors for each of these variables individually for the PLS-1 models. Both PLS-1 and PLS-2 models were explored because they yield different results, and it was not known *a priori* which type of regression would yield the best results for the system under consideration. In general when PLS-2 is used, predictions for all variables are based on the same number of LVs because the number of LVs that yield the best predictions for one variable is equal or similar to the number of LVs that yield the best predictions for the other variables when judged by RMSECV. In this work, we found from RMSECV versus LV number plots that this was not the case with the calibration sets. Because of this fact, we experimented with making predictions based on the optimum number of LVs for each response variable instead of just compromising for the PLS-2 model that used the same number of LVs to predict all four response variables.

Initially, 11 separate models were created for each growing location. Four PLS-1 models and seven PLS-2 models were necessary to test the permutations of mean centering and autoscaling the **X** and **Y** block data. Table 4.1 records the data preprocessing techniques applied for the eleven models considered. More permutations are possible for the PLS-2 models than the PLS-1 models since the PLS-1

Table 4.1: Describes the preprocessing conditions and regression type used for each of the eleven conditions sets used in the experiments.

| Model Conditions # | Regression Type | Preprocessing |
|--------------------|-----------------|-----------------------------------|
| 1 | PLS-2 | None |
| 2 | PLS-1 | None |
| 3 | PLS-2 | X-mncn^a, Y-mncn |
| 4 | PLS-1 | X-mncn, Y-mncn |
| 5 | PLS-2 | X-auto^b, Y-auto |
| 6 | PLS-1 | X-auto, Y-auto |
| 7 | PLS-2 | X-auto, Y-mncn |
| 8 | PLS-2 | X-mncn, Y-auto |
| 9 | PLS-2 | X-none, Y-mncn |
| 10 | PLS-1 | X-none, Y-mncn |
| 11 | PLS-2 | X-none, Y-auto |

^amean centered

^bautoscaled

models only look at the variables in the **Y** block one column at a time. If all seven preprocessing permutations applied to the PLS-2 models were applied to the PLS-1 models, three of those seven PLS-1 models would provide identical results to one of the remaining four models; thus, only the four unique PLS-1 models were included. Three of the seven models would provide redundant information because when the **Y** block is only comprised of one variable autoscaling the **Y** block has no effect on the model.

To determine the best method for preprocessing these data, the regression models for each of the three calibration sets were ordered from lowest total RMSECV to highest. The total RMSECV is the square root of the sum of the squared RMSECV values for the percent leaf, stem, seed coat, and hull variables in each model. Table 4.2 shows the order obtained. From these results, it can be seen that model 9 most consistently yielded the best results when minimization of RMSECV was the criterion. The regression type for model 9 was PLS-2 with no **X** block pretreatment and the **Y** block mean centered. These results reveal that in this instance a value of zero for the response variables is significant to the system, which is a logical result. Table 4.3 confirms the results from the calibration sets for the test sets. Table 4.3 orders the root mean squared error prediction (RMSEP) values for all of the model conditions for each test set from lowest RMSEP to highest. The equation for RMSEP can be considered to be the same as Eq. 1 except that predictions are for the test set instead of the excluded samples from the cross validation, and n is the number of samples in the test set. The test set data confirm that model 9 yielded the best prediction results most consistently.

Table 4.2: The model conditions numbers ranked from lowest to highest RMSECV values for the three calibrations sets A, B, and C.

| Set | Rank | | | | | | | | | | |
|-----|------|----|----|----|---|---|---|---|---|----|----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| A | 10 | 9 | 11 | 4 | 2 | 8 | 3 | 1 | 6 | 7 | 5 |
| B | 9 | 10 | 11 | 2 | 4 | 5 | 6 | 7 | 8 | 1 | 3 |
| C | 1 | 9 | 11 | 10 | 2 | 4 | 3 | 8 | 6 | 5 | 7 |

Table 4.3: The model conditions numbers ranked from lowest to highest RMSEP values for the three calibrations sets A, B, and C.

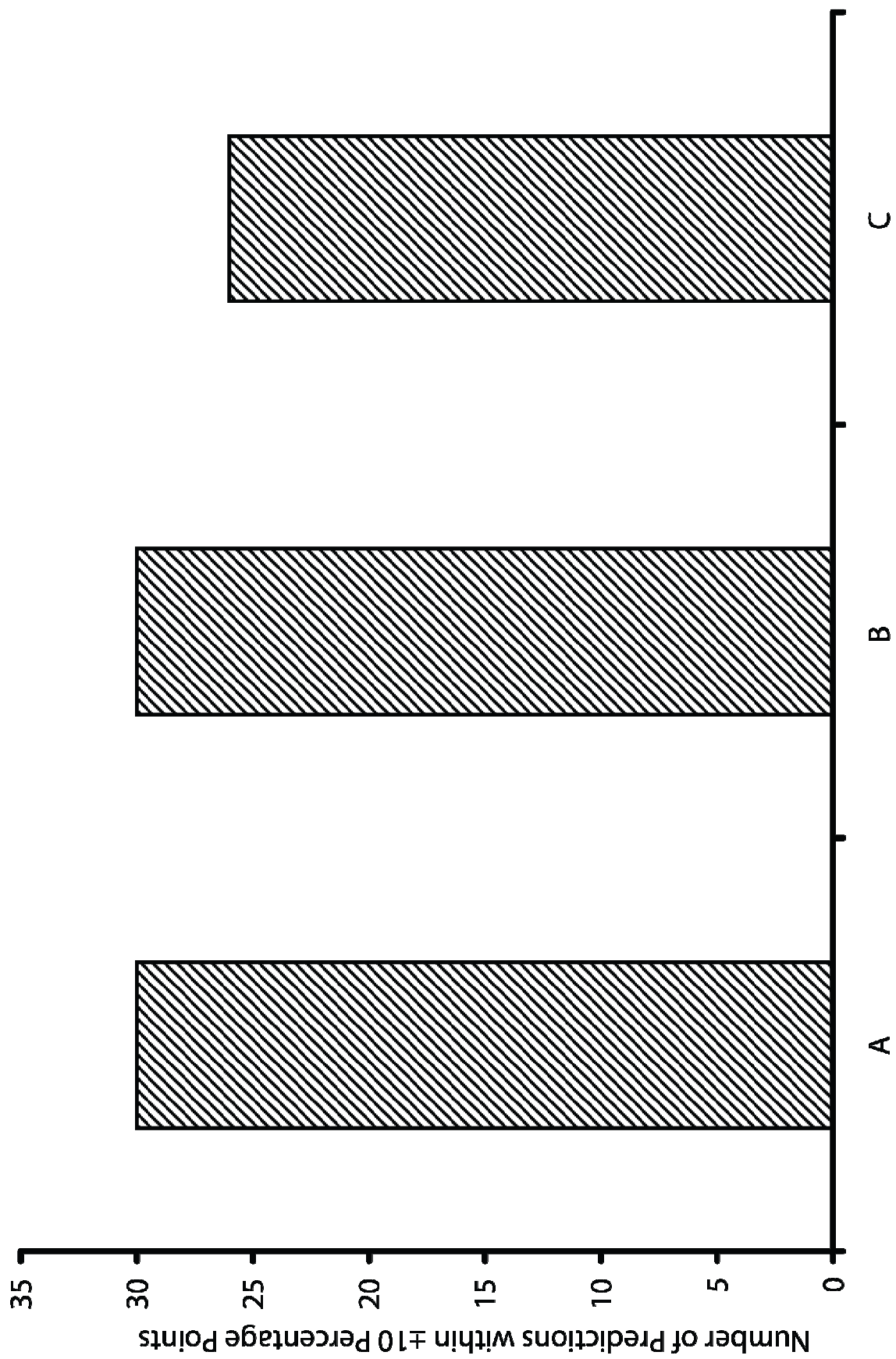
| Set | Rank | | | | | | | | | | |
|-----|------|---|---|----|---|----|----|---|---|----|----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| A | 8 | 3 | 9 | 6 | 4 | 11 | 10 | 7 | 5 | 2 | 1 |
| B | 10 | 9 | 2 | 11 | 6 | 3 | 4 | 8 | 7 | 5 | 1 |
| C | 11 | 9 | 3 | 8 | 1 | 10 | 2 | 4 | 6 | 7 | 5 |

Since the test set was not used to create the model, one expects the accuracy of the predictions to be a closer estimate to the accuracy of the model in practice than the results obtained from cross validation.

When one examines the predictions generated by using model 9, one finds that the models provide for quantitative prediction of the composition of the debris mixtures. Figure 4.3 shows the number of predictions from each model that lie within ± 10 percentage points of their true values for each test set. There were four predictions made for each of 10 samples, so there was a total of 40 predictions made for each test set. It is important to understand that although an error of ± 10 percentage points would not be suitable for many quantitative applications, it is adequate for machinery operators to determine the effects of adjustments on the composition of debris. One should note that ± 10 percentage points is a sharp cutoff, and if one were to consider the predictions just beyond this limit, improvements in prediction accuracy as judged by the criterion shown in Fig. 4.3 would be observed. Consider test set A as an example. If the criterion was ± 11 percentage points instead of ± 10 , the number of predictions falling within the desired range would increase by 3. Also, in considering the extreme spectral similarity of cotton leaves, stems, seed coats, and hulls and the chemical complexity of these natural products, the performance of these regression models is impressive; however, improvements in prediction accuracy are always beneficial if costs are not significantly increased. To this end, an iterative error correction algorithm was designed to improve prediction accuracy.

Figure 4.3: Number of predictions within ± 10 percentage points for test sets A, B, and C.

A total of 40 predictions were made for each test set.



When the predictions for the test sets were examined, several predictions of negative percent compositions were observed for mixture components in samples where those components made up only a small percent of the mixture. These observations were due to the inevitable prediction error present in the models, but physically, a component of a mixture cannot have a negative percent composition. No component can make up less than 0% of the mixture. It is also physically impossible for the sum of the percent compositions for all components to be greater than 100%. Although it is possible for the sum of percent compositions to be less than 100%, if all of the mixture components have been specified the sum should be 100%. Because of these facts, an iterative error correction algorithm was designed to take advantage of these boundary conditions. The algorithm was described earlier in the methods and materials section of this paper.

In developing the algorithm, we faced the question of how to redistribute the known error present in the predictions to improve the total prediction error. Three different methods to redistribute the error proportionally were compared. Since RMSECV is a measure of a model's accuracy, one method was based on RMSECV. Since measures of variance add linearly, the ratio of the squared values of RMSECV for a given response variable to the sum of the squared values for all response variables for a model's calibration set were used. These ratios were then used to redistribute the error in the predictions for the test set as described previously. A second redistribution method served as a null case. We tested the effect of simply splitting the error evenly among all four variables being predicted in the test set without regard to difference in magnitudes

of prediction error among different variables. Since the data revealed that the RMSECV values for percent leaf, stem, seed coat, and hull were different in all of the models, it was expected that this method would perform worse than a method that took these differences into account. The third method was a novel attempt at taking the sign of the prediction errors for individual variable's predictions into account. As described previously, this method was based on the average positive and negative prediction errors present in the predictions for the calibration set. When the sum of the predictions for a test was less than 100%, the mean negative errors were used to redistribute the prediction error, and when the sum of the predictions for a test set was greater than 100%, the mean positive errors were used to redistribute the prediction error. This method was tested because of the possibility that mean negative prediction error could be larger than the mean positive prediction error or vice versa.

The results of these experiments are shown in Fig. 4.4. The ordinate units are percent decrease in RMSEP after application of the redistribution algorithm with the use of the three different redistribution methods. One notices some variation in the results. For test sets A and C, error redistribution method 2 performed worst, but this method performed best for test set B. This discrepancy can be explained by comparison of the RMSECV values for the calibration sets. Table 4.4 shows the RMSECV for each variable in the three models. One will notice that the RMSECV values for individual variables are most similar for calibration set B. This fact means that the prediction error in this model is more evenly distributed than in the other models, explaining the better results for method 2 in set B. In fact, if one considers the variance of the four RMSECV

Figure 4.4: Comparison of the decrease in RMSEP seen for each of three test sets A, B, and C as a result of the three error redistribution methods used for the iterative error redistribution algorithm.

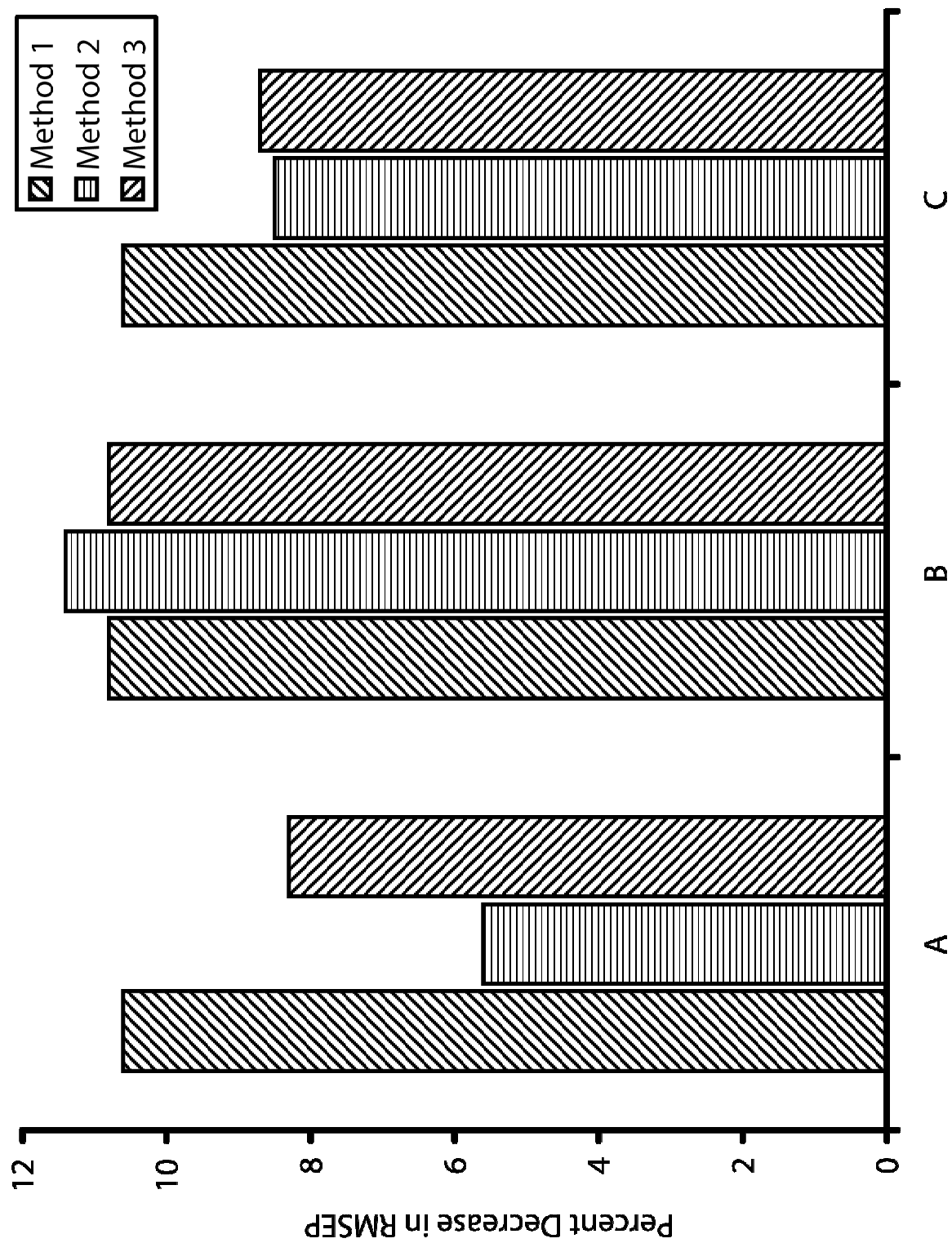


Table 4.4: RMSECV values and their variances for the three test sets.

| Set | Response Variables | | | | Variance |
|-----|--------------------|-------|-----------|-------|----------|
| | Leaf | Stem | Seed Coat | Hull | |
| A | 15.00 | 18.68 | 8.89 | 4.54 | 29.7 |
| B | 10.81 | 7.25 | 5.23 | 10.90 | 5.83 |
| C | 8.80 | 8.90 | 18.01 | 14.66 | 15.4 |

values for each model, redistribution method 2 performs better as the variance decreases. This result is expected since method 2 should perform best when the prediction errors for the different variables are equal. The results for method 3 are similar to method 2 and vary for the different test sets used. At least for this system, the concept behind method 3 does not appear to give reliable correction results.

Error redistribution method 1, which was based on the RMSECV values, gave the best results in terms of providing a dramatic and consistent decrease in the total RMSEP values for the tests sets of data. Figure 4.4 shows that for all three test sets, method 1 reduced the RMSEP values by over 10%, and these results were consistent for each test set considered. By using this method, the errors due to predictions below 0% and sum of predictions over 100% were successfully redistributed. The better performance of method 2 over method 1 for the second set of data was due to the fact that the types of prediction error being corrected for by this model are not the only model errors contributing to prediction error. By random chance, method 2 was able to yield better results than method 1 but method 1 was still the most reliable method tested.

It should be emphasized that the iterative correction algorithm developed reduced the total prediction error, but did not always reduce the RMSEP values for every variable. With the data under consideration, the algorithm sometimes slightly increased the prediction error for some variables, but the trade off in error reduction for other variables was valuable as evidenced by the reductions in total RMSEP values shown in Fig. 4.4. Table 4.5 shows the RMSEP values for the second test set before and after the correction algorithm based on method 1 of error redistribution. In this

Table 4.5: Example of RMSEP values for test set B before and after application of the iterative error redistribution algorithm.

| Response Variables | RMSEP | |
|--------------------|-------------------|------------------|
| | Before Correction | After Correction |
| Leaf | 8.40 | 8.50 |
| Stem | 13.58 | 11.06 |
| Seed Coat | 4.63 | 4.29 |
| Hull | 7.96 | 7.56 |
| Total | 18.4 | 16.4 |

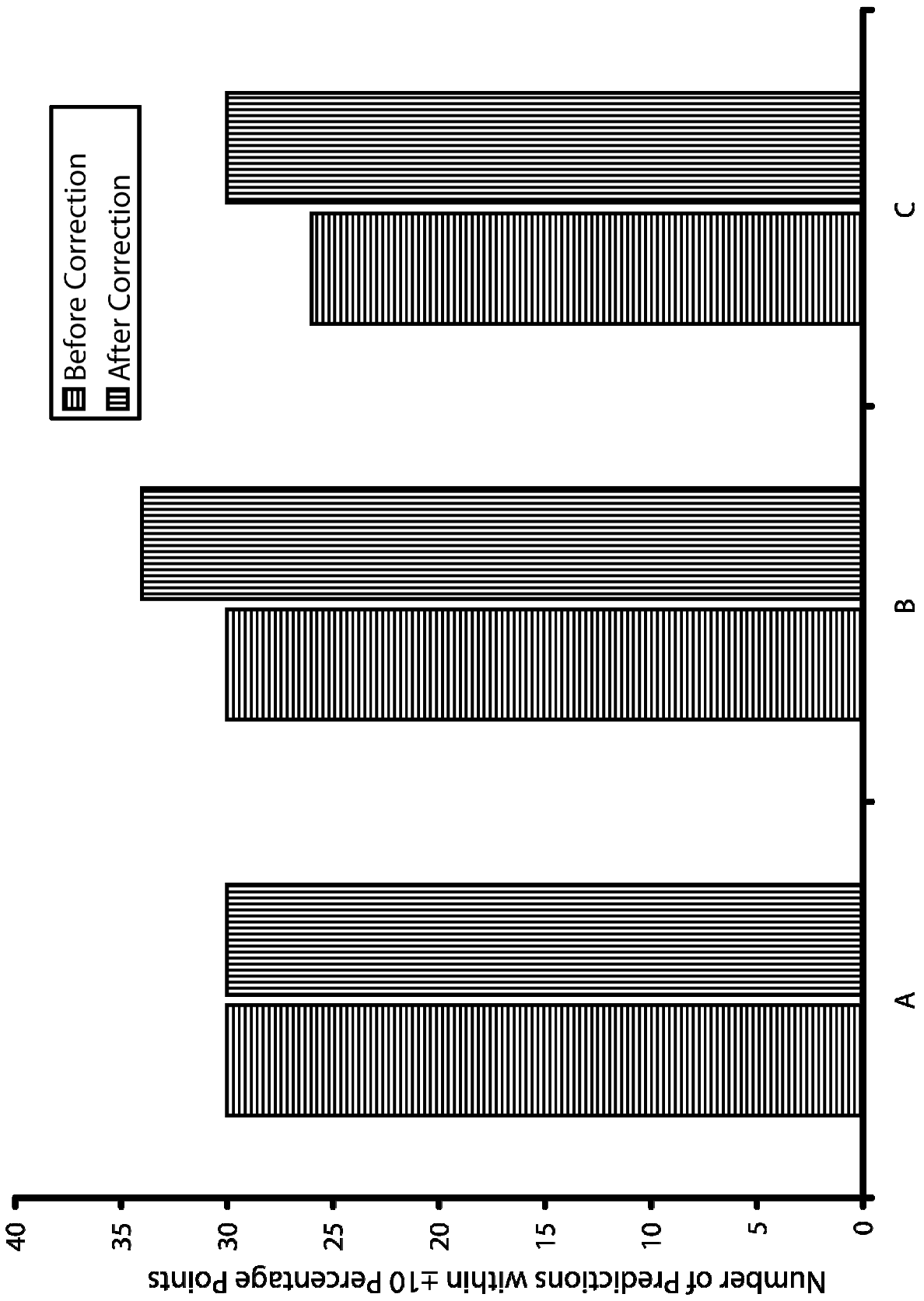
example, the RMSEP value for percent leaf slightly increased, but the RMSEP values for the other variables all decreased. Despite the leaf RMSEP increase, the fact remains that the total RMSEP was reduced by 11%.

In addition to examining the improvements in RMSEP because of the iterative error redistribution algorithm, the improvements in the number of predictions within ± 10 percentage points of their true values should also be considered. Figure 4.5 shows the effects of the error redistribution algorithm with the use of redistribution method 1. In each case but the first, a dramatic improvement in the number of predictions within ± 10 percentage points is observed. It should be reemphasized that the limit of ± 10 percentage points is a sharp cutoff, and if one were to consider the number of predictions just beyond the cutoff, the accuracy of the predictions as judged by the criterion in Fig. 4.5 would be further improved. Consider test A again as an example. If the criterion was ± 11 percentage points instead of ± 10 , the number of predictions falling within the desired range would increase by 5 predictions after the corrections were made

CONCLUSION

Through the use of PLS regression models, we were able to predict successfully the composition of mixtures of cotton leaf, stem, seed coat, and hull powders. We found that for this system, PLS-2 models performed best. The accuracy of the models' predictions were dramatically improved by the iterative error correction algorithm that was developed as evidenced by a greater than 10% improvement in RMSEP values for each test set considered. The results of these experiments demonstrate that mid-IR

Figure 4.5: Comparison of the number of predictions within ± 10 percentage points for test sets A, B, and C before and after application of the iterative error redistribution algorithm with method 1 used to redistribute the error.



spectrometry combined with chemometric methods is a capable tool for the analysis of these very complex agricultural product mixtures, and it is expected that these methods would work well in other similarly complex agricultural, industrial, or biological applications dealing with complex chemical feed stocks or sample matrices where these methods have yet to be tried.

ACKNOWLEDGEMENT

The authors sincerely thank John Foulk and Angela Allen from the USDA-ARS Cotton Quality Research Laboratory in Clemson, SC for providing the cotton plant part powders used to form the debris mixtures studied.

REFERENCES

1. Armijo, C. B.; Hughs, S. E.; Gillum, M. N.; Barnes, E. M., Ginning a Cotton with a Fragile Seed Coat [electronic resource]. *J. Cotton Sci.* **2006**, 10.
2. Baker, R. V. In *Influence of Lint Cleaning on Fiber Quality*, Proc. Beltwide Cotton Production Research Conferences 1987; 1987; pp 535-536.
3. Baker, R. V.; Price, J. B.; Rovert, K. Q., Gin and mill cleaning for rotor spinning. *Trans. ASAE* **1994**, 37, (4), 1077-1082.
4. Barger, J. D.; Garner, T. H. In *The role of seed-coat and mote-fragment neps in yarn and fabric imperfections: a survey*, Proc. Beltwide Cotton Production Research Conferences, 1988; 1988; pp 586-591.
5. Bragg, C. K.; Simpson, C. L. In *The effect of mechanical cleaning on processing efficiency and yarn quality in rotor spinning*, Proc. Beltwide Cotton Production Research Conferences, 1988; 1988; pp 584-586.

6. Bragg, C. K.; Simpson, C. L.; Brashears, A. D.; Baker, R. V., Bark effect on spinning efficiency of cotton. *Trans. ASAE* **1995**, 38, (1), 57-64.
7. Brushwood, D. E., Effects of raw cotton noncellulosic content and fiber RotorRing friction on yarn ring spinning performance. *Appl. Eng. Agric.* **2004**, 20, 407-411.
8. Foulk, J.; McAlister, D.; Himmelsbach, D.; Hughs, E., Mid-infrared Spectroscopy of Trash in Cotton Rotor Dust. *J. Cotton Sci.* **2004**, 8, 243-253.
9. Himmelsbach, D. S.; Hellgeth, J. W.; McAlister, D. D., Development and Use of an Attenuated Total Reflectance/Fourier Transform Infrared (ATR/FT-IR) Spectral Database to Identify Foreign Matter in Cotton. *J. Agric. Food Chem.* **2006**, 54, (20), 7405-7412.
10. Mangialardi, G. J., Effects of Lint Cleaning at Cotton Gins on Seed-Coat Fragment and Funiculus Distribution. *Text. Res. J.* **1969**, 39, (1), 11-14.
11. Mangialardi, G. J., Jr. In *Relationship of lint cleaning to seed coat fragments*, Proc. Beltwide Cotton Production Research Conferences, 1987; 1987; pp 535-536.
12. Perkins, H. H., Determination of seed-coat fragments in cotton by solvent-extraction and infrared spectrophotometric analysis. *Text. Res. J.* **1971**, 559-563.
13. Pilsbury, G. R. In *Eliminating bark and seed cost fragments from cotton card silver*, Proc. Beltwide Cotton Production Research Conferences, 1992; 1992; pp 1258-1263.
14. Robert, K. Q.; Blanchard, L. J. In *Fiber breakage in cotton processing. I. A model*, Proc. Beltwide Cotton Conferences, 1991; 1991; pp 894-897.

15. Simpson, J., The effect of rotor fiber groove design on trash accumulation, end breakage, and yarn properties. *Text. Res. J.* **1982**, 52, (1), 52-59.
16. Xu, B.; Fang, C., Clustering Analysis for Cotton Trash Classification. *Text. Res. J.* **1999**, 69, (9), 656-652.
17. Xu, B.; Fang, C.; Huang, R., Chromatic Image Analysis for Cotton Trash and Color Measurements. *Text. Res. J.* **1997**, 67, (12), 881-890.
18. Loudermilk, J. B.; Himmelsbach, D. S.; Barton, F. E., II; de Haseth, J. A., Novel Search Algorithms for a Mid-IR Spectral Library of Cotton Contaminants. *Appl. Spectrosc.* **2008**, in press.

CHAPTER 5

A MIXTURE GENERATOR ALGORITHM FOR GENERATION OF CALIBRATION MIXTURE STANDARDS FOR CLOSED MIXTURE SPACES¹

¹ J.B. Loudermilk, D.S. Himmelsbach, F.E. Barton II, J.A. de Haseth, To be submitted to J. Chemom.

ABSTRACT

When building a spectroscopic calibration model to predict the composition of mixtures, it is desirable to build the model with actual samples from the process being studied. These samples are analyzed via a primary method and then a calibration based on some secondary method, such as spectrometry, can be built. In some cases where mixtures are being analyzed, though, it is impossible or difficult to use actual samples from the process of interest, so simulated mixture samples must be used. In other cases, many samples from the process are available, but because of the large number of samples to select from, a representative subset of the samples may be difficult to choose. In these instances, it is important that the range of mixture samples cover the entire mixture space and that the number of mixture samples be large enough to provide adequate coverage of the mixture space. For a two component mixture, this task is straightforward, but for three or more components the task of choosing which mixtures to make becomes increasingly difficult. The practical difficulties lie in dealing with large numbers of mixtures and in the inability to visualize hyper-dimensional spaces. A mixture generator algorithm capable of determining sets of mixtures that will evenly cover the closed mixture space for three or more component mixtures has been developed. The algorithm receives as inputs a requested number of mixtures and the number of mixture components, and the algorithm outputs a representative set of mixtures that cover the concentration mixture space of interest.

Keywords: Experimental design, spectroscopic calibration, design of experiment, mixtures, quantitative analysis.

INTRODUCTION

To create a successful spectroscopic calibration model for prediction of the concentrations of mixture components requires the consideration of many factors, but one of the first and most important issues to decide on is the set of standards that will be used to build the calibration model. In many calibration problems that involve process monitoring, mixture samples from the actual process can be collected and analyzed to create a data set, but this strategy is not always possible. Sometimes the mixture calibration standards must be constructed from the individual components of the mixtures to be predicted by the finished model. There are also cases where many samples exist, but the choice of a representative subset of samples is not straightforward. These types of experiments require careful consideration to be sure that the calibration sets are representative of the mixtures that the model will be asked to predict in the future. A representative calibration set is one which contains samples that cover the full range of variation expected in future samples to be analyzed by the calibration model under construction.

Kramer¹ has discussed several important points related to the selection of representative spectral calibration sets. First, the calibration set should cover the entire concentration ranges of interest for all analytes since extrapolation of the model beyond the concentration ranges represented in the calibration set generally does not yield accurate results, even for linear models. Second, calibrations dealing with multi-

component systems require calibration sets to be designed from a multi-dimensional standpoint. Kramer points out that it is possible to span the entire concentration ranges of interest for all analytes without representing the entire region of interest if each sample contains only one analyte of interest and the concentration of each analyte is varied over the entire range of interest. This statement means that multi-component mixtures that are spaced throughout the mixture space of interest should be used in a calibration set. Finally, Kramer points out the importance of visualizing the multi-component mixture space to determine if the calibration set representatively covers the mixture space of interest.

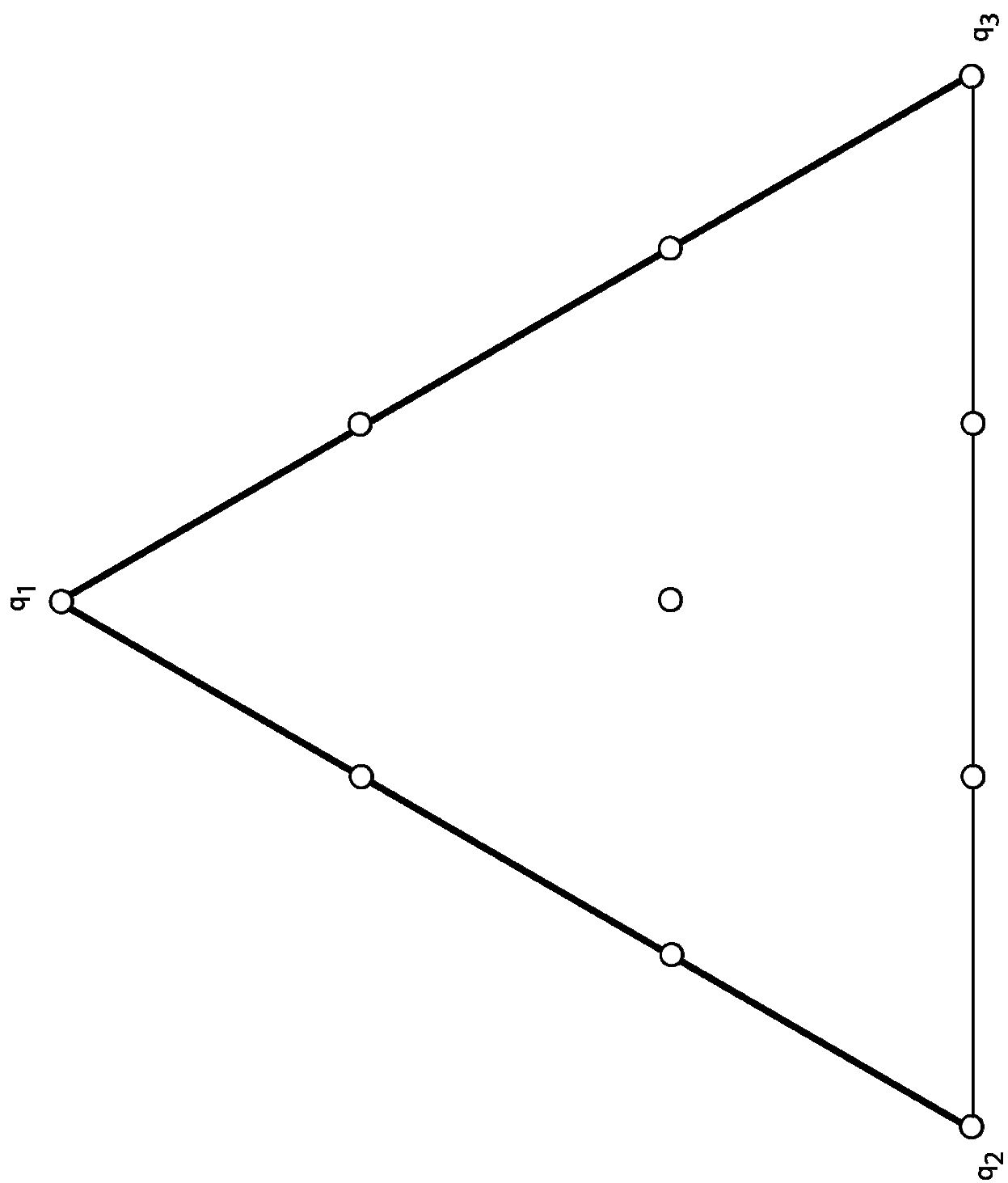
While the first and third points relate to spectral calibrations in general, the second point relates specifically to factorial experiments. A factorial experiment is only concerned with the concentrations of individual analytes. This type of system would be represented by four analytes dissolved in a solvent. In this type of mixture, the concentrations of the analytes can vary independently. In other words, this type of system allows for the concentration of a single analyte to be varied over the entire range of interest for that analyte while the concentrations of the other analytes are held at a constant level. One should note that this is the type of experimental design that Kramer is warning against in his second point. Instead, the analytes should be varied simultaneously to cover the entire mixture space in a multivariate fashion.

Many methods have been suggested to accomplish this task. The primary method is the use of a factorial design or some variant of a factorial design. The details of the factorial design have been covered in a number of references dealing with

statistics and experimental design¹⁻⁶. The two-level, two-factor, full factorial design provides a simple example of how the factorial design works. This design can be used for an experiment involving mixtures that contain two analytes that both have specified upper and lower concentration limits. For this case, the calibration set would include four mixtures. Let U_i and L_i represent the upper and lower concentration limits for the i^{th} component. For the two-level case, the mixtures would include the two components at the following concentrations: $U_1 U_2$, $U_1 L_2$, $L_1 U_2$, and $L_1 L_2$. One can see how this mixture set could be augmented by adding more than two concentration levels of each component. The factorial designs and their variants are well known and can be applied to a many different types of experiments.

Factorial designs are not directly applicable to closed mixture systems. In a closed mixture system, one is concerned not with the concentrations of analytes, but instead, with how the proportions of the system's components vary. Examples of this type of system would include solid mixtures of powders or solvent systems used in methods such as HPLC. In these closed systems if one proportion is varied, then the proportion of at least one other component in the system much change. The proportions of different components cannot be varied independently as is the case for the analytes' concentrations in a factorial experiment. Cornell⁷ has written an excellent reference on experimental design for these closed mixture systems. One of the standard experimental designs for doing these types of closed mixture experiments is called a simplex-lattice design. In this design, the mixtures or design points to be used to build the calibration are equally spaced throughout the mixture space. Figure 5.1 shows an

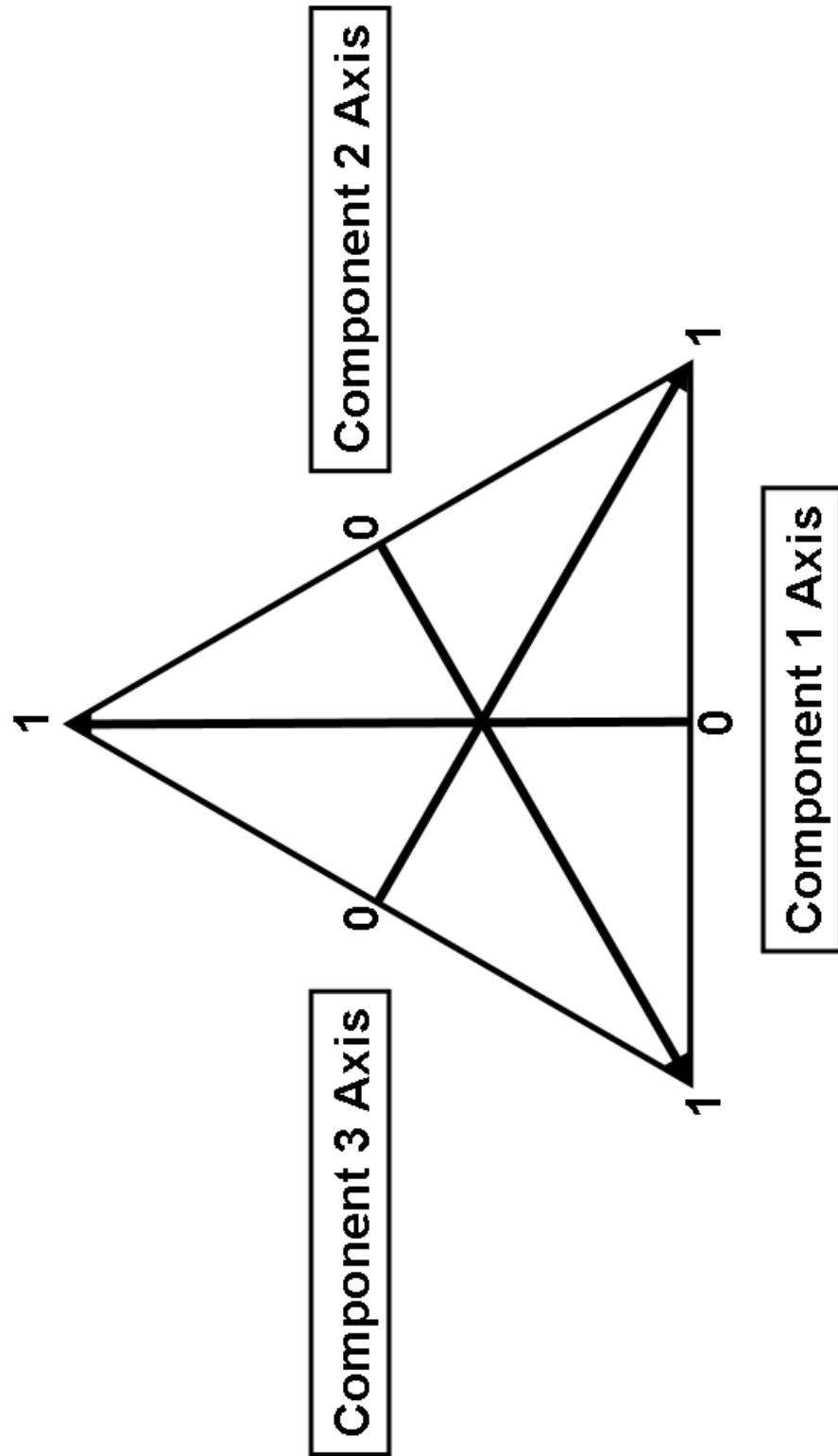
Figure 5.1: Ten point simplex-lattice experimental design for three components.



example of this type of design for a three component system. The i^{th} component is represented by q_i . In this example, the mixture space is represented by a 2-regular simplex, i.e. a triangle. Simplexes are a convenient geometrical means to represent closed mixture spaces. An m component mixture space can be represented by a regular $(m-1)$ -simplex. A 2-simplex is a triangle, a 3-simplex is a tetrahedron, and an m -simplex is the m -dimensional equivalent of a tetrahedron. It is important to realize the significance of this statement: m -component space can be represented in an $(m-1)$ -dimensional space because in a closed mixture there are only $m-1$ independent compositions that can be specified. In these simplexes, the concentration axis for each component in the mixture runs from the center of the sides or faces of the simplex to the vertices. Each axis in the simplex represents a different component of the system and has length equal to unity. At the side or face, the axis represents 0% of the specified component in the mixture, and at the vertex, the axis represents the system being entirely composed of the specified component. Figure 5.2 shows the location of the component axes for a three component system.

The simplex-lattice design can be changed to include more design points or adapted to higher component spaces or both, but a general algorithm to calculate the location of the lattice points for arbitrary numbers of mixtures and mixture components has not been reported. For small numbers of mixtures and mixture components, the calculation of the simplex-lattice design points is trivial, but calculating the design points quickly becomes a daunting task when one desires higher numbers of design points or to work with more components. In the past, simplex-lattice designs have been

Figure 5.2: Shows the location of the component axes for a three component system.



limited to relatively small numbers of mixtures because of the types of calibration models the mixtures have been used to build. Designs for closed mixture systems have primarily been used for the case of modeling systems where the component proportions are the predictor variables for a system property. One example from Cornell⁷ is to create a model to predict the general acceptance by a sensory panel of mixed beverages that contain different proportions of three fruit juices. The predictor variables for this experiment are the proportions of the juices. This type of model is very different from a spectral calibration model.

In a spectral calibration model, the factors of some dimensionally reduced form of the calibration set spectra of the mixtures will be used as the predictor variables. This means that there will generally be many more predictor variables considered than when the component proportions are the response variables; consequently, the number of samples required will be greater. Kramer¹ has discussed that in practice the number of samples required for a successful spectral calibration model ranges from as few as three times to as many as ten times the number of components in the system. When the variation present in the mixtures is extremely large, even more samples than ten times the number of system components may be required. Due to the large number of samples that must be considered, the availability of an algorithm to produce mixture designs with equally spaced design points throughout mixture spaces would allow experimentalist to easily generate potential mixture sets to be used in spectral calibrations for closed mixture systems. This work reports the development of such an algorithm. The mixture generator algorithm developed produces a set of mixtures

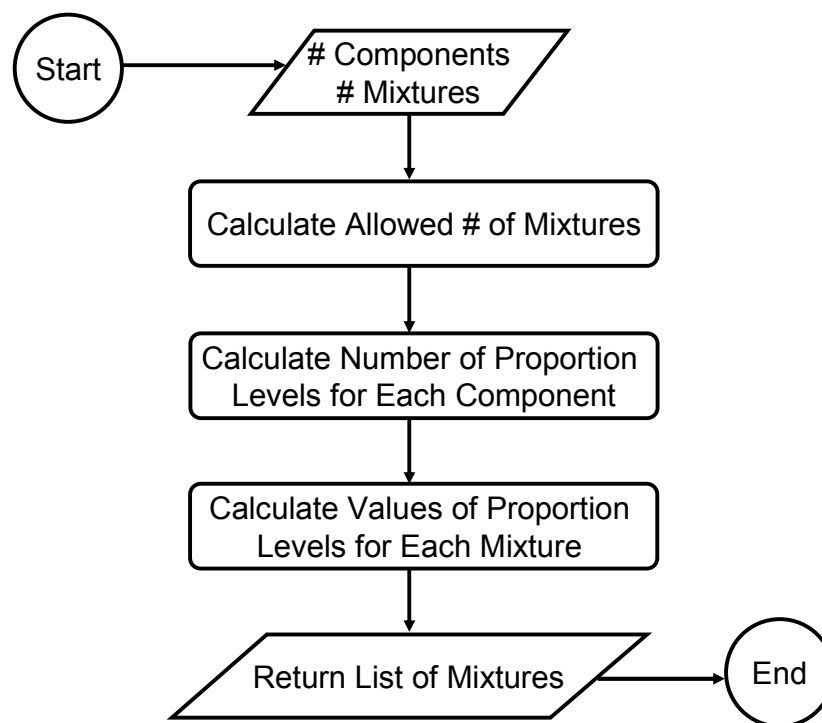
evenly distributed over any m -dimensional mixture space with the number of mixtures in the set determined by the user's input.

MATERIALS AND METHODS

The algorithm was written in the MATLAB programming language and run and tested with the use of MATLAB 7 software (The MathWorks, Natick, MA). The inputs for the algorithm are the number of components in the mixtures the user wants to create and the total number of mixtures to be created. The algorithm outputs the mixture proportions for each component of a certain allowed number of mixtures. The number of mixtures returned to the user is equal to the number of allowed mixtures closest to the number requested by the user. Only a certain number of mixtures are allowed because of the pattern followed by the algorithm to choose the individual mixtures.

The steps of the algorithm are shown in the flowchart of Fig. 5.3. The first step in the algorithm calculates the allowed number of mixtures closest to the number of mixtures requested by the user. The pattern followed by the algorithm to arrive at the allowed number of mixtures is defined by two constraints: (1) every component of the mixture system when considered separately must have its minimum and maximum proportions equal to the minimum and maximum proportions, respectively, for every other component and (2) all mixtures must be positioned at equal intervals throughout the mixture space. To calculate the allowed number of mixtures, the algorithm begins at the smallest number of mixtures allowed by constraints 1 and 2 above. The algorithm then iteratively calculates the next highest number of mixtures allowed until the number calculated exceeds the number requested. Let the final number of mixtures

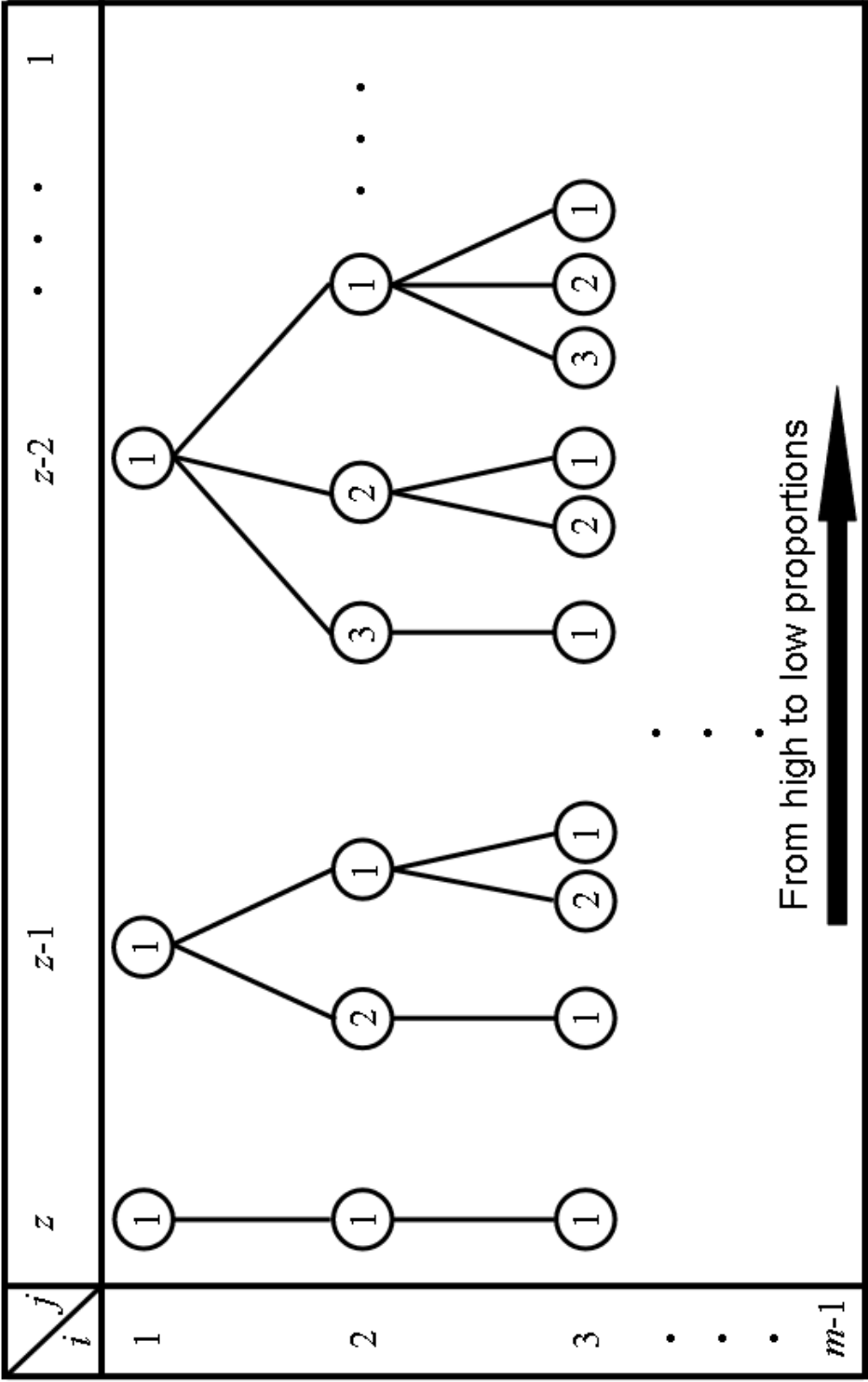
Figure 5.3: Algorithm flow chart.



calculated be the n^{th} number calculated. The algorithm then determines if the requested number of mixtures is closer to the n^{th} or $n-1^{th}$ number calculated. Whichever allowed number is closer to the requested number becomes the number of mixtures that is returned to the user in the output. The lowest number of mixtures that meets constraints 1 and 2 above is a single mixture. In this case, no matter how many components are present, the mixture will be made up of an equal part mixture of all of the components. Constraint 1 is met because the proportions of all components are equal in the single mixture and there are no other mixtures being considered. Constraint 2 is met because there are no other mixtures that this single mixture shares the mixture space with. The next higher allowed number of mixtures depends on the number of components in the system.

Figure 5.4 demonstrates the pattern that must be followed to meet the two constraints described above. In this figure, the index for the number of components is i . The total number of components is m . The index for the proportion set is j and the total number of proportion sets is z . A proportion set contains a set of mixtures that all have the same proportion of component 1. Once the proportion of component 1 is specified, all of the possible combinations of proportions for components 2 through $m-1$ are considered to belong to the proportion set for that specified level of component 1. The numbers inside the circles are the proportion levels and are represented by the index k . For instance, if one considers all the mixtures that fall under the $z-1$ proportion set, there are only two different concentrations of component 2. These two levels are labeled proportion level 2 and 1. Level 2 will be the higher proportion of component 2 and

Figure 5.4: Pattern followed by algorithm.



level 1 the lower proportion of component 2. Each unique pathway from the top of the figure to the $m-1^{st}$ row of the figure represents a completely specified mixture, i.e. once the proportions of the first $m-1$ components have been specified the proportion of the m^{th} component is known. The circles in those unique pathways represent the proportions of the components that correspond to the row of the figure the circle is found.

To calculate the allowed numbers of mixtures, the algorithm begins by setting $z = 1$. The algorithm then iteratively increases the value of z by 1 until the decision process for choosing the allowed number of mixtures closest to the requested number of mixtures is completed. At each iteration, the algorithm uses the pattern shown in the figure to calculate the number of mixtures that would be generated at that value of z . The total number of mixtures at a particular value of z would be represented by the number of circles found in the $m-1^{th}$ row of the figure. Because the algorithm uses the pattern described to calculate the allowed number of mixtures, once the allowed number of mixtures and the corresponding z value has been calculated, the algorithm has all the information needed to determine the number of proportion levels for each component within each proportion set.

From a programming sense, all of this is accomplished by implementing a series of for loops that count and record the number of proportion level divisions of each component. In Fig. 5.4 these divisions are represented by the lines connecting the circles. For each iteration, the program first counts and record the integers from 1 to z . These values can be thought of as making up a vector with z entries. Let this vector be

a. The program then writes a new vector based upon the entries of **a**. For this new vector, the program reads the first entry of **a**, and then records the integers from 1 to the first entry of **a** in the new vector that can be called **b**. The program then reads the second entry of **a** and records the integers from 1 to the second entry as the next entries in vector **b**. This pattern continues until all of the entries from vector **a** have been read and the corresponding values in vector **b** recorded. The program would create $m-2$ vectors where m is the number of components in the system. The program determines the z that corresponds to the allowed number of mixtures by summing the value of the entries of the $m-2$ vector for each z value that is tried: This sum of the entries equals the total number of mixtures that will result for that particular z value. For $z = 3$ and $m = 4$, the vectors generated would look as follows: $\mathbf{a}=(1,2,3)$ and $\mathbf{b}=(1,1,2,1,2,3)$. The entries of these vectors can be used to generate the proportion level numbers shown inside the circles in Fig. 5.4. The proportion level numbers for component 1 will always be a series of ones equal in number to the value of z . The proportion levels for the second component can be found by reading the entries of vector **a**, and then counting and recording the integers from the first entry of **A** backwards to 1, the second entry of **A** backwards to 1, and so on until all of the entries of **A** have been examined. The proportion levels for the remaining components up to component $m-1$ can be generated in this method. .

The final step of the algorithm before the mixtures are output is to calculate the value of the proportion levels for each mixture in the mixture test set. Equation 5.1 is

used to calculate the proportion values for the first $m-1$ mixture components for each individual mixture:

$$P_{ijk} = \frac{b_i p_{ijk}}{t_{ij} + m - i} \quad (5.1)$$

In Eq. 5.1, P_{ijk} is the proportion of component i in proportion set j in proportion level k . When $i = 1$, $b_i = j$, but when $i > 1$, $b_i = k$. The proportion of mixture that has not been specified for the given mixture is p_{ijk} , the number of proportion levels in the j^{th} proportion set for the i^{th} component is t_{ij} , and m is again the total number of components in the mixture system. When $i = 1$, t_{ij} is defined to be equal to z instead of 1. The proportion values for the m^{th} component are found by subtracting the sum of the proportions of the other $m-1$ components from 1. Table 5.1 shows the formulae and the proportion values for the example of $m = 3$ and $z = 3$. The numbers used in the formulae can be found by referring to Fig. 5.4.

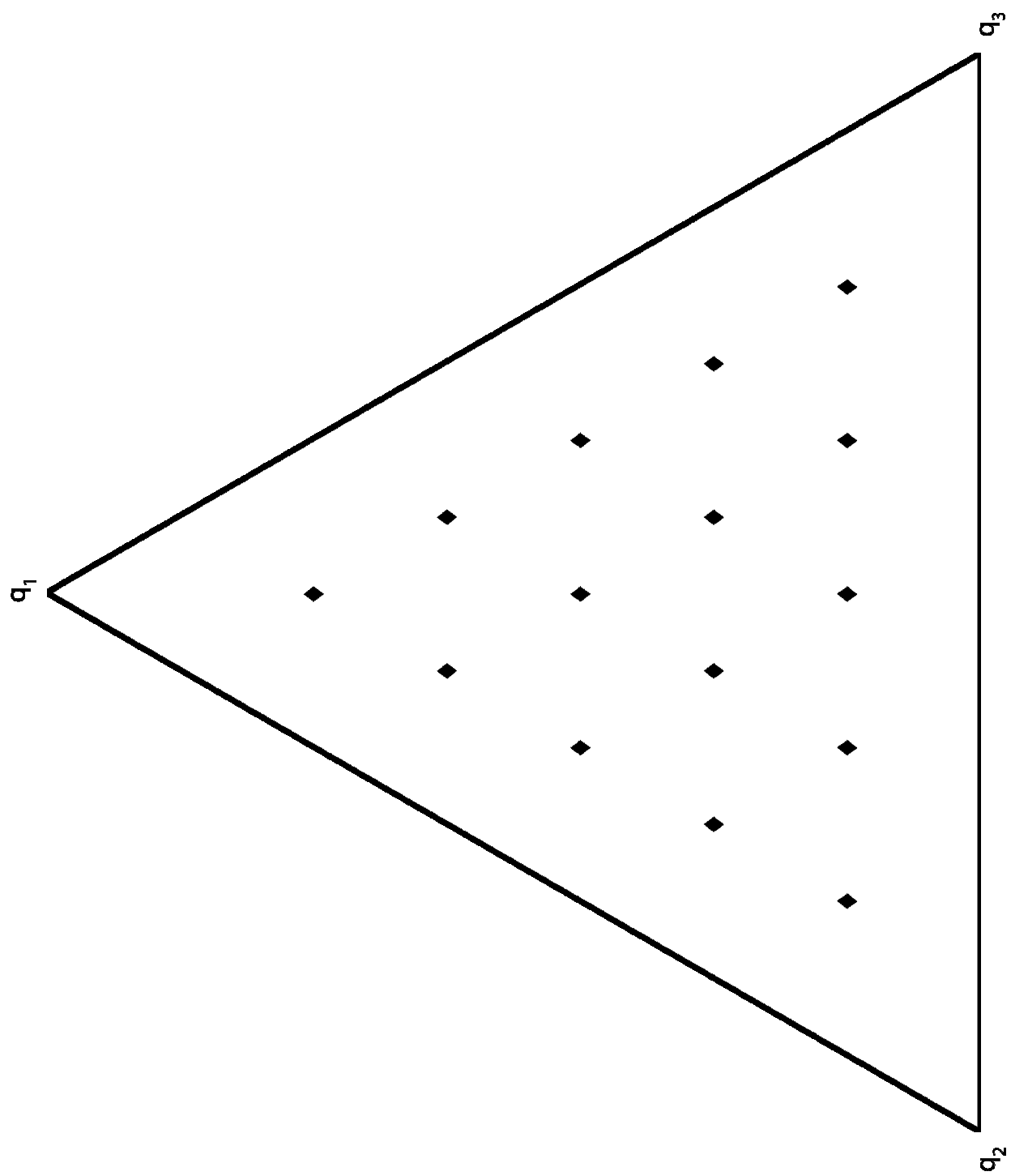
RESULTS AND DISCUSSION

Figure 5.5 shows a plot of the mixture set produced by the algorithm for a 3-component mixture space when the input to the algorithm requested 15 mixtures. One will immediately notice several facts. First, this design provides the most even coverage of the interior of the space possible with 15 mixtures. The second fact is one that has already been mentioned: the mixtures are only distributed over the interior of the space. This design is intentional and allows the experimentalist to include mixtures on the edges, vertices, and faces of the mixture space in several ways. One method would be to augment this design by using surface mixtures that would create a complete simplex-

Table 5.1: Example calculations from algorithm for three component system with $z = 3$.

| i | $z = 3$ | $z = 2$ | | $z = 3$ | | |
|-----|--|--|--|--|--|--|
| 1 | $\frac{3(1)}{3+3-1} = \frac{3}{5}$ | $\frac{2(1)}{3+3-1} = \frac{2}{5}$ | | $\frac{1(1)}{3+3-1} = \frac{1}{5}$ | | |
| 2 | $\frac{1\left(\frac{2}{5}\right)}{1+3-2} = \frac{1}{5}$ | $\frac{2\left(\frac{3}{5}\right)}{2+3-2} = \frac{2}{5}$ | $\frac{1\left(\frac{3}{5}\right)}{2+3-2} = \frac{1}{5}$ | $\frac{3\left(\frac{4}{5}\right)}{3+3-2} = \frac{3}{5}$ | $\frac{2\left(\frac{4}{5}\right)}{3+3-2} = \frac{2}{5}$ | $\frac{1\left(\frac{4}{5}\right)}{3+3-2} = \frac{1}{5}$ |
| 3 | $1 - \left(\frac{3}{5} + \frac{1}{5}\right) = \frac{1}{5}$ | $1 - \left(\frac{2}{5} + \frac{2}{5}\right) = \frac{1}{5}$ | $1 - \left(\frac{2}{5} + \frac{1}{5}\right) = \frac{2}{5}$ | $1 - \left(\frac{1}{5} + \frac{3}{5}\right) = \frac{1}{5}$ | $1 - \left(\frac{1}{5} + \frac{2}{5}\right) = \frac{2}{5}$ | $1 - \left(\frac{1}{5} + \frac{1}{5}\right) = \frac{3}{5}$ |

Figure 5.5: Visual depiction of algorithm's output for 15 mixtures of 3 components.



lattice design, or one could augment the equally spaced internal points with a different surface design. Algorithms for calculating surface design points have been reported in the literature⁷. A third choice would be simply to use the design as is and include no surface points. In any case, the algorithm provides the experimentalist with a preliminary set of evenly distributed internal mixtures to begin working with and the freedom to choose how to proceed from that point.

As stated in the introduction, Kramer's third point concerning the choice of calibration sets was that one should be able to visualize the multi-component space to be able to see how the calibration set is distributed throughout that space. Using the simplex representation of closed mixture spaces works very well for visualizing three and even four component systems, but beyond four components, visualization becomes difficult or impossible. It is certainly impossible for humans to visualize more than three orthogonal dimensions directly, so the question of how to show that the algorithm works for $m-1$ -dimensional space for any value of m arises. While specific examples for three and four component systems can be shown, the inductive argument that follows demonstrates that the algorithm is effective for higher dimensional spaces.

Earlier in this article the assertion was made that the pattern demonstrated in Fig. 5.4 along with Eq. 5.1 will yield an evenly distributed mixture set for m -component space for any value of m . Figure 5.6 shows the specific pattern used for the case of 10 mixtures in 4 components, and Fig. 5.7 shows the actual distribution of mixtures calculated by the algorithm. In a geometric sense, the last row of circles in Fig. 5.6 can

Figure 5.6: Pattern followed by algorithm to produce 10 mixtures in 4 component space.

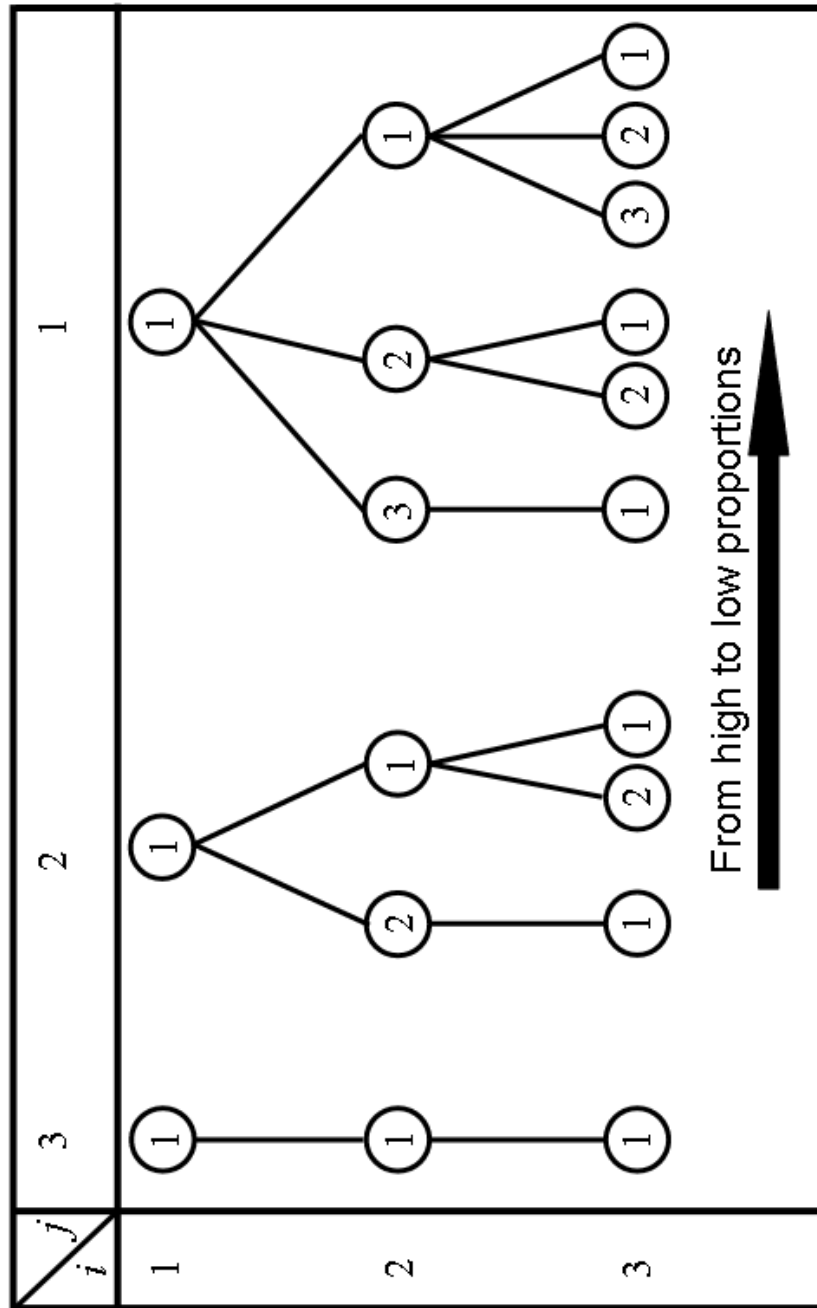
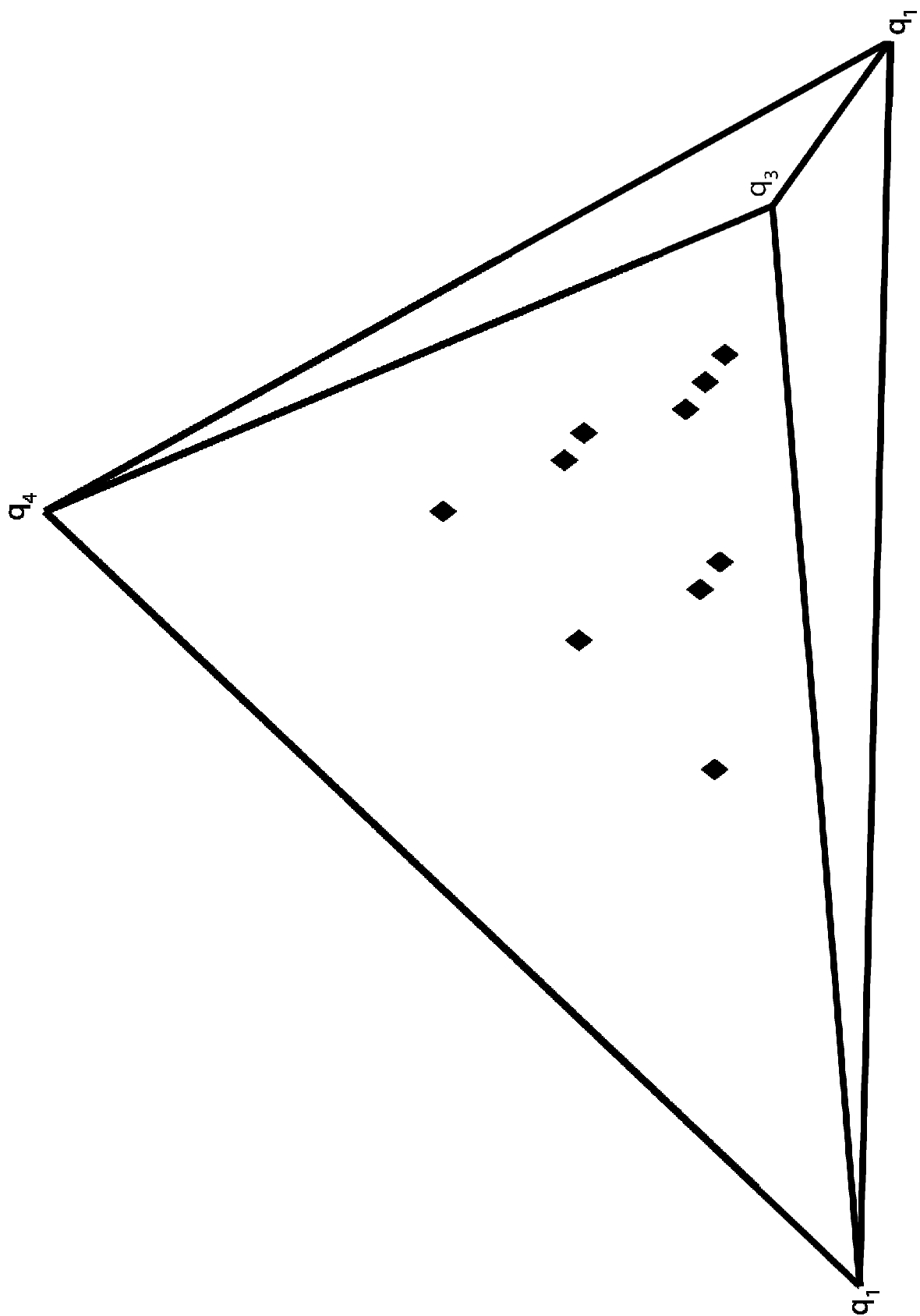
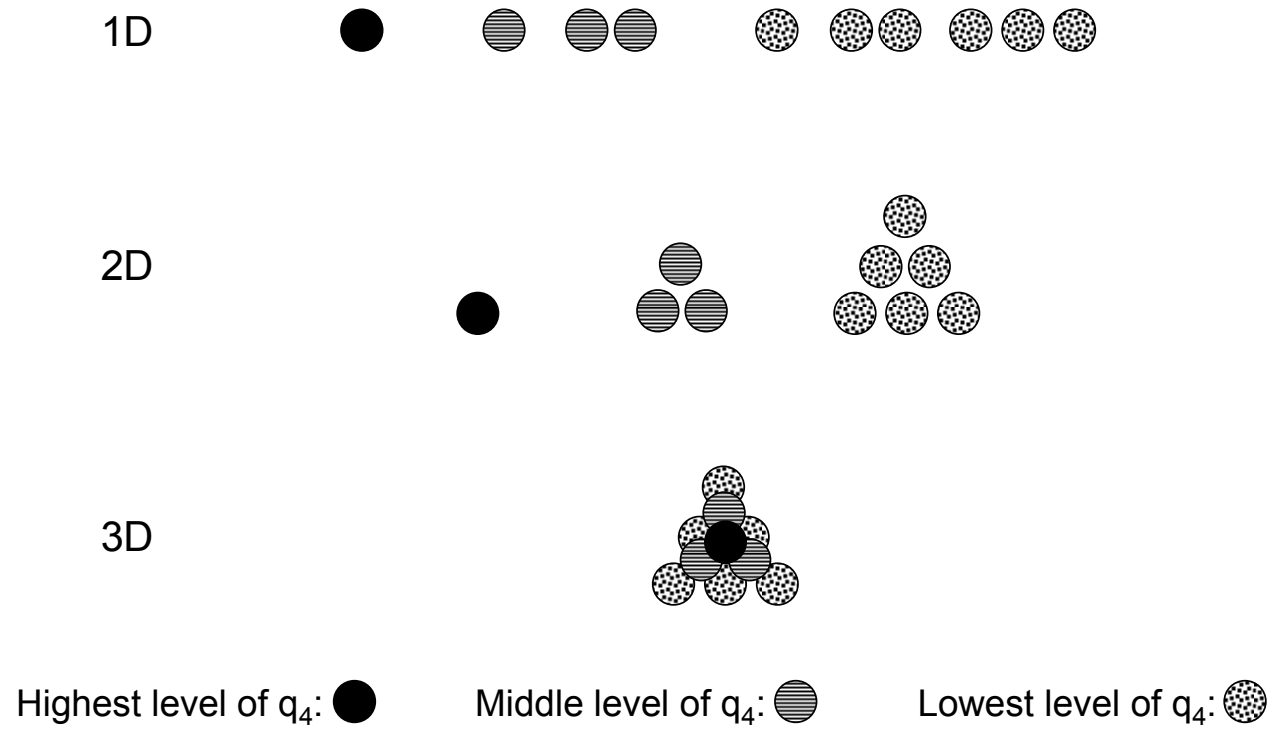


Figure 5.7: Visual depiction of algorithm's output for 10 mixtures of 4 components.



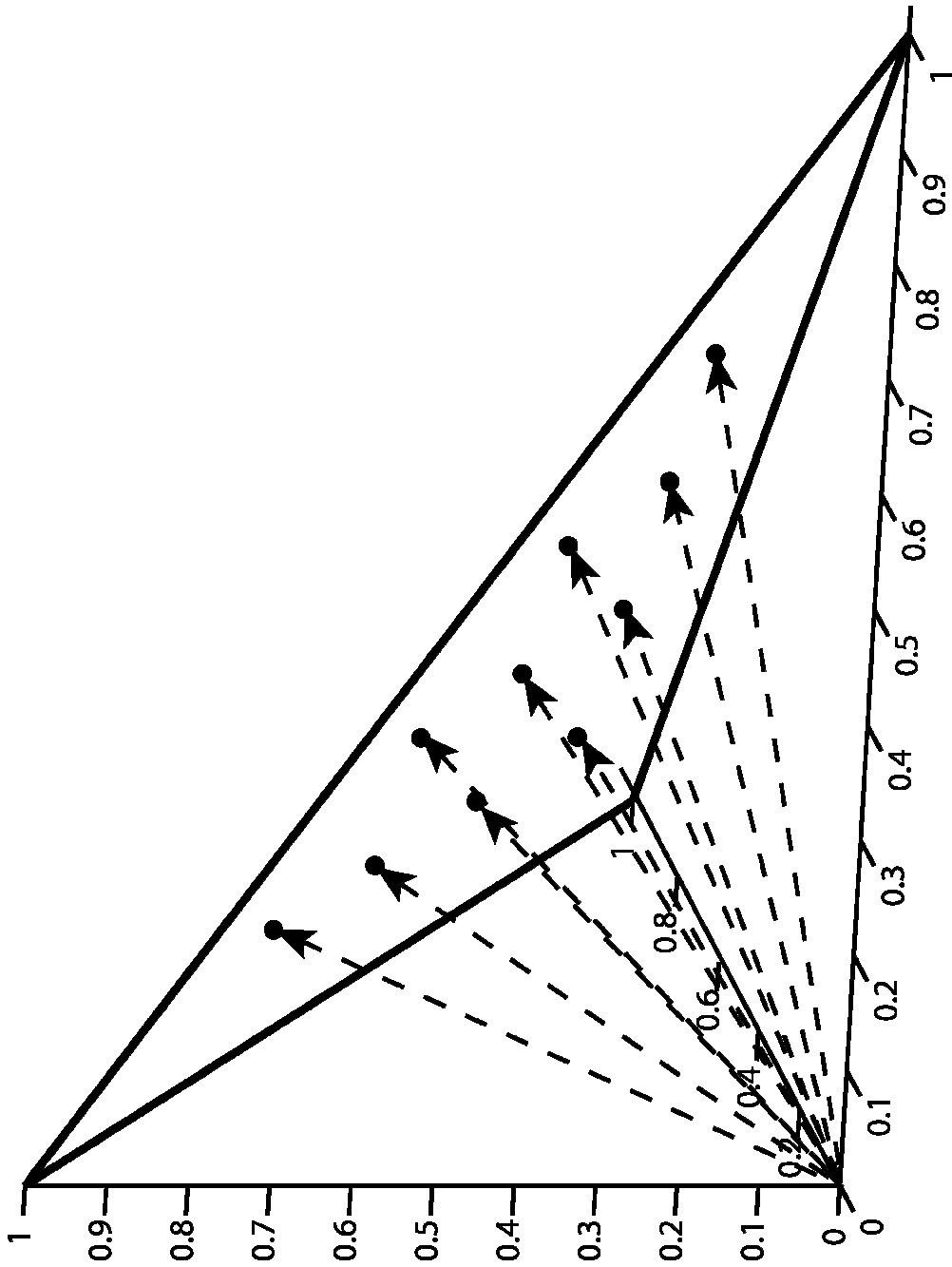
be thought of as a 1-dimensional unfolded representation of the entire mixture set. Figure 5.8 demonstrates how this 1-dimensional representation of the mixture system can be folded up into the three dimensional or four component representation shown in Fig. 5.7. In the second row of Fig. 5.8, one sees the circles folded up into a 2-dimensional representation of the system, but these triangular shaped collections of circles are really just separate three component representations similar to the one seen in Fig. 5.5. In this case though, each three component representation can be thought of as existing at a fixed value of the fourth component. The different shading patterns of the circles represent the three different levels of the fourth component. The third line of Fig. 5.8 demonstrates these three component representations folded up into a single four component representation: Each separate three component representation becomes a plane of mixtures within the four component space. (This tetrahedral collection of design points is being viewed along the q_4 axis.) It is now easy to see what happens when the case of five components is considered. The pattern that would be dictated for five components would take separate four component representations each at a different level of the fifth component and fold those separate representations up into another dimension to achieve the five component mixture set representation. Although the pattern cannot be directly visualized, one can see how this reasoning can be extended to any arbitrary dimension. In summary, the design for more complex systems can be obtained by building up the patterns known and demonstrated for three and four component systems.

Figure 5.8: Demonstrates how the pattern shown in Fig. 5.4 is an unfolded representation of the mixture sets.



In discussing this algorithm, it is important to address the concern of whether or not the mixtures generated will be linearly independent of one another. Let us consider the simplest case of a three component system. For three components, three mixtures at most can be linearly independent. For convenience, consider the case of 10 mixtures evenly distributed in 3-component space. The 2-simplex, a triangle, is used to represent a three component system, but only two orthogonal dimensions are required for the representation because the proportion of the third component in the closed system is known once the first two components have been specified. This fact means that the 2-simplex is really on a plane in 3-dimensional space. Because of this, the question of whether or not all the mixtures will be linearly dependent arises. Let each of the design points be represented by a vector with three entries equal to the three proportions of that mixture. This example is shown visually in Fig. 5.9. If there is only one linearly independent mixture, all of the vectors will lie along the same line. If there are only two linearly independent mixtures, all of the vectors will lie in the same plane. If there are three linearly independent mixtures, then three dimensions will be required to describe the set of vectors. As can be seen in Fig. 5.9, one cannot draw a single line or plane that will contain all of the vectors; thus, the system is of full rank and there will be three linearly independent mixtures. Although this example in itself is not a proof of the algorithm's ability to produce a set of linearly independent mixtures, this example could be extended to an arbitrary dimensional space to show that the algorithm produces mixture sets with the highest degree of linear independence possible.

Figure 5.9: Vector plot demonstrating linear independence of data set.



Finally, the way in which the results of the algorithm will be used should be discussed. Intuitively, if the mixtures representatively cover the sample space, one would expect the calibration model to do the best job possible of predicting the concentration of future mixtures. This might not always be the case. As Cornell⁷ has discussed, different mixtures (design points) included in the model will influence the calibration model to different extents. The leverage of each mixture is one measure of this influence. The least squares regression model is represented by Eq. 5.2:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (5.2)$$

In this equation, \mathbf{y} is the vector that contains the response variable, \mathbf{X} is the matrix of predictor variables, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is the vector of residuals. The value of the estimated regression vector, $\hat{\mathbf{b}}$, is given by Eq. 5.3:

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (5.3)$$

The model's estimate of the response variable vector, $\hat{\mathbf{y}}$, is given by Eq. 5.4:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{H}\mathbf{y} \quad (5.4)$$

In Eq. 5.4, \mathbf{H} is called the hat matrix and is defined by Eq. 5.5:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (5.5)$$

The leverage of a particular mixture on the model is measured by the corresponding diagonal entry of the hat matrix. The error vector, \mathbf{e} , contains the residuals for the predictions made by the model at the design points and is given by Eq. 5.6:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y} \quad (5.6)$$

This equation demonstrates that the matrix $(\mathbf{X}^T \mathbf{X})^{-1}$ and, thus, the mixtures or design

points chosen are related to the error in the model. Many optimization methods based on different goals for choosing the set of mixtures that will lead to the best $(\mathbf{X}^T\mathbf{X})^{-1}$ matrix and, thus, the best model with lowest errors have been reported^{6,7}. All of these optimization methods, though, require an initial pool of mixtures or design points from which to choose. The point of this work is not to explore the implications of the algorithm's results for all of these different optimization methods, but rather, the emphasis of this work is to provide the experimentalist with a tool to begin the initial experimental design process. The experimentalist may use the mixtures that are generated by the algorithm in a number of ways to develop a calibration model. The experimentalist may even choose subsets of the mixtures generated to find a representative set of mixtures for a constrained mixture space where all of the components cannot vary over the whole range from 0% to 100% of a mixture's composition.

CONCLUSION

To develop a successful spectral calibration model for a closed mixture system requires the consideration of many factors, but without first developing a representative calibration set, one cannot begin to create a successful calibration model. Until now, no general algorithm to specifically generate mixtures evenly distributed over mixture spaces has been reported. The mixture generator algorithm reported in this work fills this gap by allowing an experimentalist to generate a set of mixtures evenly distributed throughout a closed mixture space for any arbitrary number of components. The algorithm allows the experimentalist unlimited choice in the number of mixtures to be

generated. The algorithm also allows experimentalists the freedom to create a set of mixtures larger than required for the calibration set, so that subsets of the mixtures can be used for model validation and test sets. The development of this algorithm gives experimentalists a foundation from which to build successful spectral calibration models.

REFERENCES

- (1) Kramer, R. *Chemometric Techniques for Quantitative Analysis*; Marcel Dekker: New York, 1998.
- (2) Bhattacharyya, G. K.; Johnson, R. A. *Statistical Concepts and Methods*; John Wiley & Sons: New York, 1977.
- (3) Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nyström, Å.; Pettersen, J.; Bergman, R. *Chemometrics and Intelligent Laboratory Systems* **1998**, 42, 3-40.
- (4) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: a textbook*; Elsevier: Amsterdam, 1988.
- (5) Miller, J. N.; Miller, J. C. *Statistics and Chemometrics for Analytical Chemistry*; 5th ed.; Pearson Prentice Hall: Harlow, England, 2005.
- (6) Stoyanov, K.; Walmsley, A. D. In *Practical Guide to Chemometrics*; 2nd ed.; Gemperline, P., Ed.; Taylor & Francis: Boca Raton, 2006, p 263-338.
- (7) Cornell, J. *Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data*; 3rd ed.; John Wiley & Sons: New York, 2002.

CHAPTER 6

THE EFFECTS OF DIFFERENT ATR ACCESSORIES ON SPECTRAL COMPARISON IN FT-IR SPECTROMETRY¹

¹ J.B. Loudermilk, D.S. Himmelsbach, F.E. Barton II, J.A. de Haseth, To be submitted to Appl. Spectrosc.

ABSTRACT

While spectral library searching can be an invaluable tool for identification of unknowns, one must be aware of the complications that arise. Spectral differences always exist between spectra measured with different spectrometers and by different methods, e.g. transmission, attenuated total reflection (ATR), and specular reflection. All of these spectral differences must be considered when one searches a spectrum against a library when that spectrum was measured with a different spectrometer, sampling technique, or both than the instruments or techniques used to build the library. Because of differences in depth of penetration and dispersion effects, the differences between transmission and ATR spectra have received a great deal of attention, but the spectral differences encountered when comparing spectra measured on different ATR accessories has not received the same amount of attention. ATR accessories come in many different designs and incorporate several different materials for their internal reflection elements (IREs). In this study, spectra of a polyethylene terephthalate (PETE) film were measured with the use of five different ATR accessories that represented four different models of accessories from a total of three different manufacturers. Each accessory was paired with a different spectrometer except that two of the four models were paired with one spectrometer. The IREs represented were Zinc selenide (ZnSe), diamond, and Silicon (Si). Spectral differences were observed and investigated, and the ability of depth of penetration correction and a proprietary correction for ATR dispersion effects to increase the similarity of the spectra was explored. The identity of the accessories and manufacturers will remain anonymous.

Keywords: ATR, Depth of penetration, Anomalous dispersion, Spectral comparison, Spectral corrections, Baseline corrections, Internal reflection elements.

INTRODUCTION

Searching infrared spectra of unknown samples against a spectral library can be a valuable tool to determine the identity of an unknown. If spectra of the unknown or a similar substance are included in the library, the library search may give the researcher important information about the identity of the unknown, but if the spectra of the unknown in the library are not sufficiently similar to the spectrum of the unknown being searched against the library, the search may be useless or even give specious results. In addition, the rate of misidentifications will increase as the similarity of the library spectra to one another increases.

It is important to understand that differences in spectra of chemically identical substances can arise because many more factors than the chemical composition of a substance lead to the infrared spectrum of that substance. Some of the most obvious differences will arise from the method of infrared spectrometry used. For example, the spectra obtained by transmission versus ATR will be quite different¹⁻⁶. Even when using the same method to expose the sample to radiation, the response functions of different instruments will not be the same and will affect the spectra. The light source, the detector, and every part between the light source and the detector influences the response function of the instrument². The instruments' components will affect factors such as noise, spectral intensity, and alignment of the abscissa scale of the spectra. Changes in the spectral resolution, the type of apodization function used, and the

number of co-added scans will also produce differences in spectra. In short, there is any number of factors that can lead to differences in two spectra of the same substance, and we have only listed a small number of those factors here. Throughout the years, many of these issues have been explored, and the text by Griffiths and de Haseth² covers many of these points in detail.

For spectral library searching, the spectral differences between transmission and ATR spectra have been discussed at length¹⁻⁶. In transmission, the infrared radiation from the source passes through the sample where some of the radiation is absorbed on its way to the detector. In ATR, the sample absorbs infrared radiation by being in contact with an IRE that is in the beam of radiation from the source. Because the infrared radiation enters the IRE above the critical angle of the higher refractive index IRE in contact with the lower refractive index sample, the radiation beam undergoes total internal reflection inside the IRE; however, the electric field of the radiation at the interface extends into the sample allowing absorption of the radiation to take place. This electric field at the interface is known as the evanescent wave. In transmission and ATR, the absorbance of the sample is proportional to the sample thickness or depth of penetration of the evanescent wave, respectively. The absorbance for transmission mode can be calculated by Eq. 6.1 and for ATR by Eq. 6.2:

$$A = abC \quad (6.1)$$

$$A = (\log e) \frac{\eta_2 E_0^2 d_p a}{\eta_1 (\cos \theta)^2} \quad (6.2)$$

In Eq. 6.1, a is the absorption coefficient, b is the pathlength, and C is the analyte

concentration. In Eq. 6.2, η_2 is the sample refractive index, η_1 is the refractive index of the IRE, E_0 is the electric field strength of the evanescent wave at the sample/IRE interface, θ is the angle of incidence, d_p is the depth of penetration of the evanescent wave, and a is the linear absorption coefficient per unit thickness of sample.

Equation 6.2 is the key to understanding the differences between transmission and ATR spectra. The implications of this equation for ATR spectra have recently been discussed by Nishikida and Kempfert⁴, and work done in this area by Nishikida and others has resulted in the deployment of the “Advanced ATR Correction”⁷. Some of the theoretical basis for this correction is based upon work done by Plaskett and Schatz⁵ and Schatz et al.⁶ in the 1960s. From Eq. 6.2, one can see that similar to pathlength for transmission the absorbance of the sample in ATR is proportional to the depth of penetration of the evanescent wave. In transmission, the pathlength or thickness of the sample is a constant value, but in ATR the depth of penetration varies with wavelength. Equation 6.3 shows the effective depth of penetration of the evanescent wave:

$$d_p = \frac{\lambda}{2\pi\eta_1 \sqrt{\sin^2 \theta - (\eta_2 / \eta_1)^2}} \quad (6.3)$$

Equations 6.2 and 6.3 together show that the first difference between transmission and ATR spectra is that at longer wavelengths or lower wavenumber the depth of penetration will be greater than at shorter wavelengths or higher wavenumber. This change in depth of penetration means that absorption bands at the lower wavenumber end of the mid-IR spectrum will have larger absorbance values relative to the higher wavenumber end of the spectrum simply because of the increase in depth of

penetration at the lower wavenumber regions. Comparing Eqs. 6.1 and 6.2 also reveals another difference between transmission and ATR spectra: The absorbance and depth of penetration in ATR are proportional to the refractive index of the sample material. It is commonly known that sharp changes in refractive index of a material occur over absorption bands, where the refractive index is lower on the higher wavenumber side of the band and higher on the lower wavenumber side, known as anomalous dispersion, compared to the average refractive index in regions of the spectrum where no absorption occurs (normal dispersion). Anomalous dispersion tends to cause bands in ATR spectra to be shifted to lower wavenumber values compared to transmission spectra. Because the value of η_1 in the ratio η_2/η_1 will be different for each IRE material of a different refractive index, this will lead to a relative change in depth of penetration for the different IREs. This indicates that spectra measured with different IREs will experience different spectral shifts to lower wavenumber values, and the magnitude of the shift should decrease as the ratio η_2/η_1 decreases. The average refractive index for most organic materials in the infrared region of the spectrum is approximately 1.5², so as the refractive index of the IRE material increases, the spectral shifts because of anomalous dispersion become smaller.

If ATR spectra are to be successfully compared to transmission spectra, the ATR spectra must be corrected for these spectral differences. The differences caused by changes in depth of penetration due simply to changes in wavenumber can be corrected by multiplying the ATR spectrum by the wavenumber values². The spectral changes induced by anomalous dispersion require a more complex correction. Some of the

details required for such a correction have been discussed in the literature^{5, 6}. The “Advanced ATR Correction” mentioned earlier is a proprietary correction algorithm for dealing with both the dependence of depth of penetration on frequency and the spectral effects induced by anomalous dispersion. The need for these types of correction to be done before comparing ATR and transmission spectra has received a great deal of attention as evidenced by the resources cited above; however, a study that compares the spectra obtained from ATR accessories of different designs and IREs of different refractive indices has not been reported. Spectra from differently designed ATR accessories and from ATR accessories with IREs of different refractive indices will yield different spectra as evidenced by the dependence on both depth of penetration (Eq. 6.3) and absorbance (Eq. 6.2) in ATR. Since there are numerous ATR accessory designs available and ATR has become such a widely used application, it is important to consider the implications of these differences for searching a library of ATR spectra obtained with one ATR accessory against unknowns obtained with a different accessory. Any differences in ATR spectra will be even more significant when a library is composed of extremely similar spectra. The importance of unknown spectra being representative of calibration spectra for multivariate regressions based on ATR spectra must also be considered. To address some of these concerns, this work compares spectra of the same PETE film measured with different spectrometers, ATR accessories, and IREs for non-corrected spectra, depth of penetration corrected spectra, and spectra corrected with the proprietary “Advanced ATR Correction”.

MATERIALS AND METHODS

Spectrometers and ATR Accessories. Since the purpose of this study is to draw attention to the difficulties encountered in the comparison of FT-IR spectra measured with the use of different instruments and different types of ATR accessories and not to discuss the characteristics of specific models of instruments and ATR accessories, the manufacturers of the instruments and ATR accessories used in this study will remain anonymous. Four different FT-IR spectrometers representing three different manufacturers, and five different ATR accessories representing four different models and three different manufacturers were used in this study. Three of the accessories had diamond IREs, one accessory had a ZeSe IRE, and one accessory had a Si IRE. All of the ATR accessories were of single reflection design and had nominal angle of incidence of 45°. Two of the three diamond IREs were the same model ATR accessory from the same manufacturer. A set of codes will be used to identify the comparisons. The three different ATR accessory manufactures are represented by the upper case letters A, B, and C. The four spectrometers will be represented by the Roman numerals I, II, III, and IV. Note that spectrometers II and III were made by the same manufacturer. The five ATR accessories will be represented by the Arabic numbers 1, 2, 3, 4, and 5. For example, the code IC1 would represent a particular spectrometer and ATR accessory. Note that accessories 4 and 5 are the two accessories of the same model. Table 6.1 shows the combinations of spectrometers and accessories for which data were collected.

Table 6.1: Experimental design.

| Spectrometer | Accessory Manufacturer | | |
|--------------|------------------------|------------|------------|
| | A | B | C |
| I | 4(Si) | | 1(diamond) |
| II | | 3(diamond) | |
| III | 5(ZnSe) | | |
| IV | | | 2(diamond) |

Spectra. A heated and pressed PETE film was used as the sample for all the spectra measured for this study. For each spectrometer/ATR accessory combination, six replicate spectra of the film were measured and averaged to obtain the representative spectra for comparison. Between each replicate measurement, the pressure applicator on the ATR accessory was loosened, the active area of the IRE was covered with a different area of the film, and the pressure applicator was reset. The pressure applied for each ATR accessory was determined by increasing the applied pressure until no increases in absorbance could be seen in the real-time processed spectra, indicating the best possible contact between the IRE and the film. The spectra were measured at 4 cm^{-1} resolution with 256 scans co-added. The spectra were processed with the Happ-Genzel apodization function. The spectra ranged from 4000 to 650 cm^{-1} . All spectra were measured with DTGS detectors.

Spectral Comparison. Three types of spectra were compared: the original averaged spectra, the averaged spectra corrected for differences in depth of penetration with the use of the ATR correction available in the Omnic 7.1a software (Thermo Scientific, Madison, WI), and the averaged spectra corrected for depth of penetration and dispersion effects with the use of the Advanced ATR Correction available in Omnic 7.1a. The wavenumber values of the data points at which the spectra were collected were set to a reference laser wavenumber of 15798.0 cm^{-1} . This wavenumber correction was accomplished through the normalize frequency command in Omnic 7.1a. For visual comparison, the spectra were scale normalized so that the lowest point in each spectrum had an absorbance value of 0 and the highest point an absorbance value of 1.

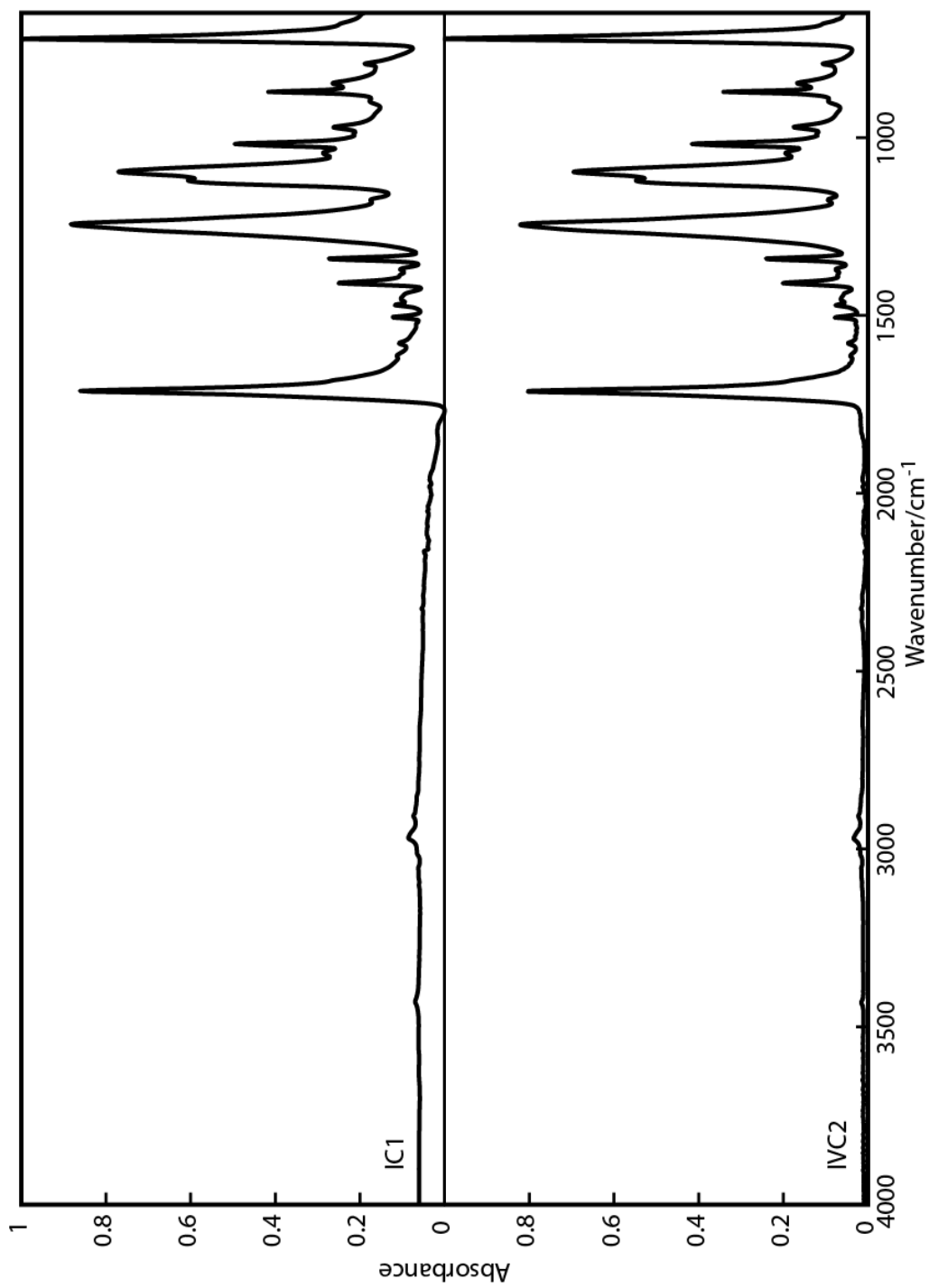
In an additional study, the averaged spectra were compared after being subjected to manual spline baseline corrections using varying numbers of correction points. These baseline corrected spectra underwent the “Advanced ATR Correction” in order to examine the effects of baseline correction on spectral similarity. Correlation coefficients among the spectra of each group were used as a comparison metric to determine how similar the spectra from different spectrometer/ATR accessory combinations were. The correlation coefficients were calculated with the use of MATLAB 7 software (The MathWorks, Natick, MA).

RESULTS AND DISCUSSION

In an ideal experimental design, each ATR accessory would have been paired with a single spectrometer, to allow for comparison of the spectral differences due solely to differences in the ATR accessories. Unfortunately, this type of experiment was impossible to conduct because of the inability to acquire all four models of ATR accessories in one laboratory; however, one can still gain a great deal of useful information about spectral comparison. Throughout this section, a series of comparisons that explore the spectral similarities and differences because of the different ATR accessories and spectrometers will be discussed.

Figure 6.1 compares the spectra measured from two ATR accessories of the same model on two different spectrometers: IC1 and IVC2. These spectra demonstrate the point discussed earlier that spectra of a sample measured with different spectrometers under the same conditions can be different. Spectrum IVC2 shows a baseline close to zero absorbance above 1770 cm^{-1} while spectrum IC1 shows a baseline substantially

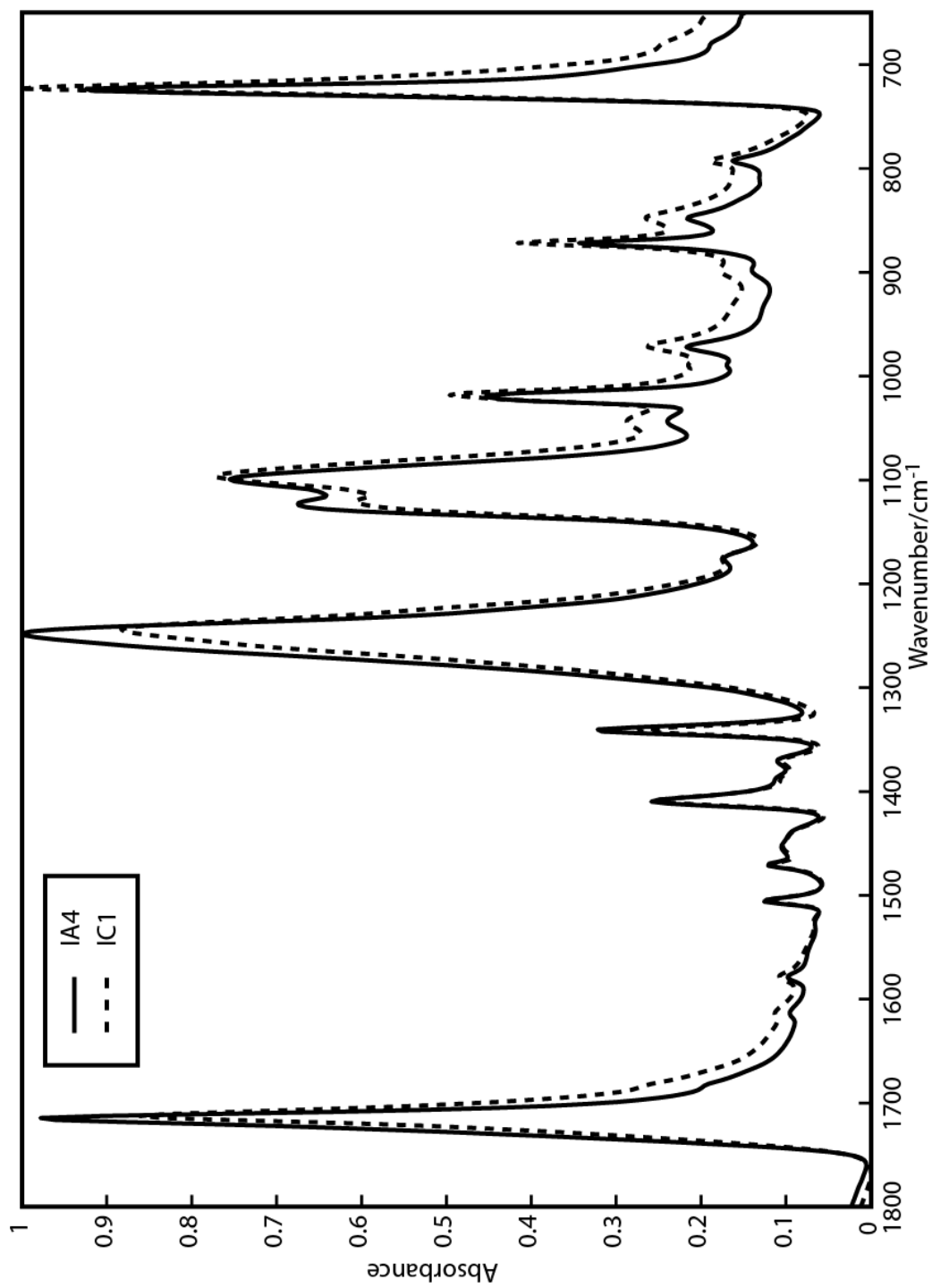
Figure 6.1: PETE spectra IC1 and IVC2.



higher than zero from 1770 to 4000 cm^{-1} . Spectrum IC1 also shows a higher baseline in the region of the spectrum below 1770 cm^{-1} as well. If one assumes that the two ATR accessories of the same model affect the spectra in the same ways, one must conclude that these spectral differences are due to differences in the spectrometers. While the reasons for these higher baselines are not known, the greater anomalous dispersion shown by spectrometer I distorts the baseline in the lower wavenumber region. It could be possible that the greater anomalous dispersion exhibited by spectrometer I is a result of a wider beam angle in spectrometer I than IV. In any case, Fig. 6.1 demonstrates that even under the same conditions different spectrometers can produce different spectra, and this fact must be considered when either qualitative or quantitative comparisons are being made.

Figure 6.2 compares two spectra measured on the same spectrometer, but measured with the use of two different ATR accessories. In this case, one can assume that the major factor that contributes to the spectral differences is the different ATR accessories. The ATR accessory for spectrum IC1 had a diamond IRE, and the accessory for spectrum IA4 had a Si IRE. As described by Eq. 6.3, the depth of penetration will vary across the spectral range for both of these accessories because of the different refractive indices of diamond ($\eta = 2.41$) and Si ($\eta = 3.41$). The effects of depth of penetration are evidenced by the intensity of the lower wavenumber bands in IC1 (diamond IRE) being greater than the same bands in IA4 (Si IRE). Recall that Eq. 6.3 says that the depth of penetration and, thus, the absorbance will be greater for diamond

Figure 6.2: PETE spectra IC1 and IA4.



than Si. One will also notice that the effect of anomalous dispersion is evidenced in these spectra by derivative-like spectral features, and the shifting of bands in spectrum IC1 (diamond, $\eta = 2.41$) to lower wavenumber relative to the band positions in IA4 (Si IRE, $\eta = 3.41$) demonstrates the dependence of anomalous dispersion on the IRE material. This shift is expected because the refractive index of diamond is less than the refractive index of Si.

Figure 6.3 shows all of the scale normalized spectra in the region from 1800 to 650 cm^{-1} . One can see the band shifting in this region of the spectrum because of anomalous dispersion. Although one must realize that more effects than anomalous dispersion are present, the bands in this region of the spectrum are generally shifted farther to the low wavenumber region as the refractive index of the associated IRE decreases from Si ($\eta = 3.41$) to diamond ($\eta = 2.41$) and ZeSe ($\eta = 2.40$).

Although these qualitative comparisons of the spectra are useful, a quantitative comparison of the spectral similarity was calculated to examine the implications of the spectral differences on library searches. The correlation coefficients between each of the absorbance scale normalized spectra were calculated for the full range from 4000 to 650 cm^{-1} and from 1800 to 650 cm^{-1} . The results are presented in Table 6.2. The correlation coefficient, r , is a commonly used metric for spectral comparison in spectral library searches, and its values can range from 0 to 1. (Theoretically, the correlation coefficient can range from a value of -1 to 1, but since the absorbance values have been confined to zero or above, the range of correlation coefficient values for this application is from 0

Figure 6.3: Scale normalized PETE spectra from 1800 to 650 cm^{-1} : a) IA4, b) IC1, c) IVC2, d) IIIA5, and e) IIB3. The vertical dotted lines draw attention to the spectral shifts.

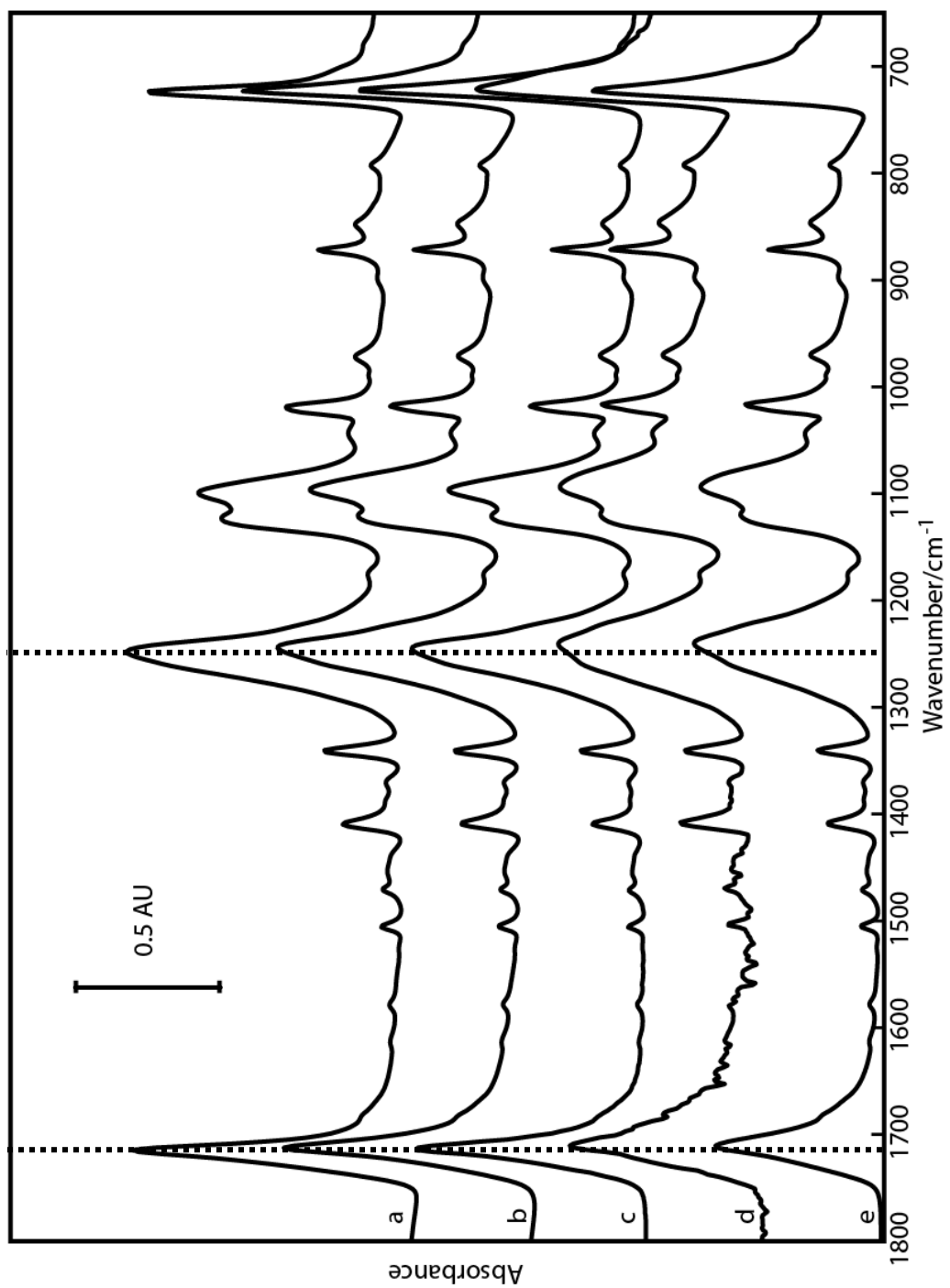


Table 6.2: Correlation coefficients for PETE spectra.

| Full Spectra | | | | | |
|--------------|--------|--------|--------|--------|--------|
| | IVC2 | IIB3 | IIIA5 | IC1 | IA4 |
| IVC2 | 1.0000 | | | | |
| IIB3 | 0.9547 | 1.0000 | | | |
| IIIA5 | 0.9126 | 0.9872 | 1.0000 | | |
| IC1 | 0.9858 | 0.9731 | 0.9509 | 1.0000 | |
| IA4 | 0.9783 | 0.9282 | 0.8955 | 0.9721 | 1.0000 |

| 1800 to 650 cm ⁻¹ | | | | | |
|------------------------------|--------|--------|--------|--------|--------|
| | IVC2 | IIB3 | IIIA5 | IC1 | IA4 |
| IVC2 | 1.0000 | | | | |
| IIB3 | 0.9384 | 1.0000 | | | |
| IIIA5 | 0.8773 | 0.9810 | 1.0000 | | |
| IC1 | 0.9826 | 0.9638 | 0.9296 | 1.0000 | |
| IA4 | 0.9717 | 0.8848 | 0.8308 | 0.9641 | 1.0000 |

to 1.) At a correlation coefficient value of 1, two spectra are identical. At a correlation coefficient value of 0, no linear relationship between the two spectra exists. Taken as a group, the values in Table 6.2 for the full spectra support the fact that the spectra measured with different spectrometers and ATR accessories produce different spectra. Even though the same sample has been measured in each case, several of the spectra pairs show correlation coefficients less than 0.95. These spectral differences cannot be attributed to inhomogeneity of the sample or human sampling error. The correlation coefficients for each pair of spectra within the replicate spectra from individual spectrometer/ATR combinations were calculated. The lowest correlation coefficient for any spectral pair within a replicate set was 0.9923. Some interesting facts can also be gleaned from looking at individual values in the table. The table reveals that the most similar spectrum to IC1 is IVC2, and the correlation coefficient for this pair is 0.9858. One expects this result because the two spectra were obtained with the same model and type of ATR accessory. One would also expect that the next most similar spectrum to either of these spectra would be IIB3 because IIB3, IC1, and IVC2 are all from diamond IREs; however, this is not the case. We see that IVC2 is more similar to IA4 than IIB3, and that IC1 is just as similar to IIB3 as to IA4. This is a puzzling result because IA4 is from a Si IRE, and one would expect spectrum IA4 to be the most dissimilar to the spectra from the diamond IREs. This result further emphasizes that the spectral differences due to spectrometers and ATR accessories cannot be ignored when spectral comparison among spectra from different spectrometers and accessories takes place.

If one considers the correlation coefficients for just the spectral region from 1800 to 650 cm^{-1} in Table 6.2, the same trends will be seen; however, the spectral differences in this region are greater, in general, than when the entire spectra are considered. This result is expected because the differences in depth of penetration should be largest at the low wavenumber end of the spectra. In one case, the correlation coefficient between IIIA5 and IA4 is 0.8308 in the region from 1800 to 650 cm^{-1} . This low value can partially be attributed to the water absorption bands present in spectrum IIIA5 and not in the other spectra. The design of ATR accessory A5 requires the inside of the accessory to be exposed to the laboratory atmosphere each time the sample is changed. This design makes proper purging of the ATR accessory difficult.

If one wanted to increase the similarity of the spectra measured with the use of different ATR accessories, the traditional thought would be to correct the spectra for the differences in depth of penetration. Table 6.3 shows the correlation coefficients for spectra that were first corrected for depth of penetration and then absorbance scale normalized. The correlation coefficients for the full spectra show that the spectral similarity actually decreased, and in some cases the decrease was substantial, compared to the non-corrected spectra. For instance, the correlation coefficient between IIB3 and IA4 was 0.6990. If one examines the correlation coefficients for the spectral region from 1800 to 650 cm^{-1} in Table 6.3, the correlation coefficients are increased over the values reported for the same region in Table 6.2. This paradox is explained by the nature of the depth of penetration correction: the larger the wavenumber the larger the linear correction term that is applied. This effect exacerbates any upward sloping baselines

Table 6.3: Correlation coefficients for PETE spectra after depth of penetration correction.

| Full Spectra | | | | | |
|--------------|--------|--------|--------|--------|--------|
| | IVC2 | IIB3 | IIIA5 | IC1 | IA4 |
| IVC2 | 1.0000 | | | | |
| IIB3 | 0.9654 | 1.0000 | | | |
| IIIA5 | 0.9199 | 0.9592 | 1.0000 | | |
| IC1 | 0.8587 | 0.8048 | 0.8729 | 1.0000 | |
| IA4 | 0.8210 | 0.6990 | 0.7532 | 0.9480 | 1.0000 |

| 1800 to 650 cm ⁻¹ | | | | | |
|------------------------------|--------|--------|--------|--------|--------|
| | IVC2 | IIB3 | IIIA5 | IC1 | IA4 |
| IVC2 | 1.0000 | | | | |
| IIB3 | 0.9599 | 1.0000 | | | |
| IIIA5 | 0.9228 | 0.9819 | 1.0000 | | |
| IC1 | 0.9856 | 0.9633 | 0.9503 | 1.0000 | |
| IA4 | 0.9652 | 0.8749 | 0.8470 | 0.9592 | 1.0000 |

that might be present in the upper wavenumber region of the spectra being processed. Figure 6.4 shows this trend in the corrected spectra. Although this unwanted side effect occurs, the similarity of the spectra in the region below 1800 cm^{-1} generally increases because the differences in depth of penetration due to the different refractive indices of the IREs are greatest at lower wavenumber. The depth of penetration correction generally succeeds in making this region of the spectra more similar.

Because it was known that anomalous dispersion also contributes to the spectral differences observed, the ability of the proprietary “Advanced ATR Correction” to increase the spectral similarity was examined. Table 6.4 gives the correlation coefficients for the spectra that were corrected and then absorbance scale normalized, and Fig. 6.5 shows the corresponding spectra. For both the full spectra and the region from 1800 to 650 cm^{-1} , most of the correlation coefficients were worse than or similar to the correlation coefficients for the spectra subjected to only the depth of penetration correction. Because the algorithm is proprietary, the cause of the decreases in similarity cannot be thoroughly explored; however, the software’s documentation warns that the correction might not work well for spectra where the baseline is not flat. Figure 6.6 shows the original non-scale normalized spectra. It is clear that in this case the baselines are not flat. This is especially true in the lower wave number region of the spectra.

Because of the software’s warning concerning sloping baselines, performing manual baseline correction on the spectra before applying the “Advanced ATR

Figure 6.4: PETE spectra after depth of penetration correction: a) IA4, b) IC1, c) IVC2, d) IIIA5, and e) IIB3.

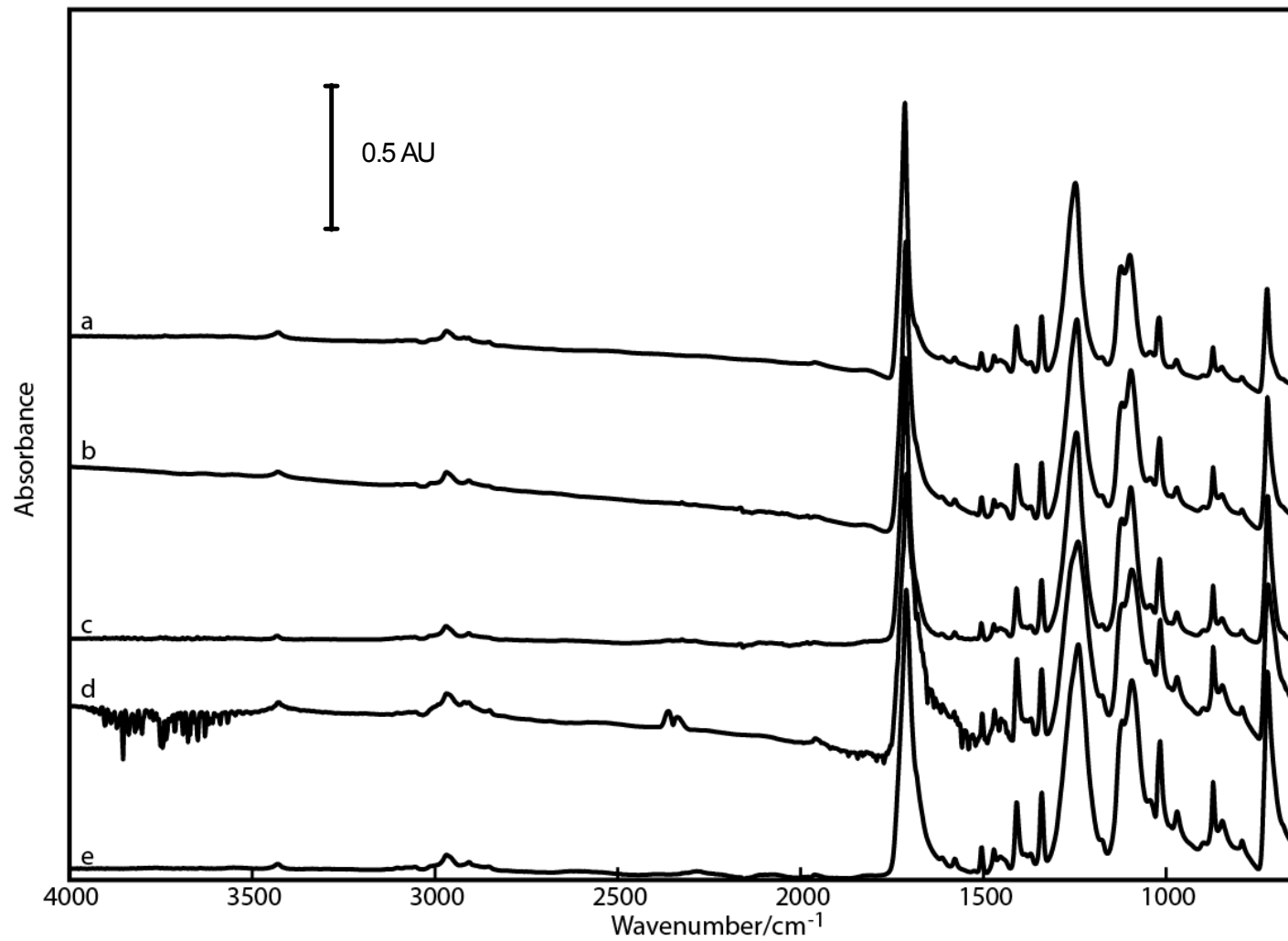


Table 6.4: Correlation coefficients of PETE spectra after “Advanced ATR Correction”.

| Full Spectra | | | | | |
|--------------|--------|--------|--------|--------|--------|
| | IVC2 | IIB3 | IIIA5 | IC1 | IA4 |
| IVC2 | 1.0000 | | | | |
| IIB3 | 0.8895 | 1.0000 | | | |
| IIIA5 | 0.8646 | 0.9594 | 1.0000 | | |
| IC1 | 0.8836 | 0.6869 | 0.7575 | 1.0000 | |
| IA4 | 0.8105 | 0.6361 | 0.7322 | 0.9754 | 1.0000 |

| 1800 to 650 cm ⁻¹ | | | | | |
|------------------------------|--------|--------|--------|--------|--------|
| | IVC2 | IIB3 | IIIA5 | IC1 | IA4 |
| IVC2 | 1.0000 | | | | |
| IIB3 | 0.8659 | 1.0000 | | | |
| IIIA5 | 0.8562 | 0.9847 | 1.0000 | | |
| IC1 | 0.9874 | 0.8099 | 0.8132 | 1.0000 | |
| IA4 | 0.9760 | 0.8492 | 0.8541 | 0.9854 | 1.0000 |

Figure 6.5: PETE spectra after “Advanced ATR Correction”: a) IA4, b) IC1, c) IVC2, d) IIIA5, and e) IIB3.

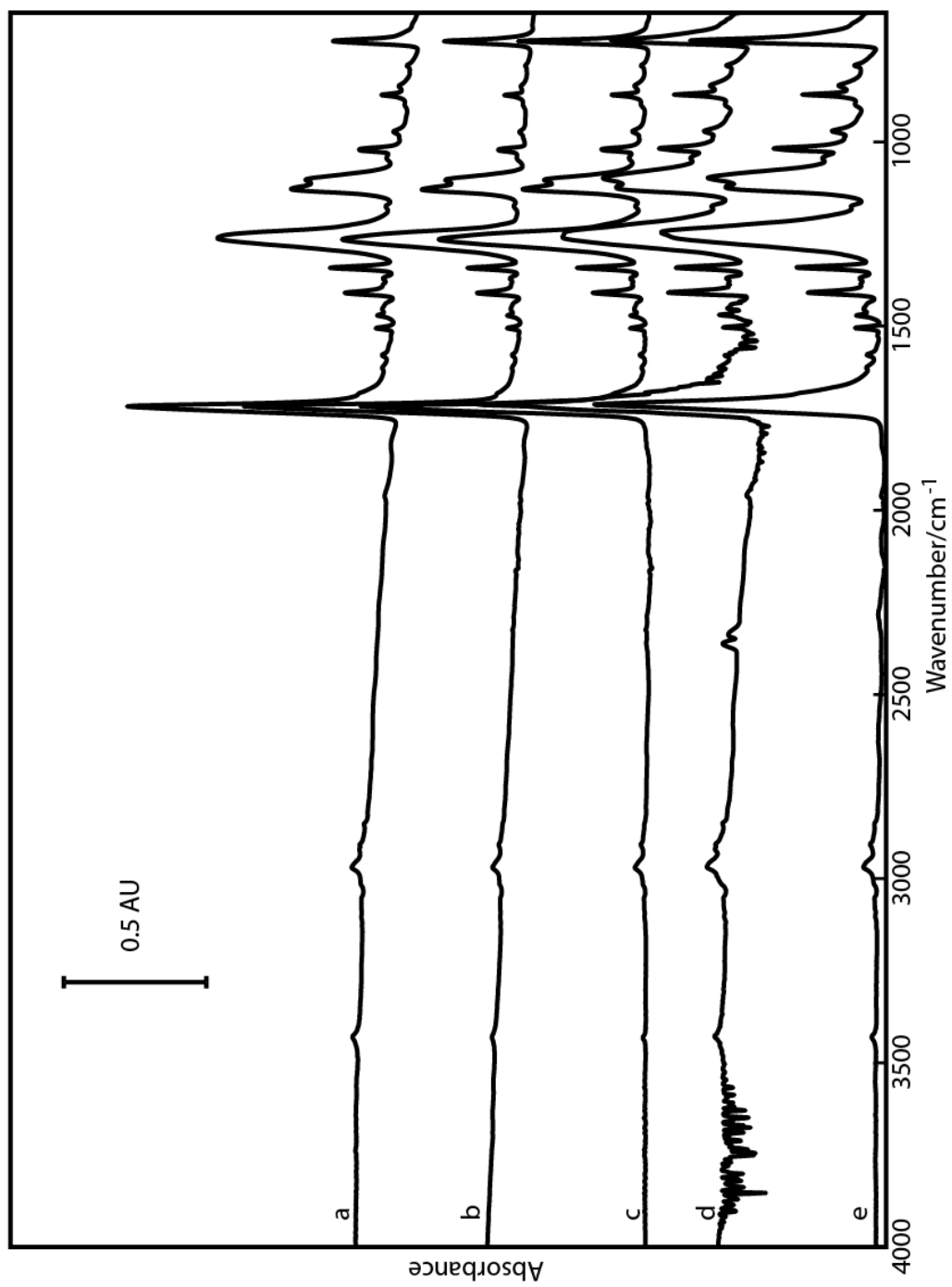
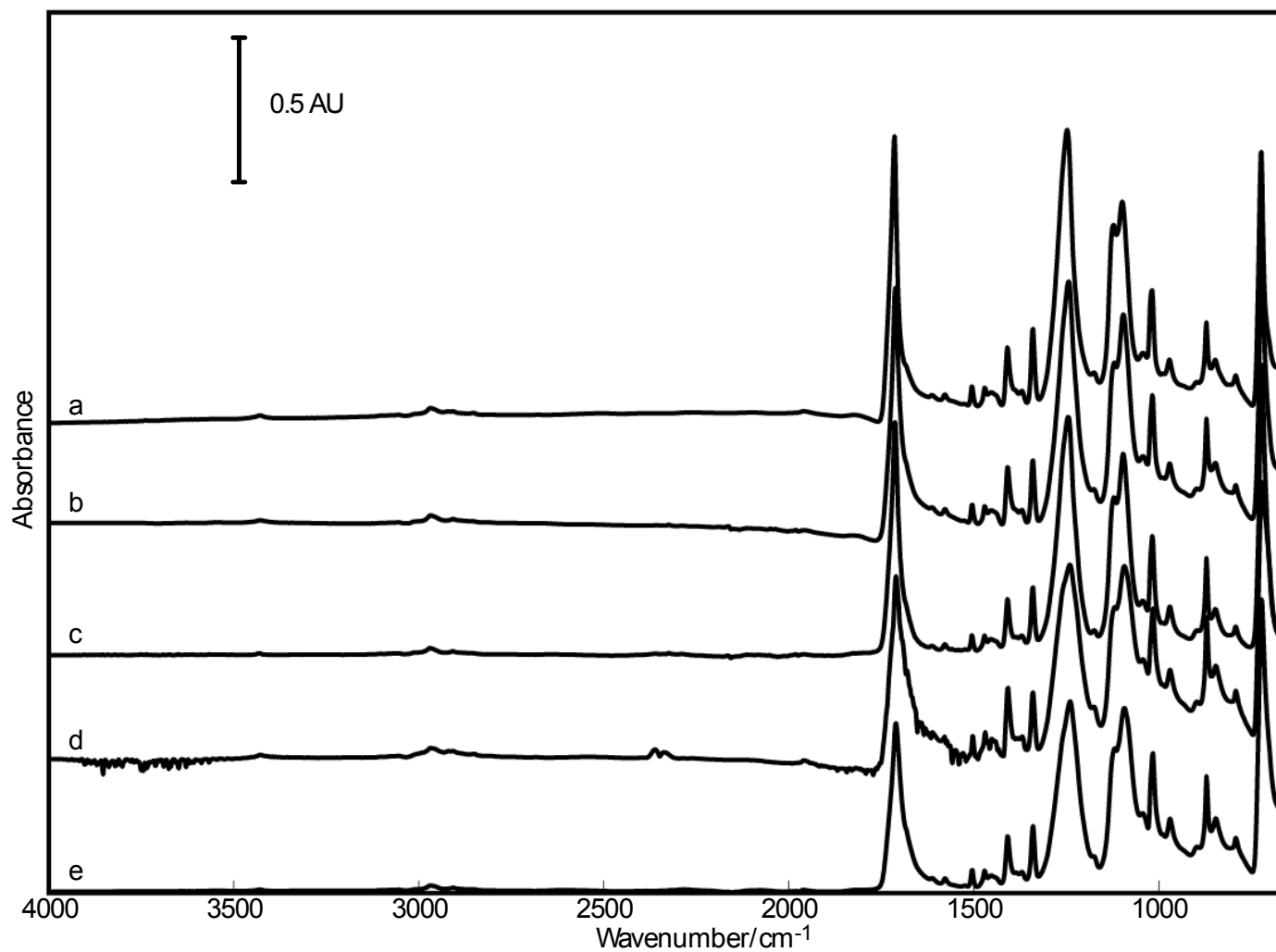


Figure 6.6: PETE spectra from 4000 to 650 cm^{-1} : a) IA4, b) IC1, c) IVC2, d) IIIA5, and e) IIB3.



correction” was tried to explore the effect on spectral similarity. Although this experiment was carried out, the validity of baseline corrections under these circumstances can be questioned². Any baseline correction is only a guess at the true baseline, and the corrected baseline may have no correspondence to the true baseline of a spectrum. The baseline correction may also introduce unwanted processing artifacts in the spectra. Before the “Advanced ATR Correction” was carried out, the region above 3500 cm⁻¹ in spectrum IIIA5 was replaced by a zero absorbance line to eliminate the effect of the water bands in that region of spectrum on the “Advanced ATR Correction”. Table 6.5 shows the correlation coefficients for the baseline corrected spectra after absorbance scale normalization. When compared with the correlation coefficients for non-baseline corrected spectra in Table 6.2, the correlation coefficients slightly decreased in many cases. This result supports the validity of the cautions issued above. Figure 6.7 shows the spectra with manually corrected baselines before the absorbance values were normalized. Table 6.6 shows the correlation coefficients for spectra that were manually baseline corrected, processed with the “Advanced ATR Correction”, and then absorbance scale normalized, and Fig. 6.8 shows the corresponding spectra. In comparison to the results in Table 6.3 for the spectra corrected for depth of penetration, the correlation coefficients for the full spectra are generally higher because the sloping baselines in the high wavenumber region have been corrected. When one considers the region from 1800 to 650 cm⁻¹, the results are mixed. In many cases, the similarities are worse after the “Advanced ATR Correction”

Table 6.5: Correlation coefficients for baseline corrected PETE spectra.

| Full Spectra | | | | | |
|--------------|--------|--------|--------|--------|--------|
| | IVC2 | IIB3 | IIIA5 | IC1 | IA4 |
| IVC2 | 1.0000 | | | | |
| IIB3 | 0.9547 | 1.0000 | | | |
| IIIA5 | 0.9162 | 0.9812 | 1.0000 | | |
| IC1 | 0.9860 | 0.9689 | 0.9559 | 1.0000 | |
| IA4 | 0.9750 | 0.9070 | 0.8814 | 0.9702 | 1.0000 |

| 1800 to 650 cm ⁻¹ | | | | | |
|------------------------------|--------|--------|--------|--------|--------|
| | IVC2 | IIB3 | IIIA5 | IC1 | IA4 |
| IVC2 | 1.0000 | | | | |
| IIB3 | 0.9384 | 1.0000 | | | |
| IIIA5 | 0.9070 | 0.9825 | 1.0000 | | |
| IC1 | 0.9866 | 0.9463 | 0.9359 | 1.0000 | |
| IA4 | 0.9635 | 0.8574 | 0.8329 | 0.9586 | 1.0000 |

Figure 6.7: Non-scale normalized PETE spectra after baseline correction: a) IA4, b) IC1,
c) IVC2, d) IIIA5, and e) IIB3.

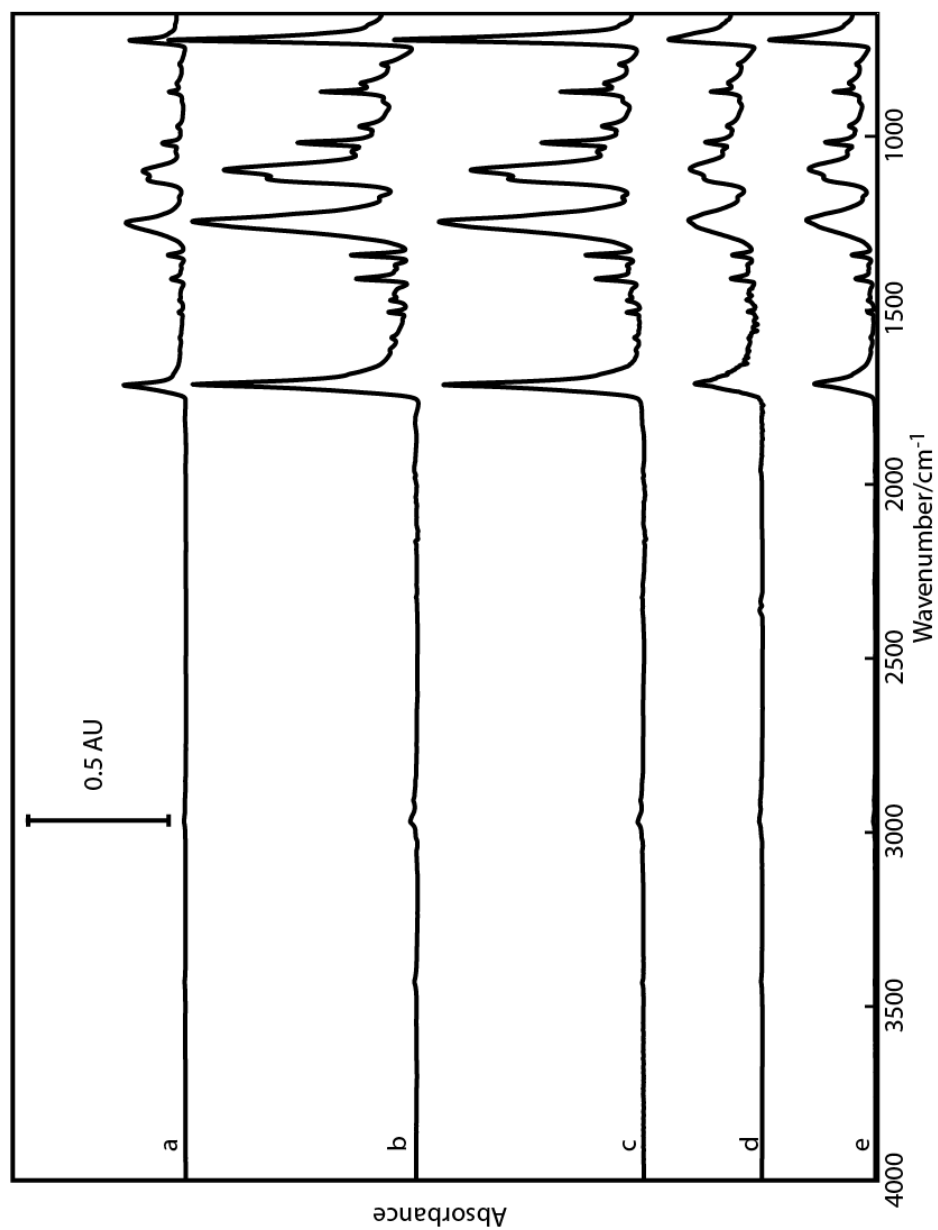
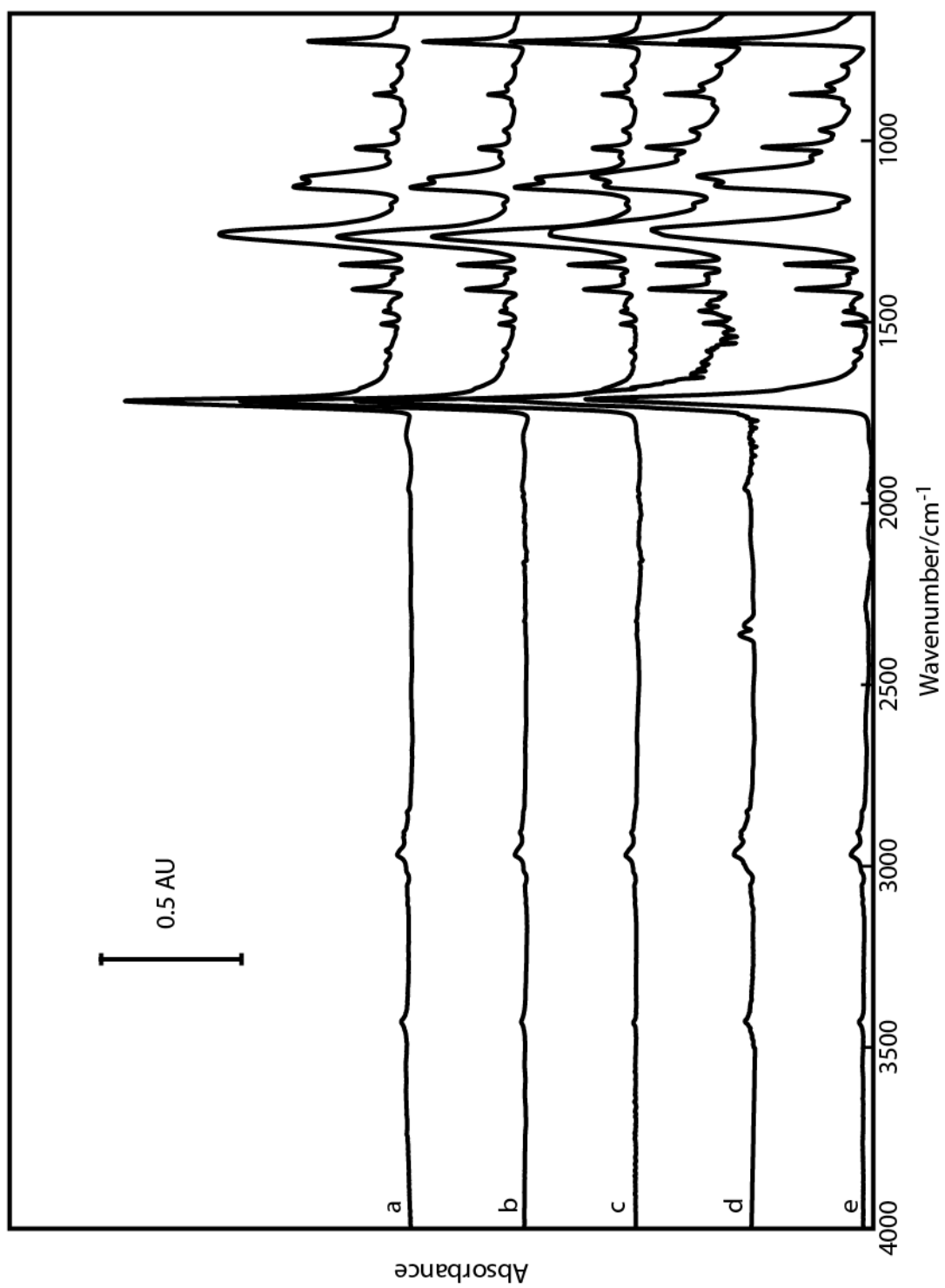


Table 6.6: Correlation coefficients for baseline corrected PETE spectra after “Advanced ATR Correction”.

| Full Spectra | | | | | |
|--------------|--------|--------|--------|--------|--------|
| | IVC2 | IIB3 | IIIA5 | IC1 | IA4 |
| IVC2 | 1.0000 | | | | |
| IIB3 | 0.8895 | 1.0000 | | | |
| IIIA5 | 0.8455 | 0.9638 | 1.0000 | | |
| IC1 | 0.9844 | 0.9038 | 0.8973 | 1.0000 | |
| IA4 | 0.9771 | 0.9312 | 0.9090 | 0.9898 | 1.0000 |

| 1800 to 650 cm ⁻¹ | | | | | |
|------------------------------|--------|--------|--------|--------|--------|
| | IVC2 | IIB3 | IIIA5 | IC1 | IA4 |
| IVC2 | 1.0000 | | | | |
| IIB3 | 0.8659 | 1.0000 | | | |
| IIIA5 | 0.8543 | 0.9775 | 1.0000 | | |
| IC1 | 0.9892 | 0.8650 | 0.8756 | 1.0000 | |
| IA4 | 0.9774 | 0.9047 | 0.9041 | 0.9877 | 1.0000 |

Figure 6.8: Baseline corrected PETE spectra after “Advanced ATR Correction”: a) IA4,
b) IC1, c) IVC2, d) IIIA5, and e) IIB3.



has been applied. However, the spectral similarity of IA4 (Si IRE) to the other spectra has increased, and it was expected that the “Advanced ATR Correction” should have the most significant effects on the spectral differences between IA4 and the other spectra. When the entire set of spectra is viewed together, the results indicate that under some circumstances the “Advanced ATR Correction”, with or without baseline correction preprocessing, will not lead to the most similar set of spectra.

In general, the experimental results taken on the whole show that for the lower wavenumber region of the spectra, where the significant absorption bands occur, the most similar spectra resulted from simply applying a depth of penetration correction, instead of also trying to correct for band shifting due to anomalous dispersion. These results demonstrate the potential danger of blindly applying correction methods before searching an unknown spectrum against a spectral library: the corrections could make the unknown spectrum less similar to the true matching library spectrum than the uncorrected unknown spectrum would be. The results reported also demonstrate that consideration of the dissimilarity of ATR spectra from different ATR accessories and spectrometers cannot be neglected when spectral comparison is performed.

CONCLUSION

In this work, the similarity of spectra of the same sample measured with different ATR accessories and spectrometers was explored. The results showed that the comparison of spectra measured with different ATR accessories and spectrometers even when the ATR accessories are of the same model is not straightforward. The analyst must consider the many factors that contribute to spectral variation, and spectral

correction should not blindly be applied. In the case of the PETE spectra being studied, it was found that simply applying a depth of penetration correction increased the similarity of many of the spectra, but that applying the proprietary “Advanced ATR Correction” that should correct for both depth of penetration and anomalous dispersion effects decreased the similarity of many of the spectra. These results do not suggest that the “Advanced ATR Correction” is an invalid correction method, but the results do reiterate the importance of not assuming that all spectral data can be handled in the same way. The overall conclusion of this study is that the analyst must use caution when comparing spectra obtained from different ATR accessories and when applying correction methods intended to correct the known effects that lead to differences among ATR spectra. This is especially true when one is comparing unknown spectra to a set of extremely similar spectra: the spectral differences due to spectrometer and ATR effects could be larger than the spectral differences due to changes in chemical composition.

REFERENCES

1. J. Grdadolnik, *Acta. Chim. Slov.* **49**, 631 (2002).
2. P. R. Griffiths and J. A. de Haseth, *Fourier Transform Infrared Spectrometry* (John Wiley & Sons, Hoboken, New Jersey, 2007), 2nd ed.
3. J.-J. Max and C. Chapados, *Appl. Spectrosc.* **53**, 1045 (1999).
4. K. Nishikida and K. D. Kempfert, "Advanced ATR Correction Algorithm for Infrared Spectroscopy", in *Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy*.
5. J. S. Plaskett and P. N. Schatz, *J. Chem. Phys.* **38**, 3, 612 (1963).

6. P. N. Schatz, S. Maeda, and K. Kozima, *J. Chem. Phys.* **38**, 11, 2658 (1963).
7. "OMNIC", (Thermo Electron Corporation, 2004).

CHAPTER 7

CONCLUSION AND FUTURE STUDIES

The five projects presented in this dissertation have both demonstrated novel analysis methods for cotton contamination and investigated the fundamentals of infrared spectral comparison. The accomplishments reported here came about through the combination of chemometrics with the inherent chemical information captured by infrared spectra. By exploring both spectrometry and chemometrics, greater advances in the understanding of spectral discrimination and identification have been made than would have been possible by focusing solely on the spectrometry or chemometrics alone.

Two projects in this work reported on improvements in qualitative spectral identification of cotton contaminants through the use of the USDA cotton contaminant library. By first exploring the performance of standard spectral comparison metrics for this library, new spectral library search algorithms capable of combining the information yielded by standard comparison metrics were developed. These voting scheme algorithms improved the identification of cotton contaminants by spectral library searching by making the process of choosing a single standard comparison metric unnecessary. The successful application of these voting scheme algorithms was seen for the case of a test set of contaminant spectra which were not represented by the growing seasons and locations or measured with the same instruments as those spectra

in the USDA cotton contaminant library. A second project demonstrated further improvements in spectral identification through the use of PLS-DA to differentiate the classes of contaminate spectra from one another better than was possible by using only the information from standard spectral comparison metrics.

One of the projects presented in this work successfully showed prediction of the percent composition of cotton contaminant powder mixtures. Although plant based cotton contaminants are extremely similar natural products, prediction of mixture composition through the use of PLS regression and an error correction algorithm was shown to be possible. This quantitative analysis project also provided the inspiration for the mixture design algorithm presented. Experimental design for mixtures is a complex and much studied topic, but this mixture generator algorithm fills the need for an easy method of generating sets of representative mixtures for any number of mixture components an experimentalist is interested in examining.

The last project presented in this work looked at some of the important implications for comparison of ATR spectra. The results of this project draw attention to the large differences that can exist among spectra of the same sample measured with the use of different spectrometers and ATR accessories. This project also demonstrated that accepted methods of correcting ATR spectra to look like transmission spectra may not always increase the similarity of ATR spectra measured with the use of different spectrometers and accessories.

In the study of cotton contamination, one of the important areas of future study that will build on the work presented will be online spectral monitoring. The projects

presented here have demonstrated that spectral discrimination of cotton contaminants is possible. The next step will be to develop spectrometers and spectrometer interfaces capable of online monitoring of cotton processing. For instance, online monitoring of cotton cleaning processes to determine the types and amounts of contaminants being removed by different cleaning machinery could improve the speed of the cleaning processes. The time savings achieved would increase profits by allowing more cotton to be processed in a shorter time and by reducing the damage done to cotton fibers by non-efficient cleaning processes that require repeated cleaning cycles to achieve the desired reduction in cotton debris levels. The realization of online contaminant monitoring will require the development of instruments, interfaces, and chemometric models that are robust to the changing samples and environmental conditions that would be experienced by the online instrumentation. Successful creation of suitable spectrometers and instruments will require careful planning and engineering to ensure that the spectrometers can obtain sufficient chemical information to allow for successful contaminant discrimination. Creation of robust chemometric models will require periodic updates of the models with new sample spectra to keep the models representative of the seasonal, varietal, and geographical changes in the chemical make-up of the cotton plants. As the projects in this work have demonstrated, effective methods of transferring chemometric models for contaminant discrimination from one spectrometer to another will also be necessary for wide spread implementation of online processing.

In summary, all of the projects included in this work should be seen as not only improving cotton contaminant identification, but as demonstrating methods of analysis for and exploring questions applicable to many types of complex samples. Many situations that deal with complex and highly similar feedstocks and products or both stand to benefit from the work presented here. Some examples include forensics, polymer production, and of course, analysis of the multitude of other natural products besides cotton. As the future work outlined above proceeds and the analysis of other complex samples takes place, the projects presented in this dissertation will provide a foundation of fundamental investigations upon which to build.