Wavelet basis selection for spectroscopic data analysis.

Thesis submitted by
David Andrew DONALD BSc(Hons)
in May 2012

for the degree of Doctor of Philosophy
in the School of Engineering and Physical Sciences
James Cook University

Statement of Access

I, the undersigned, author of this work, understand that James Cook University will make this thesis available for use within the University Library and, via the Australian Digital Theses network, for use elsewhere.

I understand that, as an unpublished work, a thesis has significant protection under the Copyright Act and;

I do not wish to place any further restriction on access to this work

_____ _____

David A Donald Date

Signed statement of sources

I declare that this thesis is my own work and has not been submitted in any other form for another degree or diploma at any university or other institution of tertiary education. Information derived from published or unpublished work of others has been acknowledged in the text and as a list of references

_____ _____

David A Donald Date

Statement on the contribution of others

Professor Danny Coomans and Dr. Yvette Everingham of the School of Mathematical and Physical Sciences, James Cook University, provided supervision, editorial assistance and imparted professional learning during these studies.

The School of Mathematical and Physical Sciences, James Cook University provided: a stipend via School of Mathematics and Physical Sciences Research Scholarship (2003-2004); a Teaching Scholarship; administrative support and; a School of Mathematical and Physical Sciences Travel Award used to attend the 12th International Conference of Near Infrared Spectroscopy in Auckland, 2005.

A travel award from the International Committee for Near Infrared Spectroscopy was awarded to attend the 12th International Conference of Near Infrared Spectroscopy, Auckland, 2005.

The Graduate Research School, James Cook University provided a stipend via a James Cook University Postgraduate Research Scholarship (JCUPRS) in 2004-2006 and professional development through graduate workshops; particularly the public speaking, negotiation skills, scientific writing and effective writing workshops. Additionally, the Graduate Research School awarded: a JCU Graduate Research International Travel Award to attend the 12th International Conference of Near Infrared Spectroscopy in Auckland, 2005; and a Doctoral Research Scheme grant used to attend the International Conference on Optimisation: Techniques and Applications in Ballarat, 2004, and visit the Australian Wine Research Institute in Adelaide and; Merit Research Grant used for computational support.

Data for Chapter 2 was provided by the Australian Wine Research Institute (AWRI) with the support of Dr. Daniel Cozzolino and Mark Gishen. The AWRI hosted a visit to the Adelaide research unit in December 2004 and assisted in professional development by inviting co-contributions in writing Grain Development Research Council milestone reports. Dr. Daniel Cozzolino assisted in the understanding and interpretations of the models development using the data provided by the AWRI. Both Dr. Cozzolino and

Mark Gishen provided editorial assistance with manuscripts and reports involving the data provided by the AWRI. Dr. Carl J. Schwarz, from the Department of Statistics and Actuarial Science, Simon Fraser University, Canada, assistance in the experimental design analysis in Chapter 2.

Chapter 3 included two data sets provided by Dr. Yvette Everingham. The seagrass data originated from Dr. Lem Aragones, previously from the Department of Zoology, James Cook University, and Dr. William Foley, from the Division of Botany and Zoology, Australian National University. The mineral data set originated from Dr. Danny Aswen, previously from the Earth Sciences Department, James Cook University.

Dr. Timothy Hancock, formerly a PhD at the School of Mathematics and Physical Sciences, JCU, co-authored the manuscript in Chapter 4 and contributed the variable selection methodology using the variable importance list generated by Random Forests. Dr. Christine Smyth, formerly a PhD candidate at the School of Mathematics and Physical Sciences, JCU, assisted Dr. Hancock in his contributions to Chapter 4. SELDI-TOF mass spectra data used in Chapter 4 was freely provided by the National Cancer Institute (of the United States of America) from their website.

Chapter 5 data originated from Brian Osborne, BRI Australia Limited, North Ryde, Australian. Code for the Metropolis search used in Chapter 5 was obtained from Professor Marina Vannucci, Department of Statistics, Rice University, Houston, Texas, USA, and subsequently modified for use in this thesis. Professor Wayne Reid, head of the School of Mathematics and Physical Sciences, provided critical review of Chapter 6.

Dr. Ian Atkinson, Dr. Wayne Mallett and Dr. Dominique Morel from the James Cook University High Performance Computing Centre provided computational support which was employed extensively for the wavelet optimisation and variable search algorithms.

Acknowledgements

I would firstly like to thank my supervisors, Danny and Yvette, for their support, encouragement and patience. They have imparted their skills and knowledge which has made me a professional researcher; which I can proudly say, has profoundly affected my career and personality in a positive way.

As the founding member of the Mathematics and Physics Students Club, I would like to thank the staff of the School of Mathematics and Physical Sciences, particularly Professor Wayne Reid, for their support for the club and encouraging young adults in their chosen academic fields. The free sausages were a bonus.

Finally a special acknowledgment to all of those who have continued to encourage me; particularly my mother Pauline and my wife, Mikayla.

Abstract

The discrete wavelet transform using adaptive wavelet bases were investigated in classification, regression and experimental design applications for spectroscopic data. Adaptive wavelets have been used previously in near infrared spectroscopy fields for classification and regression; however methods to select the parameters required in the adaptive wavelet algorithm have been largely influenced by human interaction. Methods are developed within this thesis to select parameters for adaptive wavelets along with investigating the hypothesis of using multiple wavelet bases to improve the predictability of classification and regression models.

Use of the adaptive discrete wavelet transform (ADWT) is illustrated using a repeated measures experiment. Near infrared (NIR) spectra of wine grape homogenates, from the Australian viticulture industry, underwent feature extraction via the ADWT and then modelled using penalised discriminate analysis, random forests and multiple adaptive regression splines. The correct classification rates of all three methods were substantially improved when the ADWT was applied. Scores from the ADWT penalised discriminate analysis (PDA) were analysed via multivariate analysis of variance (MANOVA) where it is reported that all main and interaction effects were significant. A bi-plot of the PDA scores illustrated the ease of which the ADWT extracted useful features from the spectra which were pertinent to the experimental design.

A method of ADWT parameter selection was derived using the Bayes' information criteria (BIC) and demonstrated in an unsupervised classification problem. Using the BIC to select ADWT parameters removed the need to for human interaction to select good, optimised, adaptive wavelets. This outcome highlighted an advantage over standard wavelet types, which gave similar unsupervised classification performances, where adaptive wavelets only need to span a relatively small set of parameters to give good models while a prohibitively large number of standard wavelet types need to be trialled.

Investigation of using multiple wavelet transforms to improve model performance - a new hypothesis in the field of chemometrics – was demonstrated in supervised classification and regression applications. In the classification example, SELDI-TOF mass spectra from a cancer study were analysed by pre-processing the spectra with a variety of standard wavelet types prior to variable elimination via a t-static and random forest approach. The retained variables were subsequently model using Treeboost where the specificity and sensitivity of the modelling process was improved by using multiple standard wavelet types compared to model using only one wavelet type alone. Models derived from wavelet processing were superior to models without pre-processing.

Further evidence supporting the multiple wavelet feature extraction hypothesis was gained in the regression application. Using a publically available and well documented NIR dataset, a Bayes Metropolis regression was modified to incorporate multiple wavelet transforms by using constrained stacking rather than Bayes model averaging as the model ensemble method. Multiple adaptive wavelets and multiple standard wavelets were trialled with the multiple adaptive wavelet approach resulting in a superior predictive regression model when compared to: all single standard wavelet models, single adaptive wavelet models, multiple wavelet standard wavelet models and models cited previously in literature for the same data set.

Methods for using adaptive wavelets, both multiple and singular wavelet bases, are outlined in this thesis with the general conclusion that the modelling process of NIR data (or juxta-positional data) can be substantially improved by the use of these wavelet transforms.

Table of Contents

Statement of Access ii
Signed statement of sources iii
Statement on the contribution of others iv
Acknowledgements vi
Abstract vii
Table of Contents ix
List of tables xii
List of figures xiii
Chapter 1 Introduction 1
1.1 Thesis outline 5
1.2 Chapter 2 5
1.3 Chapter 3 6
1.4 Chapter 4 6
1.5 Chapter 5 7
1.6 Chapter 6 7
1.7 Considerations for the NIR spectroscopy community 8
1.8 Publications resulting from thesis 10
Chapter 2 Adaptive Wavelet Modelling of a Nested 3 Factor Experimental Design in NIR Chemometrics 11
2.1 Introduction 11
2.2 Theory 13
2.2.1 Discrete wavelet transform 13
2.2.2 Penalized discriminate analysis (PDA) 15
2.2.3 Multiple adaptive regression splines (MARS) 15
2.2.4 Random Forests 17
2.3 Experimental 17
2.3.1 Data 17
2.3.2 Method 19
2.3.3 Software 20
2.4 Results and Discussion 20
2.5 Conclusions 24
2.6 Summary 26
Chapter 3 Integrated wavelet principal component mapping for unsupervised clustering on near infra-red spectra 27
3.1 Introduction 27
3.2 Theory 30
3.2.1 Principal component mapping (PCM) 30
3.2.2 Gaussian mixture models (GMM) 30
3.2.3 Wavelet transform 32
3.2.4 Adaptive wavelet matrix 35
3.3 Experimental 36
3.3.1 Data 36
3.3.2 Wavelet Principal Component Gaussian Mixture Model Mapping (WPG) 38
3.3.3 Wavelet packet transform 38
3.3.4 Principal component analysis 39
3.3.5 Gaussian mixture models 39
3.3.6 Overall WPG model selection 40

3.3.7 Adaptive wavelet optimization criterion 40
3.3.8 Software 42
3.4 Results and Discussion 42
3.4.1 Seagrass Data 42
3.4.2 Mineral Data 43
3.5 Conclusion 51
3.6 Summary 52
Chapter 4 Bagged Super Wavelts Reduction for Boosted Prostate Cancer Classification of SELDI-TOF Mass Spectral Serum Profiles 53
4.1 Introduction 53
4.2 Theory 55
4.2.1 Discrete Wavelet Transforms (DWT) – Super Wavelets 55
4.2.2 Classification and Regression Trees (CART) 56
4.2.3 Random Forests 56
4.2.4 Stochastic Gradient Boosting for CART (Treeboost) 57
4.2.5 Tree based methods for variable importance 58
4.3 Experimental 58
4.3.1 Data 58
4.3.2 Method 59
4.3.3 Benchmarking 61
4.4 Results and Discussion 62
Mean Decrease in Accuracy 63
4.5 Conclusion 66
4.6 Summary 67
Chapter 5 Joint Multiple Adaptive Wavelet Regression Ensembles 68
5.1 Introduction 68
5.2 Theory 73
5.2.1 Discrete Wavelet Transform (DWT) 73
5.2.2 Adaptive Wavelet (AW) matrix 74
5.2.3 Multivariate regression model 76
5.2.4 Variable selection 77
5.2.5 Posterior distribution of γ 78
5.2.6 Metropolis search 79
5.2.7 Stacking ensembles 80
5.3 Methodology 81
5.3.1 Near infrared spectra data 83
5.3.2 Parameter settings 84
5.3.2.1 Adaptive wavelet parameters 84
5.3.2.2 Multivariate regression model settings 84
5.3.2.3 Metropolis search settings 85
5.3.3 Computation 86
5.3.4 Analysis by previous methods 86
5.4 Results and Discussion 87
5.5 Conclusion 92
5.6 Summary 94
Chapter 6 Binomial Tree Factorization of the Matrix Polynomial Product with Shift Orthogonal Matrices 95
6.1 Introduction 95
6.2 Theory 95
6.3 Expansion of the multiple matrix polynomial product 97

6.4 Example 100
6.5 Conclusion 101
Chapter 7 Conclusion 102
7.1 Integration of adaptive wavelets 102
7.2 Adaptive wavelet optimisation criteria 105
7.3 Adaptive wavelet parameter selection 106
7.4 Multiple wavelets 109
7.5 Binomial tree algorithm for adaptive wavelets 111
7.6 Future considerations 111
Appendix 1 Beer-Lambert-Bouguer Law of Absorption 113
References 117

List of tables

Chapter 2 Adaptive Wavelet Modelling of a Nested 3 Factor Experimental Design in NIR Chemometrics 11
Table 2.1 Comparison of SNV and ANV ADWT NIRdata using PDA, MARS and RF analysis techniques 21
Table 2.2 Manova based on the PDA (1 to 4) scores from the adapted DWT. Box M statistic = 0.051, Bartletts test for sphericity statistic = 1.000. 21
Table 2.3 Manova partitioned mean squared error 21
Chapter 3 Integrated wavelet principal component mapping for unsupervised clustering on near infra-red spectra 27
Table 3.1 Parameterizations of the covariance matrix in the Gaussian model and their geometric interpretation 33
Table 3.2 Trialed standard wavelets 40
Table 3.3 Trialed values for *m, q* and *l* 40
Chapter 4 Bagged Super Wavelts Reduction for Boosted Prostate Cancer Classification of SELDI-TOF Mass Spectral Serum Profiles 53
Table 4.1 Benchmarking model performance using super wavelet 62
Table 4.2 Random Forests VIP list, cropped at the top 50 % of variables 63
Table 4.3 Benchmarking wavelet types using Random Forest performance 65
Table 4.4 Percentage false positive rates using the Random Forests on the super wavelet data. 65
Chapter 5 Joint Multiple Adaptive Wavelet Regression Ensembles 68
Table 5.1 Mean squared errors of the validation set using six calibration methods 86
Table 5.2 Re-sampled constrained stacking and Bayes model averaging (BMA) mean squared error of the validation data for each constituent using standard wavelets. 89
Table 5.3 Number of models and wavelet coefficients used in the ensembles where constrained stacking resulted in the lowest predictive MSE for each constituent. 89
Table 5.4 Re-sampled constrained stacking mean squared error of the validation data for each constituent using adaptive wavelets. 89

List of figures

Chapter 2 Adaptive Wavelet Modelling of a Nested 3 Factor Experimental Design in NIR Chemometrics 11
Figure 2.1 Nested three way design of the collected data where Variety, Storage and Homogenizer are crossed factors and the two levels of levels of replication occur within Variety and at the lowest level. Fixed effects and random effects are indicated in parenthesis as F and R respectively. 18
Figure 2.2 Sample NIR spectra of the red grape homogenates 18
Figure 2.3 Flow diagram of the adaptive DWT analysis 19
Figure 2.4 Biplot of the adapted DWT PDA 1 and PDA 2 of the combined treatments. Adapted DWT PDA 1 and PDA 2 spectra scores are represented by the scatterplot (corresponding to the bottom and left axes respectively) while the ray diagram represents the PDA 1 and PDA 2 wavelet coefficient loadings (corresponding to the top and right axes respectively). Legend: variety A - ♦, variety B - ● variety C –(▼), H1(red), H2(green), H3(blue), Frozen – solid marker, Fresh – open marker. The PDA 1 scores are represented 22
Figure 2.5 Biplot of the adapted DWT PDA 1 and PDA 3 of the combined treatments. Legend: variety A - ♦, variety B - ● variety C –(▼), H1(red), H2(green), H3(blue), Frozen – solid marker, Fresh – open marker. 23
Figure 2.6 Inverted DWT to the original NIR spectrum of the adapted DWT PDA axes. (a) PDA 1, (b) PDA 2, (c) PDA 3 25
Chapter 3 Integrated wavelet principal component mapping for unsupervised clustering on near infra-red spectra 27
Figure 3.1 Flow diagram of the proposed data mining and visualization method 28
Figure 3.2 Pictorial representation of a three band wavelet packet transform, with the discrete wavelet transform in the shaded region. With the original spectrum at the top of the pyramid, $\boldsymbol{x}^{[0]}(0)$, L the low pass filter, H_1 and H_2 the respective high pass filters 34
Figure 3.3 Sample of high pass wavelet filters (a) Daubechies 4 (b) Symmlet 7 (c) Daubechies 7 and (d) the Haar wavelet 34
Figure 3.4 Five sample spectra from each category from the Seagrass NIR data set 37
Figure 3.5 Five sample spectra from the five categories from the Mineral NIR data set 37
Figure 3.6 Seagrass adaptive WPG model scatter plot of the Bayesian information criteria (BIC) Vs classification uncertainty trimmed mean 43
Figure 3.7 Seagrass standard WPG model scatter plot of the Bayesian information criteria (BIC) Vs classification uncertainty trimmed mean 44
Figure 3.8 Adaptive WPG on the Seagrass data with adaptive wavelet parameters $m = 2$, $q = 3$, WPT band: $\boldsymbol{X}^{[1]}(8)$ 45
Figure 3.9 Standard WPG on the Seagrass data with wavelet parameters: Daubechies 2 filter on the WPT band $X^{[3]}(8)$ 45
Figure 3.10 Standard WPG on the Seagrass data with wavelet parameters: Daubechies 2 filter on the WPT band $\boldsymbol{X}^{[3]}(2)$ 46
Figure 3.11 Standard WPG on the Seagrass data with wavelet parameters: Daubechies 5 filter on the WPT band $\boldsymbol{X}^{[7]}(6)$ 46
Figure 3.12 Mineral standard WPG model scatter plot of the Bayesian information criteria (BIC) Vs classification uncertainty trimmed mean 47

Figure 3.13 Mineral adaptive WPG model scatter plot of the Bayesian information criteria (BIC) Vs classification uncertainty trimmed mean 48
Figure 3.14 Adaptive WPG on the Mineral data with adaptive wavelet parameters m = 2, q = 3, WPT band: $\boldsymbol{X}^{[1]}(8)$ 49
Figure 3.15 Optimal Gaussian mixture model on the third quadrant of Figure 3.14..... 49
Figure 3.16 Standard WPG on the Mineral data with adaptive wavelet parameters m = 2, q = 3, WPT band: $\boldsymbol{X}^{[1]}(8)$ 50
Figure 3.17 Standard WPG on the Mineral data with adaptive wavelet parameters m = 2, q = 3, WPT band: $\boldsymbol{X}^{[1]}(8)$ 50
Figure 3.18 Optimal Gaussian mixture model on the third quadrant of Figure 3.17..... 51
Chapter 4 Bagged Super Wavelts Reduction for Boosted Prostate Cancer Classification of SELDI-TOF Mass Spectral Serum Profiles 53
Figure 4.1 Examples of the different wavelet families: Daubechies 4 (a), Symlets 4 (b) and Coiflets 2 (c)..... 56
Figure 4.2 Flow diagram of the analysis..... 60
Figure 4.3 RF reduction training set CCR convergence 65
Figure 4.4 Inverse wavelet transform of the coefficients found in by Random Forests. 66
Chapter 5 Joint Multiple Adaptive Wavelet Regression Ensembles 68
Figure 5.1 Pictorial representation of a three banded (m = 3) discrete wavelet transform where the DWT has been applied twice to the original spectrum. 75
Figure 5.2 Number of wavelet coefficients in best 500 Bayes regression models generated by the Metropolis search using Coiflet 3, level 1 as the DWT 90
Figure 5.3 Constrained stacking ensemble weights for Coiflet (1) DWT level 4, (a) without resampling (b) with re-sampling..... 90
Figure 5.4 Constrained stacking ensemble weights for multiple adaptive wavelet combinations (a) without resampling (b) with resampling. Individual adaptive wavelet combinations (sets) corresponding to the rows in Table 5.4 are indicated in parenthesis 91
Figure 5.5 Adapted wavelets from different wavelet parameters used in the JAWRCS ensemble 91
Figure 5.6 Adaptive wavelet weighting resulting from two independent models within an ensemble using a similar region of the spectrum. An offset is added to one of the adapted wavelets for clarity 92
Chapter 6 Binomial Tree Factorization of the Matrix Polynomial Product with Shift Orthogonal Matrices 95
Figure 6.1 Binomial tree expansion of the projection matrices P_i used to construct the K_n matrices. 101

Chapter 1

Introduction

Study of near infrared absorption spectra is of interest for developing low cost, automated and rapid measurement systems. The near infrared (NIR) spectrum is the portion of the energy spectrum between 800nm to 2500nm where molecular dipoles absorb energy. Molecular dipoles absorb at characteristic wavelengths and the amount of absorbance relates to the concentration of the dipole.

The Beer-Lambert-Bouguer law (Appendix 1) is the most widely adopted theoretical framework to correlate molecular concentration with spectral absorbance and is particularly useful when samples have few absorbing dipoles. With a sample with few absorbing dipoles at different wavelengths, absorbance is directly proportional to concentration. However in samples that comprise of a large number of absorbing dipoles there is, as yet, no consistent theoretical framework that can be universally applied. With samples with many absorbing diploes, the measured NIR spectrum is a convolution of many NIR absorbance spectra. To overcome this obstacle, empirical methods have been developed to determine molecular concentration based upon the measured near infrared absorbance spectrum.

Projection based calibration methods such as partial least squares (PLS) [1] and principle component analysis regression (PCR) [2] have widely been used in NIR spectroscopy with considerable success to empirically correlate NIR absorbance with molecular concentrations. The idea behind projection based methods is that the NIR spectrum can be decomposed into a multitude of orthogonal spaces which can be correlated with the desired molecular concentration.

While projection based methods have been quite successful in forming empirical relationships between NIR spectra and molecular concentrations, projection methods do not utilize the physical characteristics of the NIR spectrum; particularly the juxta-positional nature of wavelengths. For example, wavelengths (or wavenumbers) can be re-ordered randomly and PLS will result in an identical model – with re-ordered PLS loadings naturally. Empirical models derived solely from projection based method can

be sensitive to the conditions in which the calibration data were collected [3]. Variants of PLS have been developed which do incorporate aspects of juxta-positioning. The most popular variant of PLS is moving window PLS (MWPLS) [4], where the spectrum is "windowed" in smaller regions. The windowing procedure incorporates some juxta-positioning information; however the portion of the spectrum within each window can still be randomly permutated to achieve the same result. Feature extraction, or signal filtering, is often used with PLS or PCR [5] to improve predictive performance as the feature extraction step incorporates physical information regarding the molecular dipole(s) spectrum.

With signal filter extraction methods, the spectrum (observed signal) is thought to consist of a superposition of underlying signals, where the signals can be characterised by a known functional form. For example, in Fourier analysis, the signals functional form is given by the sine function combined with a phase delay. Signal filters can be categorised into two classes: global and localised filters.

Fourier transforms are a classic example of a global filter where the basis function of the filter spans over the entire space of the observed signal. The Discrete Wavelet Transform (DWT) [6] and the Gabor Transform [7] are examples of localised signal filters, where the filter basis functions are localized to a small region of the observed spectrum. Most spectra consist of a superposition of overlapping signals and the desired signal, in regression applications, is widely believed to be restricted to a portion of the measured signal. With this overlapping structure, localised signal filters are ideal for feature extraction to improve modelling of spectra.

The discrete wavelet transform (DWT) has a similar structure as the spectrum superposition idea, where the DWT represents the spectrum as a superposition of scalable, localised functions. The DWT has been shown to be highly effective in improving the performance of calibration type problems in many fields of NIR spectroscopy [5]. However, unlike the Fourier transform, the DWT has a large number of basis functions to choose from and it has been demonstrated that some wavelets, used in the DWT, perform better than others in specific applications [8].
Most studies to date utilise wavelet transforms that use a mathematically derived wavelet such as a Daubechies or Morlet wavelet. These standard wavelet types have

been very successful in improving model performance particularly in the field of calibration development [9]. While Morlet and Daubechies wavelets have convenient mathematical properties, such as minimal phase distortion or maximum symmetry, they were not designed for unknown signal feature extraction for data analysis. Thus, it is more likely that a different wavelet basis, one derived for the task at hand, will more likely yield more a favourable model.

Wavelets in the DWT are functions that are fore mostly scalable and localised [7]. This criterion encompasses a broad range of functions that can be classified as wavelets. It is also possible to generate functions that fulfil the wavelet criteria. Pollen factorisation [10], Lifting [11] and Angular factorisation [12] are the most common algorithms to generate functions that meet the wavelet criteria. Additional criteria can be imposed in these wavelet generating algorithms to design wavelets specific to data analysis tasks – so called adaptive wavelets.

Adaptive wavelets are a class of wavelets which update their function frequency and phase forms to reduce a predefined optimisation criterion. The application of adaptive wavelets is quite limited in field of chemometrics with very few articles in literature [8, 12, 13]. Although the application of adaptive wavelets in literature is limited within the chemometrics field, the chemometric studies on adaptive wavelets have all indicated that adaptive wavelets are superior to standard wavelets. Nearly all of the adaptive wavelet applications in chemometrics have been on regression development [12] with only two papers on classification [8, 14].

Slow adoption of adaptive wavelets can be partially attributed to a lack of integration of adaptive wavelets into modern chemometric methods such as principle component analysis (PCA) and partial least squares (PLS). Standard wavelets have been used as a feature extraction tool for both PCA and PLS [1] chemometric applications, so it is understandable that adaptive wavelets should also be able to integrate with PLS and PCA to obtain further gains in model development. Integration of adaptive wavelets into modern chemometric methods is a key issue of this thesis, in particular how to generate the correct adaptive wavelet.

To derive the correct adaptive wavelet there are three key issues to be addressed. Firstly are the optimisation criteria; second is implementation of the adaptive wavelet algorithm and lastly selection of adaptive wavelet parameters required in the wavelet generation algorithms.

Adaptive wavelets are largely dependant on the defined optimisation criteria [7] and definition of the optimisation criteria is entirely dependant on the modelling process under investigation. Chemometric modelling of NIR spectra can take many forms, but is generally one of the following four types: (1) unsupervised classification, (2) supervised classification, (3) analysis of experimental designs and (4) regression [15]. Each of these model types has different objectives and as such has different optimisation criteria. Development of the optimisation criteria for each of the model types is outlined in this thesis and is an important issue in generating the correct adaptive wavelet.

Adaptive wavelets have also been viewed as overly complicated and so have been criticized as an unnecessary complication in the modelling process [16]. While adaptive wavelets do have complicated mathematical properties, they are no more complicated than standard wavelet types. The algorithms that give rise to standard wavelets are in fact the same algorithms that are used to generate adaptive wavelets; the only difference being for standard wavelets, predefined constraints are used [7]. With this in mind, this thesis introduces an alternative adaptive wavelet algorithm based on the more familiar concept of binomial trees.

Apart from the optimisation criteria, algorithms used to generate adaptive wavelets also contain a set of parameters that need to be defined [17]. These parameters pertain to the number of banded wavelets used in the DWT and the localisation (width) of the wavelets. Values of these parameters essentially restrict what form the resulting wavelets can take. The larger the values the more flexible the wavelets become.

An additional key issue of this thesis is wavelet homogeneity. In all applications of the DWT to spectroscopy calibration problems, a single wavelet type is used in the feature extraction process. This assumes homogeneity of underlying signals across the breath of the spectrum. However, if the underlying signals are heterogeneous along the

spectrum, different wavelet basis at different parts of the spectrum may offer further advantages in feature extraction for model development. This then leads to the main purpose of this thesis being, how to choose which wavelets to use and where to apply them.

The key issues addressed in this thesis are:

1. Integration of adaptive wavelet features within modern data analysis techniques
2. Generation of adaptive wavelet optimisation criteria for the four main types of data modelling: experimental design analysis, unsupervised classification, supervised classification and regression.
3. Automate adaptive wavelet parameter selection
4. Investigate feature heterogeneity within in a spectrum by using multiple wavelets, both adaptive and standard wavelets and,
5. To generate adaptive wavelets using a simplified binomial tree algorithm.

1.1 Thesis outline

This thesis is composed of five chapters investigating the application of wavelets, both standard and adaptive, to chemometric problems. Chapters 2 to 5 focus on incorporating adaptive wavelets with modern chemometric methods and addressing the issues related to wavelet selection, while Chapter 6 introduces a new method to generate adaptive wavelets based on a binomial tree factorisation.

Chapter 2 investigates integration of adaptive wavelets to experimental design analysis using near infrared (NIR) spectra; Chapter 3 integrates adaptive wavelets with unsupervised classification and investigates automated parameter selection for adaptive wavelets; Chapter 4 investigates heterogeneity of wavelets in building supervised classification models and; Chapter 5 focuses on multiple adaptive wavelet basis functions for regression applications and ensemble methods for adaptive wavelet parameter selection.

1.2 Chapter 2

The aims of Chapter 2 are to (i) develop adaptive wavelet optimisation criteria for experimental designs and (ii) integrate adaptive wavelets with traditional projection based methods. Chapter 2 introduces the concept of using adaptive wavelets in a

repeated measures experiment. Using an adaptive discrete wavelet transform, the method initially extracts features from the spectra that correlate with the design of the experiment. The extracted features are then mapped onto a five-dimensional hyper-plane using penalized discriminate mapping (PDM) to form PDM scores which are analysed using a multivariate mixed model (MMM) to determine if the experimental design affects the NIR spectra.

1.3 Chapter 3

Chapter 3 aims to integrate adaptive wavelets with unsupervised classification and investigate automated parameter selection for adaptive wavelets. Chapter 3 investigates a new method of unsupervised cluster exploration and visualization for spectral datasets by integrating the wavelet transform, principal components and Gaussian mixture models. This method incorporates feature extraction with model selection where the Bayesian Information Criterion (BIC) and classification uncertainty performance criteria are used to guide an automated search of commonly available wavelets and adaptive wavelets. The effectiveness of the proposed method is demonstrated in elucidating and visualizing unsupervised clusters from near infrared (NIR) spectral datasets.

1.4 Chapter 4

Chapter 4 introduces a new concept applying different wavelet transforms to different regions within the spectrum for supervised classification. Data used in Chapter 4 is not NIR spectra but SELDI-TOF mass spectra. Mass spectra (MS) and NIR spectra have similar characteristics as the data are juxta-positional so the same hypothesised data framework applies.

Features are extracted from the mass spectrum using multiple standard wavelets and incorporate into CART to develop a supervised classification model. Chapter 4 investigates the hypothesis of feature heterogeneity within the spectrum and develops methodology to use features derived from multiple wavelets simultaneously in a CART model. The method is illustrated using the publicly available prostate SELDI-TOF MS data from the American National Cancer Institute (NCI).

1.5 Chapter 5

Chapter 5 extends and combines the multiple wavelet approach to regression applications. Multiple adaptive discrete wavelet transforms were applied to NIR spectroscopic data for a multiple regression problem for the purpose of investigating the hypothesis – does the use of different wavelets, at different points, within a NIR spectrum elucidate predictive capability of regression models. This furthers the natural framework of the spectrum as different molecules exhibit different NIR signatures at different locations of the spectrum

The aims of Chapter 5 are to (i) develop adaptive wavelet criteria for regression applications, (ii) further investigate the hypothesis of feature heterogeneity within the spectrum and, (iii) develop methodology to use multiple wavelet transforms for regression. Data used in Chapter 5 is a publically available dataset pertaining to biscuit dough where sample near infrared spectra were measured by a FOSS 5000 NIR instrument and laboratory measurements were made to determine the fat, flour, sugar and moisture content.

1.6 Chapter 6

Algorithms to generate adaptive wavelets, such as Lifting [11], Quadrature Mirror Filtering [7] and Pollen factorisation [10], are complex and difficult to implement. By investigating the Pollen factorisation method, a simplified algorithm based on a binomial tree factorisation is established. The binomial method is relatively simple to implement to produce a full range of adaptive wavelets.

1.7 Considerations for the NIR spectroscopy community

Methods and techniques discussed and developed in this thesis may initially be thought to be of a passing or isotoric academic interest. However, after being actively employed in the NIR chemometric community for the previous five years, presenting at international conferences regarding NIR spectroscopy and being invited to present in industrial committees on NIR applications, there remains many issues in the fifty year old plus field that remain to be resolved. Without question, the largest issue is, and will be for some decades, measurement sensitivity of the NIR spectrum. The issue of measurement sensitivity has resulted in a general impression in the scientific community that NIR spectroscopy is a black box magic!

Near infrared spectra lack the tightly focused peak definition that is observed in all other forms of spectroscopy such as infrared, visible, ultraviolet and x-ray. The spectra of agricultural products all look the same with broad flowing mounds for peaks. Measurement sensitivity is not simply a consequence of detector sensitivity, however it does help, but measurement sensitivity in the NIR spectrum is also a product of sample presentation.

NIR energy is extremely prone to absorption, scattering and emission, so when a sample of sufficient thickness is illuminated with NIR energy, vast numbers of interactions occur and "statistically blur out." This leaves the interesting phenomena of sample presentation invariance (or close to) and peak broadening. If an incredibly thin film of a material (solid or liquid) was presented to a NIR spectrophotometer that was capable of analysing each photon and whence that photon interacted with the sample, a spectrum of clearly defined peaks would be measured. As it happens this is exactly what occurs when the NIR spectrum of gases are measured. Sadly gas NIR spectroscopy is limited and analysis of solid and liquid samples is what matters.

Methods to integrate and analyse broad flowing peaks in NIR spectra from solid and liquid samples are required. Current methods, such as PLS, utilise large portions of the measured spectra (the water absorption bands are typically ignored in most practical applications) which are mathematically used to solve Eigen vector relationships between the spectra and a measured constituent. Loadings (or regression) coefficients

from this approach rarely impart any knowledge regarding the importance of particular wavelength regions with respect to the constituent(s). Conversely, feature extraction methods utilise relatively small portions of the spectrum so a direct interpretation can be made between the spectrum and the constituent(s). Feature extraction methods almost invariably result in more predictive models than the traditional counterparts.

Feature extraction methods, such as adaptive wavelets, offer a means to resolve measurement sensitivity by de-convoluting portions of interest in the spectrum. Wavelets are still an underutilised pre-processing method in the chemometrics community partly because it involves making more choices being which wavelet to use. The field is already a flood with pre-processing techniques and introducing another which involves more complexity invokes further choice headaches.

By presenting a method which: selects/generates an appropriate wavelet, determines the portion of the spectrum to use, reduces model uncertainty and ultimately improves future predictions, the chemometrics community will develop a wider view to feature extractions methods – of which there are very few.

The question of how practical this thesis will be to the scientific community can be answered thus: Feature extraction methods illuminate localised information within the NIR spectrum which would otherwise be misinterpreted due to a lack in measurement sensitivity.

1.8 Publications resulting from thesis

Chapters 2, 3, 4 and 5 have been published in the following manuscripts respectively:

1. David Donald, Danny Coomans, Yvette Everingham, Daniel Cozzolino, Mark Gishen and Tim Hancock (2006), *Adaptive wavelet modelling of a nested 3 factor experimental design in NIR chemometrics.* Chemometrics and Intelligent Laboratory Systems, 82 (1-2). pp. 122-129.

2. David Donald, Yvette Everingham and Danny Coomans (2005), *Integrated wavelet principal component mapping for unsupervised clustering on near infra-red spectra.* Chemometrics and Intelligent Laboratory Systems, 77 (1-2). pp. 32-42

3. David Donald, Tim Hancock, Danny Coomans and Yvette Everingham (2006), *Bagged super wavelets reduction for boosted prostate cancer classification of seldi-tof mass spectral serum profiles.* Chemometrics and Intelligent Laboratory Systems, 82 (1-2). pp. 2-7.

4. David Donald, Danny Coomans and Yvette Everingham (2011), *Joint multiple adaptive wavelet regression ensembles*, Chemometrics and Intelligent Laboratory Systems, 108 (2), pp. 133-141.

Additionally, sections of this thesis contributed to a book chapter:

Donald, D.A., Everingham, Y.L., McKinna, L.W., and Coomans, D. (2009) *Feature selection in the wavelet domain: adaptive wavelets.* In: Comprehensive Chemometrics: chemical and biochemical data analysis. Elsevier, Oxford, UK, pp. 647-679.

Chapter 2

Adaptive Wavelet Modelling of a Nested 3 Factor Experimental Design in NIR Chemometrics

2.1 Introduction

Near infrared (NIR) spectroscopy, being a relatively inexpensive means of data collection is enabling many industrialists and academics the opportunity to increase the experimental complexity of their research, which in turn results in more accurate and precise information of their area of interest. An example is the comparison of the generalized randomized block design (GRBD) with the randomized block design (RBD) [18], where the GRBD is a k replicated RBD (and thus cost k times as much). The GRBD offers the opportunity to measure the effects of pseudo blocking factors, thus forming more accurate effects corresponding to the (true) fixed effects. This is not possible with the RBD. So with decreased costs for replication with NIR, GRBD experiments are becoming increasing popular and as a result of this, increasing interest (and concern) is how the experimental design affects the NIR spectrum.

Traditional methods for analysing a GRBD are ANOVA or MAVOA; however, ANOVA/MANOVA methods are ill suited to highly correlated, high dimensional data such as NIR spectra. To overcome the issue of high dimensionality, the NIR spectra are projected onto a lower dimensional, less correlated space. This is most commonly done using either a PLS [19-22] or PCR [19, 21] kernel based approach or alternatively projection via PCA alone [23].

Since the experimental design is known, PLS on the experimental design matrix, ASCA [24] or LDA [22]; would be a more appropriate projection method since this would be in effect mapping the NIR spectra onto a MANOVA space (the space that best describes the treatment factors!). In addition, while the above methods address the issue of the high dimensionality, the corresponding concern of the high variable correlation is still evident.

To overcome this issue of high variable correlation while simultaneously reducing the dimensionality and correcting for experimental design, we can employ a variety of methods such as: covariance inflation; penalized discriminate analysis (PDA) [25, 26], selection of multiple variable subsets; random forests (RF) [27] or fitting simple piece wise regressions; multiple adaptive regression splines discriminate analysis (MARS-DA) [28]. It would seem as if the problem is solved. However, PDA, MARS and RF, can become insensitive in situations where the NIR spectrum is dominated by a small fraction of the experimental design, effectively masking the effects resulting from the remainder of the experiment.

One of the main reasons for this is the NIR spectrum is composed of complex convolutions of chemical signals spanning across multiple localized wavelengths. This type of localized interactions can be difficult to detect with the above methods which focus on detecting differences arising from linear combinations of all the wavelengths simultaneously. To improve the sensitivity of PDA, MARS and RF, we focus on the localised convolutions rather than the raw wavelengths. The discrete wavelet transform (DWT) can be used as a localised convolution filter, which can be used to approximate and extract features from a NIR spectrum and has been used as such in PCR and PLS NIR regression applications [1, 29].

The wavelet transform (WT) is a projection of the spectrum onto an orthogonal basis, called a wavelet basis. This is to say that the spectrum can be represented by a set of localised, orthogonal basis functions called wavelets [6]. In this the WT has a familiar origin with the Fourier transform (FT), whose orthogonal basis functions are the sine functions. However, the DWT has a larger amount of flexibility than the FT, in the sense that the WT has an infinite choice of basis functions (wavelets) to choose from. Thus we can choose a wavelet basis that will result in good approximations of the latent features within the spectrum.

In most NIR WT applications to date, the wavelet used is selected from one of eight standard types of wavelets [7] mainly as a matter of convenience [5, 9, 12, 30-32]. However, it is possible to develop wavelets specifically for a particular application. These application specific wavelets iteratively adapt themselves towards a user defined criteria and are generally termed adaptive wavelets [8, 13, 33, 34]. It has been

demonstrated in supervised settings that adaptive wavelets – ones characteristic to the modelling process, result in higher classification rates [8] and more accurate regression models [12].

In this chapter, NIR spectra from red grape homogenates collected as part of a three way cross GRBD experimental design will be modelled using PDA, MARS-DA and RF on both the NIR spectra and the adaptive discrete wavelet transform (DWT) NIR data. Following the modelling process, the WT PDA is analysed with MANOVA to assess which fixed effect processes from the GRBD affect the spectra.

2.2 Theory

2.2.1 Discrete wavelet transform

The discrete wavelet transform (DWT) [15] like the Fourier transform, can be used to reformulate a spectrum into an alternative "feature space", by mapping the spectrum onto an analyzing function. In Fourier analysis, the analyzing functions are the set of sine function (spectra are mapped onto "frequency space"), where as for the DWT wavelets are the analyzing functions (spectra are mapped onto a "wavelet space"). The DWT is given by:

$$x(t) = \sum_{j=1}^{l} \sum_{k=0}^{2^j} c_{j,k} \psi_{j,k} \tag{2.1}$$

where $\psi_{0,0}$ is the father wavelet, from which all the other wavelets $\psi_{j,k}$ are derived from, $x(t)$ is the spectrum and $c_{j,k}$ is the wavelet coefficient calculated by the inner product between $x(t)$ and $\psi_{j,k}$.

$$c_{j,k} = \left\langle x(t) \middle| \psi_{j,k} \right\rangle \tag{2.2}$$

Unlike Fourier analysis, there are many types of analysis functions (wavelets) that can be used for the DWT – each resulting in different wavelet coefficients (mapped features), where typical (standard) wavelets used are Daubechies Symlets Coiflets. Since we do not know which wavelets will result in the best feature extraction a priori

for classification, this chapter will use Pollen's adaptive wavelets [15, 17] to extract features.

An advantage of the Pollen adaptive wavelets, is that the wavelet can be parameterized into $q+1$ normalized vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$; where $q \in Z^+$ is a smoothness parameter for the resulting wavelet. This means that we can asses the "fitness" of the wavelet as a function of the normalized vectors, which can then be iteratively updated to achieve a high "fitness". In this study, we define the fitness as the ability to discriminant between the various homogenizers, varieties and storage combinations, and to achieve this; we introduce a fitness function based on the wavelet coefficients from the DWT and the experimental design.

The fitness function is defined as:

$$\mathrm{f}\left(u_1, \ldots u_1, v\right) = \sum_{i=1}^{\mathrm{R}} \mathrm{g_i} \tag{2.3}$$

where

$$\Sigma_w^{-1}\Sigma_B\beta_i = g_i\beta_i \tag{2.4}$$

$\Sigma_{\mathbf{w}}$ is the with group covariance matrix, $\Sigma_{\mathbf{B}}$ is the between groups covariance matrix, R is the effective rank of $\Sigma_w^{-1}\Sigma_B$ and, g_i and β_i are the eigen-values and vectors of $\Sigma_w^{-1}\Sigma_B$ respectively.

The Pollen adaptive wavelets can be summarized in the following steps:

(1) Define the integer values for *m* and *q*

(2) Initialize the normalized vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$

(3) Perform the DWT and evaluate the performance of the wavelet with Eqn. (2.3)

(4) Iteratively update $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$ until a converge criteria is met.

In this study, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$ are initially assigned elements from the uniform distribution, which in previous supervised studies as shown to converge based on

similar optimization criteria detailed in eqn (2.3) [8, 13]. For a comprehensive account of the theory of the Pollen Factorization, the reader is referred to [17].

2.2.2 Penalized discriminate analysis (PDA)

Penalized discriminate analysis [26] is an extension of Fisher's linear discriminant analysis (LDA) which aims to find linear combinations of the variables that best separate the G different groups within the dataset such that the between group variability is maximised as much as possible relative to the within group variability. Here, LDA assumes that the data are drawn from G groups with K dimensional mean vectors $M_j, j = 1 \ldots G$, common within group covariance Σ_w and proportions $\pi_1, \ldots \pi_G$ of the groups in the population. Specifically, LDA finds $\beta \in \Re^K$ with $\beta^T \Sigma_w \beta = 1$ such that $f = \sum_{j=1}^{G} \pi_j \left(\beta^T M_j - \beta^T \overline{M}\right)^2$ is maximised. Here $\overline{M} = \sum_j \pi_j M_j$ is the overall population mean vector. Maximising f is identical to maximising the ratio $g = \beta^T \Sigma_B \beta / \beta^T \Sigma_w \beta$ under the constraint $\beta^T \Sigma_w \beta = 1$. Differentiation leads to the eigensystem $\Sigma_w^{-1} \Sigma_B \beta = g\beta$. In this way we can see that the eigenvectors of $\Sigma_w^{-1} \Sigma_B$ lead to the discriminate space.

In many NIR spectra situations, Σ_w is near singular due to the high correlations between adjacent wavelengths (variables), thus the eigenvalues of $\Sigma_w^{-1} \Sigma_B$ cannot be computed. To overcome this near singularity, Σ_w is replaced with $\Sigma_w' = \Sigma_w + \Omega$, where Ω is a K by K matrix such that $\beta^T \Omega \beta$ is large for undesirable β. This Ω is the central idea in PDA, where Ω penalizes the $\beta's$. We refer the reader to [26] for a detailed description of Ω.

2.2.3 Multiple adaptive regression splines (MARS)

The idea behind the MARS [28, 35] strategy is that in different areas of the sample space, different variables may have a greater or lesser contribution to the response surface via different loci. In general, the number of variables contributing significantly along one locus to any one region of the response surface will be smaller than the total number of variables. The adaptive term in MARS refers to the ability of the algorithm to select the dominant variables in each of the subregions.

The underlying MARS model can be written as:

$$y_i = \sum_j \beta_{i,j} f_j\left(X_{i,j}\right) + \varepsilon \tag{2.5}$$

where the vector y is the response vector, f_j are the various (normalized) loci, $\beta_{i,j}$ are the loci coefficients, $X_{i,j}$ are the variables (wavelengths) that significantly contribute to y_i through the loci f_j, and ε is the error in the model. The set of basis functions is called the MARS function given by:

$$f_m = \sum_j f_j \tag{2.6}$$

NIR data, which are piecewise smooth, f_j are typically multivariate polynomial regression splines [36], and the X_j are selected by trialling all permutations for X_j in f_j order to minimize a lack-of-fit (LOF) criterion described by [35]:

$$LOF\left(f_M\right) = \frac{(1/N)\sum_{i=1}^{N}\left[y_i - \boldsymbol{\beta}_i f_M\left(x_i\right)\right]^2}{\left[1 - C\left(M\right)/N\right]^2} \tag{2.7}$$

where f_m is the MARS function, C(M) is a complexity penalty function and N is the number of observations (spectra).

For the *n* spectra, there will be *n* corresponding models given by Eqn. (2.5), were the *n* models share a common MARS function, f_m, but are allowed different coefficients $\beta_{i,j}$. We can then analyze the $\beta_{i,j}$'s using LDA to differentiate between the *G* groups within the sampled spectra [36].

2.2.4 Random Forests

Random forests for classification as defined by Breiman [27] is a collection of many classification trees, each built on a unique bootstrapped (both variables and observations) sample of the data. The specific example of a RF used by Breiman [27], implements randomly selected predictor variables at each node in the building of each tree included within the bootstrapping. Breiman called this routine Forest-RI. Forest-RI randomizes during the split selection of each tree. This randomness has the effect of building new trees with different structures, increasing the variety of relationships modeled within the forest (multiple trees) which in turn improves the overall predictive performance. The classifications are the determined by a count (majority vote) of the classifications from each tree within the forest.

This strategy of randomly selecting observations and sub-sets of variables for constructing trees has a significant role in NIR data as (a) the tree approach avoids the problems associated with high wavelength correlation and (b) localized regions within the spectrum can be identified rather than a single wavelength and (c) helps to mitigate the effects of over fitting that can occur in a single classification tree.

2.3 Experimental

2.3.1 Data

Data used in this study consists of 284 near infrared spectra of red grape homogenates, which are prepared from grapes using a combination of various common sample preparation procedures. The homogenates of three red grape varieties (A, B and C) were randomly partitioned into two batches which were subjugated to one of two types of short term storage (fresh and overnight freezing). Then the homogenates were randomly prepared using one of three types of homogenisers (H1, H2 and H3). The design of the data collection is illustrated in Figure 2.1. The variety plots for A and B were replicated five times, while the C variety plots were replicated twice. Further more, each homogenate was replicate four times at the homogenizer level.

Each homogenate was scanned in a FOSS *NIRSystems6500* instrument at 2nm increments from 400nm to 2500 nm. The spectra were then truncated to 400-2448nm (1024 sample wavelengths), transformed via the log(1/R) transform and then normalized via the SNV transform [37]. Figure 2.2 shows sample spectra of the red grape homogenates.

$$\frac{\left(\begin{array}{c} Variety(F) \\ | \\ Replicates(R) \end{array}\right) \times Storage(F) \times Homogenizer(F)}{| \\ Replicates(R)}$$

Figure 2.1 Nested three way design of the collected data where Variety, Storage and Homogenizer are crossed factors and the two levels of levels of replication occur within Variety and at the lowest level. Fixed effects and random effects are indicated in parenthesis as F and R respectively.

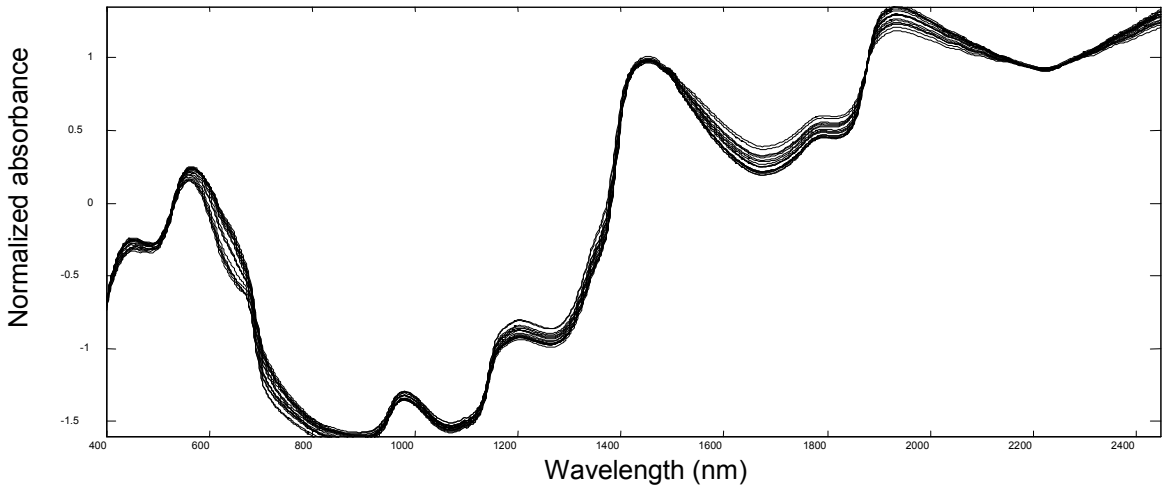

Figure 2.2 Sample NIR spectra of the red grape homogenates

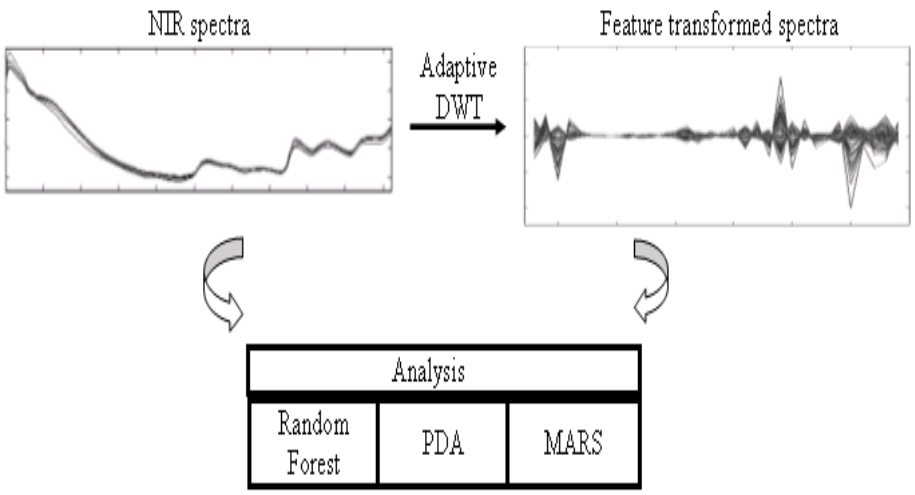

Figure 2.3 Flow diagram of the adaptive DWT analysis

2.3.2 Method

The experiment was carried out in a three step process.

1. Modeling the NIR spectra initially with Random forest (RF), PDA, and MARS-DA. Then apply RF, PDA and MARS-DA on the discrete wavelet transformed (DWT) NIR spectra using the adaptive wavelet, illustrated in Figure 2.3. For both the NIR and DWT analysis, we used the correct classification rate (CCR) as a measure of model performance. Throughout the modelling phase of the methodology, we focus on the effects of the fixed effects only.
2. Analysis of the GRBD in Figure 2.1, is performed using the adapted DWT PDA scores from (1) via a MANOVA testing for
    a. Main effects due to the fixed factors; Storage, Homogenization and Variety,
    b. Interactions between the main fixed effects,
    c. Main and interaction effects corresponding to the random effect of Variety replication.
3. Visualization of the Treatment (main and interaction) effects and their corresponding relationships to the adapted DWT coefficients are illustrated using biplots [24]. These effects (corresponding wavelet coefficients) are then mapped onto regions within the normalized NIR spectrum.

2.3.3 Software

The DWT was coded in *Matlab* [38] and the optimization function utilized for the adaptive wavelet is the unconstrained optimizer *fminu* function from the *Matlab Optimization Toolbox*® [39]. The Random Forest, PDA and MARS-DA were all generated in *R* using the modules; *randomForest* for random forest [40] and *mda* [41] for PDA and MARS. The MANOVA model was developed using the *manova* command in R [42].

2.4 Results and Discussion

Table 2.1 shows the correct classification rates for the NIR and DWT data using the PDA, MARS-DA and RF methods. Estimates for the dimensionality required for the PDA and MARS-DA models on both the adapted DWT and (SNV transformed) NIR data were taken from the effective rank of the correlation matrices, being three and four respectively.

The correct classification rates (CCR) for all three methods improved substantially when the wavelet coefficients from the adaptive DWT are analyzed rather than the original spectra. Various other Daubechies, Symlets and Coiflets wavelets were also trialed which resulted in higher CCR than the models on the (SNV transformed) NIR data, but did not outperform the adaptive wavelet.

From Table 2.1, the adaptive DWT PDA resulted in the highest CCR of 99.93%. During the MANOVA analysis of the adaptive DWT PDA, it was found that the random effect due to the Variety replication is not significant. This resulted in a simplification of the model which can be analyzed via a three factorial MANOVA.

The MANOVA model, shown in Table 2.2, on the adaptive DWT PDA revealed that all the main fixed effects, two way interactions are significant. By looking at the partitioned mean squared error (MSE) in Table 2.3, we can see that the main effects dominate the MSE for all the PDA axes (PDA1, PDA2,…,PDA4). From Table 2.3, PDA1 is largely dominated by the Variety main effect and to a lesser extent by the Homogenizer and Storage main effects. For PDA2, it is the main effects of both the

Homogenizer and Variety treatments that dominate the MSE. From Table 2.3, PDA3 is largely dominated by the Storage main effect.

Table 2.1 Comparison of SNV and ANV ADWT NIRdata using PDA, MARS and RF analysis techniques

Method	SNV treated NIR	SNV ADWT treated NIR
PDA	63.4 %	99.93%
MARS	58.6%	99.2%
RF	45.6%	76.4%

Table 2.2 Manova based on the PDA (1 to 4) scores from the adapted DWT. Box M statistic = 0.051, Bartletts test for sphericity statistic = 1.000.

Effect	Wilks' Lambda	F	Hypothesis df	Error df	Sig.
Intercept	.204	256.4	4.0	263.0	.000
Storage	.041	1550.9	4.0	263.0	.000
Homogenizer	.011	573.6	8.0	526.0	.000
Variety	.000	3656.4	8.0	526.0	.000
Storage * Homogenizer	.828	6.4	8.0	526.0	.000
Storage * Variety	.368	42.6	8.0	526.0	.000
Homogenizer * Variety	.566	10.2	16.0	804.1	.000
Storage * Variety * Homogenizer	.558	10.5	16.0	804.1	.000

Table 2.3 Manova partitioned mean squared error

PDA	**Main Effects**		
	Storage	Homogenizer	Variety
PDA1	1453.595	984.223	18137.039
PDA2	122.540	2049.847	1849.345
PDA3	4093.689	274.914	485.576
PDA4	604.739	790.966	957.461
	Two-way Interactions		
	Storage * Homogenizer	Storage * Variety	Variety * Homogenizer
PDA1	1.245	82.338	20.397
PDA2	0.704	2.754	5.729
PDA3	13.854	19.179	13.130
PDA4	10.638	77.014	3.789
	Three-way interaction		
	Storage * Variety * Homogenizer		
PDA1	18.804		
PDA2	14.692		
PDA3	5.495		
PDA4	7.281		

Biplots in Figure 2.4 and Figure 2.5 illustrate the groupings within the adaptive DWT PDA data and the relationships with the wavelet coefficients. Where in the biplots, the bottom and left axes represent the PDA scores (shown as a scatter plot), while the top and right axes are used for the PDA loadings (ray diagram of the wavelet coefficient loadings). The wavelet coefficients in the loadings plots directly relate to localized regions in the NIR spectra centered at: WC*8 + 400nm, where WC is the wavelet coefficient number.

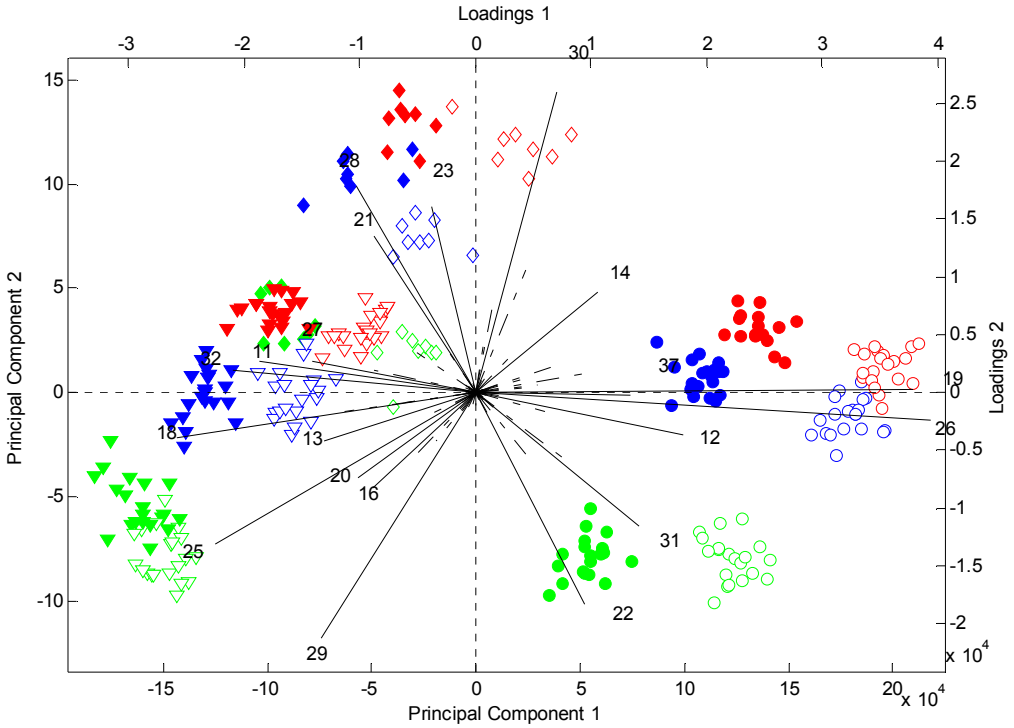

Figure 2.4 Biplot of the adapted DWT PDA 1 and PDA 2 of the combined treatments. Adapted DWT PDA 1 and PDA 2 spectra scores are represented by the scatterplot (corresponding to the bottom and left axes respectively) while the ray diagram represents the PDA 1 and PDA 2 wavelet coefficient loadings (corresponding to the top and right axes respectively). Legend: variety A - ♦, variety B - ● variety C –(▼), H1(red), H2(green), H3(blue), Frozen – solid marker, Fresh – open marker. The PDA 1 scores are represented

In the PDA1 and PDA2 biplot, Figure 2.4, there are very distinguishable groups which can be characterized by the variety/homogenizer/storage treatment combination. In Figure 2.5, the biplot of PDA1 ad PDA3, we can see that the frozen and fresh levels are separated by a downwards shift in the direction of PDA3.

Figure 2.6 shows the regions in the NIR spectrum that relate to the respective PDA axes and hence the different treatment effects. For PDA 1, which is dominated mostly by the Homogenizer treatment; we can identify four main regions: 750-810nm, 860-930nm, 980-1040nm and 1090-1140nm, that relate strongly to PDA 1. The regions that are related to PDA 2, and thus the Homogenizer and Variety main effects are: 850-860nm, 930-980nm and 1040-1085nm. For PDA 3, which is largely dominated by the Storage treatment, the NIR regions 850-980nm and 1040-1075nm were identified.

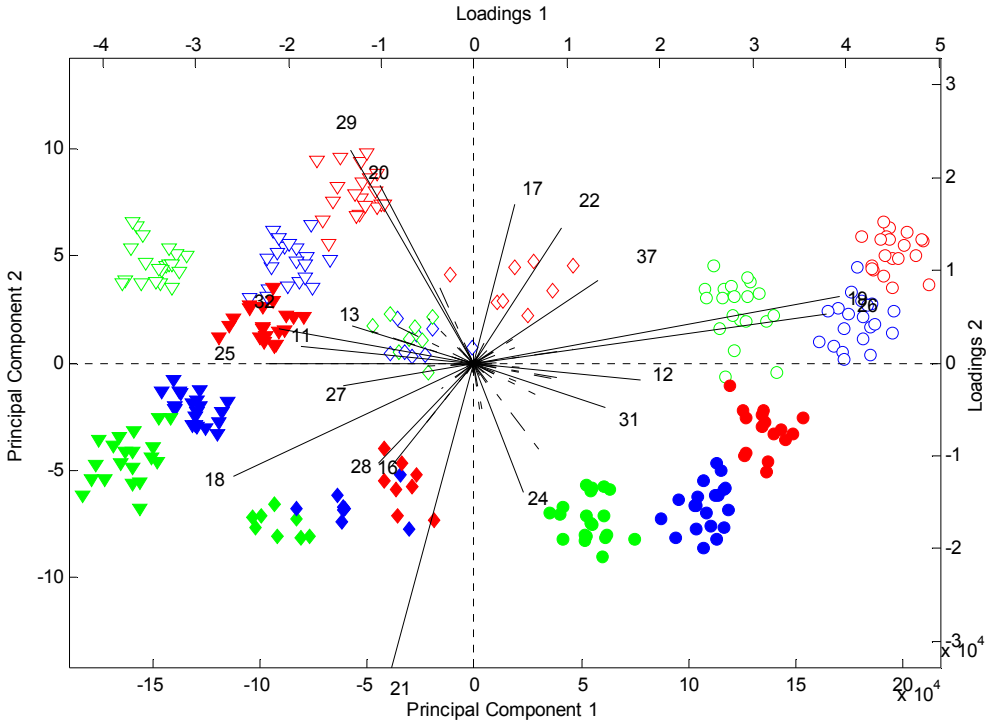

Figure 2.5 Biplot of the adapted DWT PDA 1 and PDA 3 of the combined treatments. Legend: variety A - ♦, variety B - ● variety C –(▼), H1(red), H2(green), H3(blue), Frozen – solid marker, Fresh – open marker.

The irregular appearance of the variable importance plot is due to two factors. Firstly PDA axes are typically differential over small regions and secondly, the adapted wavelet is also irregularly differential over localised regions (on the wavelength axis). The irregularity of Figure 2.6 is also compounded the auto-scaling used to obtain the relative importance scale – being the auto-scaling of the absolute value of the inverse transform of the wavelet PDA axis.

The region between 1080 and 1120 nm was likely to be an artefact created by noise in the spectra due to the change over of the detectors in the spectrophotometer, which occurs at a wavelength of 1098 nm. All the other regions affecting the PDA axes are generally attributed to OH overtones and combinations, which are most likely associated with water red grape homogenate. The treatments are therefore probably affecting the sample in a variety of ways that is manifested as changes in the interactions of water in the matrix, in particular, hydrogen bonding.

Homogenization may affect the degree of extraction of ionic species form the grapes, which in turn might affect the pH of the matrix which would be expected to affect the spectra in the region 750-860 nm. Storage might also have a similar impact. It is possible that the Variety affect observed was because of the differences in ripeness in the relatively few samples of grapes used to prepare the samples, since ripeness (i.e. sugar content) will also affect the OH absorptions in the grape spectra.

2.5 Conclusions

Using the wavelet coefficients from the adaptive discrete wavelet transform improved the correct classification rates for the random forest (RF), penalized discriminant analysis (PDA) and multiple adaptive regression splines discriminant analysis (MARS-DA) models, as compared to the models arising from the un-pre-processed NIR spectra. The best performing model was the PDA on the adaptive DWT which gave a 99.93% CCR. By analyzing the adaptive DWT DPA model with a MANOVA, we identified all main and interaction effects between the Homogenization, Variety and Storage effects as statistically significant. In analyzing the partitioned sums of squares of the MANOVA model, we were able to associate main treatment effects from the experimental design, Homogenization, Variety and Storage effects, to the respective discriminant axes from the PDA. By doing this, we were also able to identify specific regions from the spectrum that can be associated with the different treatment effects.

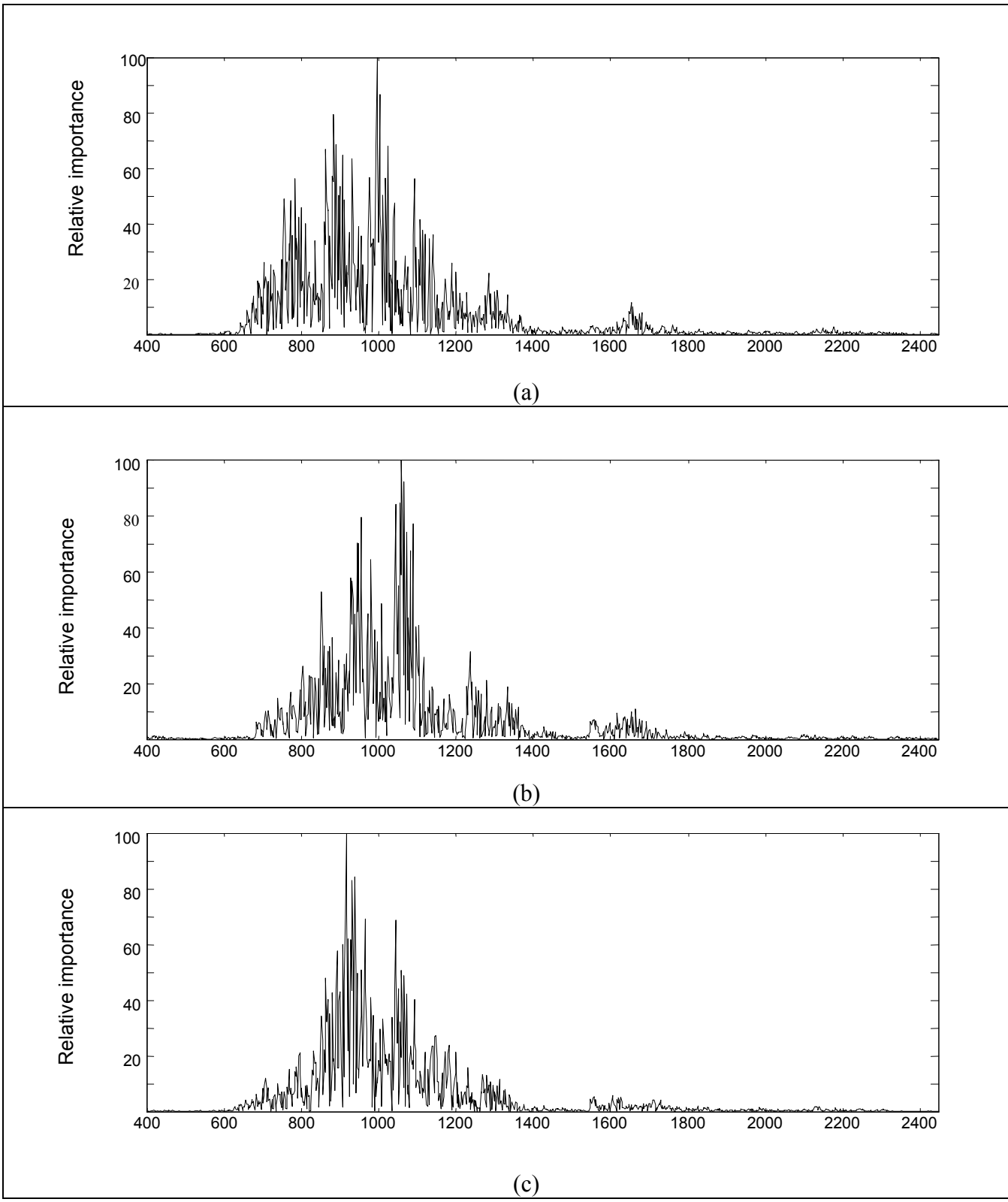

Figure 2.6 Inverted DWT to the original NIR spectrum of the adapted DWT PDA axes. (a) PDA 1, (b) PDA 2, (c) PDA 3

2.6 Summary

The objective of this study was to investigate the effects of some commonly used sample preparation procedures, including overnight freezing, and the type of homogeniser on the near-infrared (NIR) spectra of red grape homogenates. Homogenates (n = 284) of three red grape varieties were prepared using one of three types of homogenisers after one of two types of short term storage (fresh and overnight freezing) and then scanned in a FOSS *NIRSystems6500* instrument (400-2500 nm). The NIR spectral data were then analysed using various discrimination techniques, namely Penalized Discriminant Analysis (PDA), Multivariate Adaptive Regression Splines discriminant analysis (MARS-DA) and Random Forests (RF) yielding correct classification rates (CCR) of 63.4%, 58.6% and 45.6% respectively. To improve the CCR of the discrimination models, feature extraction from the NIR spectral data was performed using an adaptive discrete wavelet transformation (DWT). The DWT algorithm employs an adaptive wavelet basis function that maximizes the discrimination between the different combinations of homogenisers, storage and grape varieties. The results after adaptive DWT on the NIR spectra resulted in CCR's of 99.93%, 99.2% and 76.4% for PDA, MARS-DA and RF, respectively. Further analysis of adaptive DWT PDA via MANOVA revealed significant differences in the main and interaction effects of the three treatments, which were then associated with specific regions within the NIR spectrum.

Chapter 3

Integrated wavelet principal component mapping for unsupervised clustering on near infra-red spectra

3.1 Introduction

Calibration methods for near infrared spectroscopy (NIRS) such as partial least squares (PLS) and principal component regression (PCR) are often applied to NIR data sets based on the assumption that the spectra are uniformly homogeneous. This assumption of homogeneity is normally thought to be satisfied, especially when data has been collected in an experimental design such that the data is thought to be as homogenous as possible and as such tests for homogeneity are not typically performed. However, if unknown heterogeneities do exist then the resulting calibrations at best will be sub-optimal, or in more extreme circumstances be rendered unusable for future predictions. In this regard, the discovery of unknown heterogeneities within NIR calibration datasets provides a means of producing robust and accurate calibrations.

One type of data heterogeneity considered in this chapter is the existence of unknown Gaussian clusters, which can be investigated by using the unsupervised clustering method Gaussian mixture models (GMM). Gaussian mixture models assume that the data has been derived from several unknown Gaussian populations, which can be discovered through an automated selection process utalising the Bayes information criteria (BIC). There are three challenges associated applying GMM to NIR data being: (1) high dimensionality – typically there are more variables (wavelengths) than observations, (2) the variables are highly correlated which results in near singular covariance matrices [8, 14] and (3) visualization of the clusters are not easily seen in a two or three dimensional setting due to the high dimensionality. To overcome these challenges, the NIR spectra can be pre-conditioned via a feature selection procedure.

Recent works in NIR spectroscopy involving the discrete wavelet transform (DWT) have demonstrated redundant and superfluous information can be extract from the spectrum using the DWT reducing the dimensionality of the NIR dataset [15, 43]. In addition to reducing the dimensionality of the data, the DWT extracts large and small

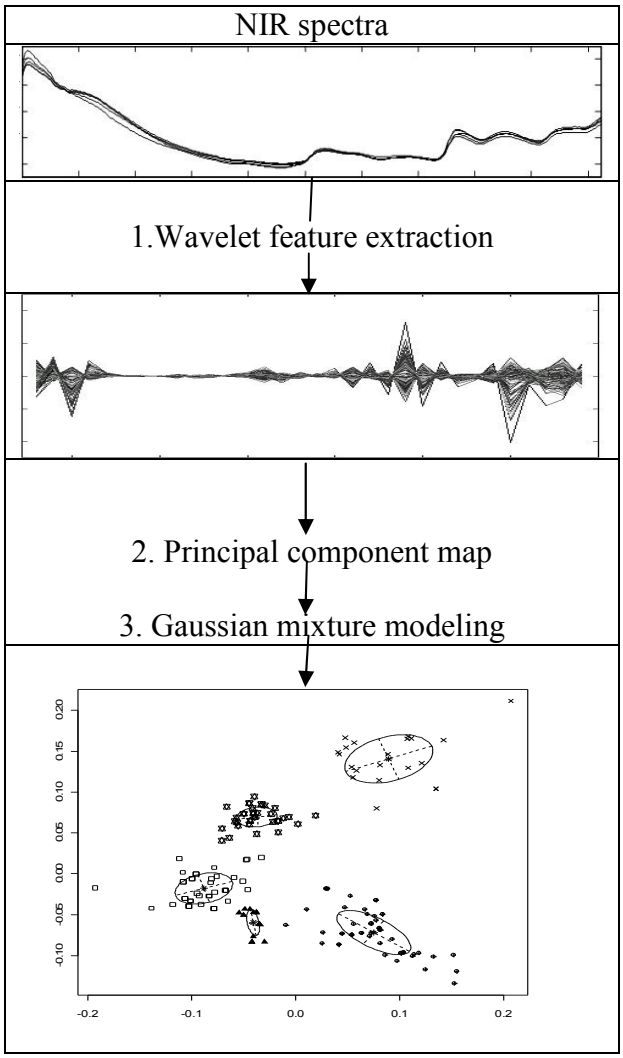

Figure 3.1 Flow diagram of the proposed data mining and visualization method

scale 'features' from the spectrum, which when analyzed, typically result in more accurate and predictive models [7, 44]. Inspired by this, we present a novel method for unsupervised clustering and visualization of NIR spectra, by integrating wavelet feature extraction, principal component mapping (PCM) and Gaussian mixture models (GMM); illustrated in Figure 3.1. The wavelet transform is used to extract features from the spectrum that can be visualized with PCM, and then analyzed with Gaussian mixture models for evidence of clusters.

In this chapter, we demonstrate the effectiveness of the proposed model on two NIR data sets and discuss possible complications of the proposed model – with accompanying solutions to these complications. Two main complications of the

proposed methodology are: (a) the choice of the wavelet to use in the DWT; there exist a multitude of wavelets, each designed to extract different features from the spectrum [7] and secondly, (b) which Gaussian mixture model to use. The latter is easily resolved by trialing a large range of possible mixture models of various numbers of candidate clusters and orientations [45-48], then using a suitable fitting criterion, such as the Bayes Information Criteria (BIC) [48-50], to select the most likely fit. The solution to the GMM problem (b), hints towards the solutions of the former problem of the wavelet choice.

The purpose of the wavelet in this instance is to extract features from the spectra that will result in group segmentation on a plane, since we are using PCM to visualize the groups. By extracting the desired features, we expect then to achieve GMM's with a high BIC and low model uncertainty [46, 47]. With this perspective, we can trial a large set of wavelets, automatically assess the GMM via the BIC and model uncertainty values, then chose to smaller subset of wavelet/GMM models for visual inspection.

The set of wavelets to be trialed raises another interesting question. In most DWT applications to date, the wavelet used is selected from one of eight standard types of wavelets [7] mainly as a matter of convenience [5, 9, 12, 30-32]. However, it is possible to develop wavelets specifically for a particular application. These application specific wavelets iteratively adapt themselves towards a user defined criteria and are generally termed adaptive wavelets [8, 13, 33, 34]. It has been demonstrated in previous settings that adaptive wavelets result in higher classification rates [8] and more accurate regression models [12] than the standard wavelets.

To address this issue, we put forward two variants of the proposed model. The first is to trial an exhaustive set of commonly used wavelets, a method which is done extensively in literature [51]. This translates to thousands of wavelet transformation trials. The second variant again uses an exhaustive search set, but using adaptive wavelets and in doing so, using adaptive wavelets in a new and novel context in an unsupervised scenario. Another favorable outcome in using adaptive wavelets is that the exhaustive search set is reduced to less than one hundred wavelets.

The chapter has been organized as follows. First we take a brief look at the theories of PCM, GMM, wavelets and then adaptive wavelets. In the experimental section, we introduce the NIR data sets then detail the proposed method, which includes a method of scrutinizing the vast sets of trialed models to produce a subset for further investigation by the researchers. Finally, a combined PCM/GMM plot of the respective data sets and model variants are presented.

3.2 Theory

3.2.1 Principal component mapping (PCM)

Principal component mapping is a projection method to visualize the variability in a dataset, which can lead to the discovery of unknown structures. In this study, we are interested in plane (2D) mappings and for demonstrative purposes only; we restrict the planes to be mapped to be derived from the first two principal components. The singular value decomposition (SVD); based on the covariance matrix, is used to extract the PCM:

$$\mathbf{Y}_{(k,n)} = \mathbf{U}_{(k,k)}\Lambda_{(k,k)}\mathbf{V}^{\mathbf{T}}_{(k,n)} \tag{3.1}$$

$$\mathbf{Y}_{(k,n)} = \mathbf{Q}_{(k,k)}\mathbf{V}^{\mathbf{T}}_{(k,n)} \tag{3.2}$$

In Eqn. (3.1), the row wise data matrix, $\mathbf{Y}_{(k,n)}$, is decomposed by the into SVD form and in Eqn. (3.2), the first two columns of the matrices $\mathbf{V}_{(k,n)}$ and $\mathbf{Q}_{(k,k)}$ are the desired PC's (principal component loadings) and the projected data points (principal component scores) respectively.

3.2.2 Gaussian mixture models (GMM)

Mixture models are useful tools for density estimation and as such are used extensively in cluster analysis applications [46, 47, 49]. The essential idea in the mixture models approach is that the dataset consists of ζ underlying probability distributions. In the case of Gaussian mixture models, the ζ probability distributions are Gaussian. Then the mixture model has the form:

$$f\left(x_j\right)=\sum_{i=1}^{\zeta}\delta_{i,j}N\left(\mu_i,\Sigma_i\right) \tag{3.3}$$

Where $\boldsymbol{N}\left(\mu_i,\Sigma_i\right)$ is the Gaussian distribution with a mean vector μ_i and covariance matrix Σ_i, $\delta_{i,j}$ is the delta function for the probability of the observation $\boldsymbol{x}_j$ belonging to the i^{th} Gaussian distribution. We refer to [46] for a more comprehensive account of GMM theory.

In situations where Eqn. (3.3) is unknown, ζ and $\boldsymbol{N}\left(\mu_i,\Sigma_i\right)$, i=1… ζ, need to be estimated from empirical data. This is done by the mixture likelihood approach that maximizes:

$$\boldsymbol{\mathcal{L}}_{\mathcal{M}}\left(\mu_1,\Sigma_1,\ldots\mu_\zeta,\Sigma_\zeta;\tau_1,\ldots,\tau_\zeta\right)=\prod_{j=1}^{n}\sum_{i=1}^{\zeta}\tau_{i,j}f\left(x_j\mid\mu_i,\Sigma_i\right) \tag{3.4}$$

where $\tau_{i,j}$ is the probability that the j^{th} observation belongs to the i^{th} Gaussian distribution and

$$f\left(x_j\mid\mu_i,\Sigma_i\right)=\frac{\exp\left\{-\frac{1}{2}\left(x_j-\mu_i\right)^T\Sigma_i^{-1}\left(x_j-\mu_i\right)\right\}}{\left(2\pi\right)^{p/2}\left|\Sigma_i\right|^{1/2}} \tag{3.5}$$

When the data are two dimensional, the clusters are ellipsoidal centered at the means μ_i while the covariances Σ_i determine other geometrical characteristics such direction and area. In calculating Eqn. (3.4), we consider the following parameterizations on Σ_i:

$$\Sigma_i=\eta_i\,\mathrm{D}_i\,\mathrm{A}_i\,\mathrm{D}_i^{\mathrm{T}} \tag{3.6}$$

where $\mathbf{D}_i$ is an orthogonal matrix containing the eigenvectors of Σ_i, A_i is a diagonal matrix whose elements are proportional to the eigenvalues of Σ_i and η_i is a scalar. The orientation of the principal components of Σ_i are determined by $\mathbf{D}_i$, while A_i determines the shape of the cluster, being either spherical of elliptical. The size, e.g.

area, the cluster is specified by η_i, which is proportional to $\eta_i^p|\mathrm{A}_i|$. Table 3.1 shows the geometric interpretations of the various parameterizations of Σ_i.

An advantage of the GMM approach is that it allows the use of approximate Bayes factors to compare models [46, 47, 49]. This gives a systematic means of selecting not only the parameterizations of the model, but also the number of clusters ζ. We refer the reader to [52] for a review and comprehensive theory of Bayes factors.

Essentially, the Bayes factor is the posterior odds for one model against the other(s) assuming neither is favored a priori. When using the mixture likelihood approach, the Bayes factor can be approximated by the Bayesian Information Criteria (BIC) [50]:

$$2\log \mathrm{p}\left(x|\mathcal{M}\right)+\text{const.} \approx 2\mathcal{L}_{\mathcal{M}}\left(\mu_1,\Sigma_1,\ldots\mu_\zeta,\Sigma_\zeta;\tau_1,\ldots,\tau_\zeta\right)-\mathrm{k}_M\log\left(n\right)\equiv \mathrm{BIC} \quad (3.7)$$

where $\mathrm{p}(\boldsymbol{x}\,|\,\mathcal{M})$ is the likelihood of the data for the model $\mathcal{M}$, $\mathcal{L}_{\mathcal{M}}(\mu_1,\Sigma_1,\ldots\mu_\zeta,\Sigma_\zeta;\tau_1,\ldots,\tau_\zeta)$ is the maximized mixture likelihood for the model from Eqn. (3.4) and k_M is the number of parameters to be estimated in the model.

The penalty term in Eqn. (3.7) is included since for mixture models, the likelihood for a mixture model can only increase with increasing k_M. Hence the likelihood cannot be used directly in comparing the various models. So the penalty term is included to mitigate this effect. Also this penalty term favors models with parsimonious parameterizations and smaller number of groups.

3.2.3 Wavelet transform

The wavelet transform (WT) enables the signal (spectrum) to be analyzed as a sum of functions (wavelets) with different spatial and frequency properties [7]. For discretely sampled spectra, several methods are available implement the WT [7]. The two most popular are the discrete wavelet transform (DWT) and the wavelet packet transform (WPT), shown in Figure 3.2.

Table 3.1 Parameterizations of the covariance matrix in the Gaussian model and their geometric interpretation

Distribution	Area:	Shape	Direction
Spherical	Equal	Equal	NA
Spherical	Variable	Equal	NA
Ellipsoidal	Equal	Equal	Equal
Ellipsoidal	Variable	Variable	Variable
Ellipsoidal	Equal	Equal	Variable
Ellipsoidal	Variable	Equal	Variable

In Figure 3.2, we see that for a discrete spectrum, the WPT and the DWT is an iterative algorithm that successively applies a series of filters on the data. These filters are called the low-pass filter, **L**, and the high-pass filter(s), **H**. The low-pass filter acts as a smoother and typically extracts low frequency information while the high-pass filter(s) are akin to difference operators; extracting high frequency information. Figure 3.3 illustrates some of the common high-pass filters (wavelets).

Two of the important properties of **L** and **H** are that they are orthogonal filters, and in the context of DWT and WPT, form a multiresolution framework [7]. This means that any combination of **L** and **H** will be orthogonal to any other different combination of **L** and **H**. This is an important result, since in Figure 3.2, we can see that the DWT is a "sub-set" of the WPT. Thus the features extracted from the un-shade bands in the WPT are unrepresented in the DWT. For this reason, we choose to work exclusively with the more flexible framework of the *m*-banded WPT. The remainder of this section, we describe how the *m*-band ($m \geq 2, m \in \mathbb{Z}$) wavelet packet transform (WPT) is calculated on discretely sampled signals of finite length. For a more comprehensive account of wavelet theory, the reader is referred to [7].

For the general *m*-band WPT, there will be one low-pass filter and *m*-1 high-pass filters. We refer to band (*l,t*) as the t^{th} band $(t \in 0,1,...,m^l)$ in level *l* of the WPT. The number of coefficients in each band will be 1/*m* of that in previous level so if *l* levels of the WPT are required, then the dimensionality of the data, *p*, should be $p = km^l$, $k \in \mathbb{N}$.

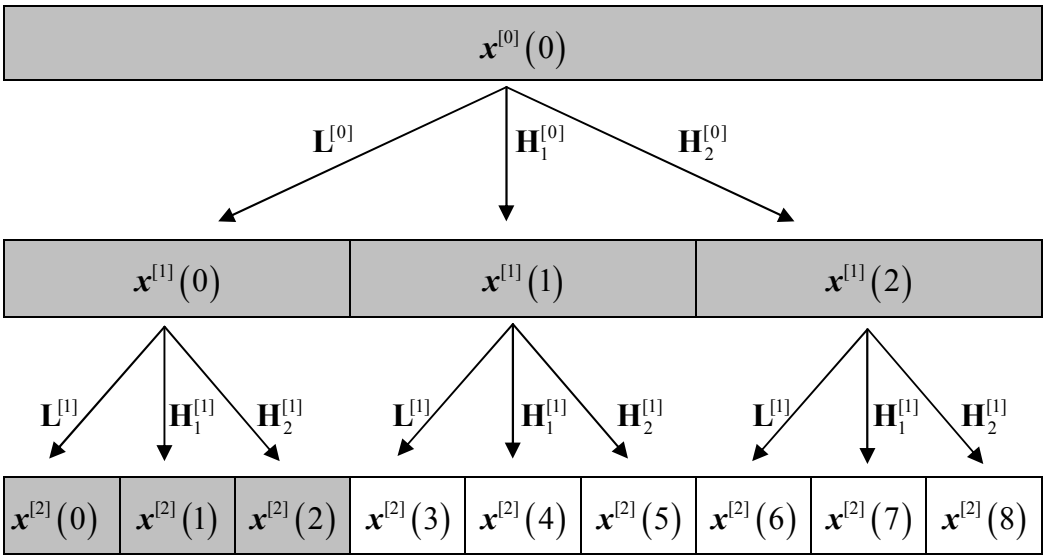

Figure 3.2 Pictorial representation of a three band wavelet packet transform, with the discrete wavelet transform in the shaded region. With the original spectrum at the top of the pyramid, $x^{[0]}(0)$, L the low pass filter, H_1 and H_2 the respective high pass filters

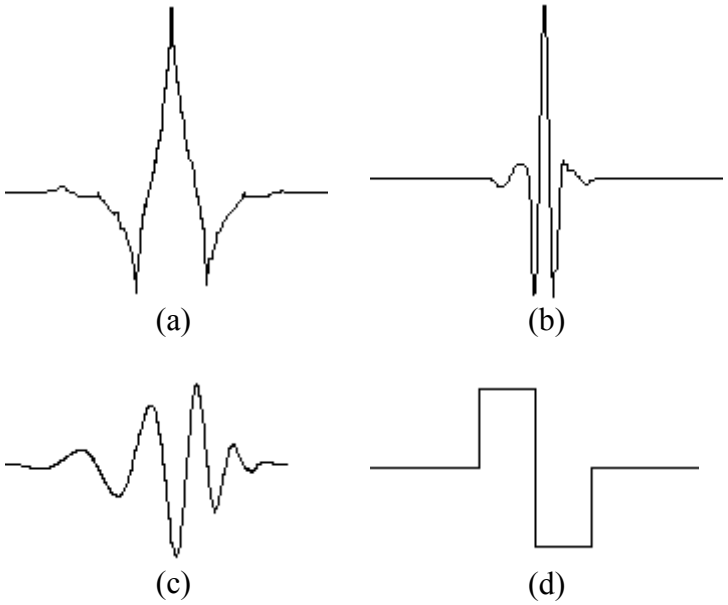

Figure 3.3 Sample of high pass wavelet filters (a) Daubechies 4 (b) Symmlet 7 (c) Daubechies 7 and (d) the Haar wavelet

The WPT is given by the following cascading algorithm until the desired level is obtained:

$$\begin{aligned}\mathbf{X}^{[l+1]}(j) &= \mathbf{W}^{[l]}\mathbf{X}^{[l]}(i); \qquad i = 0,\ldots,m^l - 1 \\ &= \left[\mathbf{X}^{[l+1]}(im)\,\mathbf{X}^{[l+1]}(im+1)\cdots\mathbf{X}^{[l+1]}(im+m-1)\right]\end{aligned} \tag{3.8}$$

where L, $\mathbf{H}_1,\ldots,\mathbf{H}_{m-1}$ are concatenate to form $\mathbf{W}$ – the wavelet matrix [7]. Also it can be seen that the resulting wavelet decomposition at level l+1 consists of m sub-bands. The inverse wavelet packet transform (IWPT) is calculated by

$$\mathbf{W}^{[l]\mathbf{T}}\mathbf{X}^{[l+1]}(j) = \mathbf{X}^{[l]}(i) \tag{3.9}$$

Since

$$\mathbf{W}^{[l]}\mathbf{W}^{[l]T} = \mathbf{I} \tag{3.10}$$

The coefficients for the objects which would lie in band (l,t) of the WPT are labeled $\mathbf{X}^{[l]}(t)$.

3.2.4 Adaptive wavelet matrix

The following section describes how the matrix $\mathbf{W}^{[l]}$, in Eqn. (3.8) is generated by an adaptive wavelet (AW) generation algorithm. There exist several wavelet generating algorithms such as Lifting [11], Angular Quadature Mirror Filtering [12], and Pollen Factorization [17], that design task specific wavelets, also known as adaptive wavelets. It is the Pollen factorization that is best suited to this particular application since it enables m-banded wavelets required for the WPT.

Another advantage of the Pollen factorization is that the m-banded wavelet matrix in Eqn. (3.8) can be parameterized into $q+1$ normalized vectors $\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_q$ and $\mathbf{v}$; where $q \in \mathbb{R}$ is a smoothness parameter for the resulting wavelet. These normalized vectors can be iteratively updated in order to extract user defined features – such as those which prove useful in unsupervised mapping. For a comprehensive account of the theory of the Pollen Factorization, the reader is referred to [17].

The Pollen factorization can be summarized in the following steps:

(1) Define the integer values for *m* and *q*

(2) Initialize the normalized vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$

(3) Construct $\mathbf{W}^{[I]}$ from $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$

(4) Perform the WPT and evaluate the performance of $\mathbf{W}^{[I]}$

(5) Iteratively update $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$ until (4) a converge criteria is meet.

In this study, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$ are initially assigned elements from the uniform distribution, which in previous supervised studies as shown to converge based on similar optimization criteria detailed in section 3.3.7 [8, 13].

3.3 Experimental

3.3.1 Data

The first data set consists of reflectance NIR signals from three different categories of seagrass, *Halophila ovalis, Halodule uninervis/pinifolia* and *Halophila spinulosa* [8]. Each species was sampled 55 times with the NIR signal sampled at 512 evenly spaced wavelengths ranging from 400nm to 2444nm. To correct for particle size effects, the standard normal variate transform (SNV) [37] was applied to the data. Five sample spectra from each species are displayed in Figure 3.4. Notably, the spectra for the three species are very similar and that the researcher was unable to correctly identify the second category into two species, *Halodule uninervis/pinifolia*, which were amalgamated into one single category.

In contrast, to the seagrass data, the second data set consists of dissimilar spectra. The second data set consists of 100 absorbance NIR spectra from five different mineral groups, Amphilolite, Calsilicate, Granite, Mica and soil [8]. Each of the spectra were transformed via the convex hull transform [53], a standard procedure for geological samples. Each category was sampled twenty times with the NIR signal being measured at 512 evenly spaced wavelengths ranging from 1478nm to 2500nm. Five sample spectra from each category are shown in Figure 3.5. For both data sets, the group categorical information is not used as a prior in the adaptive wavelet process and is only supplied for illustrative purposes.

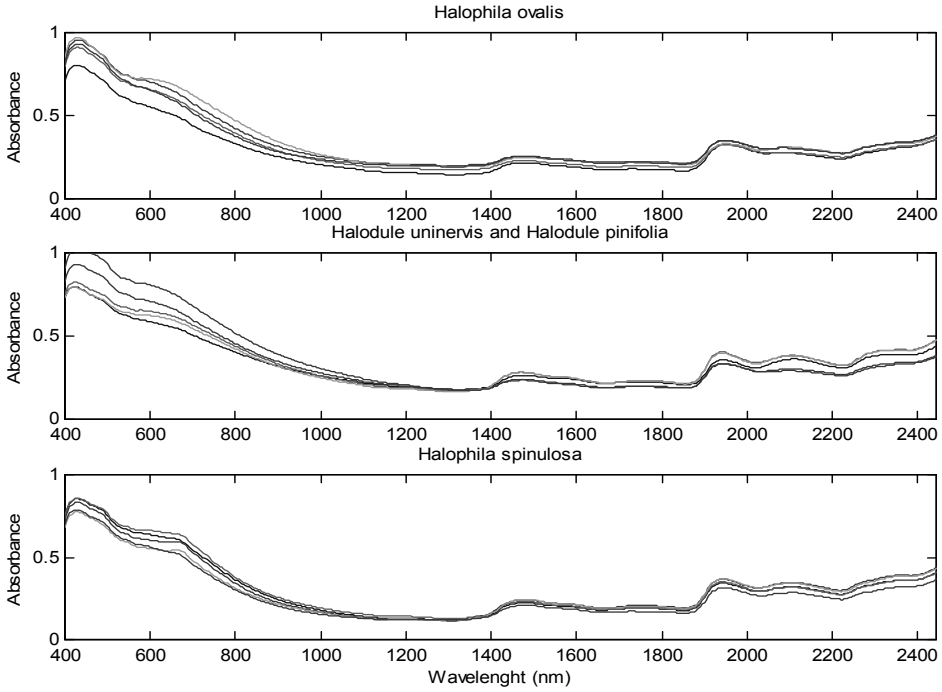

Figure 3.4 Five sample spectra from each category from the Seagrass NIR data set

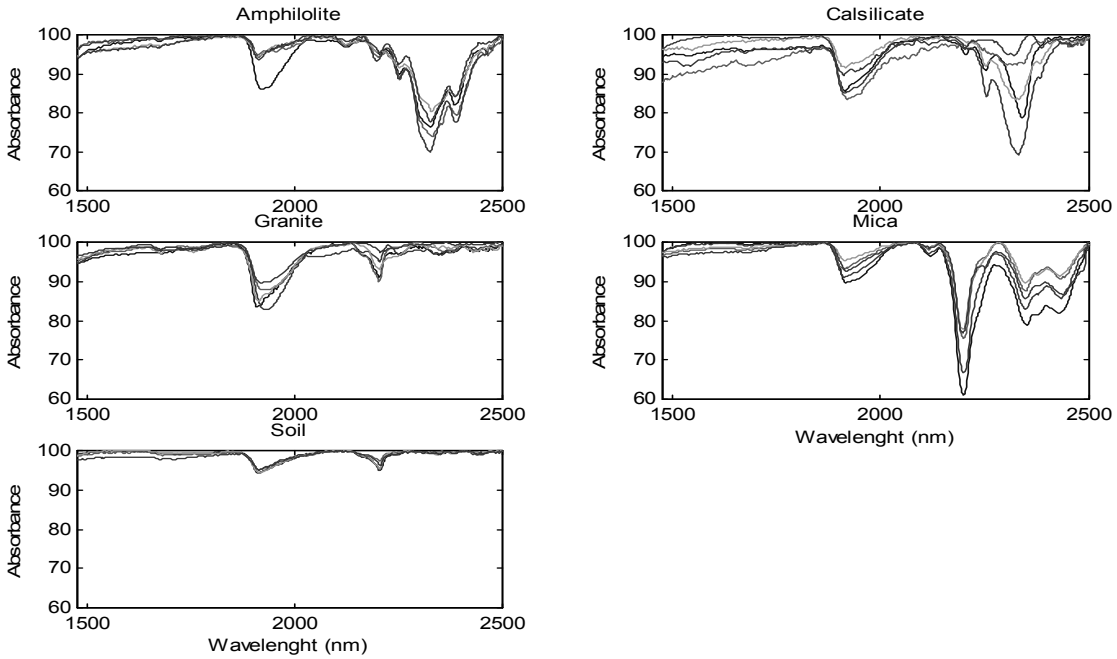

Figure 3.5 Five sample spectra from the five categories from the Mineral NIR data set

3.3.2 Wavelet Principal Component Gaussian Mixture Model Mapping (WPG)

Figure 3.1 illustrates the outline of the proposed method, where features are extracted from the spectra using the wavelet packet transform (WPT) which then is mapped onto a plane using PCA. The final process is fitting a Gaussian mixture model (GMM) on to the mapped spectra. The following details the implementation of the WPT, PCA and GMM respectively for both variants of the proposed method.

3.3.3 Wavelet packet transform

In both variants, the WPT is used to select features from the spectra and in both cases, features (wavelet coefficients) are selected from a single band in the WPT. This is done to avoid aliasing issues associated with selecting coefficients from multiple bands [7, 54-56]. The two variants differ in two aspects of how the WPT and band selection are performed.

The first variant (referred to as the standard variant), uses commonly available wavelet filters for the WPT, 35 in total, which are listed in Table 3.2. Once a wavelet has been selected, the WPT is constructed to the desired level. For both variants, l=7. So for each WPT, there are 255 possible bands to select from. However, from other works [57], analysis on the zeroth band from each level, otherwise known as the scaling bands, typically yields similar results to that using the raw spectra [57]. Thus the scaling bands are removed from the selection set.

From the 248 bands from the WPT, the wavelet coefficients from a single band are forwarded on to the PCA step. However, since no band from this set is favoured a priori, the bands are systematically selected one at a time and for each band, a WPG model is constructed. Alternative band selection methodologies could be used to incorporate wavelet coefficients from multiple bands from the WPT, such as the by variance and by scale algorithms [15], the WPT best bands selection algorithm [58]. These band selection methodologies are useful for compression of the variance of the spectrum rather than information extraction. To simplify the presented methodology and to illustrate the importance of wavelet selection, only a single band is iteratively selected from the WPT.

For the standard variant, there are 8680 WPG models to be considered. The second variant (referred to as the adaptive variant) uses an adaptive wavelet for the WPT. Restrictions on m and q in the mathematical formulation of the adaptive wavelets [17] generates 70 different adaptive wavelets that can be applied to the given data sets, given by Table 3.3.

So far, there are still 255 possible bands to select from the WPT. However one band, the optimization band is favored over all the rest. It is the wavelet coefficients from the optimized band that are forwarded to the PCA step. So for the adaptive variant, there are 70 WPG models to be considered.

3.3.4 Principal component analysis

The principal component step involves mapping the wavelet coefficients on to a plane with the largest variability. This reduces the dimensionality of the wavelet coefficients from n/m^l to 2. The PCA scores are the forwarded to the GMM step. This step is primarily performed to ais in the visualization process and as such the algorithm can be extended without difficulty by extracting k principal components; where $1 \le k \le n/m^l$.

3.3.5 Gaussian mixture models

For each set of PCA score, a range of Gaussian mixture models (GMM) are fitted. Table 3.1 lists the various parameterizations imposed on the GMM's, and for each parameterization, the number of clusters was varied from 1 through to 11. Thus 66 GMM are fitted for each set of PCA scores. From this set, the GMM with the highest BIC score is chosen as the optimal mixed model.

For the optimal GMM, the BIC value, optimal number of clusters and a 5% trimmed mean of $\tau_{i,j}$ (from Equation (4)), $\bar{\tau}$ is recorded. A trimmed mean is used in preference to the actual mean to reduce the effects of abnormal spectra which may arise from experimental errors.

Table 3.2 Trialed standard wavelets

Wavelet family	Number of filter coefficients
Daubechies	2,3,…,16
Symlet	2,3,…,16
Coiflet	1,…,5

Table 3.3 Trialed values for *m*, *q* and *l*

m	*q*	Max. level
2	2-3 4-7 8	7 6 4
4	2-6	3
8	1-3	2

3.3.6 Overall WPG model selection

There are 8680 and 70 potential WPG models for the standard and adaptive variants respectively, each model based on a different wavelet band. To identify which of the wavelet/band combinations result in interesting and informative unsupervised plot/clusters, we imposed the following criteria on each of the WPG model variants:

(a) More that one cluster

(b) Model uncertainty less than 2%, based on a 5% trimmed uncertainty mean. i.e Models with $1-\bar{\tau} < 0.02$

(c) A BIC value in the top 10%

3.3.7 Adaptive wavelet optimization criterion

In section 3.2.4, the wavelet matrix is parameterized in terms of the vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$, and through an iterative updating process, can be optimized for a specific criterion. This section details the optimization criterion used for the adaptive wavelet algorithm.

In this work, we wish to build a wavelet matrix that will be used for feature extraction prior to a 2D PCM. Thus the criterion used should:

(a) Relate to generic features of the data matrix that are likely to show the presence of groupings without prior knowledge of such groupings

(b) Contain the relevant information in two dimensions as further analysis of the wavelet coefficients will be on a plane and

(c) Optimize over a single band in the WPT.

With these requirements in mind, we formulate an optimization criterion based on the eigenvalues of the wavelet coefficients from the band $\mathbf{X}^{[l]}(t)$:

$$\frac{\lambda_1 + \lambda_2}{\sum_i \lambda_i} \tag{3.11}$$

where λ_1 and λ_2 are the two largest eigenvalues of $\mathbf{X}^{[l]}(t)$, the wavelet coefficients of band *t* at level *l*. The basis for this criterion is as follows. If there exists Gaussian clusters which can be parameterized by Eqns' (3.3) and (3.6), then the eigenvector/value structure of $\mathbf{X}^{[l]}(t)$ will be dominated by

- The differences in the cluster means and/or
- The largest eigenvector/values of the ζ covariance matrices [7, 10]

Eqn (3.11) will favour the optimization of cluster separation and/or finding variability within clusters.

To select which of the bands in the WPT to optimize the adaptive wavelet on, the following two rules were applied:

(1) The scaling $\mathbf{X}^{[l]}(0)$ is excluded from the selection set – for the reasons previously discussed in section 3.3.3

(2) The band that initially has the highest ratio from Eqn. (3.11) is kept as the optimization band.

3.3.8 Software

The optimization function utilized for the AWT is the unconstrained optimizer "fminu" function from the Matlab Optimization Toolbox® [39] and the Matlab Wavelet Toolbox® [38] is used to perform the standard wavelet packet transform using the predefined wavelet filters. Gaussian mixture models BIC/uncertainty values are generated in R using the *mclust* module [8].

3.4 Results and Discussion

3.4.1 Seagrass Data

Figure 3.6 and Figure 3.7 both show a general trend of increasing model accuracy with increasing BIC for both the adaptive and standard WPG models. Using the BIC and model uncertainty criteria, the adaptive WPG model select 6 models out of the seventy trialed combinations of m and q. While 9 were chosen for the standard WPG out of the 8680 trialed models.

Visual inspection of the three adaptive models revealed very similar plots and cluster structures, shown in Figure 3.8, with adaptive wavelet parameters $m = 2$, $q = 3$, on the WPT band $\boldsymbol{X}^{[1]}(8)$. Here we can see clear evidence of clusters with the clusters forming a "V" structure. Also we observe that the directions of the semi-major and semi-minor axes of the clusters are in the direction of the "V".

Inspection of the nine standard wavelet WPG models were not as consistent as the adaptive counterparts as the selected standard WPG models produced three main types of images (with minor variations)- shown in Figure 3.9, Figure 3.10 and Figure 3.11. In all three images, we can see clear evidence of clustering, but varying numbers of clusters between all three models. This illustrates the effect of different wavelets on the resulting image. However, in comparing Figure 3.9, Figure 3.10 and Figure 3.11, we observe a unifying feature of the orientation of the clusters – they all form a "V" structure.

3.4.2 Mineral Data

The positive trend of increasing model accuracy with increasing BIC for the standard WPG models is again evident for the Mineral NIR data set, in Figure 3.12. For the adaptive WPG, this trend is highly extenuated with an almost linear trend, Figure 3.13. Using the BIC and model uncertainty criteria, the adaptive WPG model select 12 models out of the seventy trialed combinations of *m* and *q*. While 13 were chosen for the standard WPG out of the 8680 trialed models.

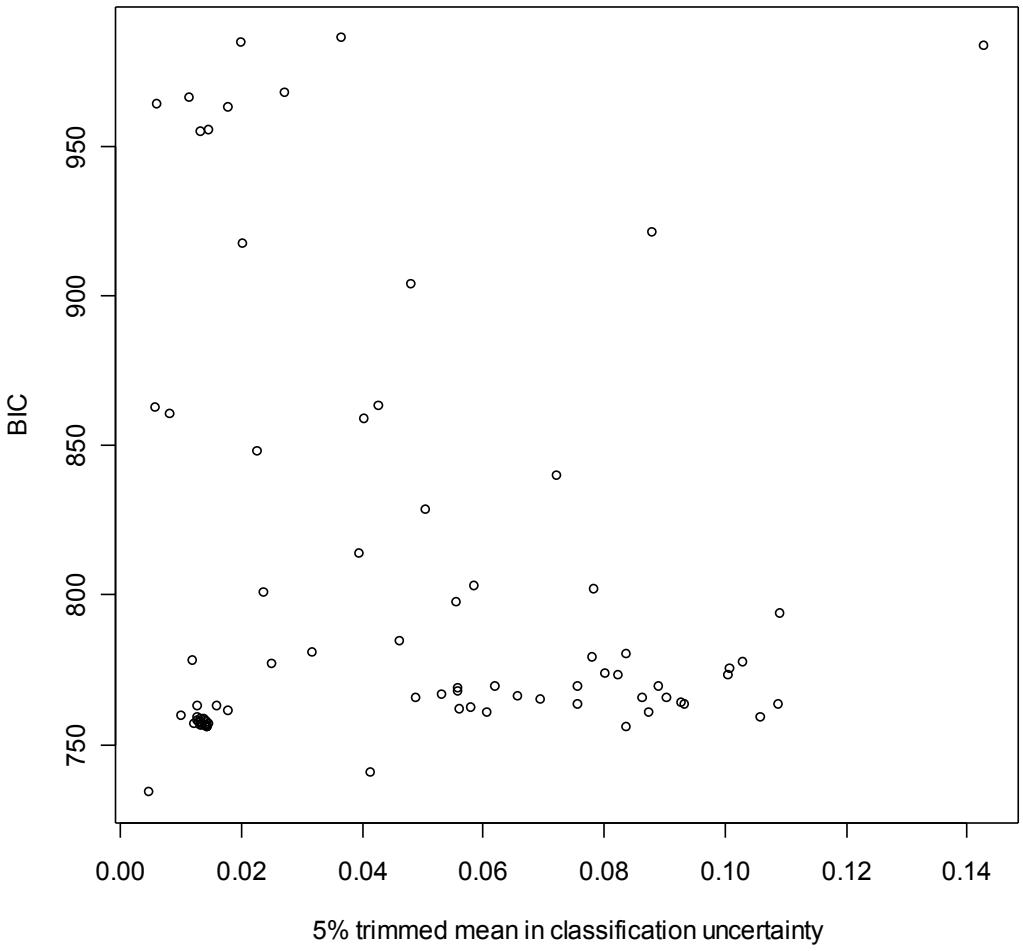

Figure 3.6 Seagrass adaptive WPG model scatter plot of the Bayesian information criteria (BIC) Vs classification uncertainty trimmed mean

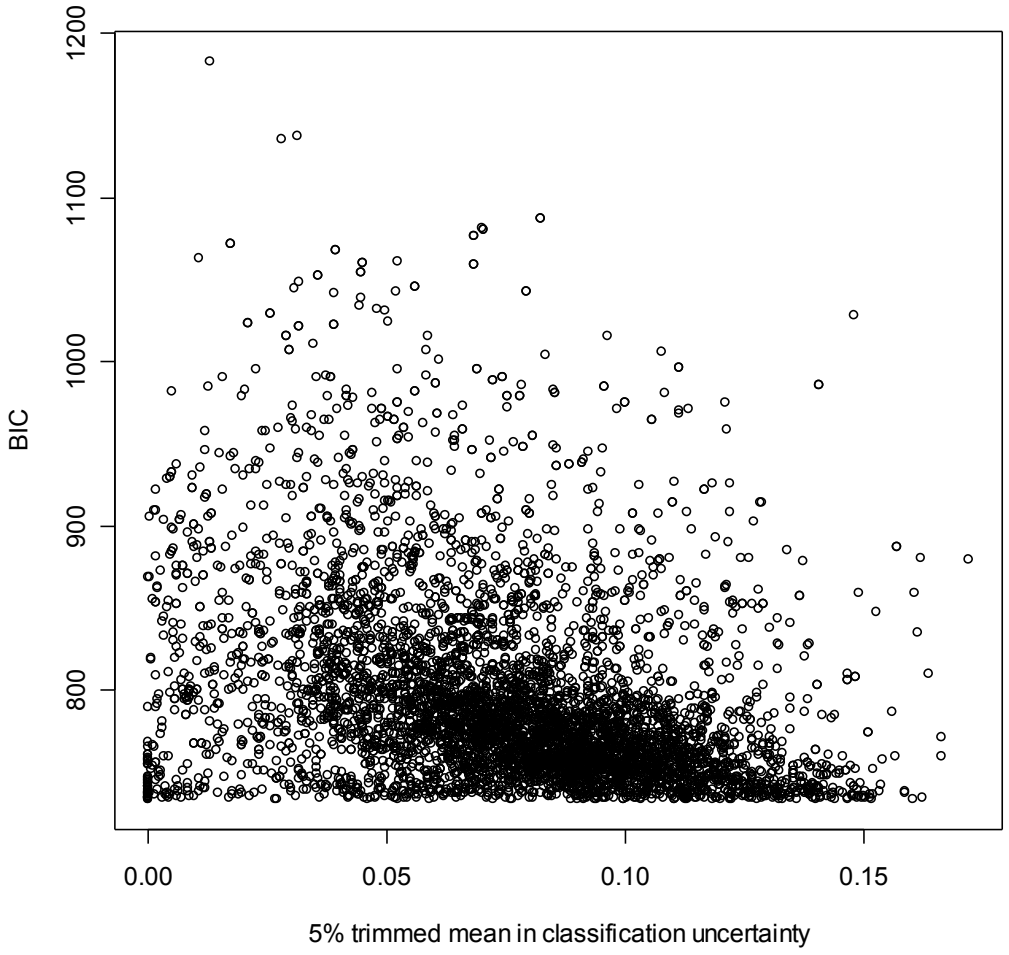

Figure 3.7 Seagrass standard WPG model scatter plot of the Bayesian information criteria (BIC) Vs classification uncertainty trimmed mean

Inspection of the adaptive models, again revealed similar PCA/GMM plots, as shown in Figure 3.14, which clearly show three clusters aligned on a "V". Further investigation of Figure 3.14 shows that the central cluster (in the third quadrant) consists of three sub-groups, as shown in Figure 3.15. This disparity in the number of clusters, arises due to the penalty term in the BIC – Eqn (3.7). The BIC favors models with fewer parameters. Ie. Favors fewer clusters with the same parameterizations such as equal area and directions. Here we can conclude that the BIC may have over penalized and that there are five clusters in Figure 3.14.

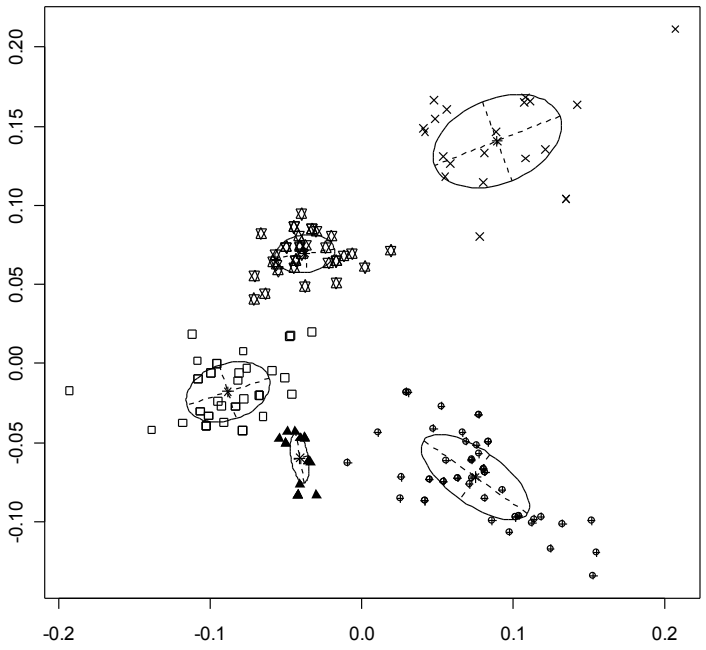

Figure 3.8 Adaptive WPG on the Seagrass data with adaptive wavelet parameters *m* = 2, *q* = 3, WPT band: $X^{[1]}(8)$

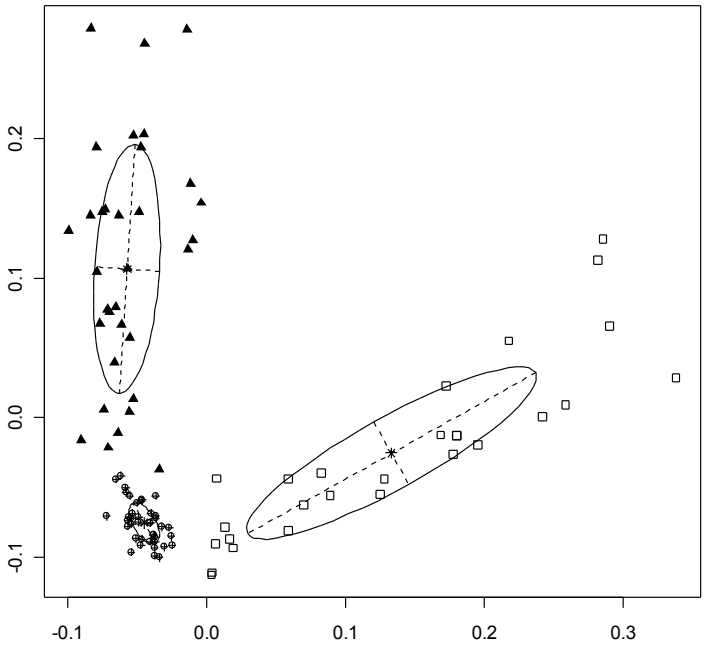

Figure 3.9 Standard WPG on the Seagrass data with wavelet parameters: Daubechies 2 filter on the WPT band $X^{[3]}(8)$

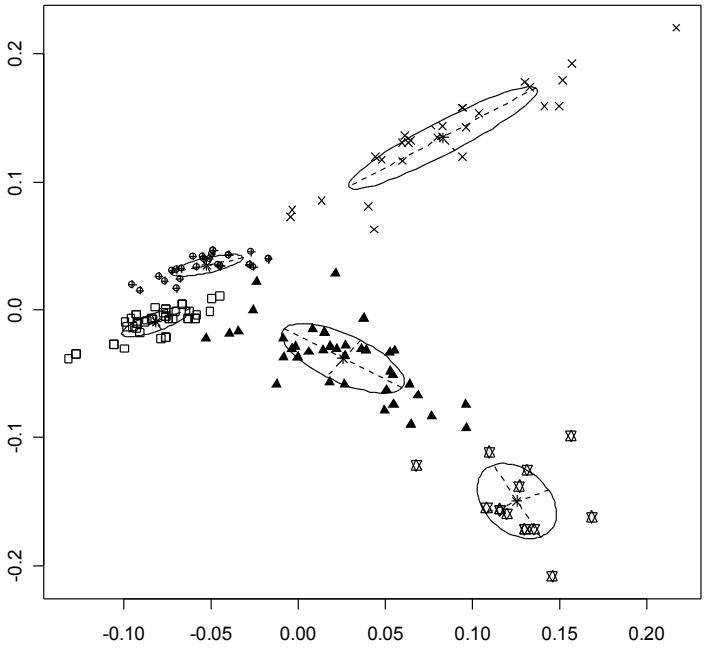

Figure 3.10 Standard WPG on the Seagrass data with wavelet parameters: Daubechies 2 filter on the WPT band $X^{[3]}(2)$

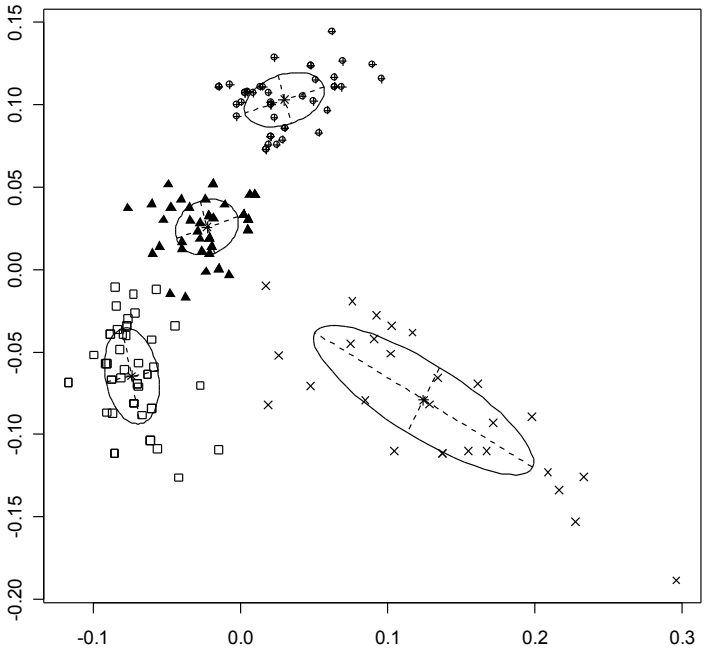

Figure 3.11 Standard WPG on the Seagrass data with wavelet parameters: Daubechies 5 filter on the WPT band $X^{[7]}(6)$

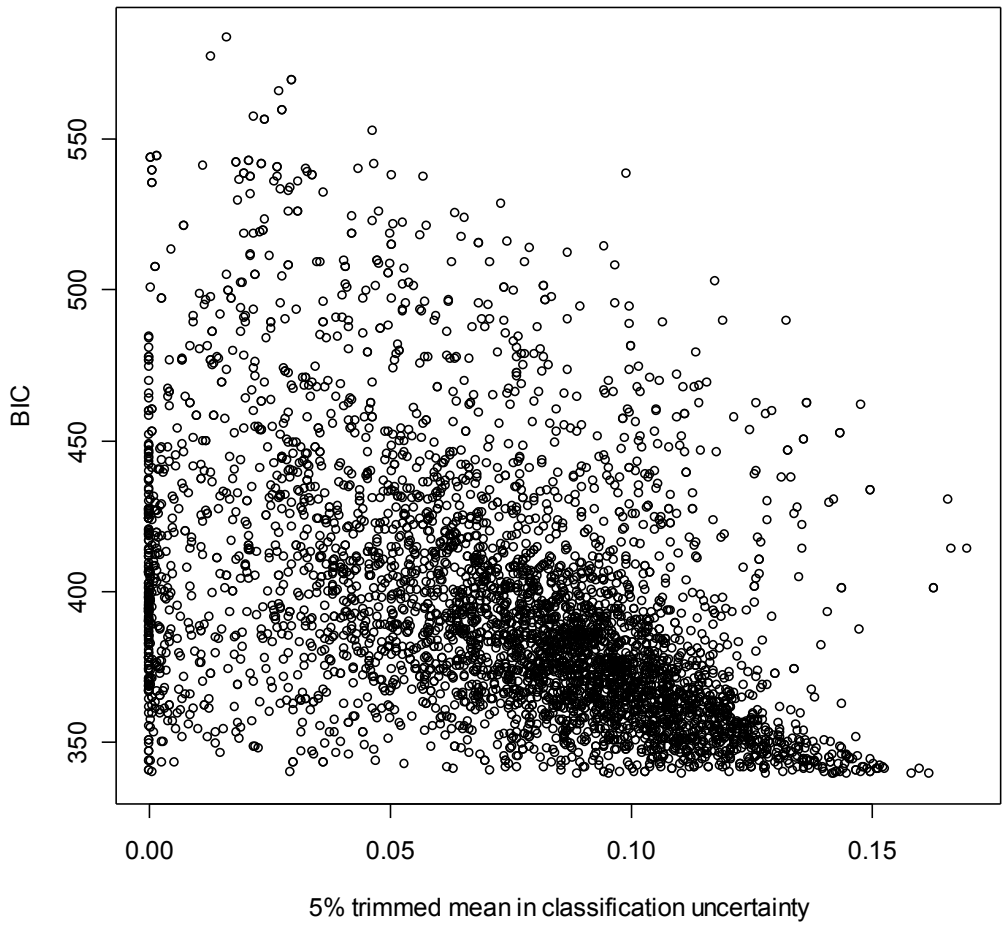

Figure 3.12 Mineral standard WPG model scatter plot of the Bayesian information criteria (BIC) Vs classification uncertainty trimmed mean

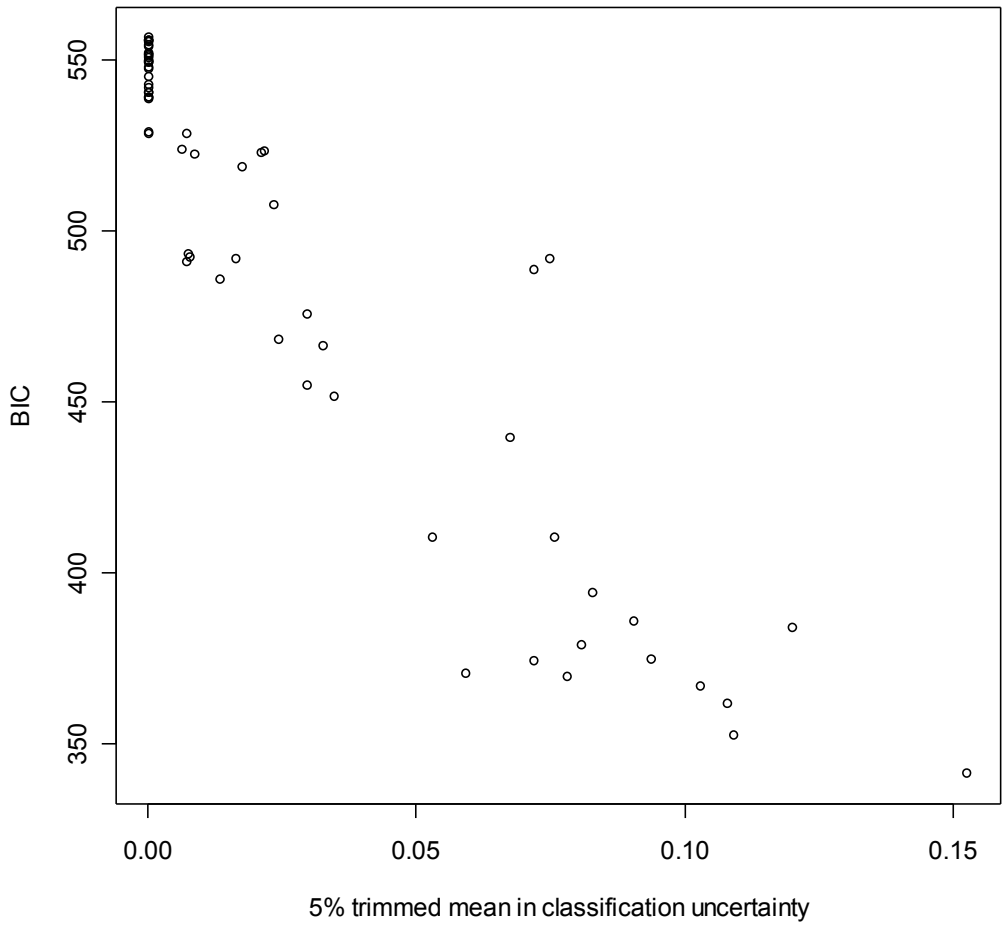

Figure 3.13 Mineral adaptive WPG model scatter plot of the Bayesian information criteria (BIC) Vs classification uncertainty trimmed mean

From the thirteen standard WPG models, six exhibit similar structures and clusters as shown in Figure 3.16, which show evidence of three clusters forming a "V". Noting the central cluster of the "V" contains over 60% of the spectra. The remaining standard WPG models, shown in Figure 3.17, resulted in a PCA/GMM plot nearly identical to the adaptive PCA/GMM models. As in the adaptive WPG, the central cluster consists of three groups, shown in Figure 3.18.

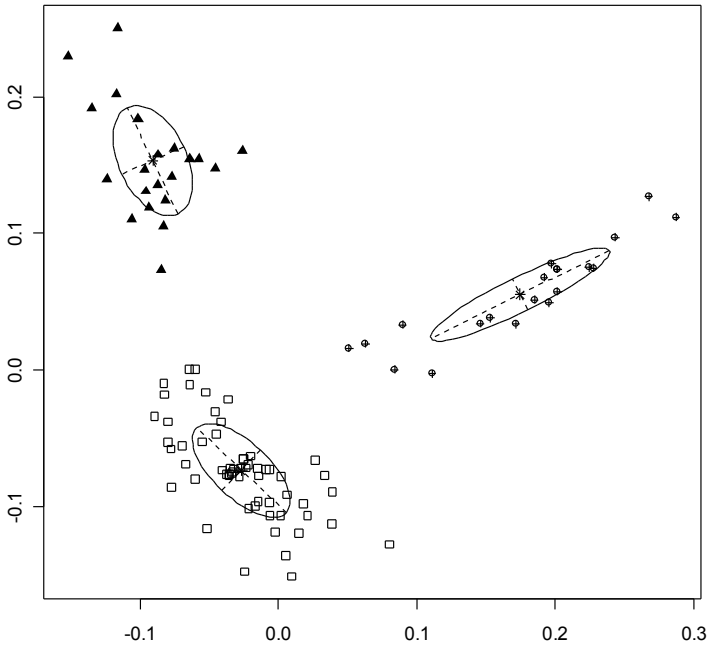

Figure 3.14 Adaptive WPG on the Mineral data with adaptive wavelet parameters $m = 2$, $q = 3$, WPT band: $X^{[1]}(8)$

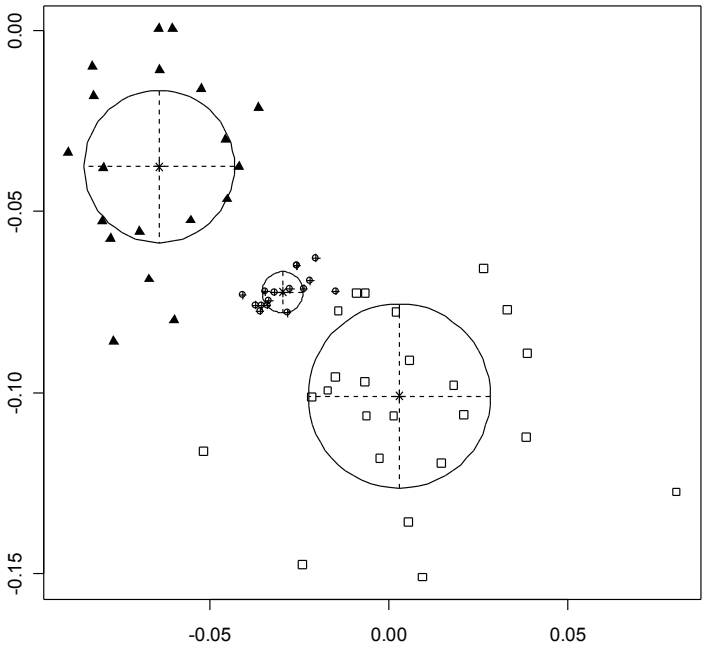

Figure 3.15 Optimal Gaussian mixture model on the third quadrant of Figure 3.14

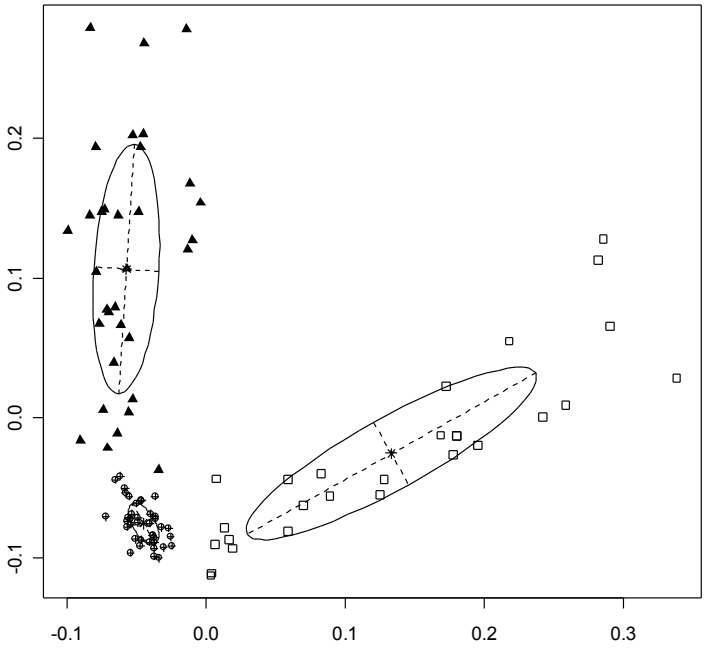

Figure 3.16 Standard WPG on the Mineral data with adaptive wavelet parameters *m* = 2, *q* = 3, WPT band: $X^{[1]}(8)$

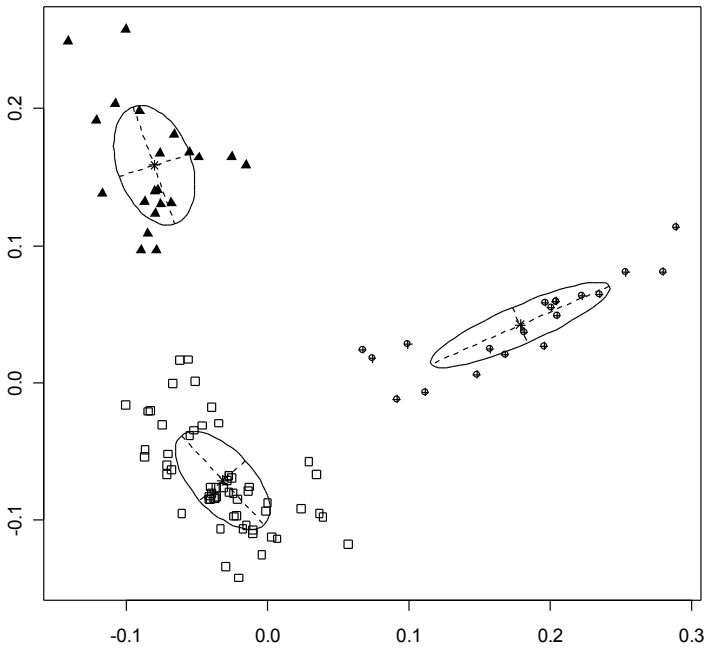

Figure 3.17 Standard WPG on the Mineral data with adaptive wavelet parameters *m* = 2, *q* = 3, WPT band: $X^{[1]}(8)$

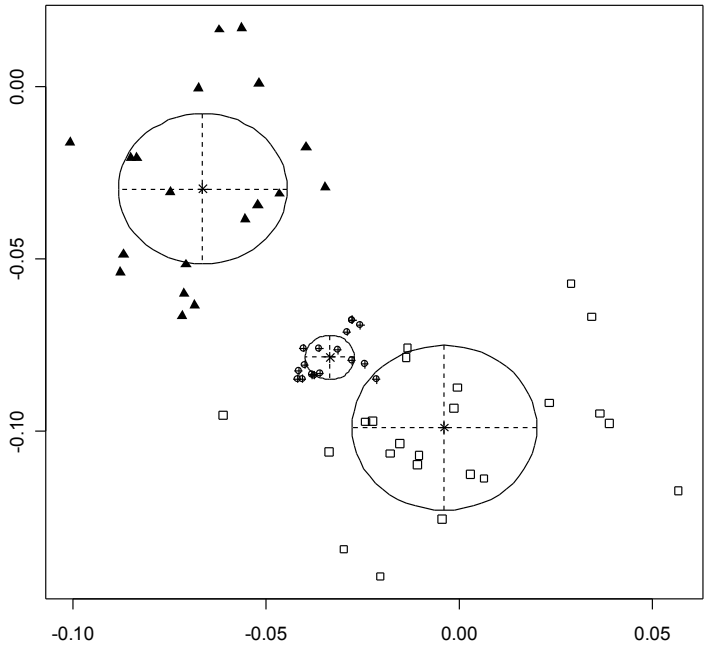

Figure 3.18 Optimal Gaussian mixture model on the third quadrant of Figure 3.17

3.5 Conclusion

The proposed method of integration of the wavelet packet transform with principal component analysis and Gaussian mixture models has been shown to elucidate unsupervised clusters from the provided NIR spectra. To address the issue of wavelet selection for the proposed method, we conducted exhaustive searches using both standard wavelets (8680 wavelets) from literature and adaptive wavelets (70 combinations of adaptive wavelet parameters).

The exhaustive search, using the BIC and model classification uncertainty as a filtering scheme, identified a small subset (<13) of wavelets for both the standard and adaptive wavelet approaches. Visual inspection of the selected wavelet models, both standard and adaptive, provided promising results in finding clusters for the presented NIR data sets. The standard wavelet method gave a range of possible clustering outcomes, with different number of clusters and different cluster orientations for different wavelets. While the adaptive wavelet method gave more consistent clusters for various combinations of m and q (adaptive parameterizations).

The consistency found using the adaptive wavelets can be view as a result of linking the optimization search criterion used to iterate the adaptive wavelet, with characteristic parameterizations of two dimensional Gaussian mixture models. So when different adaptive parameters were trialed, the features extracted from spectra would still be favorable for unsupervised Gaussian mixtures. Thus the different adaptive wavelets were extracting similar features from the spectra relevant to good group separation.

3.6 Summary

We introduce a new method of unsupervised cluster exploration and visualization for spectral datasets by integrating the wavelet transform, principal components and Gaussian mixture models. The Bayesian Information Criterion (BIC) and classification uncertainty performance criteria are used to guide an automated search of commonly available wavelets and adaptive wavelets. We demonstrate the effectiveness of the proposed method in elucidating and visualizing unsupervised clusters from near infrared (NIR) spectral datasets. The results show that informative feature extraction can be achieved through both commonly available wavelet bases and adaptive wavelets. However, the features from the adaptive wavelets are more favourable in conjunction with unsupervised Gaussian mixture models through a user specified internal linkage function.

Chapter 4

Bagged Super Wavelts Reduction for Boosted Prostate Cancer Classification of SELDI-TOF Mass Spectral Serum Profiles

4.1 Introduction

Since the development of large mass spectrum profiling; consisting of excess of tens of thousands biomarkers, modern statistical research has been focused towards distilling pertinent biomarkers relevant to diagnosable symptoms such as prostate cancer [59]. The difficulties involved in parsing such large datasets are many fold, the most general being firstly the shear size of the dimensionality of the data and secondly the unknown complexity of the relationship(s) correlating the measured mass spectrum profiles and the observed disease states.

The issue of high dimensionality and unknown model complexity has given rise to hybrid ensemble techniques such as Treeboost [60] and Random Forests [61] which are an amalgamation of a Classification and Regression Trees (CART) [62] with Boosting [63] and Bagging [61] respectively. These hybrid techniques (Treeboost and Random Forests) are universally designed to model both non-linear and linear effects, which makes them suitable as initial techniques for data exploration for biomarker discovery.

Treeboost operates by fitting a CART model to the data, then recursively fitting CART models to the residuals of the previous CART model. This translates to fitting informative linear relationships between the CART models to predict the response, which can then lead to forming linear relationships between the independent data (M/z) values and the dependant values (disease status). For moderately large number of variables, fitting Treeboost models become impractical due to the high computational cost. One method to reduce this cost is to reduce the number of variable under consideration in the Treeboost model. We use Random Forests to identify independent and weakly important variables, as a variable reduction method for Treeboost

High correlation within the spectrum profile presents another complicating issue as this often leads to numerical instabilities of the statistical model. Commonly with most forms of spectra, the juxtapositional variables (wavelengths, M/Z ratios) contain similar information usually as a result of being a measurement of the same underlying physical process. This effect can be taken advantage of in the form of feature extraction and dimension reduction, where localized features from the spectra are extracted and used as the inputs to the statistical model to predict the symptoms. In this respect the wavelet transform can be used to extract features from spectra [15].

The wavelet transform (WT) is a projection of the spectrum onto an orthogonal basis, called a wavelet basis. This is to say that the spectrum can be represented by a set of localised, orthogonal basis functions called wavelets. In this the WT has a familiar origin with the Fourier transform (FT), whose orthogonal basis functions are the sine functions. However, the DWT has a larger amount of flexibility than the FT, in the sense that the WT has an infinite choice of basis functions (wavelets) to choose from. Thus we can choose a wavelet basis that will result in good approximations of the latent features within the spectrum. However, in this investigation, the features are not known a prior; this chapter will use a combination of discrete wavelet transforms to create a super-wavelet [64] over the spectra.

This chapter investigates the practicality of the super-wavelet transform on large spectral databases. This is achieved using a data reduction heuristic using Random Forests and Treeboost to build a classification model, using SELDI-TOF mass spectrum profiles as an illustration. We also investigate wavelet selection for the proposed method by benchmarking standard wavelet types with super wavelets using random forests. Further benchmark comparisons using Random Forests and linear discriminate analysis (LDA) are provided to assess the Random Forest/Treeboost algorithm performance using the super wavelets.

4.2 Theory

4.2.1 **Discrete Wavelet Transforms (DWT)**

The discrete wavelet transform (DWT) like the Fourier transform, can be used to reformulate a spectrum into meaningful feature in another "space", by mapping the spectrum onto a n analyzing function. In Fourier analysis, the analyzing functions are the set of sine function, where as for the DWT, wavelets are the analyzing functions. The DWT is given by:

$$x(t) = \sum_{j=1}^{l} \sum_{k=0}^{2^j} c_{j,k} \psi_{j,k} \tag{4.1}$$

where $\psi_{0,0}$ is the father wavelet, from which all the other wavelets $\psi_{j,k}$ are derived from, $x(t)$ is the spectrum, *l* is the decomposition level for the DWT [15] and $c_{j,k}$ is the wavelet coefficient calculated by the inner product between $x(t)$ and $\psi_{j,k}$:

$$c_{j,k} = \left\langle x(t) \middle| \psi_{j,k} \right\rangle \tag{4.2}$$

Unlike Fourier analysis, there are many types of analysis functions (wavelets) that can be used for the DWT – each resulting in different wavelet coefficients (mapped features). Since we do not know which wavelets will result in the best feature extraction a priori for classification, this chapter will use linear combinations of wavelet functions, referred to as super-wavelets [64],to extract features. We construct two super wavelet frames using equally sized Daubechies (4 & 12), Symlets (4 & 12) and Coiflets (1 & 3) wavelets. Daubechies, Symlets and Coiflets were chosen as the analysis functions as they all have compact support, regular and high degrees of vanishing moments. The symmetry of the chosen wavelets ranges from the distinctly asymmetrical Daubechies wavelets to slightly symmetrical Symlets to the near symmetrical Coiflets [15]. An example of these wavelets is shown in Figure 4.1.

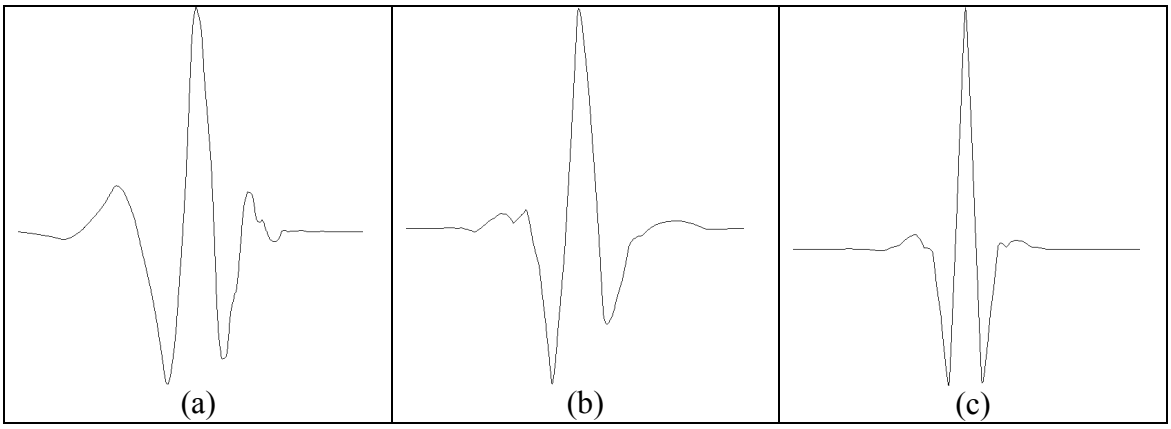

Figure 4.1 Examples of the different wavelet families: Daubechies 4 (a), Symlets 4 (b) and Coiflets 2 (c)

4.2.2 Classification and Regression Trees (CART)

Classification and Regression Trees (CART) [62] are useful tools for uncovering structure in large datasets. The algorithm partitions the data set based on a set of criteria, and from these partitions grows a binary tree. This tree is then used to predict the response. Each node within contains a splitting rule, which is determined through minimization of the relative error statistic (*RE*):

$$RE(d) = R(Left) + R(Right) \tag{4.3}$$

where *R(Left)* and *R(Right)* are the impurities for the left and right node defined at every possible decision *d* found from within a predictor variable *x*. For the classification problem, the GINI index is used to define the node impurities:

$$R(m) = GINI = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}) \tag{4.4}$$

where $\hat{p}_{mk}$ is the proportion of class *k* in node *m*. The splitting rule that minimises the *RE* is then used to construct a node in the tree.

4.2.3 Random Forests

Random forests for classification as defined by Breiman [27] is a collection of many classification trees, each built on a unique bootstrapped sample of the data. The specific example of a random forest used by Breiman, implements randomly selected predictor variables or at each node in the building of each tree included within the bootstrapping.

Breiman called this routine Forest-RI. Forest-RI randomizes during the split selection of each tree. This randomness has the effect of building new trees with different structures, increasing the variety of relationships modeled within the forest that in turn improves the overall predictive performance. The classifications are then determined by a count of the classifications from each tree within the forest.

As each tree has the same weight within a random forest, a simple proximity can be formed between the observations. This proximity is a simple count of how many times 2 cases have been classified into the same terminal node of each tree. Dividing this count by the number of trees a similarity measure s_{ij} is calculated between the two observations [27]. The implementation of random forests is the "*randomForest*" package in R.

4.2.4 Stochastic Gradient Boosting for CART (Treeboost)

Treeboost [60] is a stage wise linear combination of classification trees F_m each built from a bootstrapped sample of the data. The linear combination is built in a stage-wise manner where each new tree in grown such that it lies along the path of steepest decent given a specified loss function. This gives form new updated boosted model F_m as recurrence relation,

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; a_m) \tag{4.5}$$

where $h(x; a_m)$ is the new model to be added previous boosted model F_{m-1}, and ρ_m is the weight of the new tree in the model given by,

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, F_{m-1} + \rho h(x_i; a_m)) \tag{4.6}$$

where y is the response, F_{m-1} is the previous boosted set, and the parameters a_m of $h(x; a_m)$ are found such that $h(x; a_m)$ lies in the path of steepest decent. The predictions of boosting are then the weighted sum of the predictions for each individual tree within F_m. Treeboost was implemented using the Salford Systems package "*TreeNet*".

4.2.5 Tree based methods for variable importance

Random forests and boosting are a black box approaches to modelling as the combination of hundreds of models is too confusing to analyze individually. To aide in the interpretation of these results there are several measures of variable importance that can be used to quickly identify the most influential variables.

The CART variable importance measure is simply the reduction in impurity that a particular variable creates when it is split on. The measure is primarily dependent on where the variable is used in the tree and is defined as:

$$VIP(x) = \sum_{t \in T} RE(d) \tag{4.7}$$

where $VIP(x)$ is the variable importance of x in a node t in tree T, and $RE(d)$ is the risk as defined by Eqn (4.3). Random forests extend this VIP statistic to span over the bagged set of trees. The random forest VIP is MSE that variable induces when used to form a split within a tree within a forest.

The random forest VIP list is a useful tool for data reduction as it ranks the variables used in the forest. It should be noted that if a variable has not been used within the forest, its variable importance is zero. Therefore for a large dataset the list of important variables in the forest is considerably smaller than the number of variables within the dataset.

4.3 Experimental

4.3.1 Data

Mass spectral (MS) profiles consisting of 15154 SELDI-TOF M/Z ratios from 342 patients were collected to investigate M/Z biomarkers for the presence of prostate cancer. This data was obtained from the freely available datasets available form the American National Cancer Institute (NIC). Out of the 322 patients, 69 were diagnosed with malignant prostate cancer, 190 with benign prostate hyperplasia and the remaining 63 patients being controls [59].

Previous works on this data include [59] who, on a subset of the data ran genetic algorithms classifier to distinguish between 2 groups (control, cancer), training on 56 observations (25,31) and testing on 266 observations (212, 38) groups and obtained 95 % sensitivity and 78 % specificity. Criticisms have been expressed on the measurement design of this data [65], however, we use this data for the sole purpose of demonstrating the methodology in section 4.3.2.

Results from other authors suggest that SELDI-TOF M/Z profile can be used to distinguish the control, benign prostate hyperplasia and prostate cancerous status of patients. Qu *et al* .[66], on a different SELDI-TOF dataset, used Adaboost and boosted decision trees and stumps also to distinguish between two patient disease status; control and prostate cancer. The data used by Qu *et al* [66] consisted of a training set of 74 observations (30 control and 44 cancerous prostate) and a testing set of 88 observation (28, 66). Qu *et al* [66] achieved a sensitivity of [100 %, 93.8 %] and a specificity of [100 %, 93.8 %] respectively.

4.3.2 Method

The proposed feature extraction Treeboost methodology consists of three main phases highlighted in Figure 4.2:

1. Initial feature mapping of the MS profile is performed using the super wavelet frames
2. Variable reduction by
    a. Reduction of the SWF using t statistics
    b. Variable reduction using the VIP list from Random Forests
3. Discrimination using Treeboost on the reduced extracted features from the MS profiles.

During the first phase, the M/Z profiles are transformed using the super wavelet frame (SWF). Where the SWF consists of the Daubechies (4 and 12 tap filters), Symlets (4 and 12 tap filters) and Coiflets (4 and 12 tap filters), giving a total of six wavelets in the SWF. This then results in an expansion in the dataset size by a factor of six to approximately 60,000 variables – which is too many variables for Random Forest or Treeboost. To reduce this expansion in data size, the SWF is initially filtered using pair wise t-values between the three groups on each wavelet coefficient in the SWF.

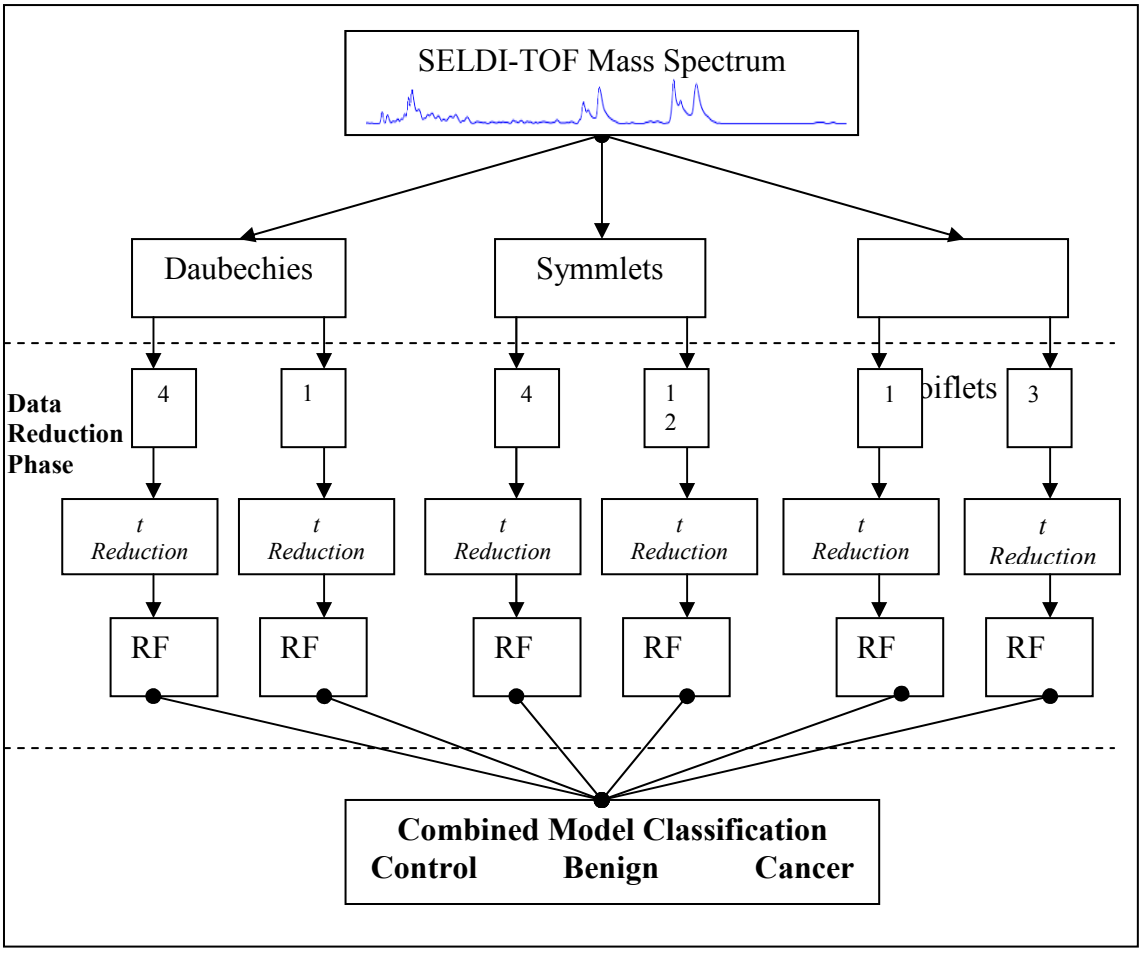

Figure 4.2 Flow diagram of the analysis

The wavelet coefficients corresponding to the largest 5% t-values are retained for further reduction using Random Forests.

Prior to the RF reduction, the data is split into a test (30%) and training (70%) sets so that cross validated correct classification rates can be assessed for the RF and TreeBoost methods. All random forests used in this investigation were grown to 200 trees sizes softly limited at a minimum terminal node size of 5.

The variable reduction phase using Random Forests was done using iterative applications of Random Forest.

1. Initially Random Forest is performed on the entire dataset passed on from the *t*-reduction phase. Using the VIP list from the initial RF, the top 30% predictive variables (wavelet coefficients) from the VIP list are removed and concatenated into a predictive dataset. The remaining variables are concatenated into a reduced dataset.
2. Random Forest is then used on the reduced dataset to form a new VIP list from which the top 30% are then removed and placed in the predictive dataset.
3. Step 2 is repeated iteratively until the predictive error using the reduced dataset plateaus. The convergence results are shown in Figure 4.3.

The motivation behind this iterative RF selection scheme is largely due to the shear size of the initial (t-reduced) dataset. As the predictor set is quite large there will be large amounts of redundancy, but also many various combinations of variables that give the same result. If only one RF were used to reduce the dataset, then the redundant but informative variables would be screened out. Successive RF's on the reduced datasets would capture most of the informative variables. Once the RF variable selection has finished, the predictive dataset is used for analysis.

4.3.3 Benchmarking

We use the results from Random Forest and linear discriminate analysis as methods to benchmark the performance of the above methodology. The dataset input to RF and LDA are the super wavelet RF reduced data that is used as the data input to TreeBoost.

Benchmarking for the super wavelet is done by comparing the Random Forest performances of the t-reduced data from the datasets generated from each of the six wavelet types composing of the super wavelet. I.e. The super wavelet RF is compared to six other RF's derived from t-reduced data using one of the six wavelets used in the super wavelet itself.

4.4 Results and Discussion

The model performances for the super wavelet reduced data are shown in Table 4.1, listing the correct classification rates (CCR) for the cancerous and benign groups for the test data and the overall CCR for the training data. The CCR for the training data is used as an indication of the overall model training performance, from which in Table 4.1, the LDA model trained best on the super wavelet data.

Table 4.1 Benchmarking model performance using super wavelet

Model	Training CCR	Test set correct classification rate		
		Control	Benign	Cancer
Treeboost	90.78 %	100 %	98.24 %	68.75 %
LDA	100 %	89.47 %	94.73 %	90.47 %
Random Forests	93.33 %	94.73 %	98.24 %	76.12 %

The CCR (cancerous and benign) for the test data are used as an indication on the robustness of the predictive performance of the model. In this setting, it is more important to correctly classify positive cancerous patient than misclassify a positive benign patient. From Table 4.1, the LDA model gave the best CCR for the cancerous patients, followed by the Random Forests model then Treeboost.

Superiority of RF over Treeboost suggests that there is high diversity between the possible trees that can be built from super wavelet basis. This diversity lends itself more to the averaging of the decision boundaries employed random forests, rather the linear combination used by Treeboost. This diversity highlights the different profiles selected by each wavelet type within the super wavelet basis. Overall however LDA performed for in the training set and for predicting the cancerous patients. However LDA required previous data reduction to achieve this result.

In investigating the role of the super wavelets; especially in exploring which wavelets are seemingly more useful in feature extraction, the Random Forest VIP list using the super wavelet, shown in Table 4.2, is analyzed. Here it is seen that the Coiflets and Symlets appear most frequently and most importantly in the VIP list. Both Symlets and Coiflets have a high degree of symmetry when compared to Daubechies wavelets.

Table 4.2 Random Forests VIP list, cropped at the top 50 % of variables

Coefficient	Mean Decrease in Accuracy
COIF3-7635	0.76
SYM4-241	0.74
SYM4-2123	0.70
SYM12-251	0.61
COIF3-2160	0.58
SYM4-2358	0.57
DB4-4051	0.56
SYM4-1885	0.55
COIF3-250	0.52
SYM12-2388	0.52
COIF1-303	0.52
DB12-1901	0.51
DB4-246	0.50
COIF1-7604	0.49
SYM12-161	0.49
COIF1-2199	0.49

When comparing the RF models arising from each individual wavelet type in

Table 4.3, the CCR for the cancerous patients is seemingly similar for all wavelet types. But when viewed jointly in the super wavelet RF model, outperforms the individual wavelet RF models for CCR for cancerous patients. This suggests that the information for the cancerous patients can be better expressed with multiple wavelets (ie a super wavelet) rather that a single wavelet.

The M/Z ratios identified by the RF VIP list for the super wavelet are shown in Figure 4.4. Of those variables selected it can be seen that most lie within the 0 to 2000 M/Z ratios. Some debate over the validity of the information within this region [65], however, other wavelet based methods on similar data have also identified M/Z ratios in this neighborhood [67].

The false positive prediction rates for the test data, using the super wavelet Random Forest, in Table 4.4 compare quite favorably to other works published on this dataset. [59] achieved false positive rates of 5% and 22% for cancerous and benign patients using a two component model (i.e. only predicting two disease states).

Table 4.3 Benchmarking wavelet types using Random Forest performance

Wavelet type	Training CCR	Test set correct classification rate	
		Cancer	Benign
Daubechies-4	87.56 %	97.29 %	78.26 %
Daubechies-12	87.11 %	96.25 %	88.23 %
Symlets-4	94.22 %	94.36 %	84.61 %
Symlets-12	89.78 %	97.29 %	82.6 %
Coiflets-1	90.22 %	95.18 %	88.23 %
Coilets-3	91.11 %	94.93 %	72.22 %
Super Wavelet	92.00 %	100 %	86.36 %

Table 4.4 Percentage false positive rates using the Random Forests on the super wavelet data.

Actual	Test misclassifications		
	Control	Benign	Cancerous
Control	NA	(1/19) = 0.052 %	0 %
Benign	0 %	NA	(1/57) = 0.017 %
Cancerous	0 %	(5/21) = 23.80 %	NA

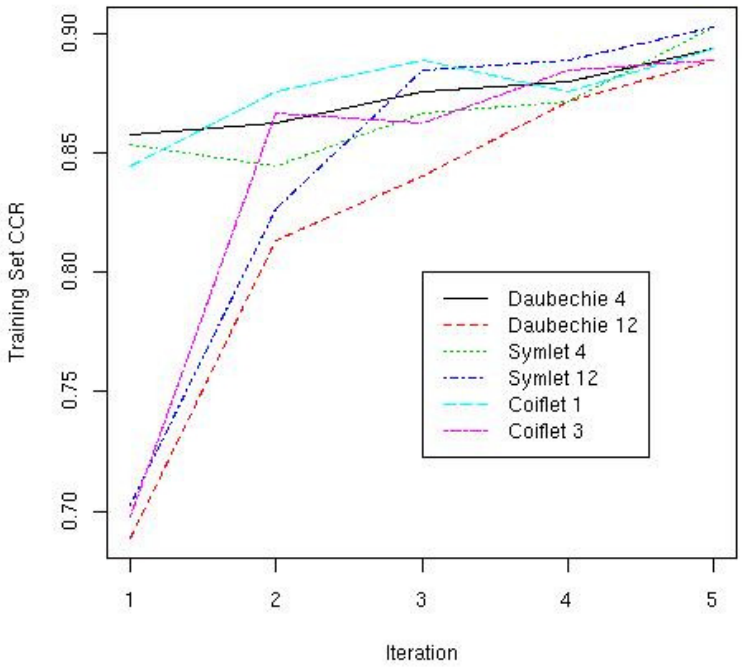

Figure 4.3 RF reduction training set CCR convergence

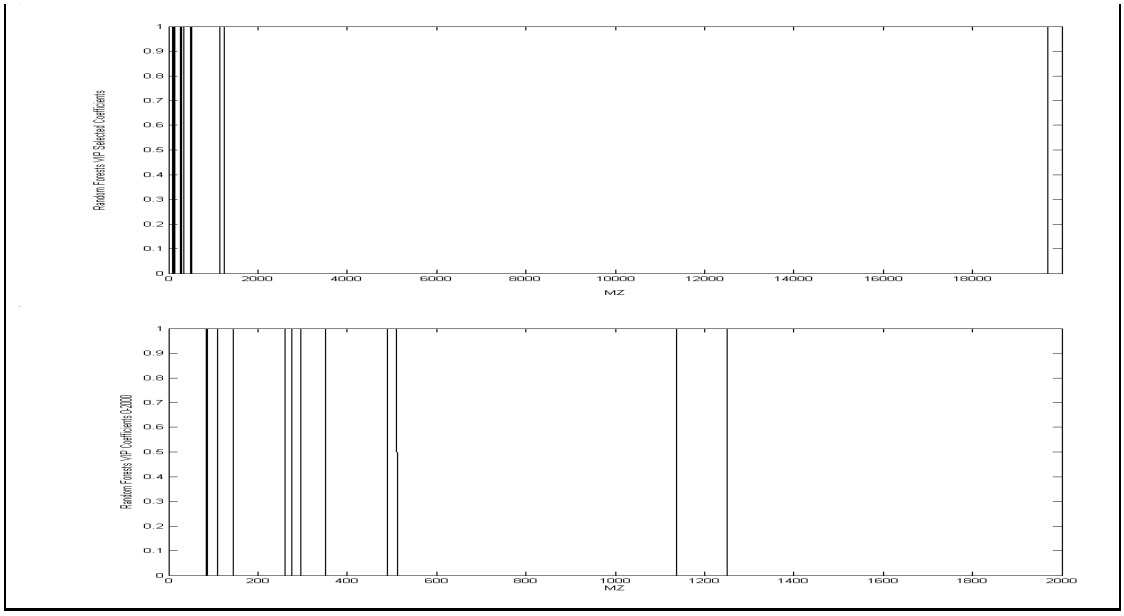

Figure 4.4 Inverse wavelet transform of the coefficients found in by Random Forests

4.5 Conclusion

We have presented a wavelet based method for classification of large datasets by employing Random Forests variable reduction and TreeBoost predictions. In the demonstration given using mass spectral profiles, it was seen that a joint analysis using multiple wavelets resulted in a lower (no) false positive prediction rates of cancerous patients when compared to the results from models using a single wavelet basis. The method of Treeboost with RF reduction, while not performing as well as RF alone for this specific data set, illustrates how variable reduction for additive CART models can be performed using other tree based methods to improve the computation speed of Treeboost.

4.6 Summary

Wavelet based analysis for mass spectrometry (MS) profiles of three groups of patients are analyzed for the purpose of developing a classification model. The first step in our model uses a DWT for feature extraction, using a linear combination of Symlets, Daubechies and Coiflets wavelet bases - collectively known as a super wavelet. Random Forests and Treeboost are then used to analyze the super wavelet coefficients to form the classification model. The method is illustrated using the publicly available prostate SELDI-TOF MS data from the American National Cancer Institute (NCI). The NCI data consists of 322 MS profiles with 15154 M/Z ratios, comprising of 69 malignant, 190 benign and 63 control patients, which we randomly divided into 70 % training and 30 % testing. From the Random Forest models, the super wavelet performed 2.7% to 5.7% better than other single wavelet types to give a 100% test set prediction rate for cancerous patients.

Chapter 5

Joint Multiple Adaptive Wavelet Regression Ensembles

5.1 Introduction

Wavelet pre-processing of spectral data has lead to increased predictability and model simplification in regression applications when compared to traditional pre-processing techniques like PCA or PLS. However, the issue of wavelet selection for pre-processing is a topic of interest since there are an infinite multitude of wavelets. Many authors have identified preferences for one type of wavelet over another for a given data set and regression method [8, 44, 68] leading to the idealism of *not all wavelets are made equal*. This chapter considers the challenge of wavelet basis selection for regression with a high number of juxta-positional explanatory variables, where the explanatory variables are in the form of near infra-red (NIR) spectra.

Modern NIR instruments measure reflectance or transmission of a substance at several hundreds of equally spaced wavelengths, typically in the range of 800nm to 2500nm. The measured NIR spectrum curve itself is comprised of a superposition of localised spectral curves, each of which is not usually directly observable. In the most simplistic case, the underlying spectral curves are non-overlapping which leads to a direct and trivial implementation of the Beer-Lambert-Bouguer law, where absorbance is proportional to concentration [69]. More realistically, the underlying signals overlap which results in a non-linear extension to the Beer-Lambert law where the signal of interest is usually masqueraded by a more dominate signal(s). Feature extraction is typically trialled to elicit the desired signal thus reverting to the trivial case.

There are two main classes of feature extraction methods which are typically used to improve spectral calibrations. The first is the factor based methods such as Principal Component Regression (PCR) [2, 70] and Partial Least Squares (PLS) [3], where the spectrum is transformed into a new set of orthogonal variables without regard to the juxta-positional nature of the spectrum. The second is the signal filter approach such as the Fourier series, where the spectrum is filtered by a *frequency analyser*. Here

frequency is meant to refer to a variable sampling frequency rather than an electromagnetic radiation frequency; the latter will be referred to by wavelengths.

With signal filter extraction methods, the spectrum (observed signal) is thought to consist of a superposition of underlying signals, where the signals can be characterised by a known functional form. For example, in Fourier analysis, the signals functional form is given by the sine function combined with a phase delay. Signal filters can be categorised into two classes: global and localised filters.

Fourier transforms are a classic example of a global filter where the basis function of the filter spans over the entire space of the observed signal. The Discrete Wavelet Transform (DWT) and the Gabor Transform are examples of localised signal filters, whose filter basis functions span a finite bandwidth which is localized to a small region of the observed spectrum [7]. Most spectra consist of many overlapping signals and the desired signal in regression applications is widely believed to be restricted to a portion of the measured signal. Due to this overlapping structure, localised signal filters are ideal for feature extraction to improve multivariate calibrations.

Unlike the Fourier transform, wavelet transforms can be created from a multitude of basis functions that range from smoothly varying wavelets (basis function) to seemingly un-wielding chaotic wavelets. Most works to date utilise wavelet transforms that use mathematically derived wavelets such as Daubechies or Morlet wavelets [6, 7]. While Morlet and Daubechies wavelets have convenient mathematical properties, such as minimal phase distortion or maximum symmetry, they were not designed for unknown signal feature extraction as is used in multivariate calibrations. Thus, it is more likely that a different wavelet basis, one derived for the task at hand, will yield a more favourable calibration.

Wavelets, as used in the DWT, have been shown to be highly effective in improving the performance of calibration type problems in many fields of NIR spectroscopy [15]. In most applications of DWT, to spectroscopy calibration problems, a single wavelet type is used in the feature extraction process. This generally assumes homogeneity of underlying signals across the breath of the spectrum. However, if the underlying signals are heterogeneous throughout the spectrum, different wavelet basis at different parts of

the spectrum may offer further advantages in feature extraction for calibration development. This then leads to the choice of which wavelets to use and where in the spectrum to apply the DWT.

Choosing wavelet types can be simplified if the underlying signal is known but this is generally not the case. It is known however, that if the correct wavelet type is chosen, the predictive performance of the model should increase. There are wavelet generating algorithms which can adapt wavelets to user definable criteria in order to help target the correct wavelet.

Adaptive wavelets are a class of wavelets which are able to traverse a large set of wavelets [7]. They iteratively update their function frequency and phase forms to match a predefined optimisation criteria. Optimisation criteria can be defined in terms of a calibration statistic, thus adaptive wavelets provide a convenient basis to search for calibration specific wavelets. Works on wavelet PLS calibrations [12], unsupervised mapping [71], clustering [8] and experimental designs [72] using spectral data have shown that adaptive wavelets outperform conventional wavelet types. In this chapter, we will use multiple adaptive wavelets to represent features from different regions in the spectrum.

In determining where to apply wavelets in the spectrum it is generally not known prior where the best predictive positions are. In regression applications it is usually unnecessary to represent all features in a spectrum to form an accurate calibration. For example, stepwise linear regression (SLR) iteratively includes and removes predictor variables so that a relatively small number of variables are used in the final predictive regression model.

Method selection techniques like stepwise linear regression (SLR) are suitable for datasets with very few predictor variables but are intractable when a large number of variables are considered such as in a NIR dataset. For example, if 700 wavelengths are used, in the first iteration of SLR, 700 models are searched with one wavelength selected. In the second iteration, 244,650 models are spanned for two selected variables, 56,921,900 models by the third iteration and a massive 991,860,000 models by the fourth iteration for four chosen variables. Modern stochastic variable selection

methods such as Random Forests [27], Classification and Regression Trees [73] and Bayesian Metropolis regression [74] offer alternative methods for discovering predictive models when there are a large number of variables relative to the number of observations.

Stochastic regression methods initially search a large range of potential models to determine an estimate of the likelihood of variable importance. The variable importance estimates are subsequently used to focus future model searches. For example in a Bayesian Metropolis regression method used by Brown *et al.* [74], the posterior probability of variable importance is estimated by trialling multiple random Markov chain Monte Carlo (MCMC) runs before the variable importance list is used in a Metropolis-Hastings search algorithm to find "good" prediction models. Typically many "good" models are found during the model search process, all of which can be used simultaneously in a model ensemble to minimise model bias and improve the overall model prediction on future samples [75].

In this chapter, the Bayesian Metropolis method developed by Brown *et al.* [74] is used as the variable selection and regression technique since the method focuses on selecting regression models with few (less than 10) variables in the final prediction models. This small model criterion was imposed as it is thought that only a small number of wavelet extracted features would be required to build useful predictors. The chosen regression method also allows for multiple constituents to be predicted simultaneously.

Multiple constituent prediction models generally result in more accurate predictors compared to multiple single constituent models [16]. Brown *et al.* [16] has previously used single wavelets in their regression method which demonstrates a substantial improvement to conventional regression techniques. The method by Brown *et al* [16] also facilitates selection of wavelet coefficients from various levels within the DWT, so band selection prior to regression is no longer necessary.

Applying adaptive wavelets with Bayesian Metropolis regression creates a problem of when to optimise the adaptive wavelets. Optimised adaptive wavelets are based on an initial random wavelet then adapted to maximise a goodness of fit criterion. Naturally the wavelet optimisation cannot be applied to the entire spectrum, so the optimisation

needs to occur *after* variable/model selection. Meaning that the variable selection is done on the features extracted using a random wavelet. This introduces another stochastic component being the initial random wavelet.

To overcome the random wavelet issue, multiple random wavelets with varying wavelet parameters are trialled. This inturn produces many more prediction models, all of which include some measure of model uncertainty being the optimised wavelets and the position within the spectrum the wavelets are applied to. The multiple optimised wavelet models can be amalgamated using ensemble methods similar to those used for stochastic regression.

Ensemble methods are used to combine a number of models in order to reduce the predictive error for future samples [76]. The basic premise for ensemble modeling is that each individual model contains uncertainties, which in turn, inflate the error of future samples. Therefore, a combination of many models will lead to an averaging out of the inflated errors of future samples.

There are many methods to form an ensemble with the most popular being: Bayes modal averaging (BMA), Bagging [61], Boosting (arcing) and Stacking [76]. Bayes model averaging combines models based on the posterior distribution of the models. During the model search of the Bayes Metropolis method by Brown *et al.*, the posterior distribution of the models had been estimated, but only for the initial random wavelet. Since the adaptive wavelets are optimized after the model search is computed, the posterior distribution estimated by the Metropolis search is longer valid.

Bagging and Boosting can overcome the limitation of the Bayes factors by using re-sampling methods to determine model variability and thus the weighting of each particular optimized model, however, the time required to undertake these methods in this application is prohibitive. Stacking is a least squares method of forming a linear combination of different predictors to arrive at an ensemble. Stacking does not rely upon posterior/prior distributions and can be used in conjunction with bootstrapping methods to mitigate over fitting on small data sets.

The methodology used in this chapter is as follows:

1. Apply a random wavelet to the spectra
2. Select regression models based on the random wavelet coefficients
3. Optimise the wavelet coefficients for the models in 2.
4. Repeat steps 1-3 to represent the initial random wavelet space
5. Form a Stacked model ensemble using the optimised wavelet models.

The following sections briefly describe the theory used in the methodology, the parameter settings and a regression example of NIR spectroscopy.

5.2 Theory

5.2.1 Discrete Wavelet Transform (DWT)

The discrete wavelet transform (DWT) [77] has become a standard tool for feature extraction, signal analysis and compression. Most applications of the DWT method use a "two-banded" system which consists of a scaling function, φ, and a single wavelet function, ψ. However, there exists a less popular "m-banded" DWT which utilizes the scaling function, φ, and m-1 wavelet functions, $\psi^{(s)}, s = 1,...,m-1$ [7]. The benefits for using the m-banded DWT include (i) the ability to use linear phase wavelets - which is not possible using the 2-banded DWT with orthogonal wavelets, (ii) increased frequency bandwidth isolation and, (iii) a larger range of possible frequencies and phase forms [7]. It is the latter reason for which the m-banded DWT is used in this investigation since the wavelet characteristics for regression are unknown and a search for appropriate wavelets is necessary.

The formulation of the m-banded DWT is similar to the 2-banded system which implements an iterative cascading algorithm. For the m-banded DWT, the cascade is described by the pair of equations:

$$\psi^{(s)}(t) = \sqrt{m} \sum_{k=-\infty}^{\infty} w_k^{(s)} \varphi(mt-k) \qquad s = 1,...,m-1 \tag{5.1}$$

$$\varphi(t) = \sqrt{m} \sum_{k=-\infty}^{\infty} \ell_k \varphi(mt-k) \tag{5.2}$$

where $w_k^{(s)}$ are the wavelet filter coefficients for the s^{th} wavelet and ℓ_k are the scaling filter coefficients. A function f is then represented by a wavelet series as

$$f(t) = \sum_{s=1}^{m-1} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_{j,k}^{(s)} \psi_{j,k}^{(s)}(t) + \sum_{k=-\infty}^{\infty} c_{j,k} \varphi_{j,k}(t) \tag{5.3}$$

with wavelet coefficients $d_{j,k}^{(s)} = \int f(t) \psi_{j,k}^{(s)}(t)\, dt$ and scaling coefficients $c_{j,k} = \int f(t) \varphi_{j,k}(t)\, dt$. Both coefficients describe features of the function f at the spatial location $m^{-j}k$ and the frequency proportional to m^j (or scale j). A pictorial example of the m-banded DWT is illustrated in Figure 5.1.

For a discretely sampled function, $\mathbf{x} = (x_1, \ldots, x_p);\ p = m^j$, with equally spaced points, the DWT is implemented as a recursive multiplication of linear filters. For illustration, this can be as:

$$\mathbf{z} = \mathbf{W}\mathbf{x} \tag{5.4}$$

with $\mathbf{W}$ an orthogonal m-banded wavelet matrix, and $\mathbf{z}$ a banded vector of scaling coefficients and wavelet coefficients. Different wavelet bands in $\mathbf{z}$ correspond to the different scales: j=1,…,J.

5.2.2 Adaptive Wavelet (AW) matrix

The following section describes how the matrix $\mathbf{W}$ in (5.4) is generated by an adaptive wavelet (AW) generation algorithm. There exist several wavelet generating algorithms that design task specific wavelets, also known as adaptive wavelets, such as Lifting [78], Angular Quadature Mirror Filtering [12, 79], and Pollen Factorization [17]. It is the Pollen factorization that is best suited to this particular application since it enables m-banded wavelets.

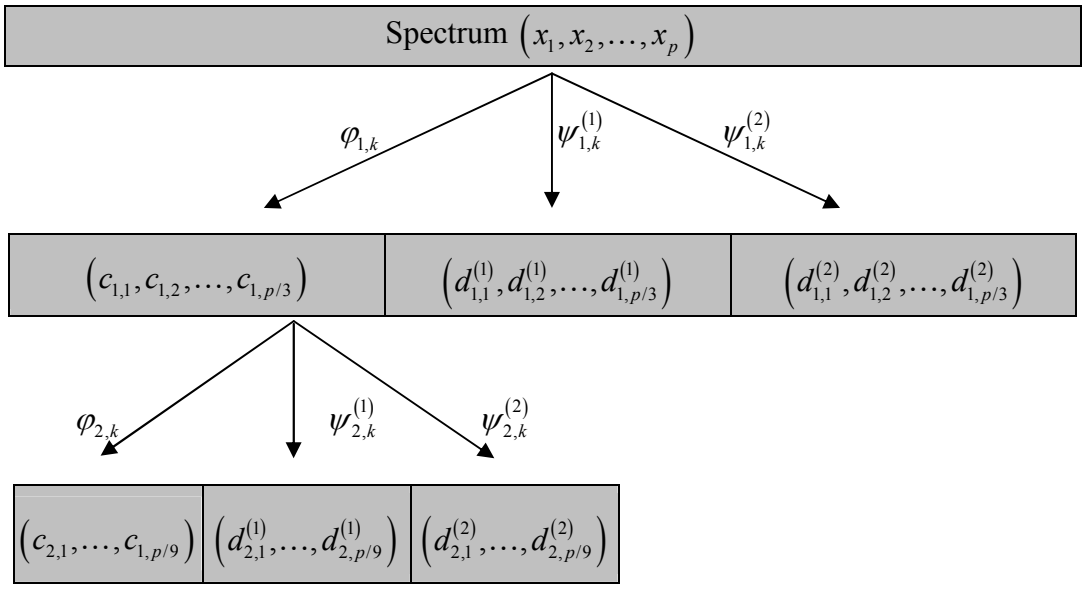

Figure 5.1 Pictorial representation of a three banded (m = 3) discrete wavelet transform where the DWT has been applied twice to the original spectrum.

Another advantage of the Pollen factorization is that the *m*-banded wavelet matrix in Equation (5.4) can be parameterized into *q+1* normalized vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$; where the number of filter coefficients in the scaling function (and the wavelet functions) is $mq+1$. These normalized vectors can be iteratively updated in order to extract user defined features. For a comprehensive account of the theory of the Pollen Factorization, the reader is referred to [17].

The Pollen factorization can be summarized in the following steps:

(1) Define the integer values for *m* and *q*

(2) Initialize the normalized vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$

(3) Construct $\mathbf{W}$ from $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$

(4) Perform the DWT

(5) Iteratively update $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$ until a converge criteria is meet.

In this study, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ and $\mathbf{v}$ are initially assigned elements from a uniform distribution, which in previous supervised studies are shown to converge based on similar optimization criteria as detailed in section 5.3 [8, 71].

5.2.3 Multivariate regression model

The basic formulation of the following multivariate regression model primarily follows Lindley [80] however was also influenced by later work performed by Brown[74]. Let $\mathbf{Y}$ denote the $n \times r$ matrix of observed values of the responses and let $\mathbf{X}$ be the $n \times p$ matrix of predictor variables. The standard multivariate normal regression model, conditional on $\boldsymbol{\alpha}, \mathbf{B}, \boldsymbol{\Sigma}, \text{and } \mathbf{X},$ has the form

$$\mathbf{Y} - \mathbf{1}_n \boldsymbol{\alpha}^{\mathrm{T}} - \mathbf{X}\mathbf{B} \sim N\left(\mathbf{I}_n, \boldsymbol{\Sigma}\right) \tag{5.5}$$

With $\mathbf{1}_n$ a $n \times 1$ vector of ones, $\boldsymbol{\alpha}$ a $r \times 1$ vector of intercepts, $\mathbf{B} = \left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r\right)$ a $p \times r$ matrix of regression coefficients and $N\left(\mathbf{I}_n, \Sigma\right)$ is the matrix-variate normal distribution [81] defined by the shape parameters: $\mathbf{I}_n$, a $n \times n$ identity matrix, and Σ, the $r \times r$ error covariance matrix . Without loss of generality the columns of $\mathbf{X}$ have assumed to have been centred by subtracting their means.

The unknown parameters are $\boldsymbol{\alpha}, \mathbf{B}, \text{and } \boldsymbol{\Sigma}$. A conjugate prior for model (5.5) is as follows [16]: first given $\boldsymbol{\Sigma}$,

$$\boldsymbol{\alpha}^{\mathrm{T}} - \boldsymbol{\alpha}_0^{\mathrm{T}} \sim N\left(h, \boldsymbol{\Sigma}\right) \tag{5.6}$$

and independently,

$$\mathbf{B} - \mathbf{B}_0 \sim N\left(\mathbf{H}, \boldsymbol{\Sigma}\right) \tag{5.7}$$

Where H is the shape parameter for matrix-variate normal distribution of $\mathbf{B} - \mathbf{B}_0$ [81]. The marginal distribution of $\boldsymbol{\Sigma}$ is then

$$\boldsymbol{\Sigma} \sim IW\left(\delta; \mathbf{Q}\right) \tag{5.8}$$

Where $IW\left(\delta; \mathbf{Q}\right)$ is an inverse Wishart distribution with a scale matrix $\mathbf{Q}$ and shape parameter δ [81].

Since little prior knowledge is known about $\boldsymbol{\alpha}$, we let $h \to \infty$ to represent a vague prior and take $\mathbf{B}_0 = \mathbf{0}$, leaving the specification of $\mathbf{H}, \mathbf{Q}$ and δ to incorporate prior knowledge of the particular application.

In applying the Discrete Wavelet Transform (DWT) to the spectra, $\mathbf{X}$, model (5.5) can be expressed as:

$$\mathbf{Y} - \mathbf{1}_n \boldsymbol{\alpha}^{\mathrm{T}} - \mathbf{X}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{B} \sim N\left(\mathbf{I}_n, \boldsymbol{\Sigma}\right) \tag{5.9}$$

Or alternatively

$$\mathbf{Y} - \mathbf{1}_n \boldsymbol{\alpha}^{\mathrm{T}} - \mathbf{Z}\tilde{\mathbf{B}} \sim N\left(\mathbf{I}_n, \boldsymbol{\Sigma}\right) \tag{5.10}$$

Where $\mathbf{Z} = \mathbf{X}\mathbf{W}^{\mathrm{T}}$ is the matrix of wavelet coefficients and $\tilde{\mathbf{B}} = \mathbf{W}\mathbf{B}$ is a matrix of regression coefficients. The DWT also affects the prior for $\tilde{\mathbf{B}}$:

$$\tilde{\mathbf{B}} \sim N\left(\tilde{\mathbf{H}}, \boldsymbol{\Sigma}\right) \tag{5.11}$$

With $\tilde{\mathbf{H}} = \mathbf{W}\mathbf{H}\mathbf{W}^{\mathrm{T}}$ [82]. The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ are unaltered by the DWT as are their prior distributions in (5.6) and (5.8).

In calculating $\tilde{\mathbf{H}}$ using a single wavelet, $\mathbf{W}$ corresponds to the DWT and the two-dimensional DWT (DWT2) can be utilize to reduce the computation time [82]. However, when multiple wavelets are used, the DWT2 method can no longer be used since $\mathbf{W}$ no longer corresponds to the typical recursive DWT.

5.2.4 Variable selection

Not all wavelet coefficients in the DWT will be useful for predictive purposes, so a method of variable selection is used to isolated potentially predictive sets of wavelet coefficients. A latent binary vector $\boldsymbol{\gamma}$ of length p indicates which predictor variables (wavelet coefficients) are to be included in the model (5.10) [74, 80]. The binary vector includes wavelet coefficients from all levels within the DWT. If the j^{th} element of $\boldsymbol{\gamma}$, γ_j is zero, then the j^{th} column of $\mathbf{Z}$ is excluded from the model.

With the assumed prior expectation of $\tilde{\mathbf{B}}$ set to zero, then

$$\tilde{\mathbf{B}}_{\gamma} \sim N\left(\tilde{\mathbf{H}}_{\gamma}, \boldsymbol{\Sigma}\right) \tag{5.12}$$

Where $\tilde{\mathbf{B}}_{\gamma}$ and $\tilde{\mathbf{H}}_{\gamma}$ are rows and columns of $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{H}}$ respectably where $\gamma_j = 1$. Rows and columns where $\gamma_j = 0$ are deleted from the matrix Under this prior, each row of $\tilde{\mathbf{B}}$ is modeled as having a scale mixture of the type [16]:

$$\tilde{\mathbf{B}}_{j,:} \sim \left(1-\gamma_j\right)\boldsymbol{\Phi}_0 + \gamma_j N\left(0, \tilde{h}_{j,j}\boldsymbol{\Sigma}\right) \tag{5.13}$$

With $\tilde{h}_{j,j}$ equal to the j^{th} diagonal element of $\tilde{\mathbf{H}}$ and $\boldsymbol{\Phi}_0$ being a distribution placing unit mass on the $1 \times r$ zero vector. Note, the rows of $\tilde{\mathbf{B}}$ are not independent.

Choosing a binomial prior distribution, $\pi(\boldsymbol{\gamma})$, for $\boldsymbol{\gamma}$ takes the elements, γ_j, to be independent with $\text{Prob}\left(\gamma_j = 1\right) = \varpi_j, \text{Prob}\left(\gamma_j = 0\right) = 1 - \varpi_j$ with the hypermeters ϖ_j to be specified. The use of mixture priors for variable selection in multivariate regressions is further detailed by Brown [74].

5.2.5 Posterior distribution of $\boldsymbol{\gamma}$

The posterior distribution of $\boldsymbol{\gamma}$ given the data, $\pi(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{Z})$, gives a posterior probability to each of the possible states for the vector $\boldsymbol{\gamma}$. This posterior arises from the combination of a likelihood, that gives a high weight to subsets explaining a high proportion of the variance in the responses, $\mathbf{Y}$, and a prior for $\boldsymbol{\gamma}$, that penalizes large subsets. The posterior distribution, $\pi(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{Z})$, is computed by integrating $\boldsymbol{\alpha}, \mathbf{B}$ and $\boldsymbol{\Sigma}$ from the joint posterior distribution. With the vague prior for $\boldsymbol{\alpha}$, $(h \to \infty)$, the parameter is essentially estimated by the mean of $\mathbf{Y}$ from the calibration set. To simplify the formulae, the columns of $\mathbf{Y}$ have been mean centred. Full details of the derivation of the posterior distribution for $\boldsymbol{\gamma}$ is given in [74] with the main result:

$$\pi(\boldsymbol{\gamma}\,|\,\mathbf{Y},\mathbf{Z}) \sim g(\boldsymbol{\gamma}) = \left\{\left|\tilde{\mathbf{H}}_{\gamma}\right|\left|\mathbf{Z}_{\gamma}^{\mathrm{T}}\mathbf{Z}_{\gamma} + \tilde{\mathbf{H}}_{\gamma}^{-1}\right|\right\}^{-r/2}\left|\mathbf{Q}_{\gamma}\right|^{-(n+\delta+r-1)/2}\pi(\boldsymbol{\gamma}) \tag{5.14}$$

with $\mathbf{Q}_{\gamma} = \mathbf{Q} + \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{Z}_{\gamma}\left(\mathbf{Z}_{\gamma}^{\mathrm{T}}\mathbf{Z}_{\gamma} + \tilde{\mathbf{H}}_{\gamma}^{-1}\right)^{-1}\mathbf{Z}_{\gamma}^{\mathrm{T}}\mathbf{Y}$, where $\mathbf{Z}_{\gamma}$ is $\mathbf{Z}$ with the columns which $\gamma_j = 0$ deleted, and $g(\boldsymbol{\gamma})$ is the relative probability of the regression for model $\boldsymbol{\gamma}$.

5.2.6 Metropolis search

Equation (5.14) gives the posterior probability of each of the 2^p different $\boldsymbol{\gamma}$ vectors , each of which represent a different subset of wavelet coefficients. Computing these posterior probabilities then allows "good" wavelet coefficients to be ascertained.

When *p* is greater than approximately 25 there are too many subsets to fully compute $\pi(\boldsymbol{\gamma}\,|\,\mathbf{Y},\mathbf{Z})$. Fortunately, simulation methods can be used to find $\boldsymbol{\gamma}$ vectors with relatively high posterior probabilities, which can then be used to identify wavelet coefficients with high marginal probabilities where $\gamma_j \approx 1$. Here a Metropolis search [83, 84] is used to find the high yielding $\boldsymbol{\gamma}$ vectors.

Since the marginal probabilities for $\boldsymbol{\gamma}$ are of interest, a broad range of $\boldsymbol{\gamma}$ vectors need to be trialed, hence the Metropolis search algorithm is employed. Metropolis searches have been successfully used in variable selection for regression applications by George and McCulloch [84], Raftery et al. [85] and Brown *et al.* [16]. Other searches which are potentially useful are simulated annealing [86] and genetic algorithms [87], where both methods were investigated in a similar regression application [88].

The Metropolis search starts from a randomly chosen $\boldsymbol{\gamma}^0$ and then moves through a sequence of further values of $\boldsymbol{\gamma}$. At each step the algorithm generates a new candidate $\boldsymbol{\gamma}$ by randomly modifying the current $\boldsymbol{\gamma}$ vector. Two types of modification are used:

1. Add or delete a component,
2. Randomly choosing one *j* in the current $\boldsymbol{\gamma}$ and inverting its value. The probability of choosing each component is ϕ.

The new candidate model $\boldsymbol{\gamma}^*$ is accepted with probability

$$\min\left\{\frac{g(\boldsymbol{\gamma}^*)}{g(\boldsymbol{\gamma})},1\right\} \tag{5.15}$$

A more probable model will always be accepted; however, the scope to include less probable models increases the scope of the search space and hence produces a more accurate simulation for the marginal probabilities for $\boldsymbol{\gamma}$.

To ensure that the Metropolis search spans a sufficiently large search space, and does not permute around a local minima, multiple starting positions are used. The multiple chains of $\boldsymbol{\gamma}$ are then concatenated to form the marginal distribution.

5.2.7 Stacking ensembles

The basic premise for model ensembles is: If f_i are the predictions from the M individual models, $i = 1$ to M, then let $\bar{f}$ be the mean of the amalgamated predictions. The $f_i's$ assumed to be identically distributed, share a common variance V and are unbiased, but not necessarily independent [89]. Therefore,

$$\begin{aligned} Var\left(\bar{f}\right) &= 1/M^2 \sum_{i=1}^{N} Var\left(f_i\right) + 2/M \sum_{i<j} Cov\left(f_i, f_j\right) \\ &= V/M + 2/M \sum_{i<j} Cov\left(f_i, f_j\right) \end{aligned} \tag{5.16}$$

If all $f_i's$ are equal then nothing is gained by averaging. If the $f_i's$ are uncorrelated then $Var\left(\bar{f}\right) = V/M$, so averaging is expected to work well if the $f_i's$ are diverse when $Cov\left(f_i, f_j\right)$ are small.

Stacking is a least squares method of forming a linear combination of different predictors to arrive at an ensemble. Stacking does not rely upon posterior/prior distributions and can be used in conjunction with bootstrapping methods to mitigate over fitting on small data sets. In the simplest form, stacking restricts the ensemble to:

$$f_\varepsilon = \sum_i \mu_i f_i \tag{5.17}$$

where the μ_i are the weights for each predictive model f_i. Here we select μ_i with the following constraints to minimize the mean squared error of the ensemble:

$$\sum_i \mu_i = 1$$
$$\mu_i > 0;\ i = 1, \ldots, M \tag{5.18}$$

In minimizing the mean squared error of the ensemble, the potential to over fit on the training data can be mitigated by re-sampling methods [76]. In this study, bootstrapping was used to generate a collection of weights for each model, $\mu_{i,j}$. The weights of each separate model were then averaged to calculate the final weight that would be used in the ensemble for each model.

The bootstrap used was to replace each model f_i by $f_{i,j}$ where $f_{i,j}$ is the 3-cross fold estimate of f_i. So for each set of cross-fold estimates, $f_{i,j}$, a constrained stacked ensemble was made to generate the weights, $\mu_{i,j}$. The final weight for the model ensemble, μ_i, was taken as the average of forty bootstrapped estimates of $\mu_{i,j}$.

5.3 Methodology

To investigate the hypothesis of improving wavelet predictions using multiple wavelets, comparisons to similar models using single wavelets for feature extraction were made. The single wavelet models all follow the same methodology described in the introduction being:

1. Feature extraction from the spectra by applying the single wavelet type using the DWT
2. Model generation using the Bayes Metropolis regression method
3. Forming a model ensemble using:
    a. Constrained stacking, with and without bootstrapping, and
    b. Bayes model averaging for standard wavelets

Single wavelet models used in the comparison were standard wavelets from literature and adaptive wavelets. The standard wavelets used were Daubechies (2 and 4 tap),

Coiflets (1 and 3 tap) and Myer wavelets. The level of decomposition in the DWT for the single wavelets was one to four, the same used for the multiple adaptive wavelets.

Bayes model averaging (BMA) [75] for the single, standard wavelets types was possible since no model optimization was performed after the Bayes Metropolis regression. Using BMA as an ensemble method for the standard wavelets gave a direct comparison to analysis of the same dataset found in literature [16] and was able to assess the effectiveness of constrained stacking. The top 500 models from the Metropolis search, with the highest likelihood, were used in the model ensemble of each BMA and constrained stacking ensemble that was derived when using a single standard or adapted wavelet.

Adaptive wavelet models were generated by applying following methodology:

1. Apply a random wavelet to the spectra
2. Select regression models based on the random wavelet coefficients
3. Optimise the wavelet coefficients in the models in 2.
4. Repeat steps 1-3 to represent the initial random wavelet space
5. Form a Stacked model ensemble using the optimised wavelet models.

The top forty models for each combination of *m, q* and *J* with the highest likelihood scores from step two were use to optimize the adaptive wavelets. For each model the non-zero elements of γ indicate an adaptive wavelet that needs to be optimized. The optimization criteria used for the adaptive wavelet algorithm minimizes the mean squared error for each of the models such that:

$$MSE_i = \sum_{\tau=1}^{n} \left(Y_{i,\tau} - Y_{\tau} \right)^2 \Big/ n \tag{5.19}$$

where

$$Y_{i,\tau} = \sum_{j} b_{i,j} d_{i,j,\tau} \tag{5.20}$$

where $Y_{i,\tau}$ is the model predictions for the i^{th} model, $d_{j,\tau}$ is the adaptive wavelet coefficient for the j^{th} non-zero component of $\boldsymbol{\gamma}$ and $\tau = 1,\ldots,n,$ with n being the number of samples in the calibration data set, and $b_{i,j}$ are the regression coefficients for the i^{th} model. For example, if the third model from step two has five non-zero components for $\boldsymbol{\gamma}$, then five adaptive wavelets are optimized jointly to minimize the MSE.

Once the wavelets for the models have been optimized, the posterior model probabilities are longer valid and cannot be used to determine a model ensemble using BMA. Consequently a constrained stacking model ensemble is formed with the forty adaptive wavelet models which have the same combinations of *m, q* and *J*.

All combinations of *m, q* and *J* were trialed for adaptive wavelets and the top forty models from each combination were then used jointly to form another constrained stacking ensemble.

5.3.1 Near infrared spectra data

The methods outlined in this chapter are implemented on a reference data set widely available for general use within the chemometrics community. The data set pertains to composition of biscuit dough and is fully described by Osborne [90]. A brief summary of the data follows.

Biscuit dough spectra were derived from a study that investigated the feasibility of using NIR spectroscopy for measuring the constituents - fat, sucrose, dry flour and water of unbaked biscuit dough. Two similar sample sets were made from a standard recipe and varied to provide a range for each of the four constituents under investigation. From each sample set, a NIR reflectance spectrum from 1100 to 2498 nm at 2 nm increments was measured on 40 dough samples. A total of, 78 spectra were recorded and divided equally into a calibration and test set.

We define $\mathbf{Y}$ and $\mathbf{Y}_f$ to represent the matrices of the response variables while rows of $\mathbf{X}$ and $\mathbf{X}_f$ represent the NIR spectra for the calibration and validation sets respectively.

5.3.2 Parameter settings

To apply various components in section 5.3, a range of parameters need to be defined for the: (i) adaptive wavelet algorithm (ii) Multivariate regression model and (iii) Metropolis search.

5.3.2.1 Adaptive wavelet parameters

Three parameters, *m, q and J,* need to be defined to implement the adaptive wavelets. Parameter *J* is the maximum number of recursive applications of the wavelet transform. Since it is unknown which wavelets are predictive, a large set of possible combinations of *m, q* and *J* are trialled. Each set of adaptive wavelet parameters are repeated four times as the initial adaptive wavelet starting vectors, $\mathbf{u}_i$ and $\mathbf{v}$, are randomized.

The range of values for *m, q* and *J* are limited by the sampling resolution of the spectrum. At each iteration of the DWT, the signal size is reduced by a factor *m*, so maximum size of *J* is defined by the minimum positive integer value of n/m^J . The number of sampling points *n* can be truncated to satisfy the integer requirement. Furthermore, the number of filter coefficients in the scaling function (and in the wavelet functions) is $N_f = mq+1$. This places an addition limit on *J* where $n/m^J \geq N_f$. Abiding by these restrictions, in this study where *n* = 700, a range possible values for *m, q* and *J* are $\{2,\ldots,8\},\{2,\ldots,8\}$ and $\{1,\ldots,6\}$ respectively. As N_f will become large for large values for *m* and *q,* which is impractical for small data sets, N_f was restricted to ≤ 10.

5.3.2.2 Multivariate regression model settings

Values for the parameters $\mathbf{H}, \delta$ and $\mathbf{Q}$ from equations (5.7) and (5.8) need to be specified, as well as the hyperparameters ϖ_j for the prior distribution of $\pi(\boldsymbol{\gamma})$. Since little information is known, vague priors are used.

For $\boldsymbol{\Sigma}$, let $\delta = 3$ as this is the smallest integer value available so the expectation of $\boldsymbol{\Sigma}$, $E(\boldsymbol{\Sigma}) = \mathbf{Q}/(\delta - 2)$, exists. The scale matrix $\mathbf{Q}$ is chosen as $\mathbf{Q} = \kappa \mathbf{I}_r$ with $\kappa = 0.05$, which is comparable in size to the expected error variances of the standardized $\mathbf{Y}$ given $\mathbf{X}$. With δ small, the choice of $\mathbf{Q}$ is not critical [16].

Choice for **H** should reflect the knowledge that the **B** coefficients are locally correlated and smooth. A first-order auto regressive process with $h_{i,j} = \sigma^2 \rho^{|i-j|}$ was used for **H**, reflecting the prior knowledge and keeping **H** in a simplified form. Integrating $\boldsymbol{\alpha}, \mathbf{B}$ and $\boldsymbol{\Sigma}$ from the joint distribution given by (5.5), (5.6), (5.7) and (5.8) for the regression on the full non-wavelet-transformed spectra, with $h \to \infty$ and $\mathbf{B}_0 = \mathbf{0}$ (ie. only mean centering on **X** and SNV transformation [37] of **Y**) results in:

$$f \propto \left|\mathbf{K}\right|^{-r/2} \left|\mathbf{Q}\right|^{(\delta+r-1)/2} \left|\mathbf{Y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{Y}\right|^{-(\delta+n+r-1)/2} \tag{5.21}$$

where

$$\mathbf{K} = \mathbf{I}_n + \mathbf{X}\mathbf{H}\mathbf{X}^{\mathrm{T}} \tag{5.22}$$

With $\kappa = 0.05$ and $\delta = 3$, equation (5.21) is therefore a function, via **H**, of σ^2 and ρ. Values of $\sigma^2 = 254$ and $\rho = 0.32$ were derived by maximizing the type II likelihood [91] of equation (5.21). Once the **H** and the wavelet(s) **W** are chosen, $\tilde{\mathbf{H}} = \mathbf{W}\mathbf{H}\mathbf{W}^{\mathrm{T}}$ is calculated.

Hyperparameters ϖ_j, for the prior binomial distribution of $\pi(\boldsymbol{\gamma})$, were set to equal a constant value, $\varpi = \varpi_j$, across all values of $j \in \{1,..,p\}$. This assumes that, initially, it was unknown which wavelet coefficients would be predictive. The value of ϖ was chosen so that small subsets (ie. $\boldsymbol{\gamma}$'s with a small number of ones) would eventually dominate by having a higher likelihood. This was chosen based on previous experiences [16] that good predictions can be done using 20 or so selected spectral points in similar regressions. Hence, ϖ in the prior for $\boldsymbol{\gamma}$ was chosen so that the expected model size was $p\varpi = 20$.

5.3.2.3 Metropolis search settings

The parameters for the Metropolis search ϕ and iteration length were set to 1/2 and 100,000 respectively. For the initial starting vector, $\boldsymbol{\gamma}^0$, four positions were trialed over four different searches. The starting vectors were (i) all even integer positions set to one, (ii) all odd positions set to one and (iii & iv) random sequences of ones derived

from a Bernoulli distribution with $\text{Prob}(\gamma_j) = 1/2$. Computation of $g(\boldsymbol{\gamma})$ was done using the QR decomposition.

5.3.3 Computation

Model development was performed on a 2.4 GHz, dual quad core Intel computer with Windows XP as the operating system. Matlab version 7 was used to implement the methodology and the Matlab Optimisation toolbox was used to optimize the adaptive wavelets.

5.3.4 Analysis by previous methods

For all the analyses reported, the spectral data and response variables were mean centered with respect to the calibration data set. The responses were also scaled to give each of the variables unit variance in the calibration set. This pre-processing the data does not influence the analysis of the previous methods, it only serves to simplify the prior specifications for the prior settings.

Table 5.1 Mean squared errors of the validation set using six calibration methods

Method	Fat	Sugar	Flour	Water
SMLR	0.044	1.188	0.722	0.221
Decision theory	0.076	0.566	0.265	0.176
Wavelet decision theory	0.059	0.466	0.351	0.047
Wavelet decision theory (Best model)	0.063	0.449	0.348	0.050
PLS	0.151	0.583	0.375	0.105
PCR	0.160	0.614	0.388	0.106

Osbourne *et al.* [90] used step-wise multiple linear regression (SMLR) on the individual constituents to form four calibrations. The mean squares of error (MSE) of the validation set is listed in Table 5.1. The quoted MSE has been converted back to the original scale the calibration set.

Brown *et al*. [92] fitted a multivariate Bayesian decision approach, row two in Table 5.1, and later improved the method with the addition of a DWT using a Daubechies

four-tap filter (row three in Table 5.1) and a Bayes model averager [16]. The best individual model from [16] is shown in row four, Table 5.1.

As a comparison to other standard methods, Brown *et al*. [16] derived calibrations using partial least-squares (PLS) and principle component regression (PCR), shown in rows five and six respectively of Table 5.1.

5.4 Results and Discussion

Coiflet, Daubechies and Meyer wavelets were trialed within the DWT Metropolis search algorithm, Table 5.2. There was no universal best wavelet type or DWT level that catered for all of the constituents, as the different wavelets at different DWT levels resulted in varying performances for each separate constituent. The best constrained stacking model mean squared error (MSE) for each constituent was 0.0322, 0.3404, 0.1816 and 0.0292 for fat, sugar, flour and water respectively. These are more favorable than the previous methods documented in Table 5.1.

Re-sampled constrained stacking (RCS) gave better predictive results than Bayes model averaging (BMA) for nearly all wavelet types which supports similar studies where BMA and stacking are compared [93]. Individual models in both the BMA and RCS models contained very few wavelet coefficients, with typically two to seven wavelet coefficients populating each Bayes regression, Figure 5.2. Re-sampled constrained stacking used fewer models and wavelet coefficients in the ensemble resulting in simpler ensembles than BMA, Table 5.3.

Re-sampling within the constrained stacking algorithm resulted in a more robust predictor, but with a more complex ensemble when compared to constrained stacking without re-sampling, Figure 5.3. Constrained stacking with and without re-sampling were shown to lower MSE, however the re-sampling constrained stacking resulted in a substantially lower MSE for the validation set (table withheld). The MSE for constrained stacking using Coiflet 1, level 4 was 0.100, 0.957, 0.579 and 0.047 for fat, sugar, flour and water respectively which is, for some constituents, almost double the MSE of the re-sampling constrained stacking, Table 5.2.

Re-sampled constrained stacking (RCS) over the entire set of standard wavelets, i.e. using multiple standard wavelets, gave a prediction worse than most single wavelet

(type) RCS ensembles, where the predictive MSE for the multiple standard wavelet RCS ensemble was 0.112, 0.817, 0.414 and 0.064 for fat, sugar, flour and water respectively. The decrease in performance for the multiple standard wavelet case is most likely due to the sheer number of individual models incorporated into the ensemble, near 10,000 in total. This problem of an over excess of models in the ensemble transcends the initial regression problem.

Adaptive wavelet RCS ensembles performed similarly to the standard wavelet types, Table 5.4, and with different adapted wavelet basis are better suited to different constituents. Each of the RSC ensembles in Table 5.4 (each row) consisted of forty individual models which made the problem of forming a multiple wavelet RCS ensemble tractable. Computation time for all of the standard wavelet models was approximately two hours and approximately six hours for all of the adaptive wavelet models.

The joint adaptive wavelet re-sampled constrained stacking ensemble (JAWRCSE) resulted in predictive MSE values of 0.0385, 0.3245, 0.2105 and 0.0280 for fat, sugar, flour and water respectively. This is currently the best single joint predictor for all the constituents and the best predictor for fat, sugar and water, with the Coiflet (1) level 1 providing slightly better predictive MSE for flour. The JAWRCSE provides a more accurate predictive ensemble than those formed from the adaptive wavelets sets listed as rows in Table 5.4, and the RCS ensembles derived from standard wavelets, Table 5.2.

Overall there are 156 adaptive wavelet regression models in the resultant JAWRCSE, coming from all of the adaptive wavelet sets in Table 5.4, Figure 5.4. Relatively few models are selected from each adapted set of wavelet parameters; however the JAWRCSE is far superior to ensembles formed from the adaptive wavelets sets listed as rows in Table 5.4.

Table 5.2 Re-sampled constrained stacking and Bayes model averaging (BMA) mean squared error of the validation data for each constituent using standard wavelets.

Wavelet	DWT Level	Constrained Stacking				BMA			
		Fat	Sugar	Flour	Water	Fat	Sugar	Flour	Water
Coifelt 1	1	0.0432	0.3404	0.1816	0.0373	0.1969	1.0325	0.4967	0.0772
	2	0.0739	0.4011	0.2502	0.0401	0.2461	1.2431	0.6481	0.0934
	3	0.0723	0.3634	0.2224	0.0392	0.2733	1.2139	0.5133	0.0814
	4	0.0774	0.4229	0.2470	0.0292	0.4150	1.0749	0.4318	0.0851
Coiflet 3	1	0.0322	0.3664	0.2140	0.0402	0.1703	0.7201	0.3258	0.0858
	2	0.0500	0.5038	0.2749	0.0461	0.2275	1.4164	0.6303	0.1042
	3	0.0502	0.5387	0.3097	0.0505	0.2424	0.9878	0.4272	0.0758
	4	0.0398	0.5846	0.3121	0.0642	0.2177	0.9798	0.3901	0.0621
Daubechies 2	1	0.0463	0.4043	0.2315	0.0456	0.1843	1.2351	0.5523	0.0831
	2	0.0543	0.6352	0.4045	0.0635	0.2415	1.6753	0.7006	0.0806
	3	0.0657	0.5008	0.3050	0.0421	0.2899	0.9731	0.3867	0.0546
	4	0.0644	0.3906	0.2398	0.0399	0.2270	0.9847	0.4722	0.0666
Daubechies 4	1	0.0488	0.3413	0.1973	0.0384	0.2009	1.3859	0.6199	0.1024
	2	0.0569	0.4886	0.2233	0.0491	0.2150	1.1851	0.5007	0.0677
	3	0.0506	0.4185	0.2014	0.0468	0.2123	0.9537	0.3964	0.0680
	4	0.0631	0.3566	0.1979	0.0528	0.2483	1.0164	0.4351	0.0597
dmey	1	0.0429	0.3825	0.2399	0.0364	0.2010	1.2969	0.5650	0.0888
	2	0.0557	0.6300	0.4601	0.0426	0.2202	1.6131	0.9360	0.0900
	3	0.0701	0.5573	0.3615	0.0501	0.2332	1.3628	0.5873	0.0779
	4	0.0579	0.5101	0.2932	0.0525	0.3475	1.0842	0.4045	0.0692

Table 5.3 Number of models and wavelet coefficients used in the ensembles where constrained stacking resulted in the lowest predictive MSE for each constituent.

Constituent	Wavelet	Constrained Stacking		BMA	
		models	wavelets	models	wavelets
Fat	Coifelt (3), level 1	341	344	500	407
Sugar	Coifelt (1), level 1	313	331	500	338
Flour	Coifelt (1), level 1	313	331	500	338
Water	Coiflet (1), level 4	312	332	500	417

Table 5.4 Re-sampled constrained stacking mean squared error of the validation data for each constituent using adaptive wavelets.

m	*q*	*J*	Constrained Stacking			
			Fat	Sugar	Flour	Water
4	2	2	0.0592	0.3718	0.2994	0.0246
3	3	1	0.0856	0.8002	0.7073	0.0353
2	4	2	0.0782	0.3859	0.3016	0.0315
2	3	3	0.0517	0.4503	0.2689	0.0428
3	3	2	0.0771	0.3870	0.2683	0.0385

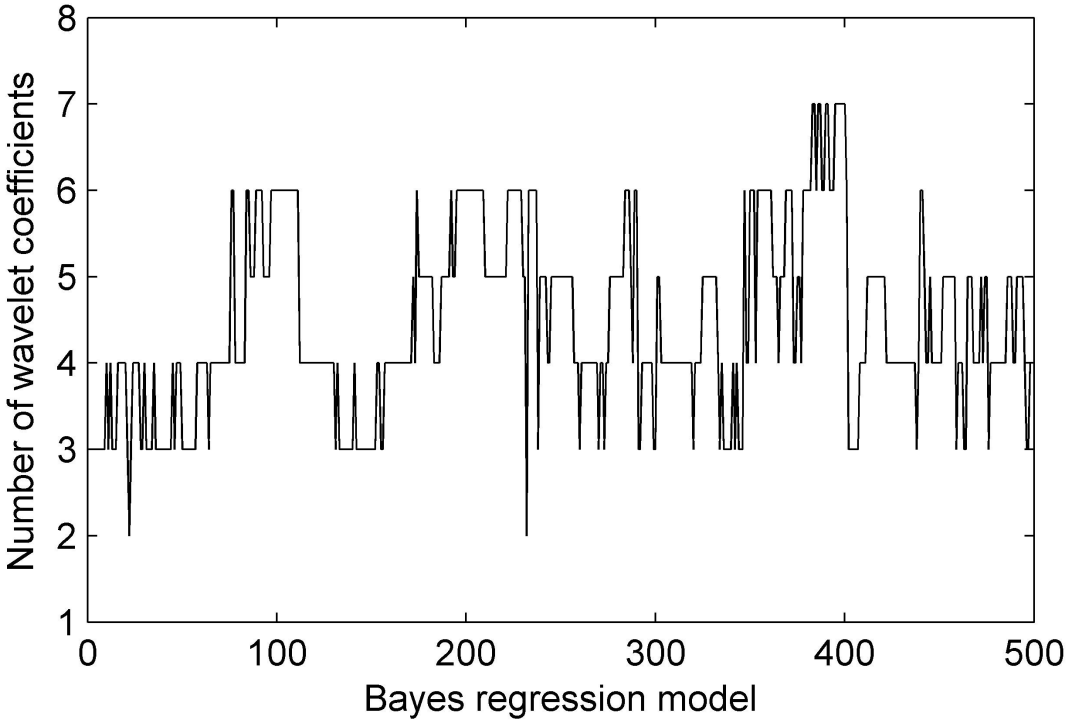

Figure 5.2 Number of wavelet coefficients in best 500 Bayes regression models generated by the Metropolis search using Coiflet 3, level 1 as the DWT

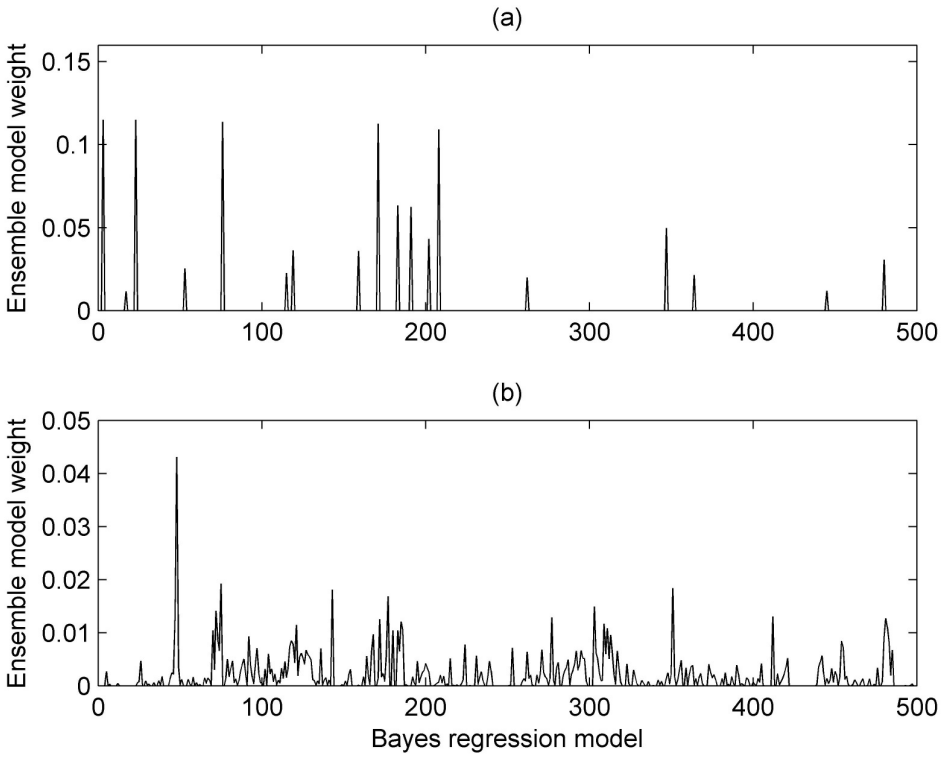

Figure 5.3 Constrained stacking ensemble weights for Coiflet (1) DWT level 4, (a) without resampling (b) with re-sampling

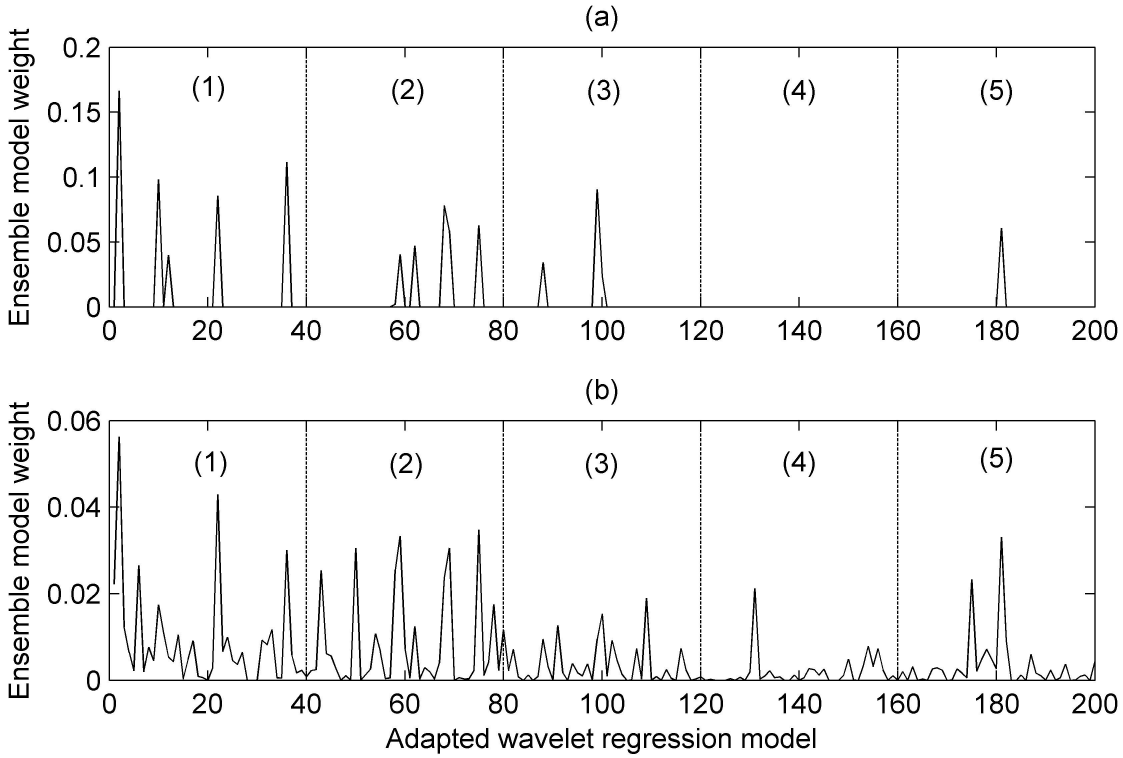

Figure 5.4 Constrained stacking ensemble weights for multiple adaptive wavelet combinations (a) without resampling (b) with resampling. Individual adaptive wavelet combinations (sets) corresponding to the rows in Table 5.4 are indicated in parenthesis

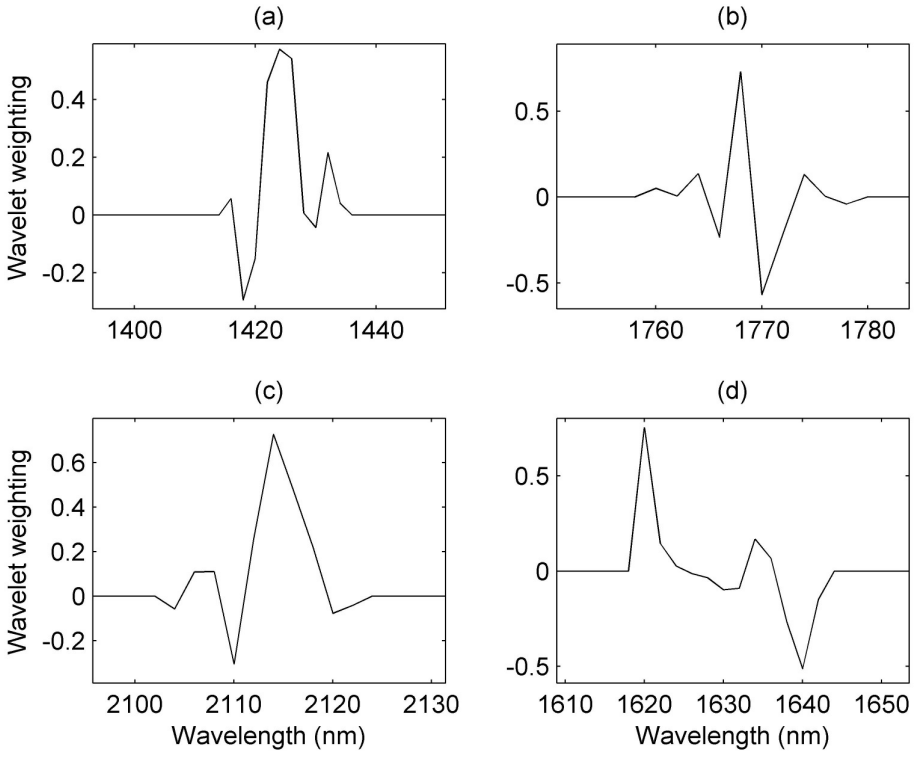

Figure 5.5 Adapted wavelets from different wavelet parameters used in the JAWRCS ensemble

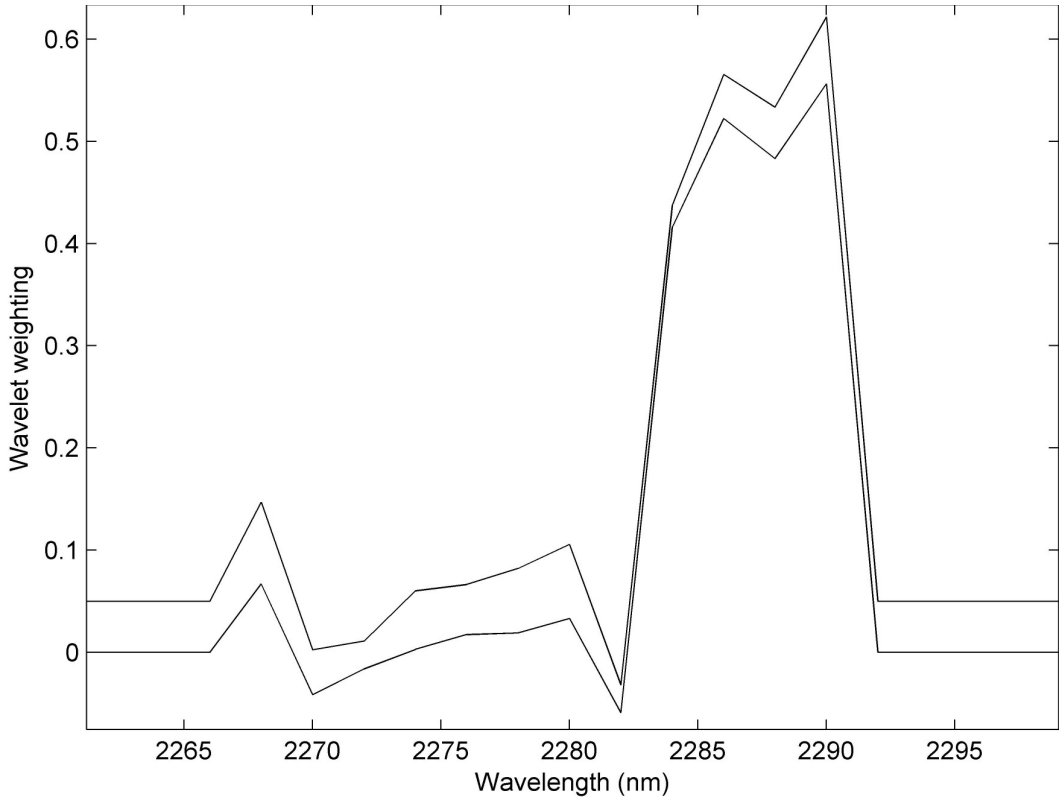

Figure 5.6 Adaptive wavelet weighting resulting from two independent models within an ensemble using a similar region of the spectrum. An offset is added to one of the adapted wavelets for clarity

Adapted wavelet ensembles in both Table 5.4 and the JAWRCS ensemble contain wavelets with varying wavelet characteristics that vary dependant upon position within the spectrum it is to be applied on, Figure 5.5. It was also observed from the JAWRCS that those models with a high ensemble model weighting typically had similar optimized wavelets within the model, Figure 5.6. This trend was observed for various wavelet filter lengths when a similar region of the spectrum was selected during the Metropolis Bayes regression search.

5.5 Conclusion

Re-sampled constrained stacking (RCS) ensembles, coupled with a discrete wavelet transform, a Bayes variable regression and a Metropolis search, were effective in producing predictive models for spectral data. The choice of wavelet within the algorithm was important as different discrete wavelet transforms (DWT) give rise to different predictive performances. There was no standard wavelet that resulted in the best RSC ensemble as was also the case with adaptive wavelets.

The use of the Bayes Metropolis search was useful in finding regions in the spectrum to use as starting points for the adaptive wavelet algorithm, however, the full usefulness of the Bayes posterior distributions approach were effectively nullified due to the optimization effect of the wavelets after the Bayes search. Alternative simpler variable selection methods such as Random Forests [27] or Classification and Regression Trees [73] could be used to form the initial point for the adaptive wavelets.

Joint adaptive wavelet RCS (JAWRCS) gave a single best multiple response ensemble with better predictive MSE than models using a single wavelet for feature extraction. The JAWRCS ensemble was composed of adapted wavelets derived from multiple sets of *m, q* and *J*. The different wavelets in the JAWRCS ensemble did utilize different information within the spectrum as the various wavelets had different characteristics, (i.e. shapes) dependant on the position within the spectrum.

A RCS ensemble using multiple standard wavelets did not result in a better ensemble compared to single standard wavelet RCS ensembles. The most likely cause for the poor performance of the multiple standard wavelet RCS ensemble was due the large amount of models (more than the number of wavelengths in the original data) considered in the RCS ensemble. This does not preclude the possibility that a permutation of standard wavelet DWT's that would give a superior RCS ensemble exists, but that the number of permutations of standard wavelets to consider is prohibitive. This is where adaptive wavelets have a definite advantage in that a large range of permutations of wavelet types can be tractably searched to produce a very good multiple wavelet, multiple response, RCS ensemble.

5.6 Summary

Multiple adaptive discrete wavelet transforms were applied during a multiple regression of spectroscopic data for the purpose of investigating the hypothesis – does the use of different wavelets, at different points, within a spectrum, elucidate predictive capability.

The model investigated was a constrained stacking regression ensemble with individual regression models chosen initially by a Bayes Metropolis search. The ensemble approach provided the ability to combine different regression models that used different types of wavelets. Models were applied to a publically available dataset, pertaining to biscuit dough, of near infrared spectra, that were measured by a FOSS 5000, and laboratory measurements of the fat, flour, sugar and moisture content.

The resultant model, which is referred to as a joint multiple adaptive wavelet regression ensemble (JMAWRE), was found to be the superior predictive model when compared to models that used standard wavelets as part of the regression ensembles. The JMAWRE was also superior when compared to other models from literature that used the same publicly available NIR dataset.

Chapter 6

Binomial Tree Factorization of the Matrix Polynomial Product with Shift Orthogonal Matrices

6.1 Introduction

High multiplicity wavelets (HMW) have highly desirable characteristics in many areas of signal analysis such as compression [7], noise reduction [7] and feature extraction [8, 71, 72]. However, due to the complexity of constructing high multiplicity wavelets, they are rarely applied with preference given to the simpler two banded wavelet.

The theory of HMW is well documented and several approaches to generate HMW have been derived, the primary algorithms being Sweldons Lifting [78], Vaidyanathan's quadrature mirror filter banks [79] and Kautsky's matrix polynomial product [17]. All three algorithms rely on the Z-transform of the polyphase wavelet matrix [7] but of the three algorithms, Kautsky's method can be reformulated into conventional matrix nomenclature with the inclusion of the matrix polynomial product. We investigate the use of the matrix polynomial product, as used by Kautsky, to further simplify generating high multiplicity wavelets.

6.2 Theory

The matrix polynomial product can be defined in both the standard matrix nomenclature and in the Z-transform notation. Initially both methods are defined, with a focus on the standard matrix notation to be used later on. The Z-transform will be used to assist defining the meaning of the matrix polynomial product.

Using standard matrix notation, the matrix polynomial product between two matrices $\mathbf{A} = \left(\mathbf{A}_0\, \mathbf{A}_1 \ldots \mathbf{A}_q\right)$ and $\mathbf{B} = \left(\mathbf{B}_0\, \mathbf{B}_1 \ldots \mathbf{B}_p\right)$, that consist of square *m* by *m* sub-matrices is

$$\mathbf{C} = \left(\mathbf{C}_0\, \mathbf{C}_1 \ldots \mathbf{C}_{p+q}\right) = \mathbf{A} \Diamond \mathbf{B} \tag{6.1}$$

with the *m* by *m* sub-matrices of $\mathbf{C}$ defined by

$$\mathbf{C}_j = \sum_k \mathbf{A}_{j-k}\mathbf{B}_k \tag{6.2}$$

The polynomial product is more readily seen using the Z-transform. We let $\Im(\mathbf{A}) = \left(z^0\mathbf{A_0}\, z^1\mathbf{A_1} \ldots z^q\mathbf{A_q}\right)$ represent the Z-transform of the poly-phase matrix $\mathbf{A}$ [7]. The Z-transform of eqn. (6.1) is

$$\Im(\mathbf{C}) = \Im(\mathbf{A} \lozenge \mathbf{B}) = \left(z^0\mathbf{A_0}\, z^1\mathbf{A_1} \ldots z^q\mathbf{A_q}\right)\left(z^0\mathbf{B_0}\, z^1\mathbf{B_1} \ldots z^p\mathbf{B_q}\right) \tag{6.3}$$

Upon expansion and equating the powers z in eqn. (6.3), we obtain the poly-phase form of $\mathbf{C}$

$$\Im\left(\mathbf{C}_j\right) = z^j\mathbf{C}_j = \sum_j z^{j-k}\mathbf{A}_{j-k}z^k\mathbf{B}_k = \sum_j z^j\mathbf{A}_{j-k}\mathbf{B}_k \tag{6.4}$$

The inverse Z-transform of eqn. (6.4) gives eqn. (6.2).

Now we wish to focus on the creation of a matrix $\mathbf{W} = \left(\mathbf{W}_0\, \mathbf{W}_1 \ldots \mathbf{W}_q\right)$ where the *m* by *m* sub-matrices satisfy the shift orthogonality conditions [17]

$$\sum_{j=0}^{q-k} \mathbf{W}_j\mathbf{W}_{j+k}^* = \rho\delta_{k,0}\mathbf{I}, \quad k = 0,1,\ldots,q \tag{6.5}$$

where $\mathbf{W}_i^*$ denotes the conjugate transpose of $\mathbf{W}_i$ and $\delta_{k,0}$ is the Kronecker delta. This means that rows of $\mathbf{W}$ all have the same norm, $\sqrt{\rho}$, are orthogonal to each other and orthogonal to themselves when shifted by a multiple of *m*. Matrices of this form are generally referred to as m-banded quadrature mirror filter banks [7], which are used extensively in signal processing and wavelet analysis.

Matrices with shifted orthogonality conditions can be factorized into a series of linear factors (symmetric projections), $\mathbf{P}_i$, using the matrix polynomial product [10, 17, 94].

$$\mathbf{W} = \mathbf{H} \lozenge \left(\mathbf{P}_1 \ \tilde{\mathbf{P}}_1 \right) \lozenge \left(\mathbf{P}_2 \ \tilde{\mathbf{P}}_2 \right) \lozenge \ldots \lozenge \left(\mathbf{P}_q \ \tilde{\mathbf{P}}_q \right) \tag{6.6}$$

$$\mathbf{W} = \mathbf{H} \lozenge_{j=1}^{q} \left(\mathbf{P}_j \ \tilde{\mathbf{P}}_j \right) \tag{6.7}$$

where $\tilde{\mathbf{P}}_i = \mathbf{I} - \mathbf{P}_i$ is the complement symmetric projection to $\mathbf{P}_i$ and $\mathbf{H}$ is an unitary matrix. The multiple matrix polynomial product term, $\lozenge_{j=1}^{q} \left(\mathbf{P}_j \ \tilde{\mathbf{P}}_j \right)$ in eqn. (6.7), leads to a binomial tree representation for eqn. (6.6), which will be shown in section 5.3.

6.3 Expansion of the multiple matrix polynomial product

Let $\mathbf{W} = \left(\mathbf{W}_0 \ \mathbf{W}_1 \ \mathbf{W}_2 \ \mathbf{W}_3 \right)$ so that the multiple matrix polynomial product is

$$\mathbf{W} = \mathbf{H} \lozenge_{j=1}^{3} \left(\mathbf{P}_j \ \tilde{\mathbf{P}}_j \right) = \mathbf{H} \lozenge \left(\mathbf{P}_1 \ \tilde{\mathbf{P}}_1 \right) \lozenge \left(\mathbf{P}_2 \ \tilde{\mathbf{P}}_2 \right) \lozenge \left(\mathbf{P}_3 \ \tilde{\mathbf{P}}_3 \right) \tag{6.8}$$

upon expansion the $\mathbf{W}_j$ terms are given as

$$\begin{aligned}
\mathbf{W}_0 &= \mathbf{H} \mathbf{P}_1 \mathbf{P}_2 \mathbf{P}_3 \\
\mathbf{W}_1 &= \mathbf{H} \left(\mathbf{P}_1 \mathbf{P}_2 \tilde{\mathbf{P}}_3 + \mathbf{P}_1 \tilde{\mathbf{P}}_2 \mathbf{P}_3 + \tilde{\mathbf{P}}_1 \mathbf{P}_2 \mathbf{P}_3 \right) \\
\mathbf{W}_2 &= \mathbf{H} \left(\mathbf{P}_1 \tilde{\mathbf{P}}_2 \tilde{\mathbf{P}}_3 + \tilde{\mathbf{P}}_1 \mathbf{P}_2 \tilde{\mathbf{P}}_3 + \tilde{\mathbf{P}}_1 \tilde{\mathbf{P}}_2 \mathbf{P}_3 \right) \\
\mathbf{W}_3 &= \mathbf{H} \tilde{\mathbf{P}}_1 \tilde{\mathbf{P}}_2 \tilde{\mathbf{P}}_3
\end{aligned}$$

This can be re-expressed as

$$\begin{aligned}
\mathbf{W}_0 &= \mathbf{H} \prod_{j \in \Theta} \mathbf{P}_j \\
\mathbf{W}_1 &= \mathbf{H} \sum_{i=1}^{3} \tilde{\mathbf{P}}_i \prod_{j \in \Theta - \{i\}} \mathbf{P}_j \\
\mathbf{W}_2 &= \mathbf{H} \sum_{i=1}^{2} \sum_{k=i+1}^{3} \tilde{\mathbf{P}}_i \tilde{\mathbf{P}}_k \prod_{j \in \Theta - \{i,k\}} \mathbf{P}_j \\
\mathbf{W}_3 &= \mathbf{H} \sum_{i=1}^{1} \sum_{k=i+1}^{2} \sum_{z=k+1}^{3} \tilde{\mathbf{P}}_i \tilde{\mathbf{P}}_k \tilde{\mathbf{P}}_z \prod_{j \in \Theta - \{i,k,z\}} \mathbf{P}_j
\end{aligned}$$

where $\Theta = \{1,2,3\}$ and $\prod_{j\in\{0\}} \mathbf{P}_j = \mathbf{I}$. The ordering priority of $\mathbf{P}_i$ still exists, however, for ease of interpretation and printability we have relaxed the order but the innermost summation can still be expanded using conventional notation by including multiple product summations.

The reasoning for this notation is so that higher $\mathbf{W}_j$ terms can be iteratively expressed in terms of $\mathbf{W}_i$ where $i < j$. An example given for $\mathbf{W}_1$. Let $\mathbf{W}_0 = \mathbf{K}_0 = \prod_{j\in\Theta} \mathbf{P}_j$ then

$$\begin{aligned}
\mathbf{W}_1 &= \mathbf{H}\sum_{i=1}^{3} \tilde{\mathbf{P}}_i \prod_{j\in\Theta-\{i\}} \mathbf{P}_j \\
&= \mathbf{H}\left(\sum_{i=1}^{3} (\mathbf{I}-\mathbf{P}_i) \prod_{j\in\Theta-\{i\}} \mathbf{P}_j \right); \qquad \tilde{\mathbf{P}}_i = \mathbf{I}-\mathbf{P}_i \\
&= \mathbf{H}\left(\sum_{i=1}^{3} \prod_{j\in\Theta-\{i\}} \mathbf{P}_j - \sum_{i=1}^{3} \mathbf{P}_i \prod_{j\in\Theta-\{i\}} \mathbf{P}_j \right) \\
&= \mathbf{H}\left(\sum_{i=1}^{3} \prod_{j\in\Theta-\{i\}} \mathbf{P}_j - \sum_{i=1}^{3} \prod_{j\in\Theta} \mathbf{P}_j \right) \\
&= \mathbf{H}\left(\mathbf{K}_1 - 3\mathbf{K}_0 \right)
\end{aligned}$$

where $\mathbf{K}_1 = \sum_{i=1}^{3} \prod_{j\in\Theta-\{i\}} \mathbf{P}_j$. Similarly $\mathbf{W}_2 = \mathbf{H}\left(\mathbf{K}_2 - 2\mathbf{K}_1 + 3\mathbf{K}_0 \right)$, with

$\mathbf{K}_2 = \sum_{i=1}^{2} \sum_{k=i+1}^{3} \prod_{j\in\Theta-\{i,k\}} \mathbf{P}_j$.

If $\mathbf{W} = \left(\mathbf{W}_0\ \mathbf{W}_1 \ldots \mathbf{W}_q \right)$, then $\mathbf{W}_j; j = 1,2,\ldots,q$ can be expressed as

$$\mathbf{W}_j = \mathbf{H}\left(a_{j,0}\mathbf{K}_j + a_{j,1}\mathbf{K}_{j-1} + \ldots + a_{j,n}\mathbf{K}_{j-n} + \ldots + a_{j,j}\mathbf{K}_0 \right) \tag{6.9}$$

where

$$\mathbf{K}_n = \sum_{i_1=1}^{q-n+1} \sum_{i_2=i_1+1}^{q-n+2} \sum_{i_3=i_2+1}^{q-n+3} \cdots \sum_{i_j=i_{j-1}+1}^{q} \prod_{j\in\Theta-\{i_1,i_2,\ldots,i_j\}} \mathbf{P}_j \tag{6.10}$$

and

$$a_{j,n} = (-1)^n \binom{q-j+n}{n} \tag{6.11}$$

Proof:

For $\mathbf{W} = \left(\mathbf{W}_0\ \mathbf{W}_1 \ldots \mathbf{W}_q \right)$, the j^{th} term can be expressed as

$$\mathbf{W}_j = \mathbf{H}\left(\sum_{i_1=1}^{q-j+1} \sum_{i_2=i_1+1}^{q-j+2} \sum_{i_3=i_2+1}^{q-j+3} \cdots \sum_{i_j=i_{j-1}+1}^{q} \tilde{\mathbf{P}}_{i_1} \tilde{\mathbf{P}}_{i_2} \ldots \tilde{\mathbf{P}}_{i_j} \prod_{k \in \Theta - \{i_1, i_2, \ldots, i_j\}} \mathbf{P}_k \right) \tag{6.12}$$

The parenthesis component of (6.12) can be expanded using $\tilde{\mathbf{P}}_i = \mathbf{I} - \mathbf{P}_i$

$$\sum_{i_1=1}^{q-j+1} \ldots\ldots \sum_{i_j=i_{j-1}+1}^{q} \left(\mathbf{I} - \mathbf{P}_{i_1}\right)\left(\mathbf{I} - \mathbf{P}_{i_2}\right) \ldots \left(\mathbf{I} - \mathbf{P}_{i_j}\right) \prod_{k \in \Theta - \{i_1, i_2, \ldots, i_j\}} \mathbf{P}_k \tag{6.13}$$

$$= \sum_{i_1=1}^{q-j+1} \cdots \sum_{i_j=i_{j-1}+1}^{q} \left[\mathbf{I} - \mathbf{P}_{i_1} - \mathbf{P}_{i_2} - \ldots \mathbf{P}_{i_j} + \mathbf{P}_{i_1}\mathbf{P}_{i_2} + \ldots + \mathbf{P}_{i_1}\mathbf{P}_{i_2} \ldots \mathbf{P}_{i_j} \right] \prod_{k \in \Theta - \{i_1, i_2, \ldots, i_j\}} \mathbf{P}_k$$

$$\begin{aligned}
&= \sum_{i_1=1}^{q-j+1} \cdots \sum_{i_j=i_{j-1}+1}^{q} \mathbf{I} \prod_{k \in \Theta - \{i_1, i_2, \ldots, i_j\}} \mathbf{P}_k \\
&+ \sum_{i_1=1}^{q-j+1} \cdots \sum_{i_j=i_{j-1}+1}^{q} \left[-\mathbf{P}_{i_1} - \mathbf{P}_{i_2} - \ldots \mathbf{P}_{i_j} \right] \prod_{k \in \Theta - \{i_1, i_2, \ldots, i_j\}} \mathbf{P}_k \\
&+ \sum_{i_1=1}^{q-j+1} \cdots \sum_{i_j=i_{j-1}+1}^{q} \left[\mathbf{P}_{i_1}\mathbf{P}_{i_2} + \mathbf{P}_{i_1}\mathbf{P}_{i_3} \ldots + \mathbf{P}_{i_{j-1}}\mathbf{P}_{i_j} \right] \prod_{k \in \Theta - \{i_1, i_2, \ldots, i_j\}} \mathbf{P}_k \\
&+ \ldots + \sum_{i_1=1}^{q-j+1} \cdots \sum_{i_j=i_{j-1}+1}^{q} \mathbf{P}_{i_1}\mathbf{P}_{i_2} \ldots \mathbf{P}_{i_j} \prod_{k \in \Theta - \{i_1, i_2, \ldots, i_j\}} \mathbf{P}_k
\end{aligned} \tag{6.14}$$

$$\begin{aligned}
&= \sum_{w=1}^{\binom{j}{0}} \sum_{i_1=1}^{q-j+1} \cdots \sum_{i_n=i_{n-1}+1}^{q-j+n+1} \cdots \sum_{i_j=i_{j-1}+1}^{q} \prod_{j \in \Theta - \{i_1, \ldots, i_j\} + {}^0\Omega_w} \mathbf{P}_j \\
&- \sum_{w=1}^{\binom{j}{1}} \sum_{i_1=1}^{q-j+1} \cdots \sum_{i_n=i_{n-1}+1}^{q-j+n+1} \cdots \sum_{i_j=i_{j-1}+1}^{q} \prod_{j \in \Theta - \{i_1, \ldots, i_j\} + {}^1\Omega_w} \mathbf{P}_j \\
&+ \sum_{w=1}^{\binom{j}{2}} \sum_{i_1=1}^{q-j+1} \cdots \sum_{i_n=i_{n-1}+1}^{q-j+n+1} \cdots \sum_{i_j=i_{j-1}+1}^{q} \prod_{j \in \Theta - \{i_1, \ldots, i_j\} + {}^2\Omega_w} \mathbf{P}_j \\
&+ \ldots + (-1)^n \sum_{w=1}^{\binom{j}{n}} \sum_{i_1=1}^{q-j+1} \cdots \sum_{i_n=i_{n-1}+1}^{q-j+n+1} \cdots \sum_{i_j=i_{j-1}+1}^{q} \prod_{j \in \Theta - \{i_1, \ldots, i_j\} + {}^n\Omega_w} \mathbf{P}_j \\
&+ \ldots + (-1)^j \sum_{w=1}^{1} \sum_{i_1=1}^{q-j+1} \cdots \sum_{i_n=i_{n-1}+1}^{q-j+n+1} \cdots \sum_{i_j=i_{j-1}+1}^{q} \prod_{j \in \Theta - \{i_1, \ldots, i_j\} + {}^j\Omega_w} \mathbf{P}_j
\end{aligned} \tag{6.15}$$

where ${}^{n}\Omega_{w}$ is a cyclic set of the $\binom{j}{n}$ permutations containing n elements of the indices $i_k; k = 1,\ldots,j$, and $0 \le n \le j$. For brevity, we introduce ${}^{j}\mathbf{Z}_{n}$ represent the term containing n permutations of the q $\mathbf{P}$ matrices.

$$\mathbf{W}_j = \mathbf{H}\left({}^{j}\mathbf{Z}_0 + {}^{j}\mathbf{Z}_1 + \ldots + {}^{j}\mathbf{Z}_j\right)$$

$${}^{j}\mathbf{Z}_n = (-1)^n \sum_{w=1}^{\binom{j}{n}} \sum_{i_1=1}^{q-j+1} \cdots \sum_{i_n=i_{n-1}+1}^{q-j+n+1} \cdots \sum_{i_j=i_{j-1}+1}^{q} \prod_{j\in\Theta-\left\{i_1,\ldots,i_j\right\}+{}^{n}\Omega_w} \mathbf{P}_j \tag{6.16}$$

In ${}^{j}\mathbf{Z}_n$ there are $\binom{j}{n}\binom{q}{j}$ elements and $\binom{q}{j-n}$ components in set $\Theta-\left\{i_1,\ldots,i_j\right\}+{}^{n}\Omega_w$. Also the union of the $\binom{j}{n}$ sets of $\Theta-\left\{i_1,\ldots,i_j\right\}+{}^{n}\Omega_w$ is equal to $\Theta-\left\{i_1,\ldots,i_{j-n}\right\}$ - which corresponds to $\mathbf{K}_{j-n}$. Additionally, due to the cyclic permutation set ${}^{n}\Omega_w$, the elements in $\Theta-\left\{i_1,\ldots,i_{j-n}\right\}$ are repeated equally across the sets $\Theta-\left\{i_1,\ldots,i_j\right\}+{}^{n}\Omega_w$. Thus, $\mathbf{K}_{j-n}$ is repeated $\binom{j}{n}\binom{q}{j}\Big/\binom{q}{j-n}=\binom{q-j+n}{n}$ times in ${}^{j}\mathbf{Z}_n$. So

$${}^{j}\mathbf{Z}_n = (-1)^n \binom{q-j+n}{n}\mathbf{K}_{j-n} = a_{j,n}\mathbf{K}_{j-n} \tag{6.17}$$

hence

$$\mathbf{W}_j = \mathbf{H}\left(a_{j,0}\mathbf{K}_j + a_{j,1}\mathbf{K}_{j-1} + \ldots + a_{j,n}\mathbf{K}_{j-n} + \ldots + a_{j,j}\mathbf{K}_0\right)$$

Analyzing eqn. (6.10), $\mathbf{K}_n$ is equal to the n^{th} row sum of the binomial tree formed by the $\mathbf{P}_i$ matrices.

6.4 Example

Consider the case where $\mathbf{W} = \left(\mathbf{W}_0\ \mathbf{W}_1\ \mathbf{W}_2\ \mathbf{W}_3\right)$ so that q equals three and $\mathbf{W} = \mathbf{H}\Diamond_{j=1}^{3}\left(\mathbf{P}_j\ \tilde{\mathbf{P}}_j\right)$. The binomial tree for this example is given in Figure 6.1.

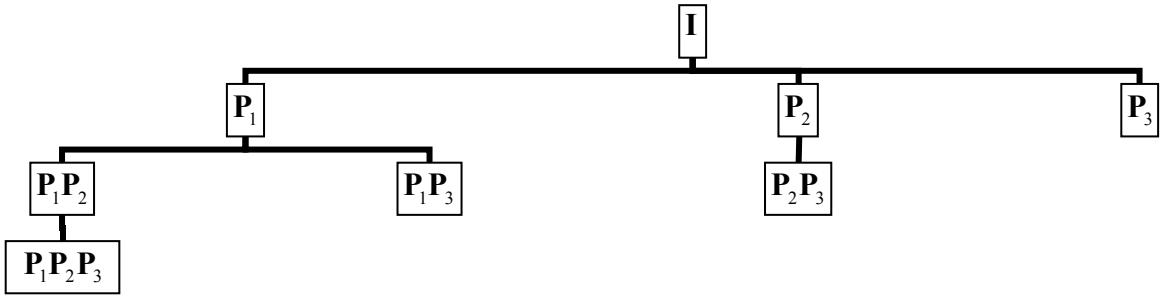

Figure 6.1 Binomial tree expansion of the projection matrices $\mathbf{P}_i$ used to construct the $\mathbf{K}_n$ matrices.

Now $\mathbf{K}_n$ is the row sum of the n^{th} level (the root node is $n=q$) and using eqn (6.11) to calculate $a_{j,n}$ we have:

$$\begin{aligned}\mathbf{K}_0 &= \mathbf{P}_1\mathbf{P}_2\mathbf{P}_3 \\ \mathbf{K}_1 &= \mathbf{P}_1\mathbf{P}_2 + \mathbf{P}_1\mathbf{P}_3 + \mathbf{P}_2\mathbf{P}_3 \\ \mathbf{K}_2 &= \mathbf{P}_1 + \mathbf{P}_2 + \mathbf{P}_3 \\ \mathbf{K}_3 &= \mathbf{I}\end{aligned} \tag{6.18}$$

And:

$$\begin{aligned}\mathbf{W}_0 &= \mathbf{H}\mathbf{K}_0 \\ \mathbf{W}_1 &= \mathbf{H}\left(\mathbf{K}_1 - 3\mathbf{K}_0\right) \\ \mathbf{W}_2 &= \mathbf{H}\left(\mathbf{K}_2 - 2\mathbf{K}_1 + 3\mathbf{K}_0\right) \\ \mathbf{W}_3 &= \mathbf{H}\left(\mathbf{K}_3 - \mathbf{K}_2 + 2\mathbf{K}_1 - 3\mathbf{K}_0\right)\end{aligned} \tag{6.19}$$

6.5 Conclusion

By investigating the properties of the multiple matrix polynomial product on matrices comprising of square symmetric projection matrix with its complement, we have developed simple recursive algorithm utilizing a binomial tree to construct m-banded quadrature mirror filter banks.

This binomial tree structure for generating adaptive wavelets is more readily understood given the familiarity of the binomial theorem in the scientific community. This then enables a wider audience the ability to generate computer code for the binomial tree factorisation, which is relatively simple compared alternative algorithms such as Lifting, Qraduature Mirror filter Banks and the original formulation of matrix polyphase multiplication.

Chapter 7

Conclusion

The Discrete Wavelet Transform (DWT) is a valuable tool for improving descriptive modelling of juxta-positional data, such as near infrared (NIR) spectra or SELDI-TOF mass spectra. A key factor in the application of DWT is wavelet basis selection. Selecting the correct wavelet basis, or wavelet bases, results in superior models compared to using a wavelet based upon convenience or random guessing. Adaptive wavelet generation algorithms can be used to target appropriate wavelets for the modelling process at hand.

Use of adaptive wavelet algorithms by the spectroscopic community has been scarce with very few applications appearing in literature. Reasons why adaptive wavelets have not been widely adopted include a perceived increase in model complexity and a general unfamiliarity with wavelet basis selection. In order to increase the use of adaptive wavelet algorithms within the spectroscopic community, this thesis investigated five key aspects of adaptive wavelet basis selection for spectroscopic data analysis:

6. Integration of adaptive wavelets with modern data analysis techniques
7. Generation of adaptive wavelet optimisation criteria for the four main types of data modelling: experimental design analysis, unsupervised classification, supervised classification and regression analysis.
8. Automation of adaptive wavelet parameter selection
9. Investigation of feature heterogeneity within in a spectrum by using both adaptive and standard multiple wavelets and,
10. Generation of adaptive wavelets using a simplified binomial tree algorithm

7.1 Integration of adaptive wavelets

A wide range of current modern data analysis techniques were integrated with adaptive wavelets in Chapters 2, 3 and 5. Techniques illustrated in this thesis include:

- Penalised Discriminate Analysis (PDA) – Chapter 2
- Random Forests (RF) – Chapter 2
- Principal component analysis (PCA) – Chapter 3

- Gaussian Mixture Models (GMM) – Chapter 3
- Multivariate Regression – Chapter 5
- Stacking – Chapter 5

In Chapters 2 and 3, a relatively straight forward method was used to integrate adaptive wavelets with each of the respective data analysis techniques. A single superlative adaptive wavelet was chosen and applied to the NIR spectrum to produce wavelet coefficients (extracted features) which were subsequently used as input for a traditional data analysis technique.

A superlative adaptive wavelet was chosen from a set of optimised adaptive wavelets which were initially random wavelets with different adaptive wavelet parameters. The random wavelets were updated to maximise an optimisation criteria. The adaptive wavelet with parameters corresponding to the highest value from the optimisation process was chosen as the superlative wavelet.

In both Chapters 2 and 3, model performance was enhanced by integrating a single superlative adaptive wavelet with the respective analysis technique. In Chapter 2, a repeated measures experimental design of wine gape homogenates was analysed via measuring the correct classification rates of penalised discriminate analysis (PDA), multiple adaptive regression splines (MARS) and random forests (RF), with and without prior transformation using the adaptive discrete wavelet transform (ADWT). The correct classification rates for all methods were substantially improved by the use of the ADWT compared to standard wavelets and traditional pre-processing methods such as the SNV transform.

Chapter 3 demonstrated an unsupervised clustering example of NIR spectra. A single superlative adaptive wavelet combined with Gaussian Mixture Models (GGM) were used to elucidate unknown clustering within the data. The number of clusters was consistent when using adaptive wavelets with high optimisation scores, whereas with standard wavelet types, the number of clusters varied depending on which standard wavelet was used.

The method of integrating a single, superlative adapted wavelet is relatively simple and enhances traditional NIR data analysis methods. Chapter 5 employed an alternative strategy for adaptive wavelet integration where the optimisation of the wavelets is a part of the analysis method rather than a strictly pre-treatment method such as in Chapters 2 and 3.

Chapter 5 illustrated how adaptive wavelets can be integrated with chemometric methods that have stochastic components such as variable selection and regression coefficient determination. Due to the stochastic nature of the methods being integrated with adaptive wavelets, an iterative approach was used to integrate adaptive wavelets with the chosen chemometric method. In Chapter 5, adaptive wavelets were combined with Baysian multivariate regression.

The method employed to combine Baysian multivariate regressing with adaptive wavelets in Chapter 5 was to apply a random wavelet basis to the data and perform a stochastic regression model search to identify predictive models that contain a small number of wavelet coefficients. The wavelet coefficients, typically less than five, were then jointly optimised by allowing assigning an adaptive wavelet to each wavelet coefficient. This iterative method differs substantially from that used in Chapters 2 and 3 where the optimisation of the wavelet basis contains all wavelet coefficients.

A joint optimisation approach was used in Chapter 5 because the stochastic regression model search identifies important interrelationships *between* the wavelet coefficients rather than important individual wavelet coefficients. A less predictive regression model was generated when wavelet coefficients are optimised individually compared to joint optimisation or even to the initial random wavelet.

The iterative approach of integrating adaptive wavelets in Chapter 5 is better suited to chemometric methods that contain heuristics which use very few variables, such as tree based methods or variable selection algorithms. The pre-treatment method used in Chapters 2 and 3 is better adapted to projection based chemometric methods that utilise all available variables (wavelet coefficients) simultaneously; methods like principal component analysis and partial least squares.

7.2 Adaptive wavelet optimisation criteria

The optimisation criteria used for adaptive wavelets typically mimic the role of the chemometric method which the adaptive wavelets are being integrated with. Three adaptive wavelet criteria were presented in this thesis representing optimisation criteria that can be used for experimental design analysis, unsupervised classification, supervised classification and regression applications.

In Chapter 2, the optimisation criteria was designed to generated wavelet coefficients that maximise differences in NIR spectra that are associated with an experimental design. This was achieved with the optimisation criteria based on the two largest eigenvalues of the matrix product between the inverse within group covariance matrix, Σ_w^{-1}, and the between group covariance matrix, Σ_B. Using two eigenvalues was important from information mapping perspective as two dimensions facilitate maximum group separation with a minimum of within group variation.

The optimisation criteria used in Chapter 2 is very versatile as Σ_B can be adapted to reflect supervised classification applications. For supervised classifications Σ_B is derived from the known groups. A simple modification can also be used for unsupervised classification, where, in Chapter 3, the optimisation criterion was to maximise the two largest eigenvalues of the covariance matrix of the discrete wavelet transformed spectra. This criterion resulted in wavelet coefficients that contained the largest amounts of variations from the spectra.

The optimisation criteria in Chapters 2 and 3 are not dependent on the modelling procedure used after application of the DWT. So, while the optimisation criteria in Chapters 2 and 3 reflect the modelling method, it is not dependent on the modelling method. Chapter 5 on the other hand, the optimisation criteria was dependant on the modelling method.

In Chapter 5 the optimisation criterion was to minimise the mean squared error (MSE) of prediction of a regression model. To determine the MSE associated with particular wavelet coefficients (or wavelets), a regression model needed to be constructed and the

MSE evaluated. In this way the optimisation of the adaptive wavelets is totally dependent on the modelling method.

A result of the dynamic relationship between the wavelet parameters and modelling methods is that there are no convenient mathematical properties of the wavelet transformed spectra, such as eigenvalues, which can be used as an optimisation function. The optimisation criteria used in Chapter 5 was a *lazy* function. A *lazy* function simply computes the score to evaluate the effectiveness of the current state. Optimisation of *lazy* functions is quite simple where the current state is perturbed then re-evaluated to determine partial derivatives required to optimise parameters.

Optimisation of *lazy* functions can lead to localisation, or sub-optimal results, and are generally slower than functions with more mathematical structural form. Localisation is not much of a problem as it can be mitigated by changing perturbation step sizes and/or initial starting values, as was done in the optimisation algorithms used in Chapter 5.

This thesis demonstrated how simple mathematical properties of the discrete wavelet transformed data, like eigenvalues, can be utilised as optimisation criteria. This type of optimisation criteria mimics the role of subsequent modelling but is independent of the modelling method. When it not possible to decouple the adaptive wavelet optimisation criteria from the modelling method, a lazy approach can be taken which evaluates the goodness of fit of the wavelet coefficients jointly with the modelling method. The lazy approach makes generating optimisation criteria extremely easy, but at the expense of speed and optimisation complexity.

7.3 Adaptive wavelet parameter selection

The adaptive wavelet algorithm investigated in this thesis has three parameters, m, q and J, along with a set of $q + 1$ unit length vectors, each containing $m - 1$ elements. The parameter m defines the number of bands used in the DWT, q defines the length of the wavelet, J is the number of iterations (or level) of the DWT and the $q + 1$ vectors define the wavelet filter coefficients (wavelets) used for the DWT. During the adaptive wavelet algorithm, m, q and J are fixed and the $q + 1$ vectors are updated to optimise some predefined criteria.

Methods for selection of the adaptive wavelet parameters in Chapters 2 and 3 are very similar where a single set of adaptive wavelet parameters were ultimately chosen, whereas is Chapter 5 an ensemble of adaptive wavelet parameters was used. In Chapters 2, 3 and 5, the parameters *m, q* and *J* were very influential on the resulting adapted wavelets. In contrast the initial choice of the $q + 1$ vectors was not critical in Chapters 2 and 3, but was important in Chapter 5.

In Chapters 2 and 3 the $q + 1$ vectors were initially randomised then updated to optimise the specified optimisation criteria in each chapter respectively. The initial staring position of vectors was not critical as several randomised starting positions typically converge to produce similar wavelet filter coefficients. This result is more an effect of modern optimisation routines as most optimisation routines check for localised minimums/maximums by introducing large perturbations then re-optimising the system. In effect, the optimisation routines used create many initial starting positions themselves, which makes the initial randomised starting vectors defined by the user less critical than previously thought.

Parameter selection of *m, q* and *J* in Chapters 2 and 3 was the critical component that determined the performance differences in the adaptive wavelet algorithm. In Chapter 2, a superlative set of parameters was chosen by trialling a set of parameters. The parameter set with the highest adapted wavelet optimisation criteria was selected as the superlative set. Chapter 3 used a similar approach with a single set of adaptive wavelet parameters being chosen by trialling approximately seventy sets of adaptive wavelet parameters. However in Chapter 3, the superlative set was chosen not by the optimisation criteria, but by using the Bayes Information Criteria (BIC) of the Gaussian Mixture Model (GMM) that the DWT data was applied to. In using the BIC, the superlative set of adaptive wavelet parameters produces the most informative GMM; which is not necessarily the same set of parameters with the best optimisation criteria score.

Chapter 2 illustrated how the optimisation criteria alone can be used to select the adaptive wavelet parameters while Chapter 3 demonstrates how a goodness of fit of the resulting model can appropriately select the wavelet parameters. In both Chapters 2 and

3, the critical parameters were *m, q* and *J.* These parameters were also important in Chapter 5 as well as the initial selection of the $q + 1$ vectors.

In Chapter 5 the initial random the $q + 1$ vectors was used to determine sets of wavelet coefficients for regression models and the wavelet coefficients were subsequently jointly optimised. Choice of the initial random the $q + 1$ vectors influenced which sets wavelet coefficients was selected. Changing the initial set of starting the $q + 1$ vectors lead to different sets of wavelet coefficients being selected; and ultimately a different adapted wavelet regression model. Because of the dependence on the initial the $q + 1$ vectors, multiple randomised starting positions were used for each set of *m, q* and *J* parameters.

Adaptive wavelet parameter selection in Chapter 5 was dependent on the full set of adaptive wavelet parameters. Additionally, the adaptive wavelet parameters used greatly influenced which wavelet coefficients were selected in subsequent regression modelling. Here, the wavelet parameters can be viewed as another stochastic component in the modelling process. So rather than chose a single set of wavelet parameters, like in Chapter 3, a stochastic approach was taken that used all of the trialled adaptive wavelet parameters simultaneously.

A re-sampled stacked ensemble was used to amalgamate and weight all the models adapted from the various trialled adaptive wavelet parameters. Using the ensemble approach, individual regression models with varying adaptive wavelet parameters were identified as being more important than other regression models with different adaptive wavelet parameter sets and initial starting (vector) positions.

Some sets of *m, q* and *J* resulted in more predictive models which could serve as a guide to further improvements for parameter selection. For example, trialling more random starting $q + 1$ vectors with *m, q* and *J* parameters that have a higher proportion of predictive models in the ensemble. Using an ensemble approach in Chapter 5 made prior selection of adaptive wavelet parameters less of a critical issue than in Chapters 2 and 3.

7.4 Multiple wavelets

Chapters 4 and 5 investigated homogeneity of underlying signals in spectra by using multiple wavelets. Multiple standard wavelets were used in Chapter 4 for a supervised classification case study of SELDI-TOF mass spectra, whereas in Chapter 5 both multiple standard wavelets and adaptive wavelets were applied in a NIR multivariate regression example. In both Chapters 4 and 5, using multiple wavelets improved the quality of data analysis compared to using a single wavelet.

Using multiple wavelet bases in both Chapters 4 and 5 posed a problem of generating an excessive amount of extracted features. Each wavelet generates p wavelet coefficients, so x wavelets will generate xp wavelet coefficients. Because of this expansion effect, data reduction methods were an integral part in the application of multiple wavelets. In Chapter 4 data reduction heuristics were used while ensemble methods were applied in Chapter 5.

Chapter 4 combined wavelet coefficients from six standard wavelets, composed of two types of Daubechies, Coiflets and Symmlets wavelets, applied to mass spectral (MS) profiles consisting of 15154 SELDI-TOF M/Z ratios from 342 patients; which were diagnosed with malignant prostate cancer, benign prostate hyperplasia or as healthy. Each application of the DWT produced 15154 wavelet coefficients. In applying the six different standard wavelets, 90924 wavelet coefficients were produced. The number of wavelet coefficients resulting from using multiple wavelet bases greatly exceeds the number of samples. A variety of data reduction techniques were applied to the multiple wavelet coefficients before data analysis using Classification and Regression Trees (CART).

Simple heuristics, pair-wise t-test and then the variable importance (VIP) list used in Random Forests, were used to reduce the large number of wavelet coefficients to a much smaller, predictive set. Simple random forests, consisting of trees with four or five branches, were then iteratively generated on the wavelet coefficients from the t-tests. Classification and Regression Trees using wavelet coefficients from multiple standard wavelets produced more favourable models than those produced with a single wavelet basis. This outcome gave some evidence to support the hypothesis that the

localised information embedded in the MS data is better approximated by different wavelets at different positions along the spectrum.

Scope of the standard wavelets used in Chapter 4 was limited to small subset of the of the three most commonly used wavelet families, Daubechies, Symlets and Coiflets. This limited sub-set of wavelets did illustrate that using multiple wavelet transforms at different positions along the spectrum does improve the performance of the modelling process compared to using a single wavelet.

Chapter 5 used both multiple standard wavelets and multiple adaptive wavelets in a multivariate regression example. As in Chapter 4, using multiple wavelets increased multiplied the number of wavelet coefficients so that some form of variable reduction was necessary. In Chapter 5 a Metropolis-Hastings search was used to produce numerous sparse regression models, which effectively reduced the number of wavelet coefficients.

The Metropolis-Hastings search generated many potentially useful regression models. Rather than select a single model, an ensemble of all potential models was formed using re-sampled constrained stacking. Re-sampled constrained stacking was useful in determining how regression models from different wavelets compare with one another.

In Chapter 5, using multiple *standard* wavelets did not improve model performance compared to models derived from a single *standard* wavelet. Re-sampled constrained stacking (RCS) over the entire set of standard wavelets, i.e. using multiple standard wavelets, gave a prediction worse than most single wavelet RCS ensembles. The decrease in performance for the multiple standard wavelet case is most likely due to the sheer number of individual models incorporated into the ensemble, near 10,000 in total.

The problem of an over excess of models in the ensemble transcends the initial regression problem, which subsequently favoured the single wavelet case. This does not preclude the possibility that a permutation of standard wavelet DWT's would give a superior RCS ensemble, but that the number of permutations of standard wavelets to consider is prohibitive. This is where adaptive wavelets have a definite advantage over standard wavelets where a large range of permutations of adaptive wavelet types can be

tractably searched to produce a very good multiple wavelet, multiple response, RCS ensemble.

Ensembles using multiple adaptive wavelets, derived from a single set of adaptive wavelet parameters, performed very comparably to the best of the standard single wavelet model ensembles. However, an ensemble using multiple adaptive wavelets that span multiple adaptive wavelet parameters was superior to any of the single standard or adaptive wavelet ensemble models. The different wavelets in the superior multiple adaptive wavelet ensemble utilized different information within the spectrum as the various wavelets had different characteristics, (i.e. shapes) dependant on the position within the spectrum.

Using multiple wavelet transforms in Chapters 4 and 5 supports the supports hypothesis of homogeneity of underlying signals within the spectrum. Multiple wavelet transforms can be used to improve feature extraction leading to gains in model development.

7.5 Binomial tree algorithm for adaptive wavelets

The Pollen factorisation of *m*–banded discrete wavelet transformed (DWT) was reformulated into a binomial tree algorithm in Chapter 6. Optimised wavelets produced the binomial formulation were identical to the previous Pollen factorised method. By recasting the adaptive wavelet algorithm in to a more widely familiar theory, it is envisioned that more independent groups can produce computer code utilising adaptive wavelet in new chemometric research.

7.6 Future considerations

Many of the methods presented in Chapters 2 – 5 are computationally intensive and involve at least one optimisation component. To this end, additional validation techniques could be used to increase the robustness and generalisation of the proposed methods. Validations techniques that could be used are (a) the use of independent validation, training and/or calibrations data sets (b) cross validation methods and (c) bootstrapping. These validation methods could be used to assist in the selection of the adaptive wavelet parameters, m, q and l, band selection and finally model development.

Other possible avenues for subsequent investigation in the topic of adaptive wavelet transforms for spectroscopic analysis include extending the methods outlined in Chapter 5, which was a regression application, to the areas of unsupervised and supervised classification as well as the analysis experimental designs. Another avenue of research is in the optimisation of adaptive wavelets.

During this thesis, the issue of which parameters to use for the adaptive wavelet algorithm arose in every chapter. A numerical, but brute force, approach was adopted in the latter chapters however a less computative solution exists in the phase forms of wavelets themselves.

Adaptive wavelets with a small number of wavelet filter coefficients are a sub-set of their longer counterparts; provided they both have the sample multiplicity (same number of m-banded wavelets). This means that when a portion of the spectrum has been analysed by a particular adaptive wavelet, then the simplex of higher order wavelets is effectively reduced. This approach would reduce the number of permutations for the ADWT parameters required and lessen the search time/space. The branching across ADWT parameter sets is then also possible; which would be useful when one set of parameters has identified a useful portion of the spectrum then further optimisation (at the same position in the spectrum) across different ADWT parameter sets would be possible – reducing the need to trials so many initial ADWT parameters.

Appendix 1 Equation Chapter 1 Section 1

Beer-Lambert-Bouguer Law of Absorption

The macroscopic description optical absorption is defined as: the decrease in intensity of a light beam per unit path length at a given position, *z*, in the absorbing medium is proportional to the instantaneous value of the intensity at that position:

$$-\frac{dI(z)}{dz} = \varepsilon(z)c(z)I(z) \tag{A1.1}$$

Where $I(z)$ is the instantaneous intensity of the light beam at position z, $\varepsilon(z)$ is the specific absorptivity at *z*, and $c(z)$ is the concentration of the absorbing medium at *z*.

For real media, composed of independent absorbing centres (molecules), Eqn (A1.1) is only valid if (i) the size of the absorbing molecules in the solution is negligible with respect to the wavelength of the monochromatic light (λ_i)(ii) the number of molecules in solution is large enough to permit the definition of a statically meaningful mean concentration of molecules per unit volume (iii) that a single molecular species is absorbing the light and (iv) the specific absorptivity, $\varepsilon(z)$, is isotropic; meaning the probability of (the mean) light absorption is invariant to the polarization of the light beam.

The concentration of the medium is dependant on two main factors (1) temperature and (2) state of the medium. Temperature plays a critical role as, when in a state of equilibrium, the distribution of the number of molecules occupying the i^{th} energetic state follows the Boltzmann distribution:

$$\frac{N_{upper}}{N_{lower}} = \exp(-\Delta E / kT) \tag{A1.2}$$

Where N_{lower} and N_{lower} is the number of molecules occupying the different energy states, $\Delta E = E_{upper} - E_{lower}$, T is the temperature in degrees Kelvin, k is Boltzmann's constant: $1.38 \times 10^{-23} \text{JK}^{-1}$. If the temperature were to increase, then where would be more molecules occupying higher energetic states which would lead to an increase absorption of the lower frequencies as changes in quantum numbers is quite typically ≤ 3 and ΔE for small changes in the quantum numbers at the higher energy levels is smaller than those experienced by the lower energy levels such as the ground state. Thus the concentration of the absorbing medium is temperature dependant.

The state (gas, liquid, solid) of the medium as influences the concentration of the absorbing medium since certain types of IR absorption are dependant on free body rotation. In the gaseous state, a molecule is able to undergo rotation-vibration interaction which results in the fine structure component in many of the fundamental frequencies, v_i. However, in the liquid phase, the rotation of the molecule can be inhibited by the presence of other molecules so that the fine structure is no longer well defined and is usually evident as a broadening of the fundamental frequencies.

In the solid phase, the rotation-vibration interaction can be inhibited completely so that only the fundamental frequencies are seen. In addition to the change in the rotation-vibration interaction with respect to the state, there is also a change in the value of the fundamental frequencies. Typically there is a change of 0-5% in the value of the fundamental frequencies, ν, where $\nu_{gas} \geq \nu_{liquid} \geq \nu_{soild}$.

For near-infrared spectroscopy, the wavelength range, λ, is in the region 100μm-1μm, where as the typical molecular radius is of the order of 1nm; approximately one thousandth the wavelength. Scattering or bifringence in the transmitted light is of no observable consequence.

The second condition regarding the distribution of particles is commonly found in biological settings where the absorbing particles (proteins, nucleic acids, porphyins, etc) are contained within organic cells such as membranes. These large particles are held in suspension in a non-absorptive media. The localized macroscopic concentration of particles within the media is continuously in flux determined by the Gibbs

thermodynamic potential. The effect of a Gibbs distribution of absorbing particles is a flattening of the absorption spectra which is wavelength dependant:

$$A^{(susp)} = A^{(sol)}\left[1 - \frac{2.303\varepsilon(1-q)}{2k\lambda^2}\right] \tag{A1.3}$$

Where ε, the specific absorptivity is assumed constant, k is a constant of proportionality, λ is the wavelength of light, q is the probability of observing a particle in a volume of size:

$$\mathrm{v} = k\lambda^2 p \tag{A1.4}$$

Where p is the optical path-length. The effect of q is to average out signals originating in v due to the finite nature of light. In near-infrared spectroscopy, λ is relatively large and the probability of finding an absorptive particle in v is nearly always equal to one. Consequently the flatting effect is not observed for molecules in suspension, however, it would be observed in systems containing large particles in suspension (as indeed would the scattering effect). Hence, $c(z)$ can be regarded as a constant, c, for NIRS of molecular sized absorption.

The effect on the absorption due to $\varepsilon(z)$ can be characterized the level of anisotropic behaviour of (a) the absorption species; being a deformation of dipole, molecular covalent bond in NIRS and (b) the statistical distribution of the polarization of the incident light beam; being either coherently polarised or unpolarised. The interaction between the molecule and an incident photon (light particle or quanta) is uniquely determined by two factors (a) the frequency of the photon and (b) the angle of incidence between the photon and the dipole. If the energy of the incident photon matches the energy required to de-form the dipole, then the dipole will absorb the photon. The probability that a matching photon will be absorbed then depends on the angle of incidence:

$$p(\theta) = \cos^2\theta \tag{A1.5}$$

Where θ is the angle of incidence an $p(\theta)$ is the probability of absorption. For the cases where either the dipole is randomly orientated or the incident light is unpolarised, then $\varepsilon(z)$ is isotropic and is a constant, ε. If the dipole is in a plane perpendicular to the unpolarised light (by means of an external electrical field), but the axes of the dipole is randomly orientated in the plane, then the specific absorptivity is again constant but greater than the aforementioned case by a factor of 3/2 [95]. However, if the dipole axes are all parallel and the incident light is in a plane perpendicular to these axes, the specific absorption is no longer constant but follows a log-linear relationship where the maximum amount of light absorbed ever exceeds 50% of the incident light.

Most applications of NIRS is done in the absence of a controlling external field (so the dipoles are randomly orientated) with either polarised or polarized light sources (lamps and lasers respectively) so that the specific absorptivity, ε, is constant throughout the analysed medium.

When both $\varepsilon(z)$ and $c(z)$ are invariant over the path-length, the optical absorption then follows the Beer-Lambert law (after integrating Eqn (A1.1)):

$$A_i = c\varepsilon_i p_i \tag{A1.6}$$

Where A_i is the absorbance of the i^{th} wavelength, c is the concentration of ζ_b, e_i is the coefficient of absorptivity and p_i is the optical path-length, $\int_0^{l_i} dz$. For fixed path-length Eqn (A1.6) is Beer's Law:

$$A_i = ce_i \tag{A1.7}$$

Beer's Law can be readily interpreted as a linear regression between the observed spectra and the concentration:

$$c = \frac{A_i}{e_i} \tag{A1.8}$$

References

1. Trygg, J. and S. Wold, *PLS regression on wavelet compressed NIR spectra.* Chemometrics and Intelligent Laboratory Systems, 1998. **42**(1-2): p. 209-220.
2. Cowe, I.A. and J.W. McNicol, *The Use of Principal Components in the Analysis of near-Infrared Spectra.* Applied Spectroscopy, 1985. **39**(2): p. 257-266.
3. Wold, S., H. Martens, and H. Wold, *The Multivariate Calibration-Problem in Chemistry Solved by the Pls Method.* Lecture Notes in Mathematics, 1983. **973**: p. 286-293.
4. Jiang, J.H., et al., *Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data.* Analytical Chemistry, 2002. **74**(14): p. 3555-65.
5. Jetter, K., et al., *Principles and applications of wavelet transformation of chemometrics*, in *Analytica Chimica Acta*. 2000. p. 169-180.
6. Daubechies, I., *Ten lectures on wavelets*. 1992, Philadelphia, Pa.: Society for Industrial and Applied Mathematics. xix, 357.
7. Strang, G. and T. Nguyen, *Wavelet and Filter Banks*. 1996, Wellesey: Wellesey-Cambridge Press.
8. Mallet, Y., et al., *Classification Using Adaptive Wavelets for Feature Extraction.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997. **19**: p. 1058-66.
9. Tian, G.Y., et al., *The application of wavelet transform in near infrared spectroscopy*, in *Spectroscopy and Spectral Analysis*. 2003. p. 1111-1114.
10. Pollen, D., *Parametrization of compactly supported wavelets.* Technical Report, AWARE Inc., 1989.
11. Sweldens, W., *The lifting scheme: A custom-design construction of biorthogonal wavelets.* Applied and Computational Harmonic Analysis, 1996. **3**(2): p. 186-200.
12. Galvao, R., et al., *Optimal wavelet filter construction using X and Y data.* Chemometrics and Intelligent Laboratory Systems, 2004. **70**(1-2): p. 1-10.
13. Coomans, D., et al. *Adaptive Wavelet Methodology for Mapping Spectral Data Sets*. in *International Conference on Computer Intelligence for Modelling Control and Automation*. 2001. Las Vegas.
14. Mallet, Y., D. Coomans, and O. deVel, *Recent developments in discriminant analysis on high dimensional spectral data.* Chemometrics and Intelligent Laboratory Systems, 1996. **35**(2): p. 157-173.
15. Walczak B (ed), *Wavelets in Chemistry*. 2000, Amsterdam: Elsevier.
16. Brown, P.J., T. Fearn, and M. Vannucci, *Bayesian wavelet regression on curves with application to a spectroscopic calibration problem.* Journal of the American Statistical Association, 2001. **96**(454): p. 398-408.
17. Kautsky, J. and R. Turcajova, *Pollen Product Factorization and Construction of Higher Multiplicity Wavelets.* Linear Algebra and Its Applications, 1994. **222**: p. 241-260.
18. Addleman, S., *The Generalized Randomized Block Design.* The Americian Statisitican, 1969. **23**(4): p. 35-36.
19. Dåbakk, E., et al., *Sampling reproducibility and error estimation in near infrared calibration of lake sediments for water quality monitoring.* Journal of Near Infrared Spectroscopy, 1999. **7**: p. 241-250.

20. Naydenova, Y. and P. Tomov, *Near infrared spectroscopy estimation of feeding value of forage perennial grasses in breeding programmes by global and specific calibrations. stimation of chemical composition and digestibility.* Joural of Near Infrared Spectroscopy, 1998. **6**: p. 153-165.
21. Roggo, Y., et al., *Sucrose content determination of sugar beets by near infrared reflectance spectroscopy. Comparison of calibration methods and calibration transfer.* Journal of Near Infrared Spectroscopy, 2002. **10**: p. 137-150.
22. Barker, M. and W. Raynes, *Partial least squares for discrimination.* Journal of Chemometrics, 2003. **17**: p. 166-173.
23. Geladi, P., H. Bergner, and L. Ringqvist, *From experimental design to images to particle size histograms to multiway analysis. An example of peat dewatering.* Journal of Chemometrics, 2000. **14**(3): p. 197-211.
24. Smilde, A., et al., *ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data.* Bioinformatics, 2005. **21**(13): p. 3043-3048.
25. Hastie, T., A. Buja, and R. Tibshirani, *Penalized Discriminant-Analysis.* Annals of Statistics, 1995. **23**(1): p. 73-102.
26. Yu, B., et al., *Penalized discriminant analysis of in situ hyperspectral data for conifer species recognition.* IEEE Transactions on Geoscience and Remote Sensing, 1999. **37**(5): p. 2569-2577.
27. Breiman, L., *Random Forests, Technical Report 421*. 2001, Department of Statistics, University of Califorina: Califorina.
28. Sekulic, S. and B.R. Kowalski, *Mars - a Tutorial.* Journal of Chemometrics, 1992. **6**(4): p. 199-216.
29. Tan, H.W. and S.D. Brown, *Dual-domain regression analysis for spectral calibration models.* Journal of Chemometrics, 2003. **17**(2): p. 111-122.
30. Walczak, B., E. Bouveresse, and D.L. Massart, *Standardization of near-infrared spectra in the wavelet domain.* Chemometrics and Intelligent Laboratory Systems, 1997. **36**(1): p. 41-51.
31. Shao, X.G., et al., *A method for near-infrared spectral calibration of complex plant samples with wavelet transform and elimination of uninformative variables.* Analytical and Bioanalytical Chemistry, 2004. **378**(5): p. 1382-1387.
32. Berry, R.J. and Y. Ozaki, *Comparison of wavelets and smoothing for denoising spectra for two-dimensional correlation spectroscopy*, in *Applied Spectroscopy*. 2002. p. 1462-1469.
33. Tokairin, T. and Y. Sato, *Design of Nonlinear Adaptive Digital Filter by Wavelet Shrinkage*, in *Electronics and Communications in Japan*. 2003.
34. Szu, H., *Adaptive Wavelet Transforms*, in *Optical Engineering*. 1994. p. 2103-2103.
35. Friedman, J., *Multivariate adaptive regression splines (with discussion).* Annals of Statistics, 1991. **19(1)**: p. 1-141.
36. Hastie, T., R. Tibshirani, and A. Buja, *Flexible Discriminant-Analysis by Optimal Scoring.* Journal of the American Statistical Association, 1994. **89**(428): p. 1255-1270.
37. Barnes, R.J., M.S. Dhanoa, and S. Lister, *Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra.* Applied Spectroscopy, 1989. **43**: p. 772-777.
38. MathsWorks, *Matlab*.
39. MathsWorks, *Matlab Optimization Toolbox*.
40. Liaw, A. and M. Wiener. *R module - random Forest*. [cited.
41. Hastie, T. and R. Tibshirani. *R module - mda*. [cited.

42. The R Development Core Team. *R*. 2003 [cited; 1.8.1:[
43. Teppola, P. and P. Minkkinen, *Wavelet-PLS regression models for both exploratory data analysis and process monitoring.* Journal of Chemometrics, 2000. **14**(5-6): p. 383-399.
44. Teppola, P. and P. Minkkinen, *Wavelets for scrutinizing multivariate exploratory models - interpreting models through multiresolution analysis.* Journal of Chemometrics, 2001. **15**(1): p. 1-18.
45. Fraley, C., *Comments on D. Coleman, X. Dong, J. Hardin, DM Rocke, DL Woodruff, Some computational issues in cluster analysis with no a priori metric, Computational Statistics & Data Analysis 31 : 1-12 (July 1999).* Computational Statistics & Data Analysis, 2000. **33**(2): p. 131-133.
46. Fraley, C. and A.E. Raftery, *MCLUST: Software for model-based cluster analysis.* Journal of Classification, 1999. **16**(2): p. 297-306.
47. Fraley, C. and A.E. Raftery, *How many clusters? Which clustering method? Answers via model-based cluster analysis.* Computer Journal, 1998. **41**(8): p. 578-588.
48. Banfield, J.D. and A.E. Raftery, *Model-Based Gaussian and Non-Gaussian Clustering.* Biometrics, 1993. **49**(3): p. 803-821.
49. Binder, D.A., *Bayesian cluster analysis.* Biometrica, 1978. **65**: p. 31-38.
50. Schwarz, G., *Estimating the dimension of a model.* The Annals of Statistics, 1978. **6**: p. 461-464.
51. Leung, A.K., F. Chau, and J. Gao, *A review on applications of wavelet transform techniques in chemical anlaysis: 1989-1997.* Chemometrics and Intelligent Laboratory Systems, 1998. **43**(1-2): p. 165-184.
52. Kass, R.E. and A.E. Raftery, *Bayes Factors.* Journal of the American Statistical Association, 1995. **90**(430): p. 773-795.
53. Bucci, O.M., A. Capozzoli, and G. D'Elia, *Determination of the convex hull of radiating or scattering systems: a new, simple and effective approach.* Inverse Problems, 2002. **18**(6): p. 1621-1638.
54. Wickerhauser, M.V., *Some problems related to wavelet packet bases and convergence.* Arabian Journal for Science and Engineering, 2003. **28**(1C): p. 45-58.
55. Wickerhauser, M.V., *Time Localization Techniques for Wavelet Transforms.* Croatica Chemica Acta, 1995. **68**(1): p. 1-27.
56. Wickerhauser, M.V., *Fast Approximate Wavelet Algorithms for Image-Processing, Classification, and Recognition.* Optical Engineering, 1994. **33**(7): p. 2225-2235.
57. Vogt, F. and M. Tacke, *Fast principal component analysis of large data sets based on information extraction.* Journal of Chemometrics, 2002. **16**: p. 562-575.
58. Walczak, B. and D. Massart *Wavelet packet transform applied to a set of signals: A new approach to the best-basis selection.* Chemometrics and Intelligent Laboratory Systems, 1997. **38**(1): p. 39-50.
59. Petricoin, E.F., *Serum Proteomic Patterns for Detection of Prostate Cancer.* Journal of the National Cancer Institute, 2002. **94**(20): p. 1576-78.
60. Friedman, J., *Stochastic Gradient Boosting.* Technical Report, Dept of Statistics, Stanford University, 1999.
61. Breiman, L., *Bagging Predictors.* Technical Report 421, Dept of Statistics, University of California, 1994.

62. Breiman, L., et al., *Classification and Regression Trees*. 1984, New York: Chapman & Hall.
63. Freund, Y. and R. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting.* Computational Learning Theory, 1995.
64. Han, D.G. and D.R. Larson, *Frames, bases and group representations.* Memoirs of the American Mathematical Society, 2000. **147**.
65. Baggerly, K.A., J.S. Morris, and K.R. Coombes, *Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments.* Bioinformatics, 2004. **20**(5): p. 777-U710.
66. Qu, Y.S., et al., *Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients.* Clinical Chemistry, 2002. **48**(10): p. 1835-1843.
67. Vannucci, M., N. Sha, and P.J. Brown, *NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection.* Chemometrics and Intelligent Laboratory Systems, 2005. **77**(1-2): p. 139-148.
68. Cocchi, M., R. Seeber, and A. Ulrici, *Multivariate calibration of analytical signals by WILMA (wavelet interface to linear modelling analysis).* Journal of Chemometrics, 2003. **17**: p. 512-527.
69. Perrin, F.H., *Whose Absorption Law?* Journal of the Optical Society of America, 1948. **38**(1): p. 72-74.
70. Jolliffe, I.T., *Principal Component Analysis*. 1986, New York: Springer-Verlag.
71. Donald, D., Y. Everingham, and D. Coomans, *Integrated wavelet principal component mapping for unsupervised clustering on near infra-red spectra.* Chemometrics and Intelligent Laboratory Systems, 2005. **77**(1-2): p. 32-42.
72. Donald, D., et al., *Adaptive wavelet modelling of a nested 3 factor experimental design in NIR chemometrics.* Chemometrics and Intelligent Laboratory Systems, 2006. **82**: p. 122-129.
73. Breiman, L., et al., *Classification and Regression Trees*. 1993, New York: Chapman & Hall.
74. Brown, P.J., M. Vannucci, and T. Fearn, *Multivariate Bayesian variable selection and prediction.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1998. **60**: p. 627-641.
75. Berk, R.A., *An Introduction to Ensemble Methods for Data Analysis.* Sociological Methods & Research, 2006. **34**(3): p. 263-295.
76. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. 2001, New York: Springer-Verlag.
77. Mallat, S.G., *Multiresolution Approximations and Wavelet Orthonormal Bases of L2(R).* Transactions of the American Mathematical Society, 1989. **315**(1): p. 69-87.
78. Daubechies, I. and W. Sweldens, *Factoring Wavelet Transforms into Lifting Steps*. 1997.
79. Vaidyanathan, P.P., *Theory and Design of M-Channel Maximally Decimated Quadrature Mirror Filters with Arbitrary M, Having the Perfect-Reconstruction Property.* IEEE Transactions on Acoustics Speech and Signal Processing, 1987. **35**(4): p. 476-492.
80. Lindley, D.V., *The Choice of Variables in Multiple Regression.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1968. **30**(1): p. 31-66.

81. Dawid, A.P., *Some Matrix-Variate Distribution-Theory - Notational Considerations and a Bayesian Application.* Biometrika, 1981. **68**(1): p. 265-274.
82. Vannucci, M. and F. Corradi, *Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1999. **61**: p. 971-986.
83. Madigan, D. and J. York, *Bayesian Graphical Models for Discrete-Data.* International Statistical Review, 1995. **63**(2): p. 215-232.
84. George, E.I. and R.E. McCulloch, *Approaches for Bayesian variable selection.* Statistica Sinica, 1997. **7**(2): p. 339-373.
85. Raftery, A.E., D. Madigan, and J.A. Hoeting, *Bayesian model averaging for linear regression models.* Journal of the American Statistical Association, 1997. **92**(437): p. 179-191.
86. Horchner, U. and J.H. Kalivas, *Simulated-Annealing-Based Optimization Algorithms - Fundamentals and Wavelength Selection Applications.* Journal of Chemometrics, 1995. **9**(4): p. 283-308.
87. Leardi, R., R. Boggia, and M. Terrile, *Genetic Algorithms as a Strategy for Feature-Selection.* Journal of Chemometrics, 1992. **6**(5): p. 267-281.
88. Vannucci, M., P.J. Brown, and T. Fearn, *A decision theoretical approach to wavelet regression on curves with a high number of regressors.* Journal of Statistical Planning and Inference, 2003. **112**(1-2): p. 195-212.
89. Efron, B., *Estimating the error rate of a prediction rule: some improvements on cross-validation.* Journal of the American Statistical Association, 1983. **78**: p. 316-331.
90. Osborne, B.G., et al., *Application of near-Infrared Reflectance Spectroscopy to the Compositional Analysis of Biscuits and Biscuit Doughs.* Journal of the Science of Food and Agriculture, 1984. **35**(1): p. 99-105.
91. Good, I.J., *The estimation of probabilities; an essay on modern Bayesian methods*. 1965, Cambridge,: M.I.T. Press. xii, 109 p.
92. Brown, P.J., T. Fearn, and M. Vannucci, *The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach.* Biometrika, 1999. **86**(3): p. 635-648.
93. Clarke, B., *Comparing Bayes Model Averaging and Stacking When Model Approximation Erro Cannot be Ignored.* Journal of Machine Learning Research, 2003. **4**: p. 683-712.
94. Heller, P., *Regular M-band Wavelets. Technical Report, AWARE Inc.* Submitted to IEEE Trans. on Signal Processing, 1992.
95. Commoner, B. and D. Lipkin, *The Application of the Beer-Lambert Law to Optically Anistropic Systems.* Science, 1949. **110**: p. 41-43.