Wavelet Based Feature Extraction Methods for the Discrimination and Regression of Spectral Data

Thesis submitted by
Yvette Lelia MALLET Bsc(Hons) *Qld*
in October 1997

for the degree of Doctor of Philosophy
in the School of Computer Science, Mathematics and Physics
James Cook University of North Queensland

In Memory of Tes Everingham

Statement of Access

I, the undersigned, the author of this thesis, understand that James Cook University of North Queensland will make it available for use within the University Library and, by microfilm or other means, allow access to users in other approved libraries. All users consulting this thesis will have to sign the following statement:

> In consulting this thesis I agree not to copy or closely paraphrase it in whole or in part without the written consent of the author; and to make proper public written acknowledgement for any assistance which I have obtained from it.

Beyond this, I do not wish to place any restriction on access to this thesis.

..... 3/3/98 .....
(Date)

Statement on Sources
Declaration

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 Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

..... 23.1.06 .....
(Date)

ACKNOWLEDGEMENTS

Foremostly, I thank my supervisors A/Prof Danny Coomans and A/Prof Olivier de Vel, for their professionalism and encouragement in the development of this thesis. I am also grateful for the many hours that they have spent revising this and other manuscripts throughout the course of my research work.

I extend my appreciation to Professor Massart for allowing me to visit his laboratory at the Free University of Brussels and providing me with the opportunity to become more familiar with NIR data. Also from the Free University, I thank Delphine Jouan-Rimbaud, Paula Fernandez, Eric Bouveresse, Wu Wen, Beate Walczak and Wim Penninckx, for their assistance.

Sincere thanks is expressed to Dr Jaroslav Kautsky at Flinders University in Adelaide, who gave up his valuable time to introduce me to wavelets. I would also like to thank Dr Bill Moran, Radka Turcajová and Pavel Turcaj for their assistance and cooperation whilst I was visiting Flinders University.

I extend my appreciation, to Professor Trevor Hastie for supplying his S-plus code and providing valuable input. Many people have provided data which I have used as part of my thesis, these people are acknowledged in Sections 7.2 and 8.2 of this thesis.

Thanks to my colleagues in the School of Computer Science, Mathematics and Physics at James Cook University for their encouragement during the final stages of this thesis. In particular to Dr Wayne Read for revising parts of this thesis. The encouragement provided by A/Prof Bob Staudte was also much appreciated.

Whilst pursuing the research summarized in this thesis, I was primarily supported by the Australian Government in the form of an Australian Postgraduate Research Award.

I wish to thank my fiancé, Andrew Everingham for his continual support and encouragement. I would also like to thank my team-mates for their patience and understanding. Finally, I thank my mother for always thinking of the little things that mean so much.

Abstract

This thesis is concerned with the application of statistical methods to spectral data. A major concern which arises from spectral data is that the number of variables or dimensionality usually exceeds the number of available spectra. This leads to a degradation in performance of traditional statistical methods. There are basically two strategies which can be implemented for overcoming such situations. It is common practice to first reduce the dimensionality of the data by some feature extraction preprocessing method, and then use an appropriate low dimensional statistical procedure. An alternative procedure is to use a high dimensional statistical procedure which is capable of handling a large number of variables. This thesis considers both approaches, and investigates the applicability of wavelets as features for statistical analyses, as well as other feature extraction procedures. The particular statistical analyses investigated are discriminant and regression analysis.

It is shown that, the wavelet based methods, particularly wavelets which have been designed to suit a particular task, perform quite adequately when compared to traditional approaches.

Contents

List of Figures

List of Tables

List of Symbols

Non-bold Lower Case Letters

- $a(\mathbf{x})$ appreciation score of $\mathbf{x}$
- $a_{\text{ccr}}(\mathbf{x})$ appreciation score equal to 1 if $P(r \mid \mathbf{x}_{i(r)}) \geq P(r \mid \mathbf{x}_i)$and zero otherwise
- $a_A(\mathbf{x})$ appreciation score equivalent to $P(r \mid \mathbf{x}_{i(r)})$
- $a_Q(\mathbf{x})$ quadratic appreciation score of $\mathbf{x}$
- a_{il} lth element in the ith principal component vector
- a parameter used in RDA which weights the pooled covariance matrix
- b parameter used in RDA which controls shrinkage of the weighted pooled covariance matrix
- b_i ith element in the vector of estimated regression coefficients $\mathbf{b}$
- $\text{band}(j,t)$ τth band $\tau \in \{0,1,\ldots,m-1\}$ at the jth level $j \in \{J, J-1, \ldots, J - \max_{\text{lev}} + 1\}$ of the DWT
- $c_{j,k}$ scaling coefficients
- $d_{j,k}$ wavelet coefficients
- f_{V_j} orthogonal projection of $f(t)$ onto V_j
- $g(\mathbf{x}, r)$ classification score
- $g_{\text{blda}}(\mathbf{x}, r)$ BLDA classification score
- $g_{\text{bqda}}(\mathbf{x}, r)$ BQDA classification score
- h_k high pass filter coefficients
- $\hbar_{ii}$ is the element along the ith diagonal of the hat matrix $\mathcal{H}$
- j_* complex number $\sqrt{-1}$

- j parameter controlling the dilation of the wavelet basis functions
- k parameter controlling the translation of the wavelet basis functions
- ℓ_k low pass filter coefficients
- m number of bands in the DWT; downsampling rate
- $\max_{lev}$ maximum number of levels in the DWT.
- n number of observational units in the training data set
- n' number of observational units in the testing data set
- n_r number of observational units from class r; rth element in the vector n
- $n_{[l]}$ number of objects in node l of CART model
- n_{levels} number of levels that an object has been transformed, in the DWT
- p dimensionality of the data set
- p_* dimensionality of the reduced data set $p_* \ll p$
- p_o number of parameters to be estimated (including the intercept) in a MLR model
- $p(\mathrm{x})$ is the class probability density of x
- q the number of sub-matrices in the filter coefficient matrix A is $q+1$
- r index for class categories
- s_o minimum of one less than the total number of classes $(R-1)$, or the dimensionality (p).
- s_* number of discriminant variables used for assigning an object to a class; $s_* \leq s_o$
- x_i ith element in the data vector x
- $x_{i[l]}$ ith object in node l of CART model
- y_i ith element in the response vector y
- y_i' ith element in the test response vector y'

- $y_{i[l]}$ response value of ith object in node l of CART model
- $\hat{y}_{-i}$ predicted value of $\mathbf{x}_i$, obtained when $\mathbf{x}_i$ is deleted from the model building process
- $\hat{y}_i$ predicted response value for object $\mathbf{x}_i$
- $\hat{y}'_i$ predicted response value for object $\mathbf{x}'_i$
- y_{ij} element in row i and column j of $\mathbf{Y}$
- $\hat{y}_{ij}$ estimate of y_{ij}
- z index for wavelet filter $z = 1, \ldots, m-1$

Non-bold Upper Case Letters

- AIC Akaike's information criterion
- CCR correct classification rate
- CCR′ correct classification rate of test set
- C_p Mallows C_p
- CVCCR cross-validated correct classification rate
- DF degrees of freedom
- DEV deviation
- $\mathcal{D}(\mathbf{x}, r)$ distance between $\mathbf{x}$ and $\bar{\mathbf{x}}_r$ in the discriminant coordinate system
- E_{cross} cross entropy measure
- E_{sym} symmetric entropy measure
- E_{tot} total symmetric entropy measure
- $\mathcal{F}_{\text{CWT}}$ continuous wavelet transform
- $\mathcal{F}_{\text{DWT}}$ discrete wavelet transform
- $\mathcal{F}_{\text{FT}}$ Fourier transform

- $\mathcal{F}_{\text{WFT}}$ windowed Fourier transform
- J highest level in the DWT ; $J = \text{ceiling}(\log p / \log m)$
- $\mathcal{J}$ criterion function applied in the adaptive wavelet or LDB algorithm
- $\mathcal{J}_{\Lambda}$ Wilk's lambda discriminatory criterion function
- $\mathcal{J}_{E}$ entropy discriminatory criterion function
- $\mathcal{J}_{\text{cvqpm}}$ discriminatory criterion function based on the cross-validated quadratic probability measure
- $\mathcal{J}_{\text{cvrsq}}$ regression criterion function based on the cross-validated r-squared measure
- $L^2(\mathbb{R})$ space of square integrable functions
- M_{ij} i,jth element in the Lawton matrix
- MCR misclassfication rate
- MSE mean square error
- N_i node identity in CART model
- N_f number of filter coefficients with nonnegative indices
- P_A average probability that an object is assigned to the correct class
- P_{QPM} quadratic probability measure
- P_{CCR} probability of correctly classifying objects
- $P(r)$ prior probability for class r
- $P(r \mid \mathbf{x})$ posterior probability that given some vector $\mathbf{x}$ it is from class r
- $P(r \mid \mathbf{x}_{i(r)})$ posterior probability for the true class of $\mathbf{x}_i$
- $P(\mathbf{x} \mid r)$ class probability density function
- $P(r \mid l)$ proportion of objects in node N_l of a CART model which are from class r

- $P_{-i}(r \mid \mathbf{x}_i)$ posterior probability for $\mathbf{x}_i$ when the covariance matrices and mean vectors in the probability density function have been calculated in the absence of $\mathbf{x}_i$
- PRESS predicted residual sum of squares
- RSS residual sum of squares
- RSS_{p_o} residual sum of squares of a MLR model with complexity p_o
- R^2 coefficient of variation (r-squared)
- R total number of class categories in a set of data
- R^* integer value less than or equal to $R-1$
- TSS total sum of squares
- V number of testing groups used in a cross-validation routine
- V_j subspace containing all the possible approximations of functions in $L^2(\mathbb{R})$ at resolution 2^j
- W_j orthogonal complement ot v_j

Bold Lower Case Letters

- $\mathbf{a}_i$ ith vector of principal component coefficients with dimension $p \times 1$
- $\mathbf{b}$ estimated vector of regression coefficients
- $\mathbf{b}_{r_{os}}$ rth column of the matrix of regression coefficients for the optimal scoring problem, $\mathbf{B}_{\text{os}}$
- $\mathbf{b}_{\text{pls}}$ estimated vector of regression coefficients from a PLS model
- $\mathbf{c}_j$ scaling coefficients at resolution (or level) j
- $\mathbf{d}_j$ wavelet coefficients at resolution (or level) j
- $\mathbf{d}_j^{(z)}$ wavelet coefficients at resolution (or level) j produced from the filter matrix $\mathbf{D}_{j+1}^{(z)}$

- $\mathrm{e}_{(r)}^{[j]}(\tau)$ class energy vector of wavelet (or wavelet packet) coefficients
- ℓ vector of low pass filter coefficients
- n $R \times 1$ vector of class sample sizes
- $\mathbf{p}$ $n \times 1$ vector containing principal component scores
- r vector of residuals in the PLS algorithm
- $\mathbf{s}$ output from low pass filtering operation
- t latent variables from PLS model
- u_i normalized vectors which are used to construct the wavelet matrix A
- v normalized vector which is used to construct the wavelet matrix A
- $\mathbf{v}$ $p \times 1$ vector of discriminant coefficients
- w_1 sums of squares and cross product between $\mathbf{X}$ and $\mathbf{y}$
- $\mathbf{w}$ output from high pass filtering operation
- $\mathbf{x}$ $p \times 1$ training data vector
- $\bar{\mathbf{x}}$ $p \times 1$ mean vector of the training data set
- $\mathbf{x}'$ $p \times 1$ testing data vector
- $\mathbf{x}^*$ $p \times 1$ column object vector from $\mathbf{X}^*$
- $\mathbf{x}^{[j]}(\tau)$ column vector containing the coefficients in band(j, τ) of the DWT
- $\mathbf{x}_{i(r)}$ $p \times 1$ data vector from class r
- $\mathbf{x}_{i(r)}^*$ ith data object from $\mathbf{X}^*$ which belongs to class r
- $\bar{\mathbf{x}}_r^*$ mean class vector from $\mathbf{X}^*$
- ${}^{o}\mathbf{x}^{[j]}(\tau)$ wavelet packet coefficients which occur at the jth level in the τth band of the wavelet packet transform

- $\mathbf{y}$ $n \times 1$ vector of training response values (regression) or class labels (discriminant analysis)
- $\hat{\mathbf{y}}$ $n \times 1$ predicted vector of response values (regression) or class labels (discriminant analysis)
- $\mathbf{y}'$ $n' \times 1$ vector of test response values (regression) or class labels (discriminant analysis)
- $\hat{\mathbf{y}}'$ $n' \times 1$ predicted vector of test response values (regression) or class labels (discriminant analysis)
- $\mathbf{z}$ $n \times 1$ discriminant variable

Bold Upper Case Letters

- $\boldsymbol{A}$ wavelet matrix
- $\boldsymbol{A}_i$ sub-matrix of the wavelet matrix $\boldsymbol{A}$
- $\mathbf{B}$ matrix of multivariate regression coefficients
- $\mathbf{B}_{\text{os}}$ optimal scoring matrix of regression coefficients
- $\mathbf{C}_j$ low pass filtering matrix at level j in the DWT
- $\mathbf{D}_j$ high pass filtering matrix at level j in the DWT
- $\mathbf{D}_j^{(z)}$ high pass filtering matrix at level j in the DWT which contains the zth set of highpass filter coefficients
- $\mathbf{D}$ diagonal matrix whose ith diagonal element is equal to $D_{ii} = 1/\sqrt{\lambda_{i_{\text{fda}}}^2(1-\lambda_{i_{\text{fda}}}^2)}$
- F_i ith factor in the wavelet matrix $\boldsymbol{A}$
- $\mathbf{L}$ low pass convolution matrix
- $\mathcal{H}$ hat matrix $\mathcal{H} = \mathbf{X}^{\mathbf{T}}(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1}\mathbf{X}$
- $\mathbf{H}$ high pass convolution matrix

- $\mathbf{P}$ matrix whose ith column contains the principal component scores vector $\mathbf{p}_i$
- $\mathbf{P}_1$ is a matrix which augments $\mathbf{1}_n$ to the first column of $\mathbf{P}$
- $\mathbf{P}_X, \mathbf{P}_{X^*}$ linear projector matrices
- Q orthogonal matrix used in contruction of the wavelet matrix A
- R projection matrix used in contruction of the wavelet matrix A
- $\mathbf{S}_B$ between covariance matrix
- $\mathbf{S}_W$ within covariance matrix
- $\mathbf{S}_{\text{pooled}}$ pooled covariance matrix
- $\mathbf{S}_r$ covariance matrix of class r
- $\mathbf{T}$ matrix whose ith column contains the ith latent vector from PLS
- $\mathbf{V}_{s_o}$ matrix whose ith column is $\mathbf{v}_i$ for $i = 1, \ldots, s_o$.
- $\mathbf{X}$ $p \times n$ training data matrix
- $\mathbf{X}_1$ training data matrix whose first row is equal to $\mathbf{1}_n^T$
- $\mathbf{X}_c$ $p \times n$ centered training data matrix
- X' $p \times n'$ testing data matrix
- $\mathbf{X}^*$ $p \times n$ data matrix which results from some feature selection/transformation procedure based on $\mathbf{X}$.
- $\mathbf{X}^{[j]}(\tau)$ matrix containing the coefficients for the objects which would lie in band(j, τ)
- $\mathbf{Y}$ $n \times R$ class indicator matrix
- $\mathbf{Z}_{s_o}$ matrix whose ith column is $\mathbf{z}_i$ for $i = 1, \ldots, s_o$

Greek Letters

- β_i ith component in the vector of regression coefficients β
- $\delta(t)$ delta function
- δ_{ij} indicator variable; $\delta_{ij} = 1$ if $i = j$, zero otherwise
- ϵ_i ith component in the vector of regression residuals ϵ
- γ_i eigenvalue corresponding to the ith principal component
- λ is a measure of the discriminant criterion $\lambda = \mathbf{v}^T \mathbf{S}_B \mathbf{v}$
- $\lambda_{i_{\text{fda}}}$ ith element of $\boldsymbol{\lambda}_{\text{fda}}$
- Λ Wilk's Lambda
- $\Lambda^{(i)}$ Wilk's Lambda at the ith iteration of a stepwise routine
- $\mho(j, \tau)$ discriminatory measure of band(j, τ) in the wavelet packet transform
- ν_i ith element in $\boldsymbol{\nu}$
- ω frequency
- $\phi(t)$ scaling function
- $\phi_{j,k}(t)$ scaling basis function; $\phi_{j,k}(t) = m^{j/2}\phi(m^j t - k)$
- $\psi(t)$ mother wavelet function
- $\psi_{j,k}(t)$ wavelet basis function; children wavelets; $\psi_{j,k}(t) = m^{j/2}\psi(m^j t - k)$
- $\hat{\rho}_{ij}$ correlation between the ith principal component and the jth variable
- $\hat{\sigma}_{\mathrm{x}_i}$ sample standard deviation of x_i
- τ band label for the DWT; $\tau \in 0, 1, \ldots, m-1$
- ϱ rank of a matrix
- $\boldsymbol{\beta}$ vector of regression coefficients

- β_{pcr} vector of regression coefficients from a PCR model
- β_{pls} vector of regression coefficients from a PLS model
- λ_{fda} vector whose elements are the eigenvalues of $\Psi^{*T}\Psi^{*}/n$
- Λ_{fda} diagonal matrix whose ith element is equal to $\lambda_{i_{\text{fda}}}$
- $\eta(\mathrm{x}^{*})$ vector of fitted values for x^{*}
- $\bar{\eta}_r$ fitted centroid of all x^{*} objects belonging to class r
- ν vector of wavelengths
- Ψ^{*} class indicator matrix used in FDA and PDA
- $\hat{\Psi}^{*}$ estimate of the class indicator matrix Ψ^{*}
- Θ matrix whose columns are the eigenvectors of $\Psi^{*T}\Psi^{*}/n$

Miscellaneous Characters

- 1_i $i \times 1$ column vector whose elements are all equal to 1
- $\downarrow m$ downsample by a factor of m

List of Algorithms

Chapter 1

Thesis Summary

1.1 Overview

This thesis investigates different strategies for performing statistical analyses on near infrared (NIR) spectra [16, 110, 124]. In recent years, the popularity of NIR spectroscopy has increased enormously, perhaps at a much faster rate than which statistical methods for analysing NIR spectra have developed. The popularity of NIR spectroscopy and indeed similar forms of spectroscopy, can be attributed to the fact that spectral methods provide a relatively efficient, non-destructive technique for analyzing chemical substances. This has many great benefits for research and can be an extremely effective method to employ for monitoring quality control procedures in industry.

Near infrared spectra are obtained by directing electromagnetic radiation with a set wavelength at some sample whose state may be a solid, liquid or gas. The amount of radiation which is reflected (or absorbed) by the sample is then measured. By changing the wavelengths of the electromagnetic radiation by constant increments and plotting the amount of reflectance (or absorption) against each wavelength, a spectrum is produced. We refer to spectra which detail the amount of radiation which has been reflected, as reflectance spectra. Likewise, absorption spectra detail how much radiation has been absorbed. Figure 1.1 shows an absorption spectrum obtained by analyzing a sample of paraxylene.

Figure 1.2 was produced to provide some indication about the near infrared region of the electromagnetic distribution. The NIR region of the electromagnetic spectrum ranges from 750 nanometers (nm) to 25 micrometers (μm). These wavelengths are longer

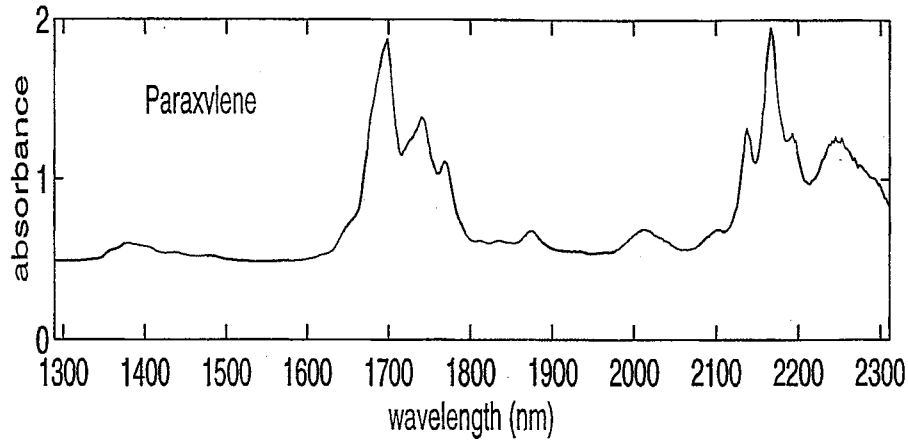
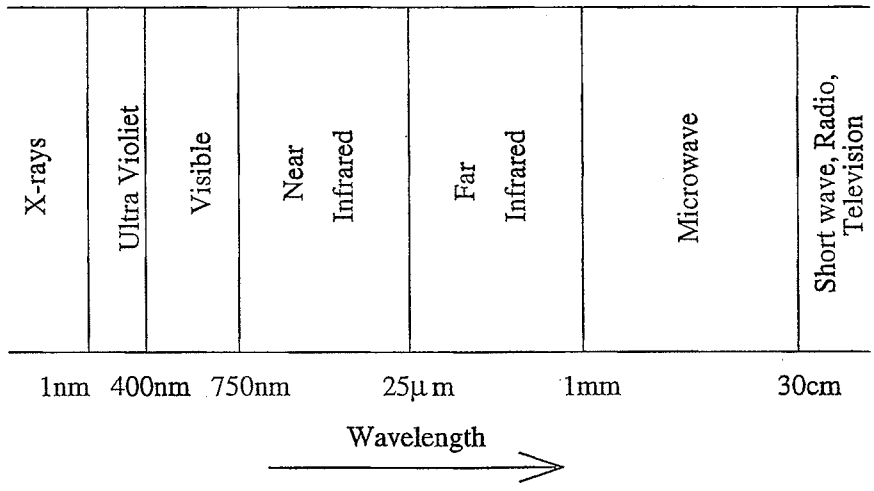


Figure 1.1: **A spectrum obtained from a sample of paraxylene.**

than the wavelengths which pertain to the visible part of the electromagnetic distribution and are much shorter than microwaves. Whilst Figure 1.2 implies that there is a cut-off point which separates the electromagnetic distribution into different regions, this is not actually the case. There is a considerable degree of overlap between the regions, and such descriptions about the electromagnetic distribution tend to vary from one text to another. The information used to produce Figure 1.2 was obtained from [142].



Figu 1.2: **The electromagnetic spectrum.**

The NIR spectra analyzed in this thesis, have wavelengths ranging from 900 nm – 2500 nm, although one data set (the seagrass data) extends into the visible region and has wavelengths incrementing from 400 nm up to 2500 nm (see Section 7.2.1).

Spectra usually vary depending on the chemical composition of the sample. This is due to molecules exhibiting different vibrational behaviours which interferes with the radiation reflected (or absorbed) for each of the wavelengths. It is quite difficult to ascertain the exact chemical composition of a substance by analyzing its NIR spectrum, but by placing particular attention on characteristics of the spectrum such as the shape, position and heights of peaks, some insight about the chemical composition of the sample may be obtained. This however, will often require the expertise of a skilled NIR analyst.

In this thesis automated statistical methods are investigated for exploring the characteristics of the NIR spectra. The statistical methods applied are discriminant analysis [102, 48] and regression analysis [29, 106].

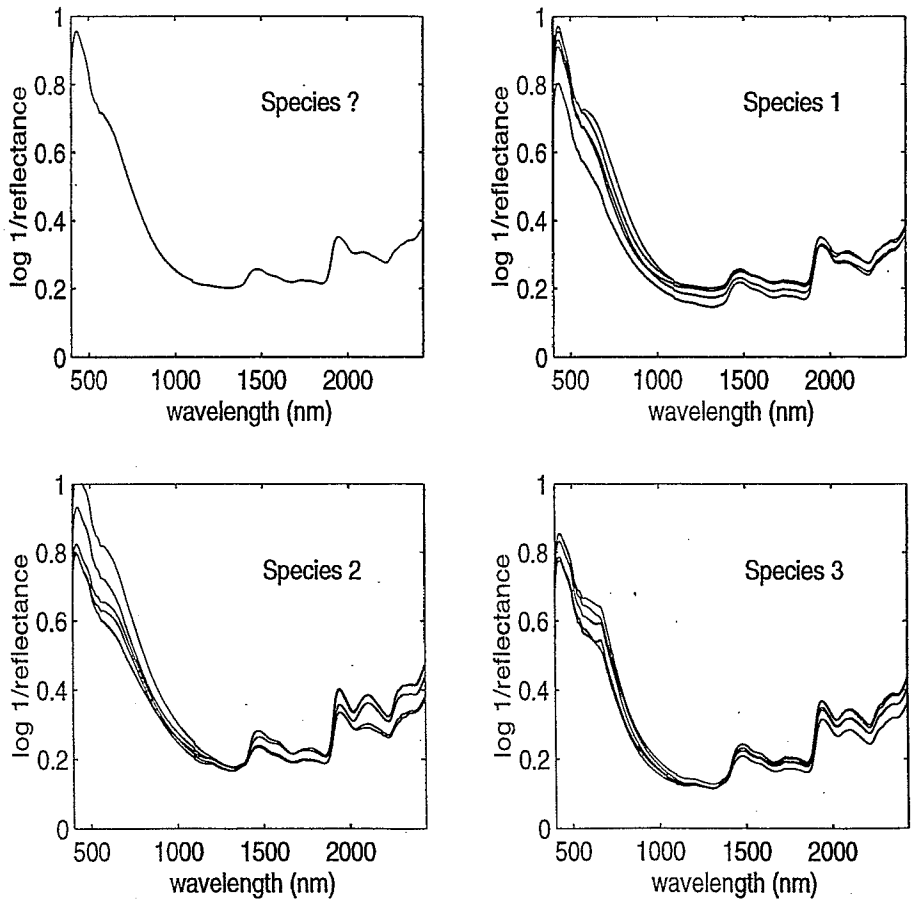


Figure 1.3: **A discriminant analysis problem.**

In the case of discriminant analysis one is interested in assigning spectra to one of several predefined categories. Figure 1.3 shows five sample spectra from three different

species of seagrasses which are referred to as Species 1, 2 and 3. The discriminant problem involves assigning the spectrum whose class identity is unknown into one of the classes (i.e. species). A simple approach is to look for similarities between the unidentified spectrum and the spectra which have been labelled. This task is not straightforward. For this data, it appears quite difficult for the human eye to detect any clue which may be able to distinguish the spectra from different classes. This problem highlights the relevance of discriminant methods for analysing spectral data.

Discriminant analysis involves trying to predict a discrete response (class label) from a set of predictor variables, which in this case are the reflectance (or absorbance) measures for each of the wavelengths. Regression analysis can be seen as an extension of discriminant analysis. For regression analysis, the response which is to be predicted (or modelled) using the predictor variables, is quantitative and may take on a continuous range of values.

A spectral data set used for performing statistical analyses will contain information about several spectra. Each spectrum represents a case or observational unit, and the wavelengths can be considered equivalent to the variables. Spectral information about the ith spectrum will be represented by the (column) data vector $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})^T$. Here p denotes the number of variables or the number of wavelengths for which the reflected (or absorbed) radiation of a sample has been measured. The symbol p may also refer to the dimensionality of the data. Each of the data vectors $\mathbf{x}_i$, for $i = 1, \ldots, n$ will be stored as columns in the $p \times n$ data matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ where n represents the number of spectra or observational units.

There are several difficulties which arise from analysing spectral data. One of the major problems is that the dimensionality p, is usually quite large, especially when compared to the number of available spectra n. Consequently the estimated parameters in the statistical models become highly variable and, in some instances, unobtainable due to numerical instabilities. This leads to a substantial performance degradation of the multivariate statistical model. Another issue is the existence of a high correlation structure in spectral data owing to the presence of a strong ordering in the variables. Such features are not limited to spectral data, and the statistical methods used in this thesis can be applied to many other forms of signals which exhibit an equivalent systematic ordering of the variables. Such ordering can for instance be made in time or space.

There are some statistical methods which have evolved in recent years with the aim of combating problems associated with high dimensionality and high correlation structures. Such techniques are referred to as high dimensional techniques and generally involve some form of regularization. High dimensional discriminant techniques include regularized discriminant analysis [44] and penalized discriminant analysis [61]. High dimensional regression methods include partial least squares [145, 41] and principal component regression [41, 37].

Techniques which begin to fail as the dimensionality becomes large when compared to the sample sizes are referred to as low dimensional techniques. Low dimensional discriminant methods include Fisher's linear discriminant analysis [34], flexible discriminant analysis [60] and the Bayesian linear and quadratic discriminant analysis [102]. The ordinary least squares multiple linear regression model is one of the most common regression methods and can be considered to be a low dimensional regression technique.

The high dimensional methods generally allow for a more automated procedure for modelling. Unfortunately though, many high dimensional methods have evolved quite recently and are therefore not as well publicised or understood by the scientific or industrial community. Also, it can be more difficult to apply high dimensional techniques since they are generally not standard procedures in most mathematical or statistical toolbox packages. Finally, most of these techniques provide few facilities for aiding the interpretation of the resulting multivariate prediction model. For these reasons, low dimensional methods are often preferred.

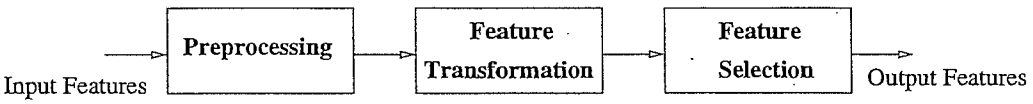


Figure 1.4: **Feature extraction model.**

Before low dimensional statistical methods are applied, some form of feature extraction should be implemented prior to the analyses. Feature extraction can consist of three main components as displayed in Figure 1.4. The first component involves preprocessing the

data. This can involve collecting the data and performing some standard data manipulations which may include transforming the data by perhaps the standard normal variate transformation [5] or the second derivative transformation [55, 5]. It may even involve subsampling the data, i.e. omitting every second or third variable. This can often be done with little loss of information due to the high correlation structure in the spectra. Once the data has been preprocessed, then it may undergo more complex variable transformations, by for example transforming the variables into orthogonal variables. This is the second component of the feature extraction model.

The third component is the feature selection algorithm which selects a subset of the transformed variables. Stepwise procedures are common feature selection algorithms. If feature selection is performed on the preprocessed data (without further transformation) then, the variable transformation can be seen as multiplying the preprocessed features with the identity matrix.

Many kinds of feature transformations have been proposed for spectral data ranging from univariate to multivariate transformations involving all the variables of the spectrum. Perhaps one of the most familiar feature transformations is principal component analysis (PCA). Principal component analysis is a multivariate technique which transforms the original variables into a new set of uncorrelated variables that are linear combinations of the original variables and are derived in decreasing order of variability. Of particular importance with spectral data is the order of the wavelengths. Unfortunately, PCA does not take advantage of the 'picture' (i.e spectrum) which is portrayed by the ordering of the variables.

The Fourier transform (FT) [88] can however be used to take into account the ordering of variables associated with a spectrum. The FT however, is a global transform and any localized changes which occur in a spectrum will be absorbed by most, if not all, of the Fourier coefficients. To avoid such global effects and to better identify localized changes, the wavelet transform [24, 128] can be quite a useful feature transformation to employ.

The wavelet transform produces a set of wavelet coefficients, which when linearly combined with a set of wavelet basis functions can be used to represent some function or signal. Wavelets are translated and dilated versions of some predefined wavelet called a 'mother wavelet'. Figure 1.5 shows some wavelet basis functions. Notice that they all have

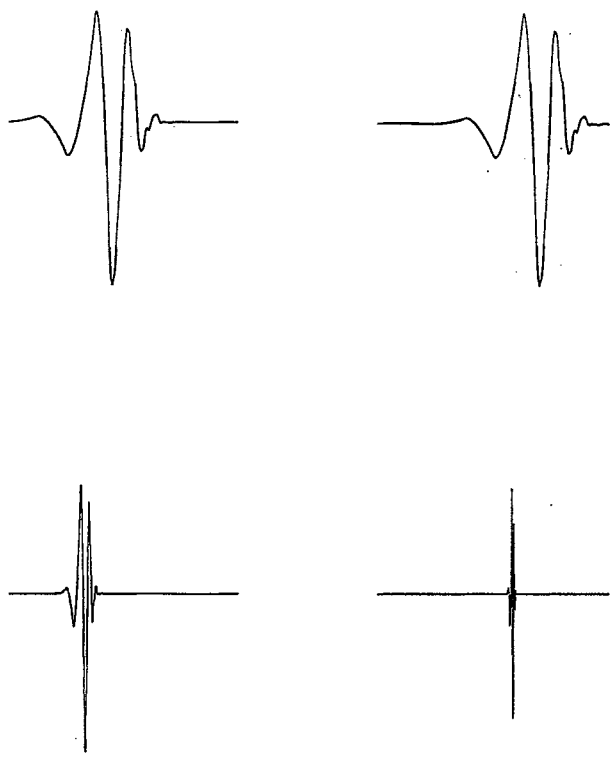


Figure 1.5: **Some wavelet basis functions.**

basically the same shape, but they differ in the amount which they are stretched (dilated) and shifted (translated) from one another.

The wavelets which we consider in this thesis are compact as seen in Figure 1.5, that is they are non-zero for a finite duration, and unlike sine and cosine waves used in the Fourier transform, they do not extend the entire horizontal axis. Since wavelets are local in space and are dilated by different amounts, the wavelet coefficients convey localized information about the frequency-like content of some function or signal. This makes the wavelet coefficients extremely useful features for representing small scale effects in spectral data. Examples which demonstrate this phenomenon are highlighted in Chapter 5.

There exists an abundant variety of wavelets and the fundamental problem to overcome is deciding which wavelet will best suit the particular application. A typical approach is to perform the wavelet transform based on a predefined (mother) wavelet from literature. The (mother) wavelet which produces the 'best' performance measures is then employed for future analyses. The performance measures will usually be based on some multivariate modelling criteria which is calculated using the wavelet coefficients produced from the

wavelet transform. Generally, a feature selection strategy will first be performed on the wavelet coefficients, before the performance measures are calculated.

We propose a new and innovative scheme which avoids the need to preselect a wavelet basis from literature. The (mother) wavelet is designed so that a specified multivariate modelling criterion is optimized. An appropriate criterion for discriminant analysis might be based on a correct classification rate, while an appropriate criterion for regression may involve the residual sum of squares.

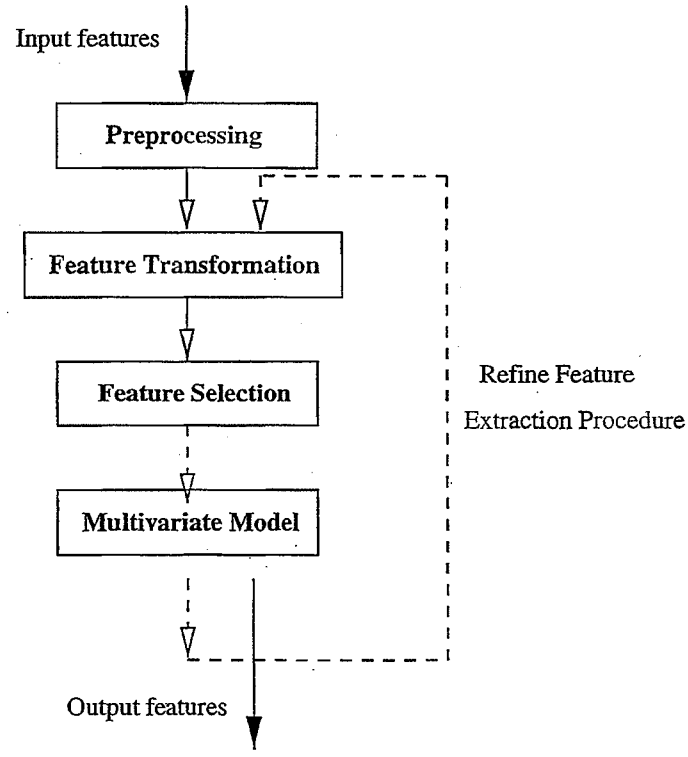


Figure 1.6: **Integrated feature extraction model.**

The wavelet gradually adapts to the application at hand, and continually updates the wavelet coefficients until the modelling criterion is optimal. The wavelet is referred to as an 'adaptive' or 'task-specific' wavelet, since it is adapting to the current task. This adaptive wavelet algorithm can be seen as an integrated feature extraction procedure. An integrated feature extraction procedure incorporates the multivariate model into the general feature extraction model as depicted in Figure 1.6

1.2 Thesis Structure and Contribution

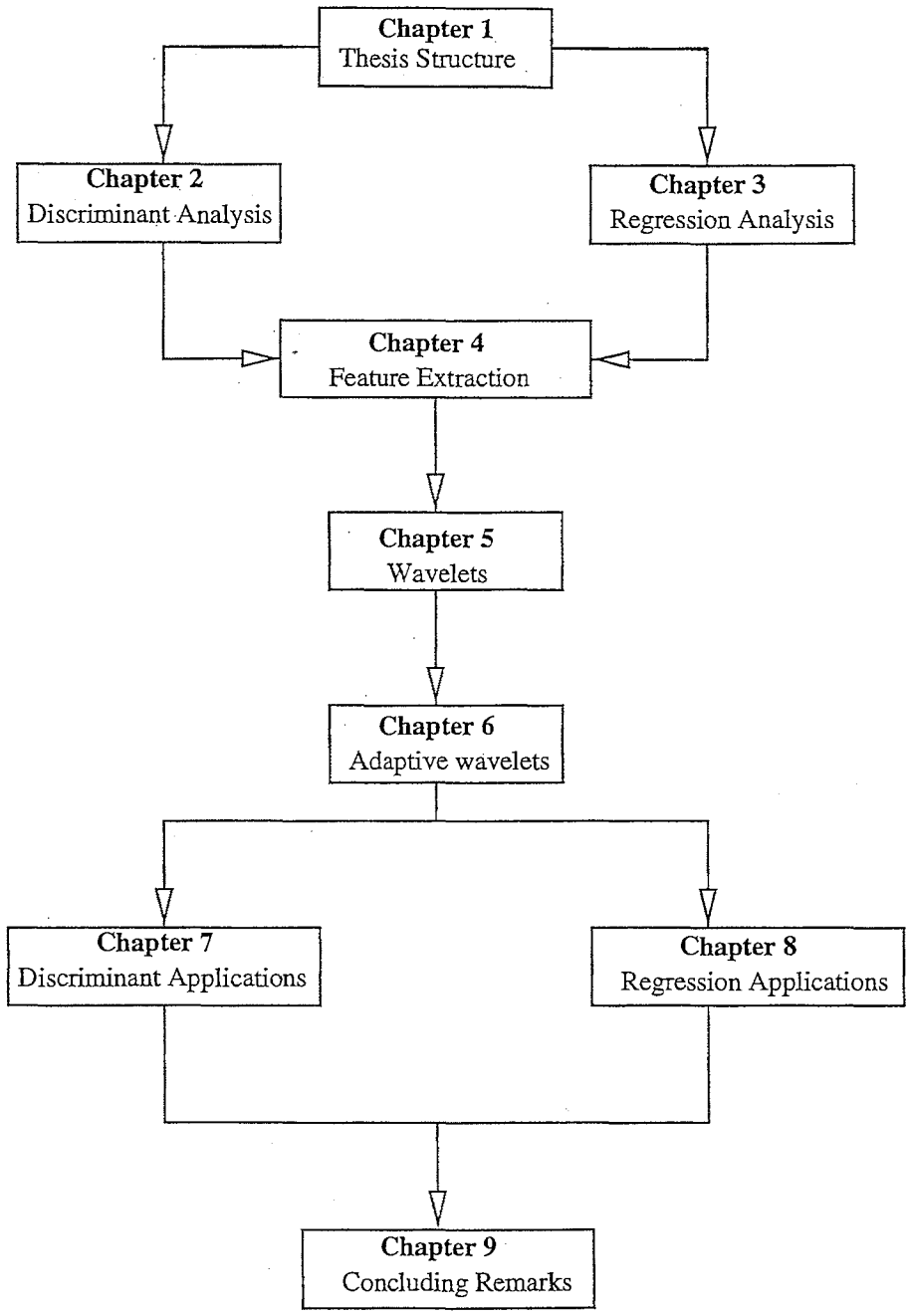


Figure 1.7: Thesis outline.

An outline of the structure of this thesis is summarized in Figure 1.7. The thesis continues from the overview by discussing discriminant analysis in Chapter 2 and regression analysis Chapter 3. The chapter on discriminant analysis is an expanded version of our

papers [97, 146]. The discriminant methods discussed are Fisher's linear discriminant analysis (FLDA), flexible discriminant analysis (FDA), penalized discriminant analysis (PDA), Bayesian linear and quadratic discriminant analysis (BLDA and BQDA) and regularized discriminant analysis (RDA). Each of these methods are applied to NIR spectral data sets in Chapter 7. The classification methods are introduced according to their origin, whether they be Fisher-based or Bayesian-based discriminant methods. Also in Chapter 2, is a discussion on different approaches for assessing the performance of discriminant models.

In Chapter 3, three regression methods are discussed — multiple linear regression (MLR) principal component regression (PCR) and partial least squares regression (PLS). This chapter together with our paper on nonparametric regression methods [93] provides a more detailed account on regression methods. Methods for assessing the adequacy of regression models are also presented in this chapter.

Chapter 4, introduces the two main approaches for feature extraction – feature selection and feature transformation. Preprocessing methods have been merged into the section on feature transformations. Of the variable transformation procedures it is mentioned that wavelet coefficients might be potentially good features to use as input to multivariate statistical techniques. Wavelets are discussed in greater detail in Chapter 5.

Wavelets have existed for many years, but it is only in the last decade that they have become increasingly popular. Much of this popularity can be attributed to Ingrid Daubechies, Yves Myer and Stéphanie Mallat. Many of the applications which utilize wavelet methodologies focus on function representation and image compression. Although there are many other applications for their use, such as solving partial differential equations, there have been relatively few applications where wavelets, or more precisely wavelet coefficients have been used as features for discriminant and regression problems, and in particular to the discrimination and regression of near-infrared spectral data. (Previous applications are documented in Chapter 4, Section 4.2.4)

Since the use of wavelet methodologies as a feature extraction procedures is quite new and remains relatively unexplored, it is important to investigate and gain further insight to their applicability of such procedures. This is one of the primary aims of this thesis.

The second aim is to investigate the potential of adaptive wavelets to discriminant and

regression problems for spectral data. Most applications of wavelets involve using standard or traditional wavelet bases which are already defined in the literature. We explore the advantages associated with designing individual wavelets to suit specific tasks. To the best of our ability, we have been unable to find references of wavelets which have been designed for discrimination and regression, that have not been based on, or linked in anyway to predefined, existing wavelets.

The adaptive wavelet methodology is based on the paper by Kautsky and Turcajová (1995) [78]. In this paper the authors describe a way in which a wavelet can be designed for removing disturbances in signals. Based on a similar algorithm, we investigate ways in which wavelets can be designed for multivariate statistical analyses. In [96] there is a detailed description about the adaptive wavelet algorithm and its applications to the classification of NIR spectral data. A summary of this paper is contained in [94]. A tutorial paper about the general application of adaptive wavelets can be found in [95].

Chapters 7 and 8 involve applications for the discrimination and regression of spectral data. Various feature extraction strategies along with several discriminant and regression methods are applied in each chapter, respectively. In conclusion, some final remarks and issues which arise from the topics presented in this thesis are discussed in Chapter 9.

Chapter 2

Discriminant Analysis

2.1 Introduction

Discriminant analysis techniques (also called classification techniques) are concerned with classifying objects into one of two or more classes. Discriminant techniques are considered to be learning procedures. Given, a set of objects whose class identity is known, a model 'learns' from the variables which have been measured for each of the objects, a procedure which can be used to assign a new object, whose class identity is unknown, into one of the predefined classes. Such a procedure is performed using a well defined discriminatory rule. One practical discriminant problem which is important to environmental scientists investigating the diets of dugongs, involves determining the species of seagrasses. The different categories or classes are formed by the various species, and the classification problem is then based on the chemical composition of the seagrasses which might be represented using spectra which measure the reflected radiation of various wavelengths.

Discriminant techniques are not necessarily used for the sole purpose of assigning objects into predefined classes. Sometimes it is of interest just to explore the group structures of the data, e.g. to visualize the positioning in space of the objects from the different classes, or, to determine which variables are important for discrimination. Thus, discriminant techniques themselves can be categorized into classes – those that:

1. are used for allocation
2. are used as exploratory procedures
3. are used for both allocation and exploratory procedures.

Fisher's linear discriminant analysis [34] (FLDA) which can be used for both allocation and descriptive purposes, is one of the traditionally favoured techniques. Typically, the discriminant analysis methods which are based on FLDA fall into the third category, whilst discriminant techniques based on probability measures such as Bayesian linear discriminant analysis (BLDA) and Bayesian quadratic discriminant analysis (BQDA) can be considered useful for allocation purposes only. It is important to pay consideration to the goal of the analysis and to choose the appropriate discriminant analysis procedure accordingly.

Discriminant techniques can be subdivided another way which is dependent upon the ratio of the number of observations (or cases) to the number of variables. Some classifiers begin to fail when the dimensionality (i.e. number of variables) becomes large compared to the number of observations. Despite what one would intuitively think, having a plentiful supply of variables does not necessarily improve the performance of the classifier. In fact, such a situation can cause the parameter estimates in the discriminant model to become highly variable (imprecise) leading to a degradation in the performance of the discriminant procedure [3, 21, 46, 69, 102, 140].

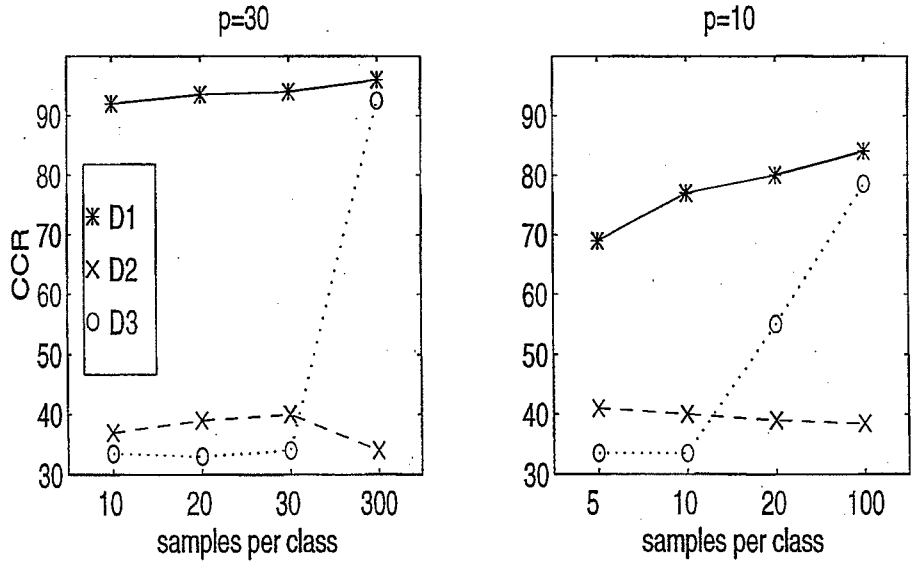


Figure 2.1: **Percentage of correctly classified objects obtained by three discriminant techniques (D1,D2 and D3) for eight combinations of dimensionality and class sample sizes.**

Figure 2.1 which is taken from [3], shows the classification performance (in terms of the correct classification rate, CCR) for three different discriminant techniques for various dimensionality and sample size settings of some simulated data. For the moment we will refer to the discriminant techniques as D1, D2 and D3. The two dimensionalities considered are 30 and 10. The class samples sizes are set at 10, 20, 30 and 300 when the dimensionality is 30. When the dimensionality is set at 10, the class sample size considered are 5, 10, 20 and 100. The data used in this example have been simulated so that there are three classes which have different circular class covariance matrices. One general observation which can be made from Figure 2.1 is that the discriminant method D1 seems to be less affected by the varying observation-to-variable ratio and consistently outperforms the discriminant methods D2 and D3. Another observation which can be made is that for small observation-to-variable ratios, the discriminant method D2 produces higher classification rates than D3. The discriminant method D3 however, produces much higher classifier rates when the class sample sizes are very much bigger than the number of variables.

We refer to classifiers which are not suited to small observation-to-variable ratios as being low dimensional classifiers. Conversely, classifiers which are suited to small ratios are referred to as high dimensional classifiers. In Figure 2.1, the method D1 is actually a high dimensional classifier, while D2 and D3 are low dimensional classifiers.

Both low and high dimensional discriminant methods consist of linear and nonlinear discriminant methods. The linear methods produce linear decision boundaries, for assigning objects into a particular class, whilst nonlinear methods will generally form nonlinear decision boundaries for performing the same task. Figure 2.2 presents a schematic overview of some modern and common discriminant methods. Two common linear low dimensional methods include Fisher's linear discriminant analysis and Bayesian linear discriminant analysis [21, 102]. (The method D2 in Figure 2.1 is BLDA). Nonlinear low dimensional methods include Bayesian quadratic discriminant analysis (BQDA) [21, 102], flexible discriminant analysis (FDA) [60], kernel density and nearest neighbour methods [21, 102] and neural networks [116]. Bayesian quadratic discriminant analysis is a nonlinear extension of BLDA which has decision boundaries of a quadratic nature. BQDA involves estimating more parameters, namely the individual class covariance matrices, in the dis-

criminant model. Generally, for BQDA to have the potential to perform satisfactorily, the ratio of the number of objects per class should be much larger (eg at least 3 times) than the dimensionality. In Figure 2.1 D3 is BQDA. One can observe that as the ratio of the class sample sizes to the dimensionality increases, then BQDA (for this data) begins to outperform BLDA. Another nonlinear low dimensional method which has developed recently is Flexible discriminant analysis [60]. Flexible discriminant analysis combines nonparametric regression methods with Fisher's linear discriminant analysis to achieve greater nonlinearity and flexibility in the decision boundaries.

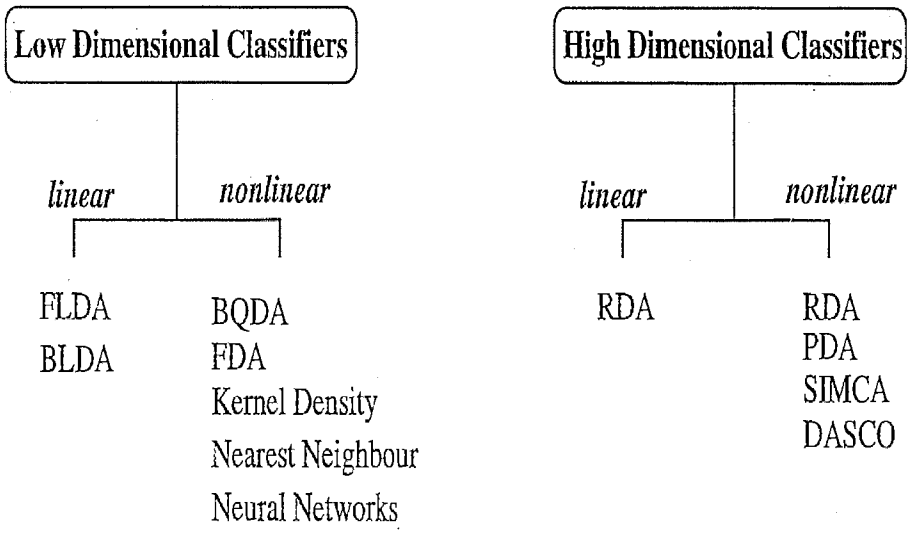


Figure 2.2: **Summary of some discriminant analysis methods.**

Due to the extensive amount of literature and wide availability of low dimensional classifiers, these methods are often the preferred candidates. If a low dimensional discriminant technique is to be used for classifying high dimensional spectral data, then it is recommended that the dimensionality of the data be reduced so that the observation-to-variable ratio becomes large. The dimensionality should be reduced with the goal of retaining as much relevant information as possible. Such a strategy is referred to as feature extraction. Feature extraction is discussed in greater detail in Chapter 4.

A distinct advantage of applying high-dimensional classifiers is the need for feature extraction can be avoided or greatly reduced. High-dimensional classifiers such as regularized discriminant analysis (RDA) [38, 44] and class modelling systems such as SIMCA [39, 38, 82, 144] and DASCO [38] are quite popular. RDA can produce linear or nonlinear

decision boundaries depending on certain parameters in the model which are dependent on the particular data set being analysed. Of late, Hastie *et. al.* [61] have developed a penalized discriminant method also capable of handling high dimensional data. Penalized discriminant analysis is based on the same principles as FDA, and thus stems from Fisher's linear discriminant analysis. The main difference between FDA and PDA is the nonparametric regression methods which are employed.

The chapter proceeds by introducing some notation and then discusses the time honoured technique FLDA. It is then shown that the low dimensional classifier, flexible discriminant analysis is a nonlinear extension of FLDA. The high dimensional classifier penalized discriminant analysis which is an extension of FLDA, derived from similar principles as FDA is also presented. The Bayesian classifiers introduced are – Bayesian linear and quadratic discriminant analysis, here BQDA is a nonlinear Bayesian extension of BLDA. The high dimensional classifier RDA is discussed next. It can be seen that RDA is a hybrid technique which is based on BLDA and BQDA. Model assessment criteria and evaluating techniques are also presented.

2.2 Notation

In many instances one will be given a set of training data consisting of n_r objects $\mathbf{x}_{i(r)}$ from class $r \in \{1, 2, \ldots, R\}$ giving a total of $n = \sum_{r=1}^{R} n_r$ objects. Each object $\mathbf{x}_i$ consists of measurements made on p variables and can be represented as a data vector of the form $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots x_{pi})^T$, where p also indicates the dimensionality of the data set. In the case of a spectral data set, each object will represent a spectrum. For each training object $\mathbf{x}_i$ the class identity $y_i \in \{1, 2, \ldots, R\}$ is known. The training objects are stored as columns in the $p \times n$ data matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ and we prefer that the class labels are stored in the $n \times 1$ column vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. The reason for defining $\mathbf{X}$ to be a $p \times n$ matrix, which is in slight contrast to the dimension of $\mathbf{y}$, is to allow for a simplification of notation when wavelets are introduced in Chapters 5 and 6.

A discriminant model which is assessed using the same training data which designed the model will usually reflect overly optimistic results. It can be appropriate to use an independent test set for assessing the validity of the model. Let $\mathbf{X}'$ define the testing data which contains n' objects $\mathbf{x}'_i$ with n'_r objects from class r such that $n' = \sum_{r=1}^{R} n'_r$ and $\mathbf{y}'$

denotes the vector of true class labels of the testing data.

2.3 Fisher's linear Discriminant Analysis (FLDA)

Fisher's linear discriminant analysis is sometimes referred to as canonical discriminant analysis due to the equivalence between FLDA and canonical correlation analysis [87]. Fisher's linear discriminant analysis seeks linear combinations of the measurement variables which separate the objects from different classes as much as possible. Factors which determine the separability of classes include the distances between groups and the compactness of each group. It then follows, that the ratio of the between-to-within variability of the transformed training data vectors (i.e. spectra) should be maximized. Equivalently, we seek the linear transformation

$$\mathbf{z} = \mathbf{X}^T\mathbf{v} \tag{2.1}$$

that maximizes

$$\mathbf{v}^T\mathbf{S}_B\mathbf{v} \tag{2.2}$$

subject to $\mathbf{v}^T\mathbf{S}_W\mathbf{v} = 1$, where $\mathbf{v} = (v_1, v_2, \ldots, v_p)^T$ is the vector of discriminant coefficients, and $\mathbf{S}_B$ and $\mathbf{S}_W$ are the between- and within-covariance matrices of the data matrix $\mathbf{X}$, respectively. These are defined by

$$\begin{aligned}\mathbf{S}_B &= \frac{1}{n}\sum_{r=1}^{R} n_r(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})^T \\ \mathbf{S}_W &= \frac{1}{n}\sum_{r=1}^{R}\sum_{i=1}^{n_r}(\mathbf{x}_{i(r)} - \bar{\mathbf{x}}_r)(\mathbf{x}_{i(r)} - \bar{\mathbf{x}}_r)^T\end{aligned}$$

where, $\mathbf{x}_{i(r)}$ is an object from class r, $\bar{\mathbf{x}}_r = \sum_{i=1}^{n_r}\mathbf{x}_{i(r)}/n_r$ is the mean vector or centroid of class r and,

$$\bar{\mathbf{x}} = \frac{1}{R}\sum_{r=1}^{R}\bar{\mathbf{x}}_r = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

is the overall mean vector. Fisher's linear discriminant analysis does not restrict the class populations to be multivariate normal, but does assume the class covariance matrices

$$\mathbf{S}_r = \frac{1}{n_r}\sum_{i=1}^{n_r}(\mathbf{x}_{i(r)} - \bar{\mathbf{x}}_r)(\mathbf{x}_{i(r)} - \bar{\mathbf{x}}_r)^T \quad \text{for} \quad r = 1, \ldots, R$$

are equal [71], that is $\mathbf{S}_1 = \mathbf{S}_2 = \cdots = \mathbf{S}_R$. The maximization problem reduces to solving

$$(\mathbf{S}_B - \lambda\mathbf{S}_W)\mathbf{v} = \mathbf{0} \tag{2.3}$$

or, assuming the inverse of $\mathbf{S}_W$ exists,

$$(\mathbf{S}_W^{-1}\mathbf{S}_B - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}. \tag{2.4}$$

Notice that there will be $s_o = \min(R-1, p)$ eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{s_o}$ and s_o corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{s_o}$ which produce s_o discriminant vectors (or variables) $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{s_o}$ such that $\mathbf{z}_i = \mathbf{X}^T\mathbf{v}_i$. The discriminant variables will have an identity within covariance.

It is convenient if the vectors $\mathbf{v}_i$ and $\mathbf{z}_i$ (for $i = 1, \ldots, s_o$) are stored as columns in the matrices $\mathbf{V}_{s_o}$ and $\mathbf{Z}_{s_o}$, that is, $\mathbf{V}_{s_o} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{s_o})$ and $\mathbf{Z}_{s_o} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{s_o})$, then $\mathbf{Z}_{s_o} = \mathbf{X}^T\mathbf{V}_{s_o}$ gives the coordinates of the objects in the s_o-dimensional discriminant coordinate system.

If Equation 2.3 is premultiplied with $\mathbf{v}^T$, we can see that $\lambda = \mathbf{v}^T\mathbf{S}_B\mathbf{v}$ is a measure of the discriminant criterion. The first discriminant variable gives the largest measure of the discriminant criterion. The second discriminant variable achieves the next largest discriminant criterion such that $\mathbf{z}_2$ is uncorrelated with $\mathbf{z}_1$, and so on for $\mathbf{z}_3, \ldots, \mathbf{z}_{s_o}$. For more details the reader is referred to Tatsuoka [132] and Lebart [87].

FLDA assigns an object $\mathbf{x}$ to the class $r \in 1, \ldots, R$, which minimizes

$$\mathcal{D}(\mathbf{x}, r) = \|\mathbf{x}\mathbf{V}_{s_*} - \bar{\mathbf{x}}_r\mathbf{V}_{s_*}\|^2, \tag{2.5}$$

where $s_* \leq s_o$ discriminant variables are used. Here, $\bar{\mathbf{x}}_r^T\mathbf{V}_{s_*}$ is the centroid for class r in the discriminant coordinate system. Thus, $\mathbf{x}$ is assigned to the class r for which the distance between $\mathbf{x}^T\mathbf{V}_{s_*}$ and $\bar{\mathbf{x}}_r^T\mathbf{V}_{s_*}$ is minimum.

Whilst FLDA can be used for predicting the class membership of future objects, it is perhaps best recognized for its graphical element. When the discriminant variables are plotted against each other, one can gain further insight to the structure of the data. Figure 2.3 plots the first two discriminant variables ($\mathbf{z}_1$ and $\mathbf{z}_2$) against each other. Since the discriminant variables are derived in order of separability, most separation among the classes will generally be observed in the first few discriminant variables. Note that in order to produce a discriminant plot prior knowledge about the class identity of the objects in $\mathbf{X}$ is required.

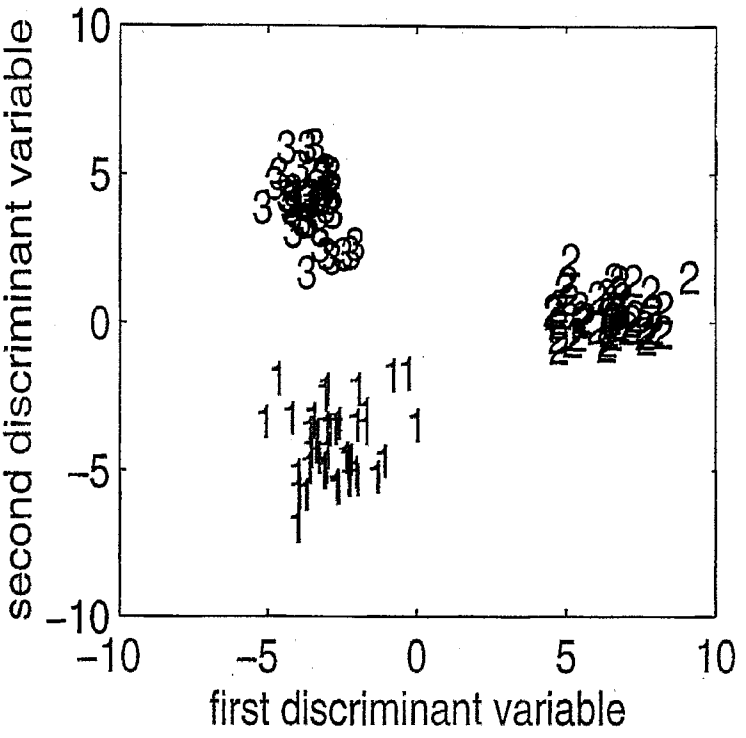


Figure 2.3: A scatterplot of the discriminant scores produced by FLDA.

2.4 Flexible Discriminant Analysis (FDA)

Flexible discriminant analysis is a nonlinear extension of FLDA which incorporates non-parametric regression methods to obtain nonlinear decision boundaries. This is achieved by first casting regression and classification into a common framework.

It is a well known fact that when $R = 2$, Fisher's discriminant coefficients are proportional to the coefficients of the multiple linear regression (MLR) model, where the variables (rows) in the data matrix X form the predictors, and the response is the vector of class labels.[1] When $R > 2$ the relationship between linear regression and classification is not so straight forward. One obvious approach is to produce a $n \times R$ class indicator matrix Y ($y_{ir} = 1$ if x_i belongs to class r and zero otherwise) and use multivariate linear regression (MVLR) [2] to predict the columns of Y. The object x_i is assigned to the class r which has

[1] This result is easily verified [87] by coding the response with the dichotomous labels 0 and 1 for classes 1 and 2 respectively, and noting that Equation 2.4 is equivalent to $(\mathrm{S}_T^{-1}\mathrm{S}_B - \lambda\mathrm{I})\mathrm{v} = 0$ with $\mathrm{S}_T = \mathrm{S}_W + \mathrm{S}_B$.

[2] The difference between MLR and MVLR is, MLR models a single response vector, whereas MVLR models several responses.

the largest value of $\{\hat{y}_{ir}\}_{i=1}^{R}$, where $\hat{y}_{ir}$ denotes the estimate of y_{ir}. The estimate, $\hat{y}_{ir}$ will not necessarily lie between 0 and 1. An alternative procedure referred to as softmax [60], assigns $\mathbf{x}_i$ to the class r which has the maximum value of $\exp(\hat{y}_{ir})/\sum_{r=1}^{R}\exp(\hat{y}_{ir}) \in [0,1]$. Hastie *et. al.* [60] suggest softmax generally does not perform as favourably as FLDA.

A more sophisticated approach for relating regression and classification is that of Breiman and Ihaka [11]. They make use of optimal scoring to establish an equivalence between linear regression and FLDA. Hastie *et. al.* [60] extend this relationship to allow for non-parametric (multivariate) linear regression methods. This technique is referred to as flexible discriminant analysis (FDA) and is described in Figure 2.4.

Flexible Discriminant Analysis

1. Construct the class indicator matrix Ψ^*.
2. Based on $\mathbf{X}$, perform a multivariate regression to predict Ψ^*. Let $\widehat{\Psi^*}$ be the predicted values of Ψ^*.
3. Calculate the eigenvectors and eigenvalues of $\left(\Psi^{*T}\widehat{\Psi^*}/n\right)$. Store the eigenvectors as columns in Θ and the eigenvalues in descending order in the vector λ_{fda}
4. Construct the diagonal matrix $\mathbf{D}$ which has $D_{ii} = 1/\sqrt{\lambda_{i_{\text{fda}}}(1-\lambda_{i_{\text{fda}}})}$. where $\lambda_{i_{\text{fda}}}$ is the ith element in λ_{fda}
5. Form discriminant variables $\widehat{\Psi^*}\Theta\mathbf{D}$.
6. Classify $\mathbf{x}^*$ into the class r which minimizes $||(\eta(\mathbf{x}^*) - \bar{\eta}_r)\mathbf{D}||$ where, $\eta(\mathbf{x}^*)$ is the predicted value of $\mathbf{x}^*$ obtained using the nonparametric regression model, multiplied with Θ. For classification based on posterior probabilities, use Equation 2.12.

Figure 2.4: **The FDA algorithm.**

The first step of the FDA algorithm involves forming a class indicator matrix Ψ^* whose columns are uncorrelated with zero mean and unit variance such that $\Psi^{*T}\Psi^* = \mathbf{I}$. If the class sample sizes are equal, then it is sufficient to have $\Psi^* = \mathbf{Y}/n_r$ as the class indicator matrix. The default procedure used in the Splus code of Hastie *et. al.* [60] constructs the indicator matrix Ψ^* by multiplying $\mathbf{Y}$ with another matrix Ψ as follows

$$\Psi^* = \mathbf{Y}\Psi. \tag{2.6}$$

The matrix Ψ, is formed by a series of steps which are described below.

1. A matrix Γ is constructed which has the form

$$\Gamma = \begin{pmatrix} 1 & -1 & -1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & -1 & -1 & -1 & \cdots & -1 \\ 1 & 0 & 2 & -1 & -1 & \cdots & -1 \\ 1 & 0 & 0 & 3 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & 0 & 0 & \cdots & R-1 \end{pmatrix}$$

   The general form of Γ is

   - the first column: $\Gamma_{i1} = 1$ for $i = 1, \ldots, R$.
   - diagonal elements $\Gamma_{ii} = i - 1$ for $i = 2, \ldots, R$
   - upper triangular elements: $\Gamma_{ij} = -1$ for $j > i$
   - remaining elements: set to zero.

2. Γ is then adjusted to account for the different class sample sizes. Let

$$n = (n_1, n_2, \ldots, n_R)^T$$

   be a column vector containing the class sample sizes, and define

$$n_{\sqrt{P}} = (\sqrt{n_1/n}, \sqrt{n_2/n}, \ldots, \sqrt{n_R/n})^T$$

   to be a column vector containing the square root of the class proportionalities. Also let

$$\Gamma_o = \Gamma \odot \left(n_{\sqrt{P}} \, 1_R^T \right)$$

   where 1_R is a $R \times 1$ column vector whose elements are all equal to 1. The symbol '$\odot$' is used to indicate a form of array multiplication across two matrices such that $\mathbf{B} = \mathbf{C} \odot \mathbf{G} \longrightarrow B_{ij} = C_{ij} G_{ij}$.

3. A QR decomposition is then performed on Γ_o so that $\Gamma_o = \mathbf{QR}$.

4. If $\mathbf{Q}_{.,-1}$ represents $\mathbf{Q}$ with the first column removed, then $\Psi = \mathbf{Q}_{.,-1} \odot n_{\sqrt{P}} 1_{R-1}^T$

Once Ψ has been formed, then $\tilde{\Psi}^*$ is calculated according to Equation 2.6.

Step 2 of the FDA algorithm involves performing a multivariate regression on the indicator matrix Ψ^*. This can be done by either the traditional multiple linear regression approach, or by using nonparametric regression methods. In either case regression is initially based on the original data matrix $\mathbf{X}$. A multivariate linear regression procedure predicts the columns of Ψ^* by

$$\widehat{\Psi^*} = \mathbf{X}^T\mathbf{B} = \mathbf{P}_X\Psi^*$$

where

$$\mathbf{B} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\Psi^*$$

and

$$\mathbf{P}_X = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}$$

is a linear projector matrix.

The nonparametric methods used in [60] produce a new predictor matrix $\mathbf{X}^*$ which is based on the original matrix $\mathbf{X}$. The matrices $\mathbf{X}$ and $\mathbf{X}^*$ both have the same number of observations n, but the dimensionality may differ for each of the matrices depending on how the nonparametric method forms the new predictor matrix. For instance, some nonparametric regression methods may actually be integrated with a feature extraction procedure to produce $\mathbf{X}^*$. A trivial example for $\mathbf{X}^*$ may be produced by augmenting the original predictor matrix with squares of the original variables. This would produce decision boundaries of a quadratic nature.

The nonparametric regression procedures used in [60] are MARS [45, 93] and BRUTO [59]. These methods use a much more creative approach for forming the new predictor matrix. MARS and BRUTO adaptively compute the predictor variables with the aim of minimizing some fitting criterion relevant to Ψ^*. MARS creates the new set of predictor variables by adaptively computing additive and interactive basis functions from regression splines. BRUTO [59] is an additive regression model which computes terms in the new predictor matrix by using smoothing splines. When FDA is applied in conjunction with the BRUTO algorithm it is possible to have a large number of predictors as the BRUTO procedure includes a variable selection method. Refer to [59, 60] for more details.

Once the new predictor matrix has been formed by the nonparametric regression methods, then the class indicator matrix is predicted, by replacing the linear operator $\mathbf{P}_X$ with

$\mathbf{P}_{X^*}$ such that

$$\widehat{\Psi^*} = \mathbf{P}_{X^*}\Psi^*.$$

For the MARS and BRUTO procedures Ψ^* can also be predicted by ,

$$\widehat{\Psi^*} = \mathrm{X}^{*T}\mathbf{B}$$

where $\mathbf{B}$ is now equal to

$$\mathbf{B} = (\mathrm{X}^*\mathrm{X}^{*T})^{-1}\mathrm{X}^*\Psi^*$$

There is little advantage in using the MLR procedure described previously, since, the end result would be equivalent to using FLDA on the original variables. The nonparametric regression procedures provide greater variability in selecting the predictor variables, which in turn allows for more nonlinearity in the decision boundaries. Or, as will be described in Section 2.5, allows a simple way to incorporate regularization into the discriminant procedure.

Step 3 of the FDA algorithm involves calculating the eigenvectors and eigenvalues of

$$\left(\Psi^{*T}\,\widehat{\Psi^*}/n\right).$$

The eigenvectors will be stored as columns in the matrix Θ and the eigenvalues will be stored in descending order in the matrix λ_{fda}. The eigen-analysis arises by formulating an optimal scoring procedure. The optimal scoring problem presented in [60] involves transforming the class indicator matrix Ψ^* in such a way that the transformed class labels (called optimal scores) are optimally predicted by linear regression on X^*. If nonparametric regression methods are not used, then prediction of the optimal scores will be based on the original predictor matrix $\mathbf{X}$. Let Θ^* denote the vector of transformed class labels which are formed by

$$\Theta^* = \Psi^*\Theta.$$

The optimal scoring problem as presented in [60] seeks the solution(s) to minimizing the average squared residual (ASR)

$$\mathrm{ASR} = \frac{1}{n}\sum_{r=1}^{R^*}\sum_{i=1}^{n}\left(\Theta^*_{ir} - \mathrm{x}_i^{*T}\mathbf{b}_{r_{os}}\right)^2 \tag{2.7}$$

subject to the constraint

$$\frac{1}{n}\,\Theta^T\Psi^{*T}\Psi^*\Theta = \mathbf{I}.$$

Here, $R^* \leq R-1$, $\mathbf{x}_i^*$ denotes the ith column or object vector in $\mathbf{X}^*$, and $\mathbf{b}_{r_{os}}$ is the rth column of the matrix of regression coefficients for the optimal scoring problem, $\mathbf{B}_{os}$. Equation 2.7 can be reformulated in terms of matrices by

$$\text{ASR} = \frac{1}{n}\,\text{trace}\left(\Theta^* - \widehat{\Theta^*}\right)^T\left(\Theta^* - \widehat{\Theta^*}\right) \tag{2.8}$$

where

$$\widehat{\Theta^*} = \mathbf{X}^{*T}\mathbf{B}_{os} \tag{2.9}$$

$$= \mathbf{P}_{X^*}\Theta^*. \tag{2.10}$$

Substituting Equation 2.9 and 2.10 into Equation 2.8 along with $\Theta^* = \Psi^*\Theta$, then the optimal scoring problem reduces to minimizing

$$\text{ASR} = \frac{1}{n}\,\text{trace}\left(\Theta^T\Psi^{*T}(\mathbf{I}-\mathbf{P}_X)\Psi^*\Theta\right). \tag{2.11}$$

When Equation 2.11 is minimized subject to $\Theta^T\Psi^{*T}\Psi^*\Theta/n = \mathbf{I}$ then one arrives at solving an eigen-equation of the form

$$\left(\frac{\Psi^{*T}\,\widehat{\Psi}^*}{n} - \Lambda_{\text{fda}}\mathbf{I}\right)\Theta = \mathbf{0}$$

where Λ_{fda} is a diagonal matrix with the ith diagonal element equal to the eigenvalue $\lambda_{i_{\text{fda}}}$. By calculating the eigenvectors of $\Psi^{*T}\widehat{\Psi^*}$ one can now compute the matrix Θ for converting the indicator matrix Ψ^* into the matrix of optimal scores Θ^*.

Step 4 computes a diagonal matrix D which has

$$D_{ii} = 1/\sqrt{\lambda_{i_{\text{fda}}}^2(1-\lambda_{i_{\text{fda}}}^2)}.$$

The diagonal matrix can then be used as part of the process for converting the regression analysis into a discriminant problem.

Step 5 of the FDA algorithm forms the discriminant variables. The key fact used in the FDA algorithm is that the columns of the matrix of regression coefficients $\mathbf{B}_{os}$ are individually proportional to the matrix of discriminant coefficients V. More specifically, they are related by

$$\mathbf{V} = \mathbf{B}_{os}\mathbf{D}.$$

The optimal scoring problem presented above does not calculate $\mathbf{B}_{os}$ directly, instead $\mathbf{B}_{os}$ is formed by converting the regression coefficients $\mathbf{B}$ used for predicting Ψ^*. This is done using

$$\mathbf{B}_{os} = \mathbf{B}\Theta$$

and follows from

$$\begin{aligned}
\mathbf{B}_{os} &= (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\Theta^* \\
&= (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\Psi^*\Theta \\
&= (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\Psi^*\Theta \\
&= \mathbf{B}\Theta
\end{aligned}$$

The discriminant variables can now be formulated by

$$\mathbf{X}^{*T}\mathbf{B}\Theta\mathbf{D}$$

or

$$\widehat{\Psi^*}\Theta\mathbf{D}.$$

Since the discriminant variables from FLDA and FDA are equivalent, then so to are the properties of the discriminant variables. That is, the discriminant variables will have an identity within covariance matrix.

The final part of the FDA algorithm is to use the model for classification. If $\eta(\mathbf{x}^*)$ denotes the vector of fitted values for $\mathbf{x}^*$ such that $\eta(\mathbf{x}^*) = \mathbf{x}^{*T}\mathbf{B}\Theta$, the coordinates of $\mathbf{x}^*$ in the discriminant coordinate system is given by

$$\eta(\mathbf{x}^*)\mathbf{D}.$$

An equivalent classification rule as that used in FLDA for assigning objects into various classes can be applied here. Assign $\mathbf{x}^*$ into the class $r \in 1, \ldots, R$ which minimizes

$$\mathcal{D}(\mathbf{x}^*, r) = \| (\eta(\mathbf{x}^*) - \bar{\eta}_r)\mathbf{D} \|^2$$

where $\bar{\eta}_r = \sum_{i=1}^{n_r} \eta(\mathbf{x}^*_{i(r)})/n_r$ is the fitted centroid of class r. Again if the regression method is based on the original predictor matrix, then $\mathbf{x}^*$ would be replaced with $\mathbf{x}$ in the above discussion.

The Splus code of Hastie *et. al.* [60] also allows for the calculation of posterior probabilities. If assignment is based on posterior probabilities, an object is assigned to the class r which has the largest posterior probability $P(r \mid \mathrm{x}^*)$ for $r \in 1, \ldots, R$. Using Bayes-optimal procedure, (see Section 2.6) the authors state that, for a Gaussian model, the posterior probability is proportional to

$$\begin{aligned} P(r \mid \mathrm{x}^*) &\propto P(r) \exp[-0.5(\mathrm{x}^* - \bar{\mathrm{x}}_r^*)^T \mathrm{S}_W^{-1} (\mathrm{x}^* - \bar{\mathrm{x}}_r^*)] \\ &\propto \exp[(-0.5\mathcal{D}(\mathrm{x}^*, r) - \log P(r))] \end{aligned} \tag{2.12}$$

where $\bar{\mathrm{x}}_r^* = \frac{1}{n_r} \sum_{i=1}^{n_r} \mathrm{x}_{i(r)}^*$ and $P(r)$ is the prior probability for class r. If the priors are equal, then classification based on posterior probabilities will be equivalent to classification based on distances $\mathcal{D}(\mathrm{x}^*, r)$.

Hastie *et. al.* [60] apply different variations of FDA against FLDA, BQDA, CART(see Section 4.1.3) and softmax on three sets of simulated data and one real data set. For these data BQDA and CART produced quite biased results and generally FLDA and FDA seemed to outperform the other techniques with the exception of one simulated data set for which BQDA did reasonably well.

2.5 Penalized Discriminant Analysis (PDA)

Penalized discriminant analysis (PDA) is a high dimensional classification technique which follows the same methodology as that presented for FDA. That is, optimal scoring provides the link between regression and classification. The main difference between FDA and PDA are the regression methods which are used in the optimal scoring procedure. Since PDA is designed with the aim of classifying highly dimensional and correlated data, then the regression methods employed by PDA should also be suited to extreme dimensionalities and somewhat resistant to multicollinearities. The regression methods which are used by PDA have some form of regularization.

PDA considers a penalized optimal scoring problem which seeks the solution(s) for minimizing

$$\mathrm{ASR} = \| \Theta^* - \mathrm{X}^T \mathrm{B}_{\mathrm{os}} \|^2 + \mathrm{B}_{\mathrm{os}}^T \Omega \mathrm{B}_{\mathrm{os}} \tag{2.13}$$

where $\mathrm{B}_{\mathrm{os}}^T \Omega \mathrm{B}_{\mathrm{os}}$ is the penalty term. The discriminant variables and classification is then as for the FDA algorithm.

The penalized optimal scoring problem is made equivalent to a penalized linear discriminant analysis which seeks the matrix of discriminant coefficients $\mathbf{V}$ such that

$$\mathbf{V}_{\text{pda}}^T \mathbf{S}_B \mathbf{V}_{\text{pda}}$$

is maximized subject to the constraint

$$\mathbf{V}_{\text{pda}}^T (\mathbf{S}_W + \Omega) \mathbf{V}_{\text{pda}} = \mathbf{I}.$$

Hastie *et. al.* [61] trial different versions of PDA on vowel and digit recognition data. Various regression methods including improper splines and generalized ridge regression were incorporated into the testing procedure. This allowed for different forms of the penalty matrix to be used. Their analysis highlights the vast improvement gained in applying PDA as opposed to FLDA which had the tendency to overfit.

2.6 Bayesian Classifiers

Unlike methods stemming from Fisher's linear discriminant analysis, Bayesian classifiers are not based on discriminant variables. Some Bayesian classifiers are based on the assumption that the class probability densities $p(\mathbf{x} \mid r)$ follow a multivariate normal distribution. That is,

$$P(\mathbf{x} \mid r) = (2\pi)^{-\frac{p}{2}} \mid \mathbf{S}_r \mid^{-0.5} \exp[-0.5(\mathbf{x} - \bar{\mathbf{x}}_r)^T \mathbf{S}_r^{-1} (\mathbf{x} - \bar{\mathbf{x}}_r)]. \tag{2.14}$$

Commonly, the class covariance matrices $\mathbf{S}_r$, and the class mean vectors $\mathbf{x}_r$, are calculated using the maximum likelihood estimates

$$\begin{aligned} \mathbf{S}_r &= \frac{1}{n_r} \sum_{i=1}^{n_r} (\mathbf{x}_{i(r)} - \bar{\mathbf{x}}_r)(\mathbf{x}_i - \bar{\mathbf{x}}_r)^T \\ \bar{\mathbf{x}}_r &= \frac{1}{n_r} \sum_{i=1}^{n_r} \mathbf{x}_{i(r)}. \end{aligned}$$

Bayesian classifiers are then based on Bayes decision rule which assigns an object $\mathbf{x}$ to the class r, which maximizes the posterior probability

$$P(r \mid \mathbf{x}) \quad \text{for } r = 1, \ldots, R. \tag{2.15}$$

By performing a direct application of Bayes theorem, the posterior probability in Equation 2.15 can be written as

$$P(r \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid r) P(r)}{p(\mathbf{x})}. \tag{2.16}$$

Here, $P(r)$ is the *a priori* probability of belonging to class r and $p(\mathbf{x})$ is the probability density of $\mathbf{x}$. The classification problem can be reformulated as—assign object $\mathbf{x}$ to the group r, which maximizes the classification score

$$g(\mathbf{x}, r) = p(\mathbf{x} \mid r)P(r) \quad \text{for } r = 1, \ldots, R. \tag{2.17}$$

Since $p(\mathbf{x})$ is independent of r, it is not considered in Equation 2.17.

2.6.1 Bayesian Linear Discriminant Analysis (BLDA)

For BLDA, the class covariance matrices $\mathbf{S}_r$, are assumed to be equal and are replaced with the pooled covariance matrix

$$\mathbf{S}_{\text{pooled}} = \frac{1}{n}\sum_{r=1}^{R} n_r \mathbf{S}_r = \mathbf{S}_W$$

in Equation 2.14. Taking the natural logarithm of Equation 2.17 and ignoring the constants, the following classification rule for BLDA results

$$g_{\text{blda}}(\mathbf{x}, r) = -0.5(\mathbf{x} - \bar{\mathbf{x}}_r)^T \mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_r) + \ln P(r). \tag{2.18}$$

Given homogeneous class covariance matrices and equal priors, the equivalence between FLDA using s_o discriminant variables and BLDA can be established since

$$\mathcal{D}(\mathbf{x}, r) = (\mathbf{x} - \bar{\mathbf{x}}_r)\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_r)^T.$$

For more details concerning this relationship the reader is referred to Johnson and Wichern [71](page 549). Fearn [35] also gives an excellent discussion about the relationship between Mahalanobis distance and FLDA for the two group case ($R = 2$).

If Equation 2.18 is expanded and the constants which are the same for each $g_{\text{blda}}(\mathbf{x}, 1), \ldots, g_{\text{blda}}(\mathbf{x}, R)$ are disregarded then one arrives at the linear function

$$g_{\text{blda}}(\mathbf{x}, r) \propto \mathbf{x}^T \mathbf{S}_{\text{pooled}}^{-1} \bar{\mathbf{x}}_r - 0.5\bar{\mathbf{x}}_r^T \mathbf{S}_{\text{pooled}}^{-1} \bar{\mathbf{x}}_r + \ln P(r).$$

One can then understand that the decision boundary which partitions objects from classes is also linear.

2.6.2 Bayesian Quadratic Discriminant Analysis (BQDA)

If the class covariances are not equal, then the class probability densities in Equation 2.14 remain unchanged. The quadratic discriminant rule, results when the class probability densities in Equation 2.17 are replaced by Equation 2.14. Taking the natural logarithm and ignoring constants, the quadratic classification rule can be written as

$$g_{\text{bqda}}(\mathbf{x}) = -0.5(\mathbf{x} - \bar{\mathbf{x}}_r)^T \mathbf{S}_r^{-1}(\mathbf{x} - \bar{\mathbf{x}}_r) - 0.5 \ln \mid \mathbf{S}_r \mid + \ln P(r). \tag{2.19}$$

If Equation 2.19 is expanded and the constants which are the same for each $g_{\text{bqda}}(\mathbf{x}, 1), \ldots, g_{\text{bqda}}(\mathbf{x}, R)$ are disregarded then one arrives at the quadratic function

$$g_{\text{bqda}}(\mathbf{x}, r) \propto -0.5\, \mathbf{x}^T \mathbf{S}_r^{-1} \mathbf{x} + \mathbf{x}^T \mathbf{S}_r^{-1} \bar{\mathbf{x}}_r - 0.5\, \bar{\mathbf{x}}_r^T \mathbf{S}_r^{-1} \bar{\mathbf{x}}_r - 0.5 \ln \mid S_r \mid + \ln P(r).$$

Now the decision boundaries which partition the objects from different classes are quadratic.

Both BLDA and BQDA are parametric discriminant methods, since the class probability densities $p(\mathbf{x} \mid r)$ were assumed to follow a particular distribution namely, a multivariate normal distribution. Some nonparametric discriminant methods have focused on "distribution free" estimates for the class probability densities. Kernel density and nearest neighbour methods [21, 102] are two examples of discriminant methods which relax the normality assumption about $p(\mathbf{x} \mid r)$.

2.7 Regularized Discriminant Analysis (RDA)

Regularized discriminant analysis [44], is a high dimensional classifier which introduces regularization into the covariance matrix. RDA differs from penalized discriminant analysis, in that, RDA stems from the Bayesian classifiers, while PDA is a Fisher-based approach. Both techniques involve performing some form of regularization to the covariance matrix, but PDA does this in a regression context.

Regularizing the covariance matrix can produce substantial improvements in classification [3], particularly in ill- and poorly-posed settings when the estimates of the class covariance matrix become highly variable [44]. At the expense of increasing bias in the parameter estimates, regularization of the covariance matrix reduces the variance of the estimates. A simple form of regularization occurs when reverting from BQDA to BLDA.

Replacing the class covariance matrix with the pooled covariance matrix means fewer parameters are required to be estimated. This form of regularization has the significant effect of reducing the variance of the estimates, thus producing enhanced classification results despite the differences in the class covariance matrices.

Friedman's regularized discriminant method initially replaces the class covariance matrix with a linear combination of the class and pooled covariance matrices

$$\mathrm{S}_r(a) = (1-a)\mathrm{S}_r + a\mathrm{S}_{\text{pooled}}. \tag{2.20}$$

The covariance matrix $\mathrm{S}_r(a)$ is further adjusted in order to under estimate the larger eigenvalues and over estimate the smaller eigenvalues. This is achieved by shrinking $\mathrm{S}_r(a)$ to a multiplier of the identity matrix $\mathbf{I}$,

$$\mathrm{S}_r(a,b) = (1-b)\mathrm{S}_r(a) + b\,\mathrm{trace}\left(\mathrm{S}_r(a)\right)\mathbf{I}/p, \tag{2.21}$$

here, the multiplier is the average eigenvalue of $\mathrm{S}_R(a)$.

The parameter $a \in [0,1]$, controls the degree to which the pooled covariance matrix should be used. The value of $b \in [0,1]$ determines the degree to which $\mathrm{S}_{(r)}(a)$ is shrunken toward a multiplier of the identity matrix. A grid of a and b values ranging between 0 and 1 are trialled. The pair of values which produce the minimal risk of misclassification are used. If more than one pair of values produce the same number of misclassified objects, then the chosen a and b parameters are determined by the largest b corresponding to the largest value of a [115]. Rayens and Greene [115] have developed a procedure based on an empirical Bayes formulation to estimate the degree to which the covariance matrices should be pooled.

Several articles have been written which compare the performance of RDA particularly against BLDA and BQDA. Friedman [44] presented simulations to help identify situations when RDA is likely to outperform its predecessors BLDA and BQDA. Frank and Friedman [38] presented applications of RDA compared with BLDA, BQDA, SIMCA and DASCO. Six simulated data sets were generated each of dimension 6 and 40, in addition four real data sets were tested. It was concluded there exists several striking advantages of RDA that make it an extremely useful and worthwhile technique to apply. Other articles which involve applications of RDA, BLDA and BQDA include [2, 97, 146].

2.8 Assessment of Model Performance

Before classifying new spectra whose true class identity in not known, it is important to assess how well the discriminant model actually works. There are two items which should be addressed when assessing the performance of a discriminant model. Firstly, consideration should be given to the assessment criterion. Once the assessment criterion has been selected, it is then necessary to choose a test set for evaluating the assessment criterion.

2.8.1 Assessment Criteria

The correct classification rate (CCR) or misclassification rate (MCR) are perhaps the most favoured assessment criteria in discriminant analysis. Their widespread popularity is obviously due to the ease in interpretation and implementation. Other assessment criteria are based on probability measures. Unlike correct classification rates which provide a discrete measure of assignment accuracy, probability based criteria provide a more continuous measure and reflect the degree of certainty which assignments have been made.

Correct Classification Rates (CCR)

In the descriptions to follow we speak of correct classification rates when misclassification rates (MCR=1-CCR) would equally suffice. A correct classification rate can be interpreted as the probability of assigning an object to the correct class. The correct classification rate is typically formulated as the ratio of correctly classified objects (from a testing set) with the total number of objects in the test set. More formally, let $\mathbf{y}$ denote the vector of true class labels and $\hat{\mathbf{y}}$ the vector of predicted class labels with $y_i, \hat{y}_i \in 1, ..., R$. The correct classification rate can then be expressed as follows

$$\text{CCR} = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i,\hat{y}_i} \ .$$

Here δ is an indicator variable such that $\delta_{y_i,\hat{y}_i} = 1$ if $y_i = \hat{y}_i$ and zero otherwise. For an interesting documentation involving error-rate estimation procedures to simulated data, the reader is referred to [65].

A correct classification rate is a discrete measure whose calculation is based upon which side of a decision boundary the observations lie. It does not reflect how "close" or "far

away" the observations lie from the decision boundary and hence how clear the assignments are made. An advantage of using probabilistic based classification methods such as those based on Bayes decision rule, is that it is possible to obtain more information than just the correct classification rate. Probabilistic measures provide information about the assignment accuracy, but they also reflect the degree of certainty which assignments have been made. We now consider other probabilistic measures which assess the trustworthiness or distinctness of the class predictions.

Probabilistic Measures

Most probabilistic discriminatory measures have the basic form

$$\mathrm{P} = \frac{1}{n}\sum_{i=1}^{n} a(\mathbf{x}_i) \tag{2.22}$$

where $a(\mathbf{x}_i)$ is an appreciation function which produces an appreciation score for $\mathbf{x}_i$. The correct classification rate of a Bayesian discriminant method such as BLDA, can for example be expressed in terms of a probabilistic measure. This would require the appreciation function having the very simple form

$$a_{\mathrm{ccr}}(\mathbf{x}_i) = 1 \qquad P(r \mid \mathbf{x}_{i(r)}) \geq P(r \mid \mathbf{x}_i) \tag{2.23}$$

$$= 0 \qquad \text{otherwise.} \tag{2.24}$$

$P(r \mid \mathbf{x}_{i(r)})$ denotes the posterior probability for the true class of $\mathbf{x}_i$. The correct classification rate is then written as

$$\mathrm{CCR} = \frac{1}{n}\sum_{i=1}^{n} a_{\mathrm{ccr}}(\mathbf{x}_i) = \mathrm{P}_{\mathrm{ccr}}.$$

Another simple probabilistic measure results when the appreciation score is

$$a_A(\mathbf{x}_i) = P(r \mid \mathbf{x}_{i(r)}).$$

The associated probabilistic measure is the average probability that an object is assigned to the correct class:

$$\mathrm{P}_A = \frac{1}{n}\sum_{i=1}^{n} a_A(\mathbf{x}_i).$$

The quadratic appreciation score which is used in Chapter 7 is formulated as follows,

$$a_Q(\mathbf{x}_i) = \frac{1}{2} + P\left(r \mid \mathbf{x}_{i(r)}\right) - \frac{1}{2}\sum_{r=1}^{R} P\left(r \mid \mathbf{x}_i\right)^2. \tag{2.25}$$

The quadratic probabilistic measure is then defined

$$\text{QPM} = \frac{1}{n}\sum_{i=1}^{n} a_Q(\mathbf{x}_i) = \text{P}_{\text{QPM}}.$$

The quadratic probability measure is related to the Brier quadratic score, which is a loss function for comparing two probability vectors, and is used for the elucidation of probabilities [13, 21].

Probabilistic measures based on appreciation functions other than the a_{ccr}, are less variable than correct classification rates, especially when there are relatively few observations. On the downside, most probabilistic measures and appreciation functions are generally more difficult to interpret than correct classification rates. This can be due to the fact that appreciation functions are less frequently encountered. As a general rule – a higher probability measure implies objects have been assigned to their respective groups with a greater degree of certainty. The assessment criterion applied in future chapters are based on the correct classification rate and the quadratic probability measures. These methods are subsequently referred to in the next section when procedures for choosing an evaluation set are discussed.

2.8.2 Choosing the Evaluation Set

Careful consideration should be given to choosing an evaluation set. Based on some assessment criteria, the evaluation set determines how well (or how poorly) a discriminant model actually performs. There are several procedures for selecting an evaluation set. This section will describe four approaches, namely the resubstitution, holdout, leave-one-out cross-validation and bootstrapping methods.

Resubstitution Method

The resubstitution method is quite simple. Here the evaluation (or testing) set, is exactly the same as the training set which designed the discriminant model. This approach is generally not preferred, since the results are often overly optimistic, giving somewhat of a 'false' insight to the true performance of the classifier. This is a consequence of the parameters in the discriminant model being estimated from the same data which are later used to assess the model. Several articles have been written which demonstrate this phenomenon, see for example [83, 101].

Holdout Method

The holdout method attempts to reduce the overly optimistic results obtained when the testing data are identical to the training data. With the holdout method, the sampled observations are divided into two separate sets of data – the training and testing data. Here, the training data designs the classifier, and the testing data is used for determining how well the classifier works. The classification performance using the holdout method is likely to be slightly pessimistic. It can be a useful exercise to calculate the assessment criterion on the training and testing data set, since this can provide some indication about the bounds of the assessment measure [46].

If the test set $\mathbf{X}'$ contains n' objects $\mathbf{x}'_i$ with n'_r objects from class r such that $n' = \sum_{r=1}^{R} n'_r$ and $\mathbf{y}'$ denotes the vector of true class labels of the testing data and $\hat{\mathbf{y}}'$ is the corresponding vector of predicted class labels with $y'_i, \hat{y}'_i \in 1, ..., R$, then the correct classification rate of the testing data can be expressed as follows

$$\text{CCR}' = \frac{1}{n} \sum_{i=1}^{n'} \delta_{y'_i, \hat{y}'_i}$$

where $\delta_{y'_i, \hat{y}'_i} = 1$ if $y'_i = \hat{y}'_i$ and zero otherwise. The QPM based on the testing data set is then

$$\text{QPM}' = \sum_{i=1}^{n'} a_Q(\mathbf{x}'_i).$$

where

$$a_Q(\mathbf{x}'_i) = \frac{1}{2} + P\left(r \mid \mathbf{x}'_{i(r)}\right) - \frac{1}{2} \sum_{r=1}^{R} P\left(r \mid \mathbf{x}'_i\right)^2. \tag{2.26}$$

The parameter estimates such as the covariance matrices and mean vectors associated with the calculation of the posterior probabilities are calculated using the training data.

An issue which arises from the holdout method concerns the sample sizes of the training and testing data. The interested reader is referred to [46, 47, 114, 113, 68] for more detailed information on such topics.

In some circumstances, it may not be viable to have a test set. For instance, there may only be a sufficient number of samples available to build the discriminant model and not to test its adequacy. Cross-validation is a mechanism which can be used for assessing the predictive ability of a classifier when the holdout method is not suitable.

Leave-out-one Cross-Validation Method

Cross-validation [83, 135] is the procedure for deleting objects from the training data, building the model in absence of the deleted objects, and then assessing the performance of the classifier based on the deleted objects. The deleted objects are then replaced, and another set of objects are deleted, (these objects can not have been deleted before). Again, the discriminant model is built in absence of these objects and the performance is assessed based on the deleted objects. Once all of the objects have been deleted a performance rate based on all of the deleted objects can be measured.

If the number of objects in each of the deleted groups is one, then the procedure is referred to as leave-one-out cross-validation. If $\hat{y}_{-i}$ is the predicted value of $\mathbf{x}_i$, obtained when $\mathbf{x}_i$ was deleted from the model building process then, define the leave-out-one cross-validated correct classification rate to be

$$\text{CVCCR} = \frac{1}{n}\sum_{i=1}^{n}\delta_{\hat{y}_{-i},y_i}$$

where $\delta_{\hat{y}_{-i},y_i} = 1$ if $\hat{y}_{-i} = y_i$ and zero otherwise. Similarly, define the leave-out-one cross-validated quadratic probability measure to be

$$\text{CVQPM} = \frac{1}{n}\sum_{i=1}^{n} a_Q(\mathbf{x}_i, -i) \tag{2.27}$$

where

$$a_Q(\mathbf{x}_i, -i) = \frac{1}{2} + P_{-i}\left(r \mid \mathbf{x}_{i(r)}\right) - \frac{1}{2}\sum_{r=1}^{R} P_{-i}\left(r \mid \mathbf{x}_i\right)^2$$

with $P_{-i}(r \mid \mathbf{x}_i)$ being the posterior probability for $\mathbf{x}_i$ when the covariance matrices and mean vectors in the probability density function have been calculated in the absence of $\mathbf{x}_i$.

When several objects are 'left out' the procedure is referred to as V-fold cross-validation [12 Here, V refers to the number of testing groups created. It is preferable to have the same number of objects in each testing group with an equal distribution of objects from different classes.

Leave-one-out cross-validation can be a time consuming operation. It is possible however to make use of fast updating formula (see for example [2, 44]) which can dramatically speed up the leave-one-out procedure.

Cross-validation is not limited to classification, nor are the resubstitution and holdout methods for that matter. These methods can also be used in other statistical applications such as regression analysis. When the least squares linear regression model is being applied, then an explicit formulae can be used for calculating a leave-one-out cross-validated measure of the predictive residual sum of squares. This formula does not require any updating formula or the actual deletion of observations (see Section 3.6).

Bootstrapping Method

The bootstrapping method [30, 31, 32] samples (with replacement) n objects from the original data set. These samples are referred to as the bootstrap samples. The bootstrap samples can be randomly generated from the original data, or can be generated by some artificial generating process. The bootstrap samples are then used to build the discriminant model. The discriminant model is assessed twice, using the bootstrap samples and the original samples. The difference CCR_{Δ} between these two estimates is averaged over several runs (eg 10-200) to produce an estimate of the bias in the resubstitution method. The *actual classification rate* which is the estimate obtained by classifying future unknown objects is then estimated by substracting CCR_{Δ} from the optimistic classification rate based on the original training set.

Recommendations

The way in which the evaluation set is chosen depends mostly on the number of available samples. If there are sufficient samples to warrant an independent test set then the holdout method is generally preferred. If there is not enough training data to have a training and separate test set, then some form of cross-validation or bootstrapping method should be implemented. Both these methods are computationally expensive, although for some parametric models, fast updating formula can be implemented to make the cross-validation procedures less burdensome.

Chapter 3

Regression Analysis

3.1 Introduction

The previous chapter discussed methods for predicting discrete response values based on a set of predictor variables. The response values were the class labels of the objects in the data set. This chapter is also interested in predicting response measurements based on a set of predictor variables, but now, the response values may take on a continuous range of measurements as opposed to discrete values.

An example of a regression application which is considered further in Chapter 8 is to predict the amount of fibre present in sugar cane samples. NIR spectra are obtained for several samples of sugar cane, and the reflectance measures for the NIR wavelengths, form the set of predictor variables. The predictor variables along with the fibre (response) measurements for the samples constitute the training data. Based on the training data, the regression model is designed with the aim of fitting and predicting the data responses adequately.

It is important to mention that it is possible for a model to fit the data well and give a good prediction of the training responses, but be very poor at predicting future (unseen) samples. It is therefore necessary to determine how well the model predicts by for instance implementing a cross-validation routine or by the use of an independent test set.

Regression methods when applied to spectral or other forms of highly dimensional data, are susceptible to similar problems which are encountered by discriminant methods, that is highly variable parameter estimates which can degrade the performance of the model.

Biased regression methods which are suited to low observation-to-variable ratios, reduce the variance of the parameter estimates in the regression model, at the expense of increasing bias. It is hoped that this bias / variance tradeoff will reduce the expected squared error of the parameter estimates and produce a more stable model for predicting future samples. Some biased regression methods which are commonly used on spectral data include principal component regression [37, 41], partial least squares [37, 41, 51, 52, 67, 145] and ridge regression [66, 119].

Using the same terminology adopted in earlier chapters, biased regression methods can be considered high dimensional methods since they can be applied to situations where the observation-variable ratio is quite small. Likewise, low dimensional regression methods are better suited to high observation-variable ratios.

The least squares multiple linear regression model [29, 106] is undoubtedly the most widely applied regression model and can be considered to be a low dimensional technique. In high dimensional settings, some form of feature extraction is highly recommended if this technique is to be employed.

This chapter proceeds by first introducing some notation and the regression methods – multiple linear regression, principal component regression and partial least squares regression. Following these discussions, some model selection criteria are introduced.

3.2 Notation

In this chapter we follow much of the same notation as that presented in Chapter 2. The response vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ which may contain continuous values is a $n \times 1$ column vector. The $n \times p$ predictor matrix also remains unchanged, but will be augmented with a $1 \times n$ row vector of ones $\mathbf{1}_n^T$ to allow for an intercept term in the regression procedure. The variables in $\mathbf{X}$ will be referred to as predictors or independent variables, and the response vector $\mathbf{y}$ may also be referred to as the dependent variable.

3.3 Multiple Linear Regression (MLR)

The general form of the multiple linear regression model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{ip} + \epsilon_i.$$

Here, y_i is the response measurement for the ith object $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})^T$, ϵ_i is the residual or prediction error for the ith observation, the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the regression coefficients and β_0 is also referred to as the (y-)intercept,

The multiple linear regression model can also be described in terms of matrices as follows

$$\mathbf{y} = \mathbf{X_1}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ and $\mathbf{X_1}$ is the matrix which augments 1_n^T with the matrix of predictor variables. In practice, the vector of regression coefficients $\boldsymbol{\beta}$, is usually unknown and is typically estimated by the least squares method. The least squares method calculates regression coefficients so that the residual sum of squares $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ is minimized. The least squares solution is

$$\mathbf{b} = (\mathbf{X_1}\mathbf{X_1}^T)^{-1}\mathbf{X_1}\mathbf{y}$$

where $\mathbf{b} = (\mathrm{b}_0, \mathrm{b}_1, \ldots, \mathrm{b}_p)^T$ is the estimate of the true regression coefficients $\boldsymbol{\beta}$. The estimated response is then

$$\hat{\mathbf{y}} = \mathbf{X_1}\mathbf{b}$$

and

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}}.$$

The MLR model assumes the residuals are independent and $\epsilon_i \sim N(0, \sigma^2)$. If the predictor and dependent variables are centered, so that they sum to zero, then it is not necessary to have a constant vector of ones in the predictor matrix, since, the intercept term will be zero.

3.4 Principal Component Regression

Principal component regression is simply MLR performed on the principal components, that is the predictor variables are now the principal components. Section 4.2.2 discusses principal component analysis in greater detail, but briefly, the principal components $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_p$ which are stored as columns in the $n \times p$ matrix $\mathbf{P}$, are obtained by multiplying the data matrix $\mathbf{X}$ with a set of eigenvectors such that

$$\mathbf{p}_i = \mathbf{X}^T \mathbf{a}_i \quad \text{for } i = 1, \ldots, p$$

where $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ are the eigenvectors of the total covariance matrix of $\mathbf{X}$, that is $(\mathbf{S}_T - \gamma\mathbf{I})\mathbf{a} = \mathbf{0}$.

The multiple linear regression model is then

$$\mathbf{y} = \mathbf{P}_1 \boldsymbol{\beta}_{\text{pcr}} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta}_{\text{pcr}}$ are the regression coefficients estimated by the method of least squares and $\mathbf{P}_1$ is a matrix with the first column equal to $\mathbf{1}_n$ and remaining columns equal to $\mathbf{P}$.

Of course there is still a total of p principal components, and in the case of spectral data $p \gg n$ hence a subset of the principal components should be selected. If the first p' principal components are used then, this is called a top down approach. There is no guarantee however that the first p' principal components will be best for regression analysis since the principal components are formed without any consideration of the dependent variable $\mathbf{y}$. Partial least squares regression however, uses components called latent variables which are constructed with consideration given to $\mathbf{y}$.

3.5 Partial Least Squares Regression

Partial least squares (PLS) can be used for modelling single or multiple response vectors. If one response is being modelled than the procedure is generally referred to as PLS1, whilst PLS2 is commonly used to explore the relationship between several response vectors. In this section we consider the partial least squares algorithm for predicting a single response vector.

Partial least squares (PLS) regression forms a new set of predictor variables called latent variables $t_1, t_2, \ldots, t_p$. These latent variables are stored as columns in the matrix

$\mathbf{T} = (t_1, t_2, \ldots, t_{p_*})$. The response is then modelled using

$$\mathbf{y} = \mathbf{T}\beta_{\text{pls}} + \epsilon$$

where the estimate of β_{pls} is calculated by

$$\mathbf{b}_{\text{pls}} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y}$$

The PLS predictor variables or latent variables are in fact linear combinations of the original variables, i.e., $\mathbf{T} = \mathbf{X}^T\mathbf{W}$ where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_{p_*})$. This implies a relationship of the form

$$\mathbf{y} = \mathbf{X}^T\mathbf{W}\beta_{\text{pls}} + \epsilon$$

and hence a PLS regression estimator $\tilde{\mathbf{b}}_{\text{pls}} = \mathbf{W}\mathbf{b}_{\text{pls}}$.

The latent variables are determined sequentially, such that each new predictor variable or latent variable has maximal covariance with the response, and is uncorrelated with previously constructed latent variables. Generally the number of latent variables p_*, is chosen such that $p_* \ll p$. There are several ways of choosing p_* in practice. In Chapter 8 p_* was initially set to 16. Then for $p_* = 1, \ldots, 16$ the predictive residual based on leave-one-out cross-validation was calculated. Finally the value of p_* was chosen as the one which minimized the predictive residual sum of squares.

The PLS algorithm which has been applied in Chapter 8 is that of Denham [26], which is a slight modification of the PLS algorithm due to Helland [62]. The algorithm implemented by Denham [26] is summarized in Figure 3.1.

The first step in the PLS algorithm simply centers the response vector by subtracting the mean $\bar{y} = \sum_{i=1}^{n} y_i/n$ from the uncentered response. Likewise, the second step in the PLS algorithm centers the predictor matrix by subtracting the mean object vector $\bar{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i/n$ from each observation. Step 3 calculates the sums of squares and cross product (SSCP) between $\mathbf{X}$ and $\mathbf{y}$, this is denoted by $\boldsymbol{w}_1$. Step 4 calculates the first latent vector and Step 5 uses this latent vector to calculate residual vector $\boldsymbol{r}$. The algorithm then enters a loop where the vectors $\boldsymbol{w}$ and $\boldsymbol{t}$ are calculated for each iteration of the loop using the updated residual vector $\boldsymbol{r}$. Unless specified otherwise, this procedure continues p_* times.

Partial least squares regression can be used when the observation-to-variable ratio is low and in situations when the predictor variables are highly correlated. This makes PLS

Partial Least Squares Algorithm

1. $\mathrm{y} \leftarrow \mathrm{y} - \bar{y}\mathrm{1_n}$
2. $\mathrm{X} \leftarrow \mathrm{X} - \bar{\mathrm{x}}\mathrm{1}_n^T$
3. $w_1 = \mathrm{X}\mathrm{y}$
4. $t_1 = \mathrm{X}w_1$
5. $r = \mathrm{y} - t_1(t_1^T t_1)^{-1} t_1^T \mathrm{y}$
6. FOR $i = 2, \ldots, p_*$
7. $\quad w_i = \mathrm{X}r$
8. $\quad t_i = \mathrm{X}^T w_i$
9. $\quad r = \mathrm{y} - \mathrm{T}_i(\mathrm{T}_i^T \mathrm{T}_i)^{-1}\mathrm{T}_i^T \mathrm{y}$
10. END
11. $\mathrm{b_{pls}} = (\mathrm{T}_{p_*}^T \mathrm{T}_{p_*})^{-1}\mathrm{T}_{p_*}^T \mathrm{y}$
12. $\tilde{\mathrm{b}}_{\mathrm{pls}} = \mathrm{W}_{p_*}\mathrm{b_{pls}}$

where $\mathrm{T}_i = (t_1, \ldots, t_i)$ and $\mathrm{W}_i = (w_1, \ldots, w_i)$

Figure 3.1: **Partial least squares algorithm.**

a popular regression technique to employ for spectral data and is indeed quite popular in the field of chemometrics. There exists many variations to PLS algorithms, see for example [25, 57, 62, 40, 98, 145].

3.6 Assessment of Model Performance

When comparing different regression models produced using the same data set, some criterion must be specified that gives a measurement defining how 'good' one model is relative to another. The word *good* is usually meant to reflect two properties of a model—how well it predicts, and how well it fits the data. Some of the most common assessment criteria are now discussed.

3.6.1 Assessment Criteria

RSS and R^2

The residual sum of squares (RSS) and the R-squared (R^2) criteria both measure how well the model fits the data. The residual sum of squares and R^2 are defined respectively to

be

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \hat{\epsilon}^T\hat{\epsilon} \tag{3.1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}}. \tag{3.2}$$

The RSS measures the sum of squared deviations between the actual and predicted values of the response. Typically, a lower measure of the RSS is preferred. The R^2 criterion measures the variation explained by the model. If response values were always predicted to be the sample mean of the data, then the residual sum of squares would be equal to the total sum of squares (TSS) and hence $R^2 = 0$. It is hoped that the RSS will be much less than the TSS so that an R^2 measure closer to one will be obtained. The rank order of the models in terms of the residual sum of squares and R-squared values will be the same for all candidate models.

It is important to note that a low measure of RSS and hence a high R^2 value could simply be a reflection of overfitting. The more terms in a regression model the lower the RSS and the higher the R^2 value. It is for this reason why the RSS and R^2 should not be used to compare models of different complexities.

MSE, R^2_{adj}, C_p, AIC

Unlike the RSS and R^2, the mean square error (MSE), adjusted R^2 (R^2_{adj}), Mallow's C_p and Akaike's Information Criterion (AIC) can be used to compare the performance of models with different complexities. By incorporating a term into these formula which accounts for the varying degrees of freedom, these criterion will deteriorate if marginally important variables are included in the model. The symbol p_o will be used to denote the number of parameters estimated (including the intercept) in the model. The degrees of freedom (DF) is then equal to $n - p_o$.

The mean square error is simply the ratio of the residual sum of squares to the degrees of freedom and is written as

$$\text{MSE} = \frac{\text{RSS}}{\text{DF}} = \frac{\text{RSS}}{n - p_o}.$$

The adjusted R-squared is typically formulated (see for example [106]) as follows

$$R^2_{adj} = 1 - \frac{\text{RSS}/(n - p_o)}{\text{TSS}/(n-1)}.$$

The rank order of the models in terms of the MSE and adjusted R-squared values will be the same for all candidate models as was the case between the RSS and R-squared criteria. This is more clearly seen by representing the adjusted R-squared in terms of the mean square error as follows

$$R^2_{adj} = 1 - \frac{\text{MSE}(n-1)}{\text{TSS}}.$$

Mallows C_p is derived by taking into consideration both the biases and variance in a regression model. Biases result from regression models which suffer from a lack of fit, and high variability is a consequence of overfitting. A derivation for Mallows C_p based on these principles is given in [106] where it is shown that

$$C_p = \frac{\text{RSS}_{p_o}}{\text{MSE}} + 2p_o - n$$

where RSS_{p_o} denotes the RSS of the model with complexity p_o, and the MSE is based on the largest postulated model. If values of C_p are plotted against p, candidate models will lie close to the line $C_p = p$.

The AIC score is the deviance (DEV) of the model plus twice the degrees of freedom times a dispersion parameter (φ), that is

$$\text{AIC} = \text{DEV} + 2\,\text{DF}\,\varphi$$

where, the deviance is twice the difference between the log-likelihood of the full model $\ell(b_{max}; y)$ and the log-likelihood of the actual model $\ell(b; y)$

$$DEV = 2[\ell(b_{max}; y) - \ell(b; y)]$$

A model with a low AIC score is preferred to one that has a high AIC score.

3.6.2 Choosing the Evaluation Set

The same methods described in Section 2.8.2 for choosing an evaluation set for discriminant analysis can also be used in regression analysis. The methods previously discussed were the resubstitution, holdout, cross-validation and bootstrapping methods.

If one is interested in assessing how well the model fits the data, then the resubstitution method could be applied. That is, assessment is simply made on the original data which

built the regression model. If however, we are interested in determining how well the model will predict response values for a new set of observations, then it is necessary to base the assessment criteria on an independent test set which differs from the data which designed the regression model.

The holdout method allows for an independent testing set. As was the case with discriminant analysis, there are no strict rules for formulating an independent test set, and much controversy can surround this topic. The interested reader is referred to Myers [106] and Snee [120] for more details.

In this thesis the performance of the regression methods will be compared on the basis of an independent test set. It has been decided to formulate an R^2 measure for the test set which is denoted by R^2_{test}

$$R^2_{\text{test}} = 1 - \frac{\text{RSS}_{\text{test}}}{\text{TSS}_{\text{test}}}.$$

The residual and total sum of squares for the testing data are defined respectively to be

$$\text{RSS}_{\text{test}} = \sum_{i=1}^{n'} (y'_i - \hat{y}')^2 \tag{3.3}$$

$$\text{TSS}_{\text{test}} = \sum_{i=1}^{n'} (y'_i - \bar{y}')^2 \tag{3.4}$$

$$\tag{3.5}$$

where $\mathbf{y}'$ is the response values of the independent test set, $\hat{\mathbf{y}}'$ are the predicted test response values, and n' is the number of objects in the test data set.

A criterion which is not a function of degrees of freedom has been chosen because for some of the biased regression methods such as partial least squares it is not clear how the degrees of freedom would be formulated. Of course if an independent test set is unavailable then measures of predictive performance may be obtained by using cross-validation or bootstrapping method as previously described in Section 2.8.2.

Recall that cross-validation is the procedure which involves deleting a group of observations from the training data, building the regression model in the absence of these 'pseudo' testing objects and then calculating the prediction performance based on the deleted objects. This procedure is repeated until each object from the training set has been removed once. If a single object is deleted at each iteration of the cross-validation procedure, then this is referred to as leave-one-out cross-validation. In the case of linear estimators such

as multiple linear regression, it is possible to calculate a predictive measure of the residual sum of squares (PRESS) without actually having to build a new model each time an observation is deleted. Instead it is necessary to construct a single model only.

Define the PRESS statistic to be

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{-i})^2.$$

Here, $\hat{y}_{-i}$ is the predicted value for y_i, but object $\mathbf{x}_i$ was 'left out' when estimating the parameters in the regression model. Another way of calculating the PRESS statistic is simply by using

$$y_i - \hat{y}_{-i} = \frac{y_i - \hat{y}_i}{1 - \hbar_{ii}} \tag{3.6}$$

where, $\hbar_{ii}$ is the element along the ith diagonal of the hat matrix $\mathcal{H} = \mathbf{X}^{\mathbf{T}}(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1}\mathbf{X}$. This avoids the need to leave out observations in turn. A leave-out-one cross-validated R-squared score could then be formulated as

$$\text{CVRSQ} = 1 - \text{PRESS}/\text{TSS} \tag{3.7}$$

The formulation of Equation 3.6, makes the leave-out-one method of cross-validation quite a useful and relatively inexpensive procedure to employ.

Chapter 4

Feature Extraction

It has been discussed in earlier chapters that some form of feature extraction should be implemented prior to performing multivariate analyses using low dimensional statistical methods. Besides improving the performance of the statistical analyses, having fewer variables often means results can be obtained with reduced computational and economical expense. Another reason for extracting features may simply be that the features are more meaningful than the raw data and thereby enhance the interpretability of the data.

Feature extraction is a dimension reducing procedure which selects a subset of variables p_* from a much larger set of p variables. The variables can be selected from the original data, or, from data which has been preprocessed or transformed in some other way. Typically, the dimensionality p_* of the subset of feature variables is less than p and usually very much less than the number of observations, or spectra n. The feature extraction procedure aims to retain as much of the information as possible, whilst simultaneously eliminating redundant features which do not contribute or have an adverse effect on the statistical procedure.

Feature extraction can consist of three modules – a preprocessing module, a feature transformation module and a feature selection module. We first introduce some feature selection strategies and then consider some feature transformation methods. It was decided to make preprocessing methods a part of the section on feature transformation, since preprocessing methods often involve some form of transformation. There are many ways to perform feature extraction. In this chapter we discuss some modern and customary approaches used for feature extraction.

4.1 Feature Selection

As previously mentioned, feature selection involves selecting a smaller set of variables from a bigger set on the basis of some criteria. The criteria used for feature selection should reflect the goal of the statistical analysis. For example, if the goal is to assign an object to a particular class as accurately as possible, then the criterion for variable selection could involve a misclassification rate. Or, if the goal is to select variables which are used to predict some response such as the concentration levels of chemical substances, then feature selection could be based upon the predictive residual sum of squares.

When feature selection is based on the original data which may or may not have been preprocessed, then the procedure is sometimes called variable selection as opposed to feature selection. In this thesis we allow feature selection to be a selection of either the transformed data or, the original data. When feature selection is from the original data, the feature transformation is via the identity matrix.

One item which needs to be addressed is the number of features to be selected. Of course this will often depend on the kind of statistical method which is implemented. For example it is possible to have more variables for Bayesian linear discriminant analysis than Bayesian quadratic discriminant analysis. Some examples of references which address the topic in a classification perspective include [36, 47, 114, 113]. While references which address this problem and related issues with respect to regression analysis include [106, 120].

We now consider feature selection strategies separately for discriminant analysis and regression problems.

4.1.1 Feature Selection Strategies for Discriminant Analysis

In this section we present a brief overview of some of the many feature selection strategies that can be applied in discriminant analysis. There are two main goals of discriminant analysis. The first is to accurately predict group memberships of unclassified objects and the second is to observe and understand the spatial separation of objects whose group identity has been established. These two goals are quite different, and it is recommended by McLachlan [102] that the criteria should reflect the aim of the discriminant procedure. As mentioned above, if the goal is assignment, then the variable should be selected on

the basis of an error rate. If however, one is purely interested in observing some spatial separation, then the variable selection criterion should be based on a measure of separation. Of course it is still possible, in some cases, to obtain favourable classification rates from variables selected by measures of separation. Some separation criteria include the ratio of between-to-within-variability, Mahalanobis distance [7, 99], Euclidean distance and Wilk's Lambda [102]. Assignment or allocation criteria usually involve classification rates or posterior probabilities.

Wu *et. al* [148] employ some of the above criteria when implementing univariate feature selection techniques. One feature selection method involves selecting variables which produce high values of the ratio of the between-to-within variability. That is, for each variable, the between-variability divided by the within-variability is calculated and the variables which produce large values of this quotient are used as input to the classifier. This measure of between-to-within-variability is sometimes referred to as the Fishers criterion. A second scheme chooses variables that are most correlated with the (ordinal) response which, in this case, is the vector of group labels. Another feature selection method utilized by the same authors involves searching for non-overlapping regions of the spectra from different classes. For each variable the range of response values for each of the groups are individually determined. If the response ranges for some variable, for each group are non-overlapping (distinct), then that particular variable is a likely candidate to be selected. Ideally, the variables for which the response ranges (for each of the groups) are most separated will be extracted and used for classification.

Consideration should also be given to selecting combinations of variables. For example two features chosen separately may produce less favourable results than two features chosen in combination. With high dimensional data it is not realistic to perform an exhaustive search of every possible combination of variables. In this case, stepwise methods can be considered. A forward stepwise search sequentially incorporates variables into the discriminant procedure that contribute to the discrimination power of the model. Generally, variables cease to enter the model when little or no change to the performance of the discriminant model is registered. Conversely, a backward stepwise search removes variables from the model until the performance of the model begins to deteriorate. In the presence of high dimensional data, it is not advisable to use a backward stepwise

search, since initially all the variables will be fed to the classifier which will lead to ill- and poorly-posed situations, not to mention the computational expense and the numerical instabilities which are likely to arise. Instead, a backward selection scheme could be used following the implementation of a forward stepwise selection method.

One possible forward selection scheme is as follows; first select the variable which gives the largest value of Wilk's Lambda Λ. Calculate Λ for every remaining variable (in combination with the first selected variable) and enter the variable which gives the largest value of Λ. If $\Lambda^{(i)}$ equals Λ at the ith iteration, then the routine continues until $\Lambda^{(i+1)} - \Lambda^{(i)} \leq \Delta$ where Δ is prespecified. If the purpose would be to find a set of variables which classifies accurately the same procedure could be employed, but with a cross-validated correct classification rate replacing Λ. Since stepwise techniques involve repetitive calculations, it is worthwhile to make use of fast updating formulae [2]. This avoids the need to completely recalculate parameters such as the covariance matrix and dramatically reduces the computational burden.

The SAS [122] and SPSS [109, 108] statistical computing packages have an option for performing stepwise discriminant analysis. SAS allows for a forward, backward, and forward/backward combination which enters/removes variables according to the Wilk's Lambda criterion [132]. SPSS, also has a stepwise procedure which combines the features of a forward and backwards selection procedure. There are several selection criterion made available for entering and removing variables from the model. These criterion are based on Wilk's Lambda, Rao's V, Mahalanobis distance, between groups t-statistic and the sum of unexplained variance. The reader is referred to [108] for more details about these criterion.

Another multivariate variable selection method is the branch-and-bound technique. This technique becomes very inefficient when there are 30 or more variables [48, 58].

4.1.2 Feature Selection Strategies for Regression Analysis

One of the most standard feature selection procedures applied in regression analysis are stepwise approaches. Stepwise procedures for regression analysis can be forward, backward, or a combination of forward and backward procedures. As is the case with stepwise procedures for discriminant analysis, backward procedures are generally not recommended

as the initial procedure to apply for highly dimensional data, since many instabilities are likely to arise.

A simple criterion which is used for the entry of variables into a regression equation is the residual sum of squares. A stepwise procedure which is applied in Chapter 8 is similar to that outlined in Draper [29]. For each variable the R^2_{train} of the model was calculated. The variable which gave the largest increase in R^2_{train} entered the model. At each iteration all the variables in the current model were tested for removal. Variables were removed if their t-statistic for testing if the regression coefficient is significantly different from zero, was less than some prespecified amount. The procedure stopped when no more variables were retained in the model, or, until there were p_* variables in the model which ever came sooner. Here, p_* was also prespecified.

In earlier discussions about model selection criteria it was noted that as more terms are added into a model the residual sum of squares (or R-squared) value will decrease. It may seem contradictory to have stepwise methods continually incorporating variables into a model based on this criterion. In practice, one typically plots the residual sum of squares against the iteration of the stepwise procedure. Small changes in the residual sum of squares from one iteration to the next usually imply that the drop in residual sum of squares is a consequence of adding variables into the model, and not from the variables providing more useful information. The t-statistic which is calculated at each iteration for testing if the coefficients are zero at each iteration can also help form a safeguard against this problem. This t-statistic is also quite useful since as more terms are incorporated into the model, variables which entered the equation in the early stages of the stepwise procedure can be found to be less useful, as other terms enter the model.

Stepwise techniques usually come standard in many statistical packages for regression, these include SAS [122], SPSS [109] and S-Plus [126] for example. Splus (version 3.2) allows for these three stepwise procedures to be performed. The forward stepwise procedure begins to include variables into the model which give the largest decrease in residual sum of squares, while the backward procedure begins to delete variables (initially from the full model) which give the smallest decrease in the residual sum of squares. Efroymson's stepwise method combines a forward entry and backwards deletion scheme (where necessary). This is similar to a forward procedure, except each time a new variable is incorporated

into the model, partial correlations [29, 106] between the independent variables and the response are considered to determine if any variables should be removed from the model. Splus also allows for an exhaustive procedure, where the smallest residual sum of squares of all possible variable subsets of a specified size are calculated. The final model has the smallest residual sum of squares.

SPSS (version 6.1) allows for forward and backward stepwise procedures as well as a combination stepwise method which tests if any variables should be removed from the model at each iteration of a forward procedure. The criteria which SPSS allows the user to specify for variables to enter or exit a model are a t-to-enter/exit or the probability of t-to-enter/exit. Note that criteria based on partial correlations, t-statistics or f-statistics can be considered equivalent.

SAS (version 6.03) also allows for forward, backwards and the combination method as well as forms of best subset searches of a specified size. The forward/backwards and combination stepwise procedures are based on the f-statistic criterion. The best subset method searches for the subset of variables which give the most suitable R-squared, adjusted R-squared, or Mallows' C_p, which ever is specified by the user.

For the above packages, we have mentioned that forms of exhaustive searches are available. For spectral data however, such searches can be very computationally expensive and inappropriate, as was the case with exhaustive methods for discriminant analysis.

The branch-and-bound technique [48, 58] could also be used for regression, but is not recommended as a feature selection technique for spectral data. Perhaps more appropriate are genetic algorithms [89, 134]. Genetic algorithms have previously been used as a feature selection method for regression see for example [74, 84, 85]. In [85], the genetic algorithms are also used for outlier detection. Other variable selection strategies which are performed in the presence of outliers include that discussed in Sommer *et. al.* [121].

Another simple feature selection strategy is to select the variables which are most correlated with the response as described in [75]. The wavelengths which are most correlated could then be used in a multiple linear regression model for example.

4.1.3 Classification and Regression Trees (CART)

CART is a nonparametric method which can be used for discriminant or regression analysis. CART [10] is based on a recursive partitioning scheme and is an extension of the work performed by [6, 9, 54, 64, 90, 103, 43]. For a thorough account about the CART algorithm, the reader is referred to Breiman et. al. [10]. Other useful references include [17, 117, 149].

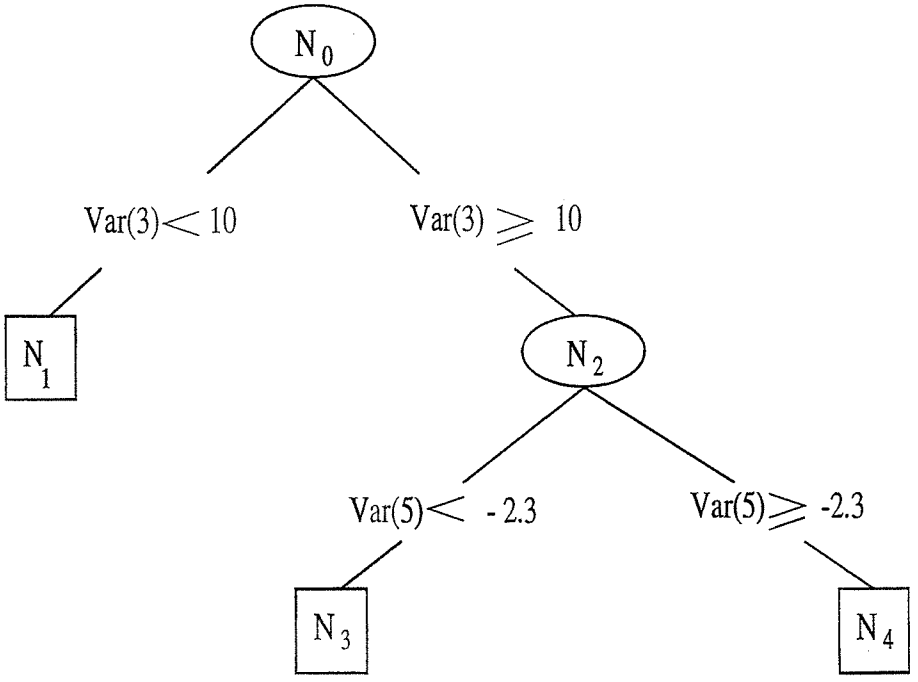


Figure 4.1: **A CART model.**

CART recursively splits the datasets into homogeneous subsets. Figure 4.1 shows a simple CART model which has a binary tree structure and contains 5 nodes. At the top of the tree there is a single node which is called a root node, and is labelled by N_0. At the next level, the nodes are referred to from left to right as N_1 and N_2. Similarly, at the next level the nodes are N_3 and N_4. Nodes which do not have any descendants are referred to as terminal nodes, thus N_1, N_3 and N_4 are terminal nodes and are indicated by the square boxes.

Initially, all the observations are stored in the root node. In Figure 4.1 the objects in N_0 are split depending on their measurements for var(3), where var(3) denotes the third variable in the data set. All objects for which var(3)<10 move into N_1, and all objects for which var(3)≥10 move into N_2. Node 1 is not split any further. Node 2 splits the

objects based on variable 5. The objects in N_2 that have var(5)<-2.3 move into N_3 and the objects for which var(5)≥ -2.3 move into N_4. The objects in N_3 have var(3)$\geq$10 and var(5)<-2.3. The objects in N_4 have var(3)$\geq$10 and var(5)≥ -2.3.

If Figure 4.1 was a classification tree, then each terminal node would be assigned a group label. For example N_1=group 2, N_3=group 1 and N_4=group 2. (It is possible for terminal nodes to have the same class label). If Figure 4.1 was a regression tree, then each terminal node would be assigned the mean response value for the objects in that node. For example if $\mathbf{x}_{i[l]}$ denotes the ith object in node N_l, and $n_{[l]}$ denotes the number of objects in node N_l, then the terminal nodes in Figure 4.1 would have the values $N_1 = \sum_{i=1}^{n_{[1]}} y_{i[1]}/n_{[1]}$, $N_3 = \sum_{i=1}^{n_{[3]}} y_{i[3]}/n_{[3]}$, and $N_4 = \sum_{i=1}^{n_{[4]}} y_{i[4]}/n_{[4]}$, where $y_{i[l]}$ is the response value of $\mathbf{x}_{i[l]}$. A response measure for a new object can be predicted by observing which terminal node it would lie.

The CART algorithm splits the data with the aim of obtaining terminal nodes which contain objects whose properties are similar. For classification, the goal is to split the data so that terminal nodes contain objects from the same class. For regression, the goal is to have the objects with similar response measurements in each of the terminal nodes. The CART algorithm searches each potential split point in the data set and chooses the split point which minimizes some impurity measure of the node.

The impurity measure used for classification is entropy (see also Section 6.3.1). Here, the impurity measure will be highest if the node has equal portions of objects from each class, and will be minimal when the node contains objects from a single class only. The entropy impurity measure is defined as

$$-\sum_{r=1}^{R} P(r|l) \log P(r|l)$$

where $P(r|l)$ is the proportion of objects in node N_l which are from class r. The impurity measure used for regression is simply the residual sum of squares.

The variables selected by CART may be good features to use with other multivariate techniques. Alternatively, CART could be used as a stand alone method for classification or regression.

The CART algorithm as implemented in Splus ceases to split a node if there are too few observations in the node, or when the impurity measure reaches a certain threshold. In

the S-Plus statistical package, the minimum number of observations and threshold criteria can be specified by the user. The reader is referred to [10, 17, 138] for more details.

4.2 Feature Transformation

Examples of spectral features which maybe quite informative include the heights, positions or shapes of peaks for instance. Other features commonly used include principal components [33, 53] and Fourier coefficients [150, 147], while more recently, wavelet coefficients [8, 86, 118, 131] have been explored. The procedure for calculating a new set of features is called feature transformation.

In this section we discuss some common and modern feature transformation methods. First, we describe some preprocessing methods which can be applied to spectral data. It can be considered slightly unusual to discuss preprocessing methods as part of feature transformations, or indeed feature extraction. Since preprocessing methods usually involve some transformation procedure, it has been decided to describe preprocessing methods as part of the feature transformation section.

4.2.1 Preprocessing Methods and Transformations

For some spectral data sets, it may be appropriate to preprocess the data before performing statistical analyses. This can be done for several reasons. The obvious reason is that one may be able to obtain improved results from spectral data which has been preprocessed. Another reason for transforming the data is so that the spectra can be appropriately aligned. For example, if spectra representing samples of gasoline have been collected on different days and it is obvious that some effect is present purely because the spectra may have been obtained on different days, then a transformation could be used to counteract this misalignment. Thus, some transformations can be used to 'align' the spectra, so that a fair assessment can be made. This section discusses some preprocessing transformations which are commonly applied to spectral data.

Standard Normal Variate Transformation

The SNV transformation [5, 148] produces spectra of similar shape and slope as the original (untransformed) spectra. If $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})^T$ denotes the ith spectrum, then the

mean value of $\mathbf{x}_i$ is calculated by

$$\bar{x}_i = \frac{1}{p}\sum_{l=1}^{p} x_{li}.$$

The sample standard deviation of the ith spectrum is then

$$\hat{\sigma}_{\mathbf{x}_i} = \sqrt{\frac{\sum_{l=1}^{p}(x_{li} - \bar{x}_i)^2}{p-1}}.$$

The SNV transformation of $\mathbf{x}_i$ is then defined as

$$\frac{\mathbf{x}_i - \bar{x}_i \mathbf{1}_p}{\hat{\sigma}_{\mathbf{x}_i}}$$

where $\mathbf{1}_p$ is a $p \times 1$ column vector of ones.

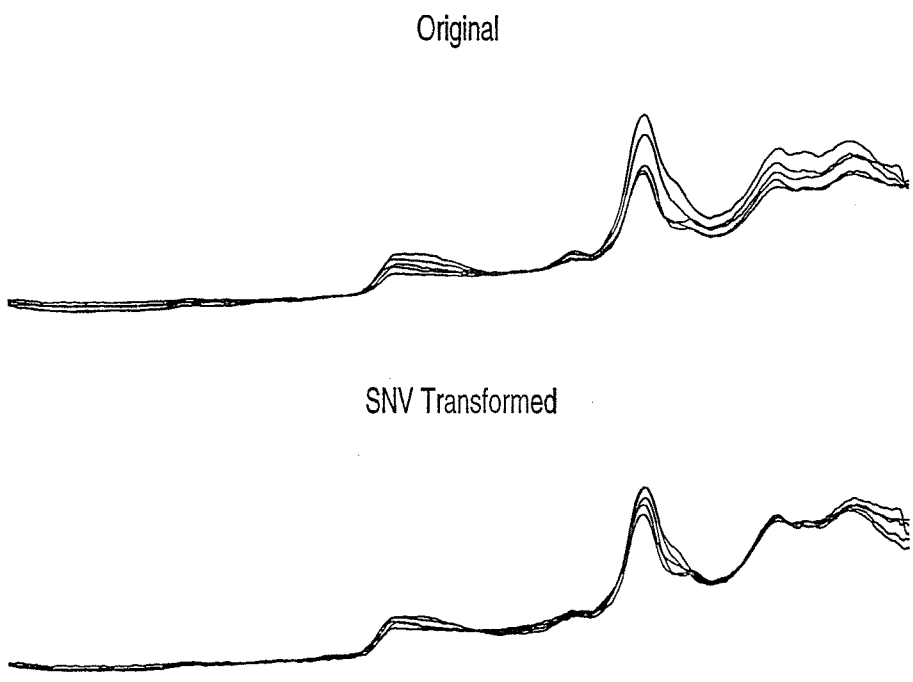


Figure 4.2: **Demonstration of the SNV transformation.**

Figure 4.2 shows the effect of performing the SNV transformation on five sample spectra. It is seen that the SNV transformed spectra have a similar shape and slope as that for the original spectra, but the variability between spectra is reduced.

Detrending

A single spectrum $\mathbf{x}_i$ may be detrended using a second degree polynomial as follows. Let $\boldsymbol{\nu} = (\nu_0, \nu_1, \ldots, \nu_{p-1})^T$ be a column vector of wavelengths which is to be regressed against

$\mathbf{x}_i$ using a second degree polynomial. That is

$$\mathbf{x}_i = \mathrm{b}_0 + \mathrm{b}_1\nu + \mathrm{b}_2\nu^2 + \hat{\epsilon}_i.$$

Here, $\hat{\epsilon}_i$ is the residual vector corresponding to the ith data vector, and $\mathbf{b} = (\mathrm{b}_0, \mathrm{b}_1, \mathrm{b}_2)^T$ are the regression coefficients which are found by

$$\mathbf{b} = \left([1_p \;\; \nu \;\; \nu^2]^T [1_p \;\; \nu \;\; \nu^2]\right)^{-1} [1_p \;\; \nu \;\; \nu^2]^T \mathbf{x}_i.$$

The matrix $[1_p \;\; \nu \;\; \nu^2]$ contains a $p \times 1$ vector of ones in the first column, ν in the second column and ν^2 in the third column. The detrended spectrum is obtained by

$$\mathbf{x}_i - \hat{\mathbf{x}}_i$$

where

$$\hat{\mathbf{x}}_i = \mathrm{b}_0 + \mathrm{b}_1\nu + \mathrm{b}_2\nu^2 \, .$$

Figure 4.3 shows the effect of detrending and performing the SNV transformation on five spectra. Detrending has the tendency to remove the baseline trend and curvilinearity, but there is still some variability within the spectra. By taking the SNV transformation on the detrended spectra, then this variability can be reduced.

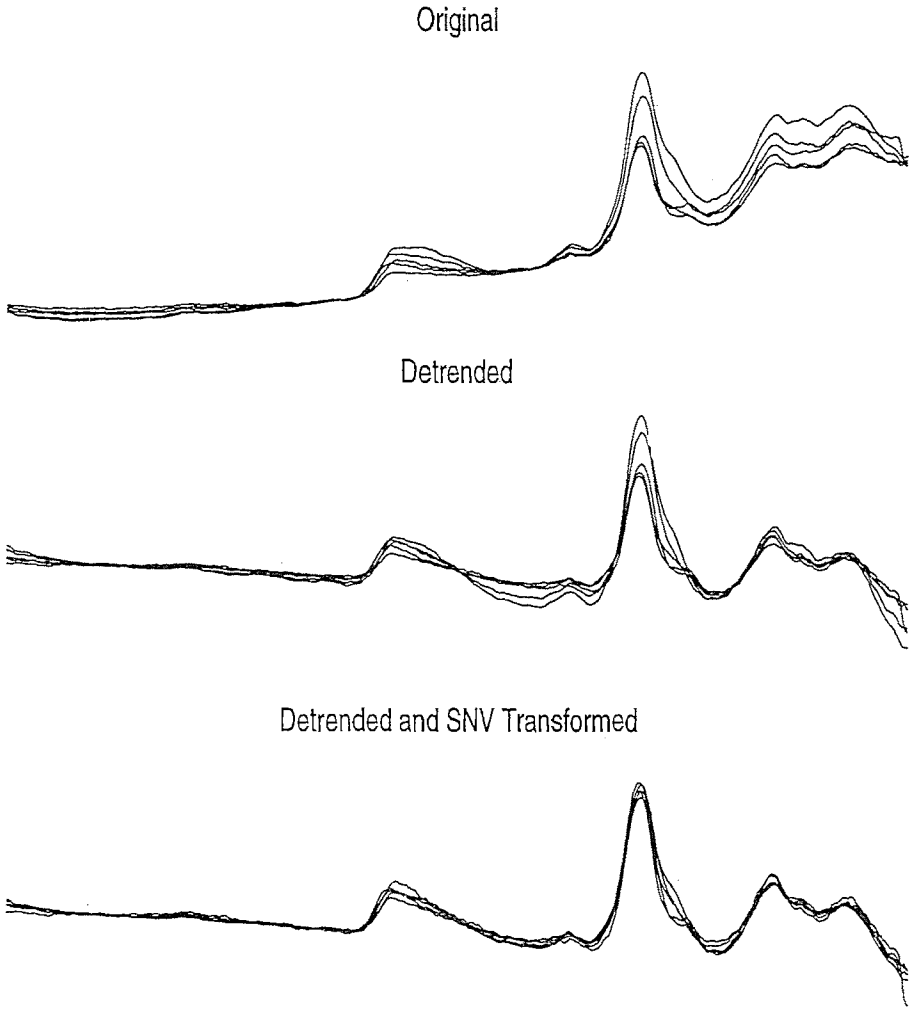


Figure 4.3: **Demonstration of detrending combined with the SNV transformation.**

Hull Quotient

Another way of removing the baseline effects is obtained by taking the hull quotient. This is commonly used in geology and is a standard transformation option in the spectral package PIMAVIEW [111]. The hull quotient does not have an implicit mathematical formula but it is obtained by finding the ratio of the spectrum to the lowest convex curve lying above the given spectrum. Figure 4.4 provides a pictorial interpretation as to how the hull quotient of a spectrum is obtained. In the subplot at the top of Figure 4.4 there are two lines drawn. The thick line is the lowest convex curve lying above the spectrum (also called the hull), the spectrum is represented by the thinner line. By taking the ratio

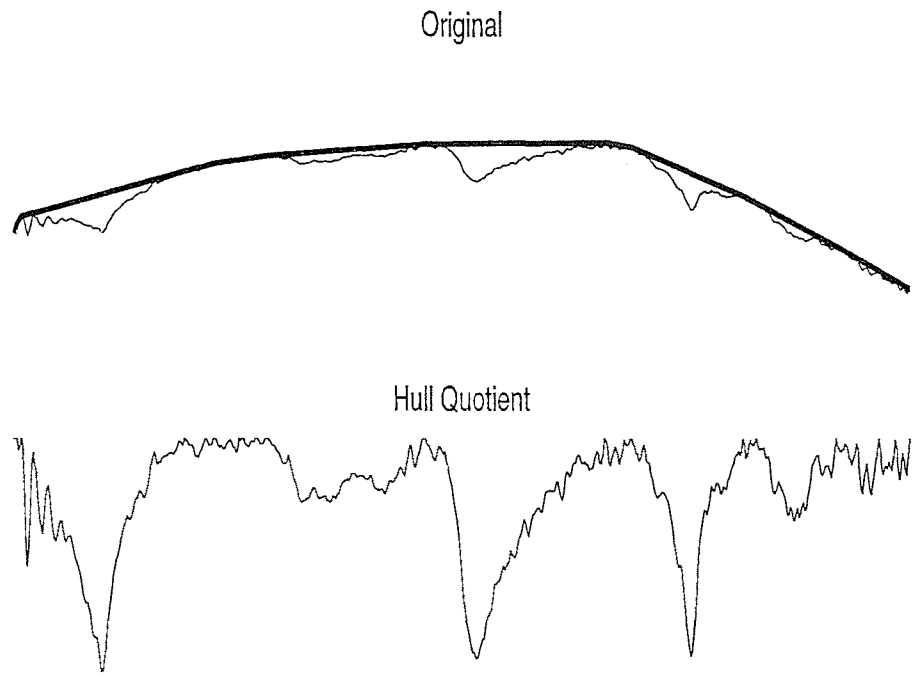


Figure 4.4: **Demonstration of the hull quotient.**

of these two lines at each point, the hull quotient spectrum is obtained. This is shown in the subplot at the bottom of Figure 4.4.

Second Derivative

The second derivative transformation (2D) removes the slope and parallel shifts in the spectra [5]. There are several complications which arise from the 2D transformation. One is that the 2D transformed spectra look quite different to the original spectra. This makes interpretation quite difficult.

Different packages may calculate the second derivative spectra in different ways. Typically, the 2D transformation consists of a differentiation operation and a fitting procedure. The fitting procedure involves fitting a model to the data, and the differentiation procedure then differentiates the fitted model at some point. Some packages perform an additional smoothing procedure to the differentiated data.

We provide an example of a very simple procedure for calculating the second derivative transformation of a spectrum, the results of which are displayed in Figure 4.5. The smoothing and differentiation procedure is based on a moving window which contains 17 data points.

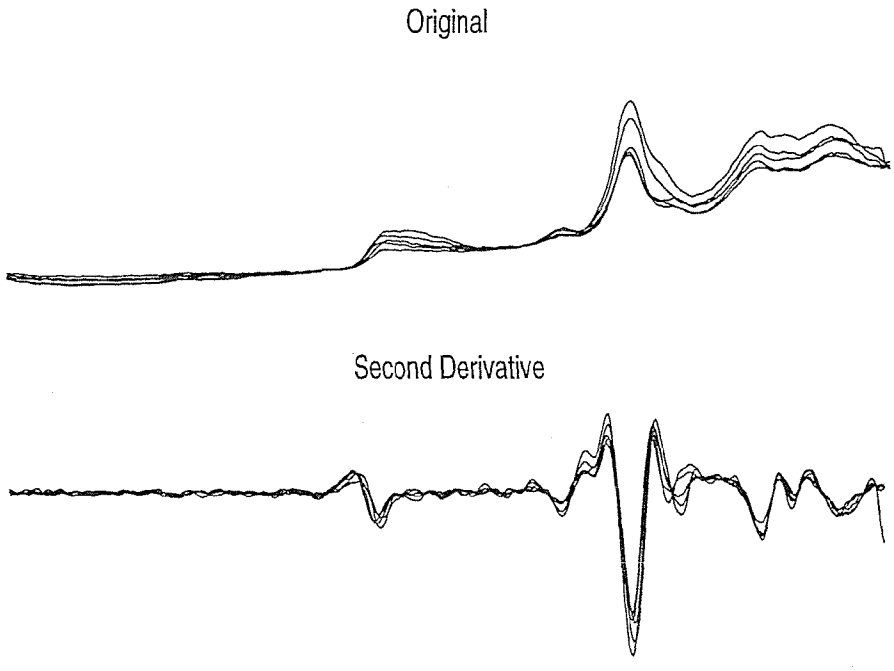


Figure 4.5: **Demonstration of the second derivative transformation.**

Second Derivative Transformation

1. $\mathrm{x} = (x_1, x_2, \ldots, x_p)$
2. n_w = odd number of data points in each window
3. $n_{\mathrm{lr}} = (n_w - 1)/2$
4. $\mathrm{j} = (-n_{lr}, \ldots, -1, 0, 1, \ldots, n_{lr})^T$
5. $\mathrm{x}_{2d} = (\ \ )$
6. FOR $i = 1 + n_{lr}$ to $p - n_{lr}$
7. $\quad \mathrm{x}_w = (x_{i-n_{lr}}, \ldots, x_{i+n_{lr}})^T$
8. $\quad (\mathrm{b}_0, \mathrm{b}_1, \mathrm{b}_2)^T = \left([1_{n_w}\ \ \mathrm{j}\ \ \mathrm{j}^2]^T\,[1_{n_w}\ \ \mathrm{j}\ \ \mathrm{j}^2]\right)^{-1}[1_{n_w}\ \ \mathrm{j}\ \ \mathrm{j}^2]^T\,\mathrm{x}_w$
9. $\quad \mathrm{x}_{2d} = (\mathrm{x}_{2d}, 2\mathrm{b}_2)$
10. END

Figure 4.6: **A simplified procedure for performing the second derivative transformation.**

The algorithm used for performing the second derivative transformation which produced Figure 4.5 is described for a single spectrum in Figure 4.6. Steps 1 – 4 are simply initialization procedures. The spectrum $\mathrm{x} = (x_1, x_2, \ldots, x_p)$ which contains p evenly spaced points is to be transformed using the second derivative transformation. The notation n_w indicates the number of points for which a second degree polynomial will be fitted, that is, the number of points in the window. The second derivative will be calculated for the point in the middle of the window. The midpoint has $n_{lr} = (n_w - 1)/2$ points to the left and right of itself. Step 4 constructs the independent variable whose values are indices which range from $-n_{lr}$ to n_{lr}. The transformed spectra will be labelled as x_{2d}. In Step 5, x_{2d} is initialized as a vector with no components. Step 6 begins the FOR loop, which is indexed from $i = 1 + n_{lr}$ to $p - n_{lr}$. This index range was chosen since it avoids any complications which may arise at the end points. These values for i represent the indices of points for which the second derivative will be calculated. If it is necessary to calculate the second derivative at the end points we refer the reader to Gorry [55] for further details. Step 7 extracts the data points from x which are in the current window, and will be fitted using a second degree polynomial. A second degree polynomial will be fitted to x_w as follows

$$\hat{\mathrm{x}}_w = \mathrm{b}_0 + \mathrm{b}_1\,\mathrm{j} + \mathrm{b}_2\,\mathrm{j}^2$$

for which the second derivative is equal to $2b_2$. Step 8 calculates the coefficients of the second degree polynomial and Step 9 stores the second derivative information in $\mathbf{x}_{2d}$.

A more sophisticated approach for calculating second derivative data can be based on the work performed in Gorry [55]. Gorry [55] describes a least squares smoothing and differentiation procedure which is based on the Savitzky-Golay convolution method [123]. The method of Gorry fits a series of Gram polynomials to windows of the spectrum. Their method also allows for derivatives of a higher order to be calculated.

Mean Centering

Mean centering is quite a convenient and common transformation to apply to spectral data which is to be used for regression analyses. This is especially the case when partial least squares and principal component regression are the regression methods being applied. If

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

is the mean (column) spectrum of the spectral data set $\mathbf{X}$, then the centered spectra are obtained as follows

$$\mathbf{x}_i - \bar{\mathbf{x}}$$

for $i = 1, 2, \ldots, n$. Equivalently, the data matrix can be mean centered by

$$\mathbf{X}_c = \mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^T.$$

As can be seen in Figure 4.7, when the data is mean centered each of the variables sum to zero. This results from the data being spread evenly around the horizontal axis.

Subsampling

Subsampling refers to the procedure of omitting every lth data point (or variable) $l \in Z$. For example, if the reflectance of a spectrum has been measured for the wavelengths 400,401,...,2200 nm, then a subsampled spectrum may consist of reflectance measurements for every second wavelength i.e. 400,402,404,...,2200 nm. If the spectra have been obtained by measuring the reflected (or absorbed) radiation for consecutive wavelengths, then in many cases little information will be destroyed by subsampling by small factors of l, eg $l = 2, 3, 4$.

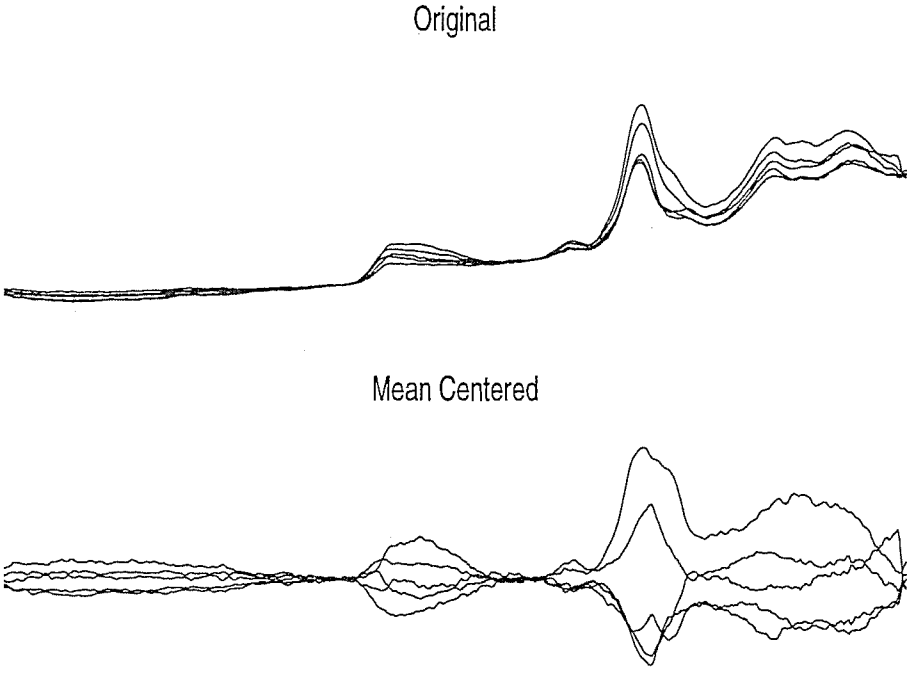


Figure 4.7: **Demonstration of mean centering.**

4.2.2 Principal Component Analysis (PCA)

Principal component analysis (PCA) also known as the Karhunen-Loéve method is mostly recognised as a dimension reducing technique in both statistics and engineering. In statistics, PCA is often used for reducing the dimensionality of a data set. It can also be used as an exploratory technique to help identify relationships among variables or even to help identify outliers or spurious points.

Principal component analysis seeks linear combinations of the original variables, such that the variance of the transformed objects is maximized. Equivalently, we seek the linear transformation

$$\mathbf{p} = \mathbf{X}^T\mathbf{a}$$

which maximizes

$$\mathbf{a}^T\mathbf{S}_T\mathbf{a}$$

subject to $\mathbf{a}^T\mathbf{a}$=1, where $\mathbf{S}_T$ is the total covariance matrix of $\mathbf{X}$. The vector $\mathbf{p}$ is the principal component which contains n principal component scores (one score for each object), and $\mathbf{a} = (a_1, a_2, \ldots, a_p)^T$ is the vector of principal component coefficients. The

maximization problem reduces to solving

$$(\mathbf{S}_T - \gamma\mathbf{I})\mathbf{a} = \mathbf{0}. \tag{4.1}$$

Notice that there will be p eigenvalues $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_p$ and p corresponding eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ where each $\mathbf{a}_i = (a_{1i}, a_{2i}, \ldots, a_{pi})^T$ produces p corresponding principal components $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_p$, which is the new coordinate system. Premultiplying Equation 4.1 with $\mathbf{a}^T$ it can be seen that the eigenvalues $\gamma_i = \text{variance}(\mathbf{p}_i) = \mathbf{a}_i^T\mathbf{S}_T\mathbf{a}_i$. The first principal component has the largest variance. The second principal component has the next largest variance such that it is uncorrelated with $\mathbf{p}_1$ and so on for $\mathbf{p}_3, \ldots, \mathbf{p}_p$. We will store the principal components as columns in the matrix $\mathbf{P}$. For more details the reader is referred to [72, 71].

Principal components can also be derived by using a singular value decomposition of the column centered matrix $\mathbf{X}_c^T$. The singular value decomposition of $\mathbf{X}_c^T$ can be written

$$\mathbf{X}_c^T = \mathcal{S}\mathcal{D}\mathcal{V}^T$$

where $\mathcal{S}$ is a $n \times n$ matrix, $\mathcal{D}$ is a $n \times p$ and $\mathcal{V}$ is a $p \times p$ matrix. The squared elements along the ith diagonal of $\mathcal{D}$ are proportional to the eigenvalues γ_i of the covariance matrix, that is

$$\frac{\mathcal{D}_{ii}^2}{n} = \gamma_i.$$

The orthogonal unit eigenvectors $\mathbf{a}_i$ for $i = 1, \ldots, p$ of the covariance matrix are stored in the ith corresponding column of $\mathcal{V}$. The principal component matrix can then be calculated by

$$\mathbf{P} = \mathbf{X}^T\mathcal{V},$$

and the centered principal components are formed by

$$\mathbf{P}_c = \mathbf{X}_c^T\mathcal{V} = \mathcal{S}\mathcal{D}.$$

The discussion so far has focused on finding the eigenvalues and eigenvectors of the covariance matrix of $\mathbf{X}$. The principal components can also be calculated from the correlation matrix of $\mathbf{X}$. Generally this approach is recommended if the units of the variables in $\mathbf{X}$ are measured on scales with widely differing ranges or if the measurement units differ. There is some mathematical simplifications which occur if the correlation matrix

is analysed. One useful result is that the principal component coefficients a_{ij} which have been calculated from the correlation matrix are directly proportional to the correlation between the ith principal component and the jth variable, or more formally

$$\hat{\rho}_{ij} = a_{ij}\sqrt{\text{variance}(\mathbf{p}_j)}.$$

Provided the data are correlated, which is true of spectral data, then most of the variability in the entire data set is accounted for in the first few principal components. This is an attractive feature of PCA. When PCA is being applied in conjunction with regression or discriminant analysis one item frequently overlooked is that, while most of the variability in the data can be accounted for in the first few principal components, these components are not necessarily the best for discrimination [102] or calibration [75]. This is because the criterion in which the principal components are constructed is not related to a regression or discrimination criterion, i.e, the data matrix is decomposed without any reference to the response vector (regression) or to the vector of group labels (classification).

For this reason, it can prove to be advantageous to use PCA followed by a feature selection technique. For classification purposes, SIMCA [38, 144] which is a popular chemometrics method, skillfully takes advantage of PCA. Here, the principal components are used to model the data from each class, and an object is assigned to the appropriate class depending on the distance between the object and the class model. Another approach for classification is to select the principal components based on a stepwise strategy [28]. Similar stepwise approaches have been used for calibration [75]. When principal components are used for regression, the procedure is typically called principal component regression (PCR). Instead of using the principal components as features for the statistical techniques, PCA can be used as a feature selection technique for selecting the original variables. Jouan-Rimbaud *et. al.* [74] select the wavelengths for calibration which have a high loading on the principal components. These particular principal components have previously been selected for regression by use of a stepwise procedure.

There is yet another way in which PCA can be used to provide information about which variables are important for the particular statistical procedure. This is achieved by using biplots which are a graphical representation of data. Biplots can be considered as an overlay of two scatterplots. The first plot could be the scatterplot of the first principal component versus the second principal component which shows the n principal component

scores. The second plot gives the relative positions of the p variables by plotting lines such that the lengths and direction of the lines for each of the wavelengths provides an indication as to which variables are important for each component. Thus if we know that the second principal component is useful for regression, and if the pth variable lies a fair distance along and near to this axis, one can then presume that this variable may be important for regression. Basically, biplots are a visual tool for investigating the loadings of the variables, and providing visual information about the correlation structure between the variables and objects.

A disadvantage of PCA is that if one spectrum is changed, then so do all the principal components. Another item of somewhat importance is that PCA does not take into consideration the particular ordering of variables (i.e. shape) of a spectrum. By interchanging the wavelengths so that the spectra become completely rearranged, the principal components will not be altered. The Fourier transform which we discuss next, does take into account the ordering of a spectrum. Also, if one spectrum is altered, then only the Fourier coefficients pertaining to that spectra vary.

4.2.3 Fourier Transform (FT)

A transformation often used with high dimensional data, particularly infra-red and near infra-red spectra is the Fourier transform (FT). This transformation was initially used in spectroscopy as a way of increasing the signal to noise ratio. The FT is also useful for providing visual interpretations of spectra, so useful in fact that many instruments perform this transformation automatically.

The FT [112] represents a function or discrete signal as a linear combination of complex exponential basis functions. Let $\mathbf{x} = (x_0, x_1, \ldots, x_{p-1})$ denote a discrete object such as a spectrum, then the discrete FT is

$$\omega_l = \sum_{i=0}^{p-1} x_l \exp(j_* 2\pi i l/p),$$

and the inverse discrete FT is

$$x_i = \frac{1}{p} \sum_{l=0}^{p-1} \omega_l \exp(-j_* 2\pi i l/p) \text{ where } j_* = \sqrt{-1}.$$

The Fourier coefficients convey information about the underlying frequency content of a spectrum, that is, they represent the weight to which a basis function of a particular

frequency contributes to the fit of the spectrum. Since the Fourier coefficients are complex numbers, the magnitude of each ω_l, denoted by $\mid \omega_l \mid$, is typically used for further multivariate analyses.

One of many references which have applied Fourier methods in classification is that by Young and Calvert [150], who mention that the frequency spectrum obtained by a FT of signal in the time domain can be quite valuable for the classification of speech signals. Wu *et. al.* [147] have also demonstrated that the Fourier transform is a useful feature extraction method to apply for classification.

While the FT does take into consideration the ordering of a spectrum, the Fourier coefficients are not localized. If one part of a spectrum or signal is changed slightly, then all the Fourier coefficients change as a result. To avoid such global effects, wavelet coefficients which are produced from the wavelet transform can be used. The wavelet coefficients are able to convey localized frequency information of a spectrum.

4.2.4 Discrete Wavelet Transform (DWT)

The wavelet transform has mostly been used for data compression and denoising. Some examples mentioned by Vidaković and Muller [139] in their tutorial paper include compressing fingerprint images and denoising old sound recordings. It is only recently that the DWT has been considered as a feature extraction method for discriminant analysis [8, 86, 96, 118, 130, 131, 133, 141]. Wavelet coefficients have also been used in regression analysis, but are mostly used for function estimation purposes, or in situations when there is a single independent variable.

The advantage associated with the discrete wavelet transform is that the output (or wavelet coefficients) convey localized frequency information about a signal. The localization is achieved by using basis functions (or basis vectors for discrete data) which are dilated and translated by different amounts.

Tate *et. al.* [131] have used the DWT for classifying magnetic resonance spectra (MRS). They performed a PCA on the wavelet coefficients from the DWT which were correlated with the class labels. The principal components were then used as features for classification. The authors noted that while their results were slightly better than typical methods for analyzing MRS data, the use of the DWT allowed for a more automated procedure

and reduced the amount of pre-processing.

Bos and Vrielink [8] found that the classification accuracy was improved when they chose to supply sets of wavelet coefficients to the classifiers, as opposed to supplying the full spectra. The classification results were based on a linear classifier and a non-linear neural network classifier. They also stress that computational expense was somewhat lessened by training the non-linear neural network on data of reduced dimensionality.

Saito and Coifman [118] provide an automated approach using the wavelet packet transform (WPT) for discriminant analysis. The wavelet packet transform is an extension to the DWT where the wavelet coefficients are organised in a binary tree structure. They also make use of the best-basis algorithm due to Coifman and Wickerhauser to select an orthonormal basis for signal classification [20]. Their application on two simulated data sets of dimension 32 and 128 highlights the potential of wavelet coefficients as features for discriminant analysis. The classifiers FLDA and CART were applied using the full data set and the reduced set of wavelet packet coefficients. It is clearly noted that less biased results were obtained with the wavelet packet coefficients. Both FLDA and CART on the original data had the tendency to overfit. The method of Saito and Coifman, which is referred to as the LDB algorithm for local discriminant bases is discussed in greater detail in Section 5.14.2.

Learned and Willsky [86] have also used the WPT for classification. The energy of the nodes in the WPT is calculated by the mean sum of squared coefficients in each of the nodes. A selection of these measures of energy are then used for classification. Walzcak *et. al* have also selected features from the wavelet packet transform using univariate feature selection strategies and the LDB algorithm. The features were used for classifying NIR spectra.

A new and innovative technique based on adaptive wavelets, which aims to reduce the dimensionality and optimize the discriminatory criterion is presented in Mallet *et. al* [96, 94]. The discrete wavelet transform is utilized to produce wavelet coefficients which are used for classification. Rather than using one of the standard wavelet bases, they generate the wavelet which optimizes specified discriminant criteria. The application of adaptive wavelets has also been extended to regression analysis.

Previous applications involving the optimization of wavelets for classification include the

work performed by Telfer *et. al.* [133] and Szu *et. al.* [130]. Telfer *et. al.* [133] consider optimizing the shift and dilation parameters of the discretization of a chosen wavelet transform, while Szu *et. al.* [130] sought the optimal linear combination of predefined wavelet bases for the classification of speech signals. The adaptive wavelet method is made distinct because the wavelet is designed from its humble beginnings. It also allows for the general m-band wavelet transform to be utilized, as opposed to the more common 2-band wavelet transform.

Adaptive wavelets are presented in Chapter 6. This follows a general overview of wavelet theory which is presented next.

Chapter 5

Wavelets

In the previous chapter we mentioned that wavelet coefficients might be good features to use as input to multivariate statistical techniques. Wavelet coefficients are potentially good features because they are able to detect changes which occur rapidly in a signal (or spectrum) as well as changes which occur over a longer duration in the signal. More importantly, wavelets have the ability to detect when the changes occur, unlike Fourier coefficients.

Consider the following example which demonstrates the ability of wavelet coefficients to capture local events. Figure 5.1 plots the function $\sin(2t)$ which has been sampled 512 times in $[-\pi, \pi]$. The sine curve on the right has a small disturbance at $t \approx 1.5$. Below each of the sine curves are the Fourier coefficients and the wavelet coefficients. Since the Fourier coefficients are complex, the magnitude of the coefficients is shown. Two plots of the Fourier coefficients have been shown. When the first half of the Fourier coefficients are displayed, it is difficult to detect any change in the Fourier coefficients produced for the original and disturbed signal. This is due to the large coefficient at the second index which reflects the period of the sine curve. When the magnitude of the 3rd- 19th Fourier coefficients are considered, then one can note the difference in Fourier coefficients produced from the two signals. Whilst the Fourier coefficients are different for the disturbed signal, (compared to the original signal) the small disturbance at $t \approx 1.5$ is absorbed across most of the Fourier coefficients. However, in the case of the wavelet coefficients most of the disturbance has been absorbed by only a few of the coefficients. What is also appealing is that the change in wavelet coefficients occurs in approximately the same region as the disturbance in the sine curve.

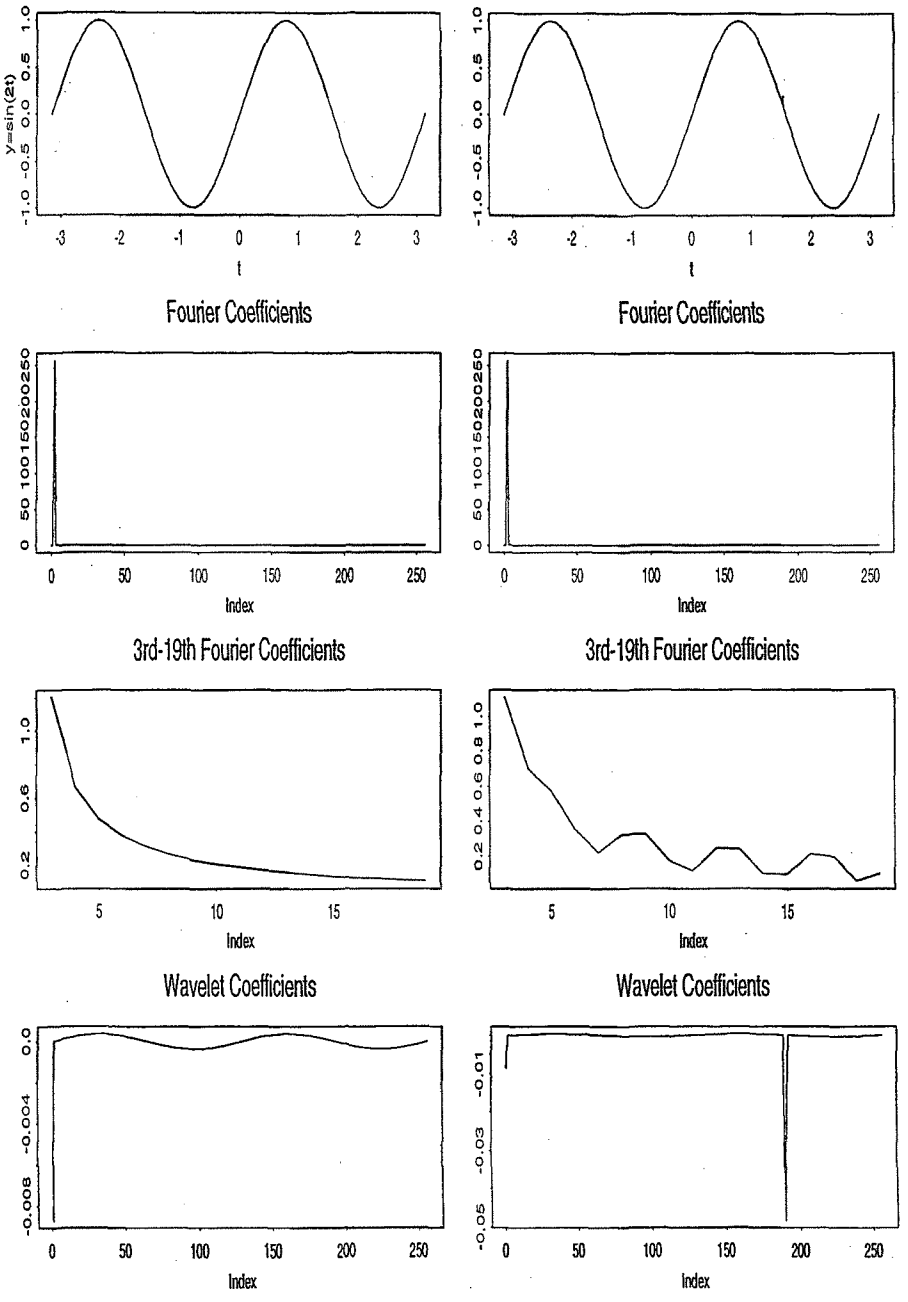


Figure 5.1: **Fourier and wavelet coefficient of a sampled sine signal, with (right) and without (left) a small disturbance.**

It should be mentioned that the relatively large disturbances occurring at the 0th index for the wavelet coefficients (left and right) can most likely be attributed to end effects. In this example we have endeavoured to provide some motivation for the use of wavelets in

statistics, and in particular, highlight some of the favourable properties they possess for feature extraction.

5.1 Introduction

Essentially, the wavelet transform allows us to view signals through different 'windows'. Some windows provide high frequency information and some windows provide low frequency information. The wavelet coefficients shown in Figure 5.1 are from a high frequency window.

We now set out to discuss in more detail the theory of wavelets. To avoid confusion, it should be stated that much of the theory of wavelets has evolved from continuous functions, so wavelets are initially explained in this chapter by using functions which are continuous. Following this, the wavelet transform for discretely sampled data is presented.

Wavelets form a set of basis functions which can be used to represent a function which is from the class of square integrable functions $L^2(\mathbb{R})$. The set of basis functions are derived by translating and dilating one basic wavelet, called a mother wavelet. The dilated and translated wavelet basis functions are called children wavelets. The coefficients in the expansion of the wavelet basis functions are calculated by the wavelet transform, and the coefficients are referred to as wavelet coefficients. The wavelet coefficients convey information about the weight that a wavelet basis function contributes to the function. Since the wavelet basis functions are localized and have varying scale, the wavelet coefficients therefore provide information about the frequency-like behaviour of the function (e.g. Figure 5.1).

To gain a better understanding of wavelets and their special characteristics, we first recall some details about Fourier analysis. The traditional Fourier transform provides information about the overall frequency content of a signal. The windowed Fourier transform (also called the short time Fourier transform) was introduced so that the frequency information about a signal could be localized. This is done by analysing pieces of a signal using a windowing function. In many instances, the procedure for determining the width of the windowing function is not straight forward. If a window width is too small or too large, then important information may still remain undetected or become distorted. The wavelet transform differs from the windowed Fourier transform, in that it allows us to

view the signal through windows whose widths vary in size.

The Fourier transform and windowed Fourier transform are briefly introduced in Sections 5.2 and 5.3, respectively. For a more complete account of Fourier theory see for example [19]. The continuous and discrete wavelet transforms (both of continuous functions) are introduced in Sections 5.4 and 5.5. Multiresolution analysis is then described in Section 5.6. Multiresolution analysis provides a neat framework for better understanding wavelets, what they represent and also leads to a fast algorithm for estimating the discrete wavelet transform. This is referred to as the fast wavelet transform and is discussed in Section 5.7. Higher multiplicity wavelets are discussed in Section 5.8. Each of the sections outlined above make reference to continuous functions (also referred to as continuous signals). Although our spectral data is discrete, we have decided to first discuss continuous signals, because it provides a historical account of wavelet theory, which then allows us to draw analogies between the wavelet transform for continuous functions and the wavelet transform for discrete data.

The discrete wavelet transform of discrete signals is then introduced in Section 5.9. If the reader wishes to avoid much of theory of wavelets in order to obtain a practical account of wavelet transforms, then they might like to advance to this section. The discrete wavelet transforms of (discrete) signals is introduced using ideas from filtering processes. The traditional wavelet transform is then extended to the more general m-band wavelet transform in Section 5.10 for a single object, and in Section 5.11 for an entire data set. Filter coefficient conditions are discussed in Section 5.12, and Section 5.13 gives a brief account of some boundary related issues before the idea of wavelet transforms is extended to wavelet packet transforms in Section 5.14. Wavelet packet transforms have a tree based structure with parent and children nodes.

In this thesis we apply wavelets which are orthogonal and have compact support, that is, they are non-zero over a finite interval only. Much of the literature is perhaps biased towards discussions on orthogonal wavelets because they are convenient and simple to implement. However, we feel that is necessary to make the reader aware that wavelets need not be orthogonal and that wavelets with other properties can be quite useful too. Briefly, when using an orthogonal basis it is not straight forward to obtain a wavelet which has symmetrical properties [24, 80] and allows for an exact reconstruction. That is of course

with the exception of the trivial Haar wavelet. Biorthogonal wavelets relax the assumptions of orthogonality, and allow for a perfect reconstruction with symmetrical wavelets. There are also semiorthogonal wavelets which are slightly more restrictive than biorthogonal wavelets, but may also be worthy of consideration. This thesis does not want to discuss in great detail other forms of wavelets, but simply wishes to mention their existence, and directs the reader to [14, 24, 70, 128, 105] for more information. Strang [128] presents a section on the symmetry and orthogonality issue and suggests some alternative approaches which can be considered if both symmetry and orthogonality is required. Turcajová [136] provides an excellent discussion on the application of higher multiplicity wavelets as an alternative approach to using wavelets with symmetrical properties. Besides the fact that many applications utilize wavelets which are orthonormal, we prefer to discuss orthogonal wavelets because it provides a convenient frame in which to design wavelets, as will be discussed in Chapter 6.

5.2 Fourier Transform

Let $f(t)$ represent a signal from the $L^2(\mathbb{R})$ class of functions, that is $\int_{-\infty}^{\infty} f^2(t)dt < \infty$. The continuous (integral) Fourier transform of $f(t)$ is then written

$$\mathcal{F}_{\mathrm{FT}}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j_*\omega t}dt, \tag{5.1}$$

where $t \in \mathbb{R}$ and $j_* = \sqrt{-1}$. Equation 5.1 states that in order to obtain information about a single frequency ω, it is necessary to integrate over the entire signal. Thus, any isolated frequency changes in the signal is averaged with the frequencies across the remainder of the signal. We would like to extract information pertaining to short bursts of high frequency activity from a signal. This leads to the windowed Fourier transform [49, 76] which was designed to provide localized frequency information about a signal.

5.3 Windowed Fourier Transform

The windowed Fourier transform of $f(t)$ is defined as

$$\mathcal{F}_{\mathrm{WFT}}(\omega, b) = \int_{-\infty}^{\infty} f(t)G(t-b)e^{-j_*\omega t}dt \qquad t, b \in \mathbb{R} \tag{5.2}$$

for some window function $G(t)$. The windowing function should have a finite integral and be non-zero over a finite interval. In general, window functions place more weight on the

function which is central to the window, and less weight as the function nears the border of the window. Equation 5.2 is essentially performing the Fourier transform on weighted blocks of $f(t)$ in an attempt to acquire localized frequency information about $f(t)$. The Fourier coefficients are now a function of two variables, ω and b. The parameter b controls the translation of the window function.

There are some drawbacks associated with the windowed Fourier transform. The precision with which the localized frequency information is obtained is limited by the size of the window. Choosing a window width too small may obscure effects of a slightly larger scale, and vice versa. There exists a tradeoff between time and frequency localization which is dependent on the window size. This tradeoff may be less apparent, if the size of the window could be adjusted. That way, we would be able to obtain information about high frequency events, which change quickly in time, as well as low frequency events, which change slowly over time. This is what wavelets set out to achieve. Wavelets are windowing functions which, as well as being translated in time, are also dilated by varying amounts in scale.

5.4 Continuous Wavelet Transform

Wavelets are translated and dilated versions of a single wavelet, called a mother wavelet. Figure 5.2 displays some translated and dilated wavelet basis functions from the Daubechies family [24, 22]. Mathematically, the windowing function $G(t-b)$, is replaced with a window function of the form $G\left(\frac{t-b}{a}\right)$, where a is the dilation parameter. The windowing function in the continuous wavelet transform (mother wavelet) is often denoted by $\psi(t)$, and the children wavelets are then $\psi\left(\frac{t-b}{a}\right)$. The continuous wavelet transform

$$F_{\mathrm{CWT}}(a,b) = \mid a \mid^{-\frac{1}{2}} \int_{-\infty}^{\infty} f(t)\psi\left(\frac{t-b}{a}\right) dt \qquad a, b \in \mathbb{R},\ a \neq 0$$

is an inner product of the signal $f(t)$ with the wavelets. Notice that the frequency parameter ω has been replaced by the dilation or scale parameter a. The factor $\mid a \mid^{-\frac{1}{2}}$ is included so that the rescaled wavelets all have equal energy, that is, $\| \psi\left(\frac{t-b}{a}\right) \| = \| \psi(t) \|$.

The original signal can be reconstructed using

$$f(t) = \frac{1}{c} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{\mathrm{CWT}}(a,b) \mid a \mid^{-\frac{1}{2}} \psi\left(\frac{t-b}{a}\right) \frac{da\, db}{a^2} \tag{5.3}$$

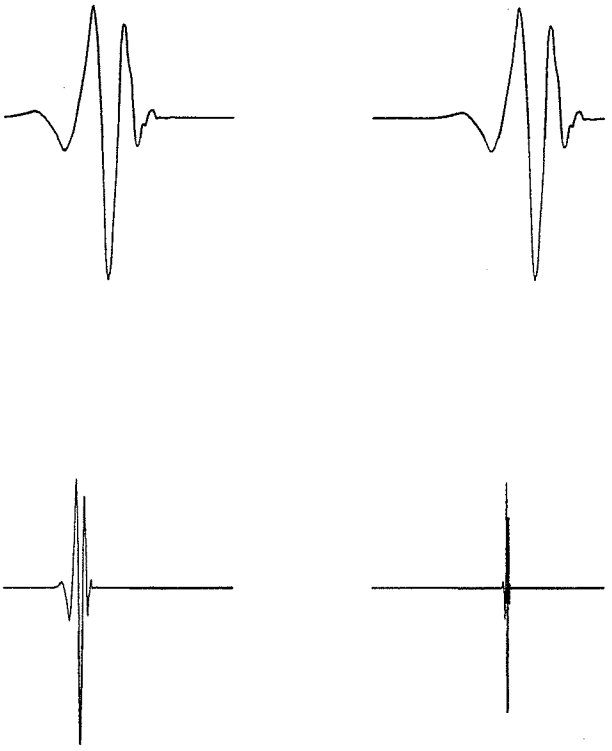


Figure 5.2: **Some wavelet basis functions from the Daubechies family.**

where c is a constant calculated by $c = 2\pi \int \mid \psi_{\text{ft}} \mid^2 d\omega / \mid \omega \mid$. For c to be finite the Fourier transform of the mother wavelet should equal zero, i.e., $\psi_{\text{FT}}(0) = 0$ and $\psi(t)$ oscillates so that its integral is zero. A decaying function $\psi(t)$ with $\int \psi(t) = 0$ is a suitable wavelet for the continuous wavelet transform [128].

It is not necessary to perform the continuous wavelets transform for all values of a and b, since $f(t)$ can be reconstructed from a much sparser set of (a, b) values. In fact, it is possible to obtain an analysis which is just as accurate, and more efficient, by using discrete values for the parameters a and b. This leads to the discrete wavelet transform (of a continuous signal).

5.5 Discrete Wavelet Transform

Restricting the parameters a and b to represent the discrete measures

$$\begin{aligned} a &= m^{-j} \\ b &= m^{-j}k \end{aligned}$$

where j, $k \in \mathbb{Z}$, $m \geq 2$, $m \in \mathbb{Z}^+$, then the discrete wavelet transform is defined

$$F_{\mathrm{DWT}}(j,k) = m^{j/2} \int_{-\infty}^{\infty} f(t)\psi(m^j - k)\, dt \qquad j,\, k \in Z\,.$$

Typically, m is set at two [42, 23, 22], in which case the mother wavelet is stretched or compressed by factors of two. Wavelets with $m > 2$ are sometimes referred to as higher multiplicity wavelets – these are discussed in Section 5.8. Our immediate discussion will however assume that $m = 2$ unless otherwise stated.

5.6 Multiresolution Analysis

Multiresolution analysis (MRA) [24, 91, 92] provides a concise framework for explaining many aspects of wavelet theory such as how wavelets can be constructed [128, 70]. MRA provides greater insight to the representation of functions using wavelets and helps establish a link between the discrete wavelet transform of continuous and discrete signals. The MRA also allows for an efficient algorithm for implementing the discrete wavelet transform. This is called the fast wavelet transform and follows a pyramidal scheme. Of course it should be stated that MRA still exists in the absence of wavelets, and that wavelets need not be associated with a multiresolution. However, the wavelets which we prefer to use, i.e. those with compact support, will, in most instances be generated from a MRA. For these reasons it is desirable to have wavelets which satisfy the properties of a multiresolution.

Multiresolution analysis allows us to represent functions at different resolutions, which can be likened to wavelets analysing functions through different size windows. A multiresolution divides the space of all square integrable functions $L^2(\mathbb{R})$, into a nested sequence of subspaces $\{V_j\}_{j\in\mathbb{Z}}$. Each subspace corresponds to a particular scale, and this provides the key for representing functions from $L^2(\mathbb{R})$ at different resolutions. The reason being, given some function $f(t) \in L^2(\mathbb{R})$, then $f(t)$ has pieces in each subspace. Let f_{V_j} denote the piece of $f(t)$ deposited in V_j, then f_{V_j} is an approximation of $f(t)$ at resolution 2^j.

There is something special about f_{V_j}, it is not just any approximation of $f(t)$ at resolution 2^j, it is the closest approximation to $f(t)$ at this resolution. That is,

$$\forall\, g(t) \in V_j, \quad ||g(t) - f(t)|| \geq ||f_{V_j} - f(t)||,$$

hence, f_{V_j} is an orthogonal projection of $f(t)$ onto V_j. The subspace V_j contains all the possible approximations of functions in $L^2(\mathbb{R})$ at resolution 2^j.

For the subspaces to generate a multiresolution, they must satisfy some conditions. It has already been mentioned that the subspaces are nested, this means that $\forall j \in \mathbb{Z}$, $V_j \subset V_{j+1}$. That is, a function at a lower resolution can be represented by a function at a higher resolution. Another condition is $\lim_{j\to-\infty} V_j = \bigcap V_j = \{0\}$. Since information about a function is lost as the resolution decreases, eventually the approximated function will converge to zero. Conversely, as the resolution increases the approximated function gets progressively closer to the original function, thus, $\lim_{j\to\infty} V_j = \bigcup V_j = L^2(\mathbb{R})$.

Where do these subspaces come from? The subspaces $\{V_j\}$ can be generated from each other by scaling the approximated functions in the appropriate subspace such that,

$$g(t) \in V_j \Leftrightarrow g(2t) \in V_{j+1} \quad j \in \mathbb{Z}.$$

It can also be stated that integer translates of the approximated functions, remain in the same subspace:

$$g(t) \in V_j \Leftrightarrow g(t-k) \in V_j \quad j,k \in \mathbb{Z}.$$

Summarising, the sequence of subspaces $\{V_j\}_{j\in\mathbb{Z}}$ is a multiresolution of $L^2(\mathbb{R})$ if the following conditions are satisfied:

1. $V_j \subset V_{j+1}$, $\bigcap V_j = \{0\}$, $\bigcup V_j = L^2(\mathbb{R})$
2. $g(t) \in V_j \Leftrightarrow g(2t) \in V_{j+1}$
3. $g(t) \in V_j \Leftrightarrow g(t-k) \in V_j$

If $\{V_j\}_{j\in\mathbb{Z}}$ is a multiresolution of $L^2(\mathbb{R})$, then there exists a unique function $\phi(t) \in L^2(\mathbb{R})$, called a scaling function such that $\{\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)\}$ is an orthonormal basis of V_j [92]. This then implies that any function in V_j can be represented by a linear combination of the $\{\phi_{j,k}(t)\}$. Hence, the orthogonal projection of $f(t) \in L^2(\mathbb{R})$ into V_j can be expressed as

$$f_{V_j} = \sum_{k=-\infty}^{\infty} c_{j,k}\phi_{j,k}(t).$$

the coefficients $c_{j,k}$ are called scaling coefficients. Since $V_0 \subset V_1$,

$$\phi(t) = \sqrt{2}\sum_{k=-\infty}^{\infty} \ell_k \phi(2t-k). \tag{5.4}$$

So how do wavelets enter the picture? Wavelets are basis functions which can be used to represent the information lost in approximating a function at a lower resolution. This difference is called the detailed part of the function. We prefer that this error lie in the orthogonal complement of the V_j's. Consider the difference between approximating a function at resolution 2^j and at 2^{j+1}. This difference will lie in the orthogonal complement of V_j which is denoted by W_j such that,

$$V_{j+1} = V_j \oplus W_j. \tag{5.5}$$

In terms of the functions in the subspaces, then

$$f_{V_{j+1}} = f_{V_j} + f_{W_j} \tag{5.6}$$

where f_{W_j} is the orthogonal projection of $f(t)$ into W_j. Further decomposing f_{V_j} produces

$$\begin{aligned} f_{V_{j+1}} &= f_{V_{j-1}} + f_{W_{j-1}} + f_{W_j} \\ &= \sum_{i=-\infty}^{j} f_{W_i} \end{aligned}$$

Then for some function $f(t)$ we have

$$\begin{aligned} f(t) &= f_{V_j} + \left(f(t) - f_{V_j}\right) \\ &= f_{V_j} + \sum_{i=j}^{\infty} f_{W_i} \end{aligned}$$

and one can then understand how a multiresolution allows us to represent a function at various resolutions.

Next, consider how we can represent each f_{W_j}. In order to represent the orthogonal projection of $f(t)$ into W_j, it is convenient if we have an orthonormal basis for W_j, just as we had an orthonormal basis for V_j. It can be shown [92] that provided $\{\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)\}$ is an orthonormal basis for V_j then there will exist a wavelet basis $\{\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)\}$ which spans W_j.

Since $W_0 \subset V_1$, an expression for the wavelets can be obtained from a linear combination of the scaling functions in the space V_1. That is

$$\psi(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} h_k \phi(2t - k). \tag{5.7}$$

The detail of the function obtained by decreasing the resolution from 2^{j+1} to 2^j is

$$f_{W_j} = \sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(t).$$

Since $L^2(\mathbb{R}) = \oplus_{j=-\infty}^{\infty} W_j$, every function in $L^2(\mathbb{R})$ can be represented as a linear combination of wavelet basis functions

$$f(t) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(t). \tag{5.8}$$

Thus we have arrived at the wavelet series representation of $f(t)$. Alternatively, one could write $f(t)$ as a linear combination of scaling and wavelet basis functions as follows

$$f(t) = \sum_{k=-\infty}^{\infty} c_{j_o,k}\phi_{j_o,k}(t) + \sum_{j=j_o}^{\infty} \sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(t)$$

The $c_{j,k}$ are referred to as scaling coefficients and the $d_{j,k}$ are the wavelet coefficients as described previously. Chan [18] shows that the $d_{j,k} = F_{DWT}(j,k)$.

Due to the orthogonality of the scaling and wavelet functions the scaling coefficients can be calculated by the inner product

$$c_{j,k} = \int f(t)\phi_{j,k}(t)\, dt$$

and the wavelet coefficients can be calculated by

$$d_{j,k} = \int f(t)\psi_{j,k}(t)\, dt.$$

The orthogonality conditions on the scaling and wavelet coefficients as presented in Strang [128] are summarized as follows:

1. The scaling functions $\phi(t-k)$ are orthonormal to each other:

$$\int_{\infty}^{\infty} \phi(t-k)\phi(t-k')\, dt = \delta(k-k').$$

2. The scaling functions are orthogonal to the wavelets:

$$\int_{\infty}^{\infty} \phi(t-k)\psi(t-k')\, dt = 0.$$

3. The wavelets $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$ are orthonormal at all scales:

$$\int_{\infty}^{\infty} \psi_{j,k}(t-k)\psi_{j',k'}(t-k')\, dt = 0.$$

In many cases $\phi(t)$ and $\psi(t)$ will not have a closed form, and are not straight forward to calculate. Strang [128] discusses various procedures for calculating $\phi(t)$ and $\psi(t)$. If the calculation of $\phi(t)$ and $\psi(t)$ can be a tedious matter, concern may arise regarding how the scaling and wavelet coefficients will be calculated. In the next section, we show that the wavelet coefficients can be obtained without actually having to construct $\phi(t)$ and $\psi(t)$, using the properties of the MRA.

5.7 Fast Wavelet Transform

The fast wavelet transform provides an efficient algorithm for computing the discrete wavelet transform. We will show that provided we know some function f_{V_j}, then the scaling and wavelet coefficients can be calculated in the absence of the scaling and wavelet functions. An expression for the scaling coefficients will be derived first, an expression for the wavelet coefficients then follows.

Lets assume that we know f_{V_j}, which is expressed as follows

$$f_{V_j} = \sum_{k=-\infty}^{\infty} c_{j,k}\phi_{j,k} \, .$$

Since the scaling basis functions in V_j are orthonormal to their translates,

$$c_{j,k} = \int_{-\infty}^{\infty} f_{V_j}\phi_{j,k} \, dt =< f_{V_j}, \phi_{j,k} > \, . \tag{5.9}$$

Equation 5.9 requires some formulation of $\phi_{j,k}$ and hence $\phi(t)$ which may be difficult to obtain. It is desirable that an expression for the scaling $c_{j,k}$ and wavelet coefficients $d_{j,k}$ be attainable without the need to construct $\phi(t)$ or $\psi(t)$. We now set about doing this. First, write

$$\begin{aligned} f_{V_j} &= \sum_{k=-\infty}^{\infty} c_{j,k}\phi_{j,k} \\ &= \sum_{k=-\infty}^{\infty} c_{j-1,k}, \phi_{j-1,k} + \sum_{k=-\infty}^{\infty} d_{j-1,k}, \psi_{j-1,k} \, . \end{aligned}$$

This is an expression for f_{V_j} which has projections in V_{j-1} and W_{j-1}. Now an expression for the scaling coefficients can be written as

$$\begin{aligned} c_{j-1,k} &= < f_{V_j}, \phi_{j-1,k} > \\ &= < \sum_{k=-\infty}^{\infty} c_{j,k}\phi_{j,k} \, , \, \phi_{j-1,k} > \, . \end{aligned}$$

Using the fact that $\phi_{j-1,k} = 2^{(j-1)/2}\phi(2^{j-1}t - k)$ and $\phi(2^{j-1}t - k) = \sum_{k=-\infty}^{\infty} \ell_{k-2i}\phi_{j,k}(t)$, then the following expression for the scaling coefficients is obtained

$$c_{j-1,i} = \sum_{k=-\infty}^{\infty} \ell_{k-2i} c_{j,k} \ . \tag{5.10}$$

Essentially we are just using the scaling coefficients at the higher resolution to calculate the scaling coefficients at the next resolution. This is sometimes referred to as the pyramidal algorithm [92, 104]. A similar procedure is performed for obtaining the wavelet coefficients, leading to the following expression

$$d_{j-1,i} = \sum_{k=-\infty}^{\infty} h_{k-2i} c_{j,k} \ . \tag{5.11}$$

Provided we know the scaling coefficients at some resolution level j, the remaining scaling coefficients and wavelet coefficients can be found by the pyramidal filtering algorithm without even having to construct a wavelet or scaling function. We need only work with the coefficients ℓ_k and h_k. In the sections to follow, h_k will be referred to as high pass filter coefficients, and the ℓ_k will be referred to as low pass filter coefficients. It will also be shown in Section 5.12 that conditions can be placed on the filter coefficients and independently of $\phi(t)$ and $\psi(t)$ so that a MRA and associated wavelet basis exists.

5.8 Higher Multiplicity Wavelets

In the discussion so far, we have rescaled wavelets by a factor of $m = 2$. In some situations it may be advantageous to rescale by some integer $m > 2$. When $m > 2$, wavelets are referred to as higher multiplicity wavelets [79, 81, 127, 63]. For higher multiplicity wavelets, there exists a single scaling function defined by

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

which generates $m - 1$ wavelets

$$\psi^{(z)}(t) = \sqrt{m} \sum_k h_k^{(z)} \phi(mt - k) \qquad z = 1, \ldots, m-1$$

with $m - 1$ corresponding sets of high pass filter coefficients, $h_k^{(z)}$. The constant $\sqrt{m}$ is used so that the wavelets form an orthonormal basis.

We first consider redefining a multiresolution to cater for situations when functions are rescaled by a general factor $m \geq 2$ and then show how the fast wavelet transform (or pyramidal algorithm) is performed for higher multiplicity wavelets.

The sequence of closed subspaces $\{V_j\}_{j\in\mathbb{Z}}$ is a m-multiresolution of $L^2(\mathbb{R})$ if the following conditions are satisfied [136]:

1. $V_j \subset V_{j+1}, \bigcap V_j = \{0\}, \bigcup V_j = L^2(\mathbb{R})$
2. $g(t) \in V_j \Leftrightarrow g(mt) \in V_{j+1}$
3. $g(t) \in V_j \Leftrightarrow g(t-k) \in V_j$

The subspace V_j contains all the possible approximations of functions in $L^2(\mathbb{R})$ at resolution m^j. The orthogonal projection of some function $f(t) \in L^2(\mathbb{R})$ into V_j is written as

$$f_{V_j} = \sum_{k=-\infty}^{\infty} c_{j,k}\phi_{j,k}(t)$$

and

$$f_{W_j} = \sum_{z=1}^{m-1}\sum_{k=-\infty}^{\infty} d_{j,k}^{(z)}\psi_{j,k}^{(z)}(t)$$

is the orthogonal projection of $f(t)$ into W_j. Notice that the wavelet coefficients $d_{j,k}^{(z)}$ are also indexed by z. Here,

$$\phi_{j,k}(t) = m^{j/2}\phi(m^j t - k)$$

where

$$\phi(t) = \sqrt{m}\sum_k \ell_k\,\phi(mt-k)$$

and

$$\psi^{(z)}(t) = \sqrt{m}\sum_k h_k^{(z)}\phi(mt-k) \qquad z = 1,\ldots,m-1.$$

The function $f(t)$ can also be completely described by the wavelet basis functions as follows.

$$f(t) = \sum_{z=1}^{m-1}\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} d_{j,k}^{(z)}\psi_{j,k}^{(z)}.$$

A pyramidal algorithm can also be used for calculating the scaling and wavelet coefficients for higher multiplicity wavelets. The procedure is similar to the case when $m = 2$.

That is, the scaling coefficients at some resolution are used to produce the scaling and wavelet coefficients at the next (lower) resolution. This is done as follows

$$c_{j-1,i} = \sum_{k=-\infty}^{\infty} \ell_{k-mi} c_{j,k} \ . \tag{5.12}$$

$$d^{(z)}_{j-1,i} = \sum_{k=-\infty}^{\infty} h^{(z)}_{k-mi} c_{j,k} \ . \tag{5.13}$$

5.9 The Discrete Wavelet Transform of Discrete Data

The previous sections, have made reference to continuous signals $f(t)$. We now diverge, and begin to discuss the discrete wavelet transform for discretely sampled signals. There are many similarities between the DWT of continuous signals and the DWT of discrete signals. The most notable feature is that the wavelet and scaling coefficients are calculated in the same way. That is the scaling coefficients $\mathbf{c}_j = \{c_{j,k}\}$ at some resolution or level j, are used to produce the scaling coefficients $\mathbf{c}_{j-1} = \{c_{j-1,k}\}$ and the wavelet coefficients $\mathbf{d}_{j-1} = \{d_{j-1,k}\}$ at the next lower level $j-1$.

The DWT of discrete signals can be likened to filtering procedures. There is one low pass filter (**L**) and one high pass filter (**H**). The low pass filter acts as a smoother, which produces a smoothed version of the data sequence which it is filtering. The high pass filter acts as a differencing operator which extracts the high frequency components of the signal that the low pass filter did not capture. The wavelet coefficients $\mathbf{d}_j = \{d_{j,k}\}$ are the outputs of the high pass filters and the outputs of the low pass filters are the scaling coefficients $\mathbf{c} = \{c_{j,k}\}$. This filtering procedure is related to the DWT with $m = 2$. For any general $m \geq 2$ the filtering operations would have one low pass filter, and $m-1$ high pass filters. For the moment we relate our discussion to the $m = 2$ case only. Section 5.10 relates the filter procedure to the DWT with $m \geq 2$.

We now proceed to mathematically describe the filtering operations used for obtaining the scaling and wavelet coefficients of some discrete data vector. To make the jump from continuous functions to discrete data vectors less impacting, we initially consider data vectors which have infinite length. The low pass filtering procedure is first introduced, and is then followed by the high pass filtering procedure.

A general filtering operation transforms a vector x into another vector s by

$$s_i = \sum_{k=-\infty}^{\infty} \ell_k x_{i+k} \ . \tag{5.14}$$

Here, $\ell = (\ldots, \ell_{-1}, \ell_0, \ell_1, \ldots)$ is the vector of low pass filter coefficients, which also has infinite length. Let the vector of filter coefficients be stored as rows in the matrix **L**, such that in the second row each element in ℓ has been shifted to the right by one position, and so on for successive rows. Then, Equation 5.14 can conveniently be described as a product of a low pass convolution matrix **L** and the data vector x, as follows

$$\mathrm{s} = \begin{pmatrix} \vdots \\ s_{-1} \\ s_0 \\ s_1 \\ \vdots \end{pmatrix} = \begin{pmatrix} & \vdots & \vdots & \vdots & \\ \cdots & \ell_0 & \ell_1 & \ell_2 & \cdots \\ \cdots & \ell_{-1} & \ell_0 & \ell_1 & \cdots \\ \cdots & \ell_{-2} & \ell_{-1} & \ell_0 & \cdots \\ & \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} \vdots \\ x_{-1} \\ x_0 \\ x_1 \\ \vdots \end{pmatrix} = \mathbf{L}\mathrm{x}. \tag{5.15}$$

A filter as described in Equation 5.14 is a linear shift-invariant operator. This means if our input vector x is shifted by some amount, then the output vector s is shifted by the same amount. Another consequence of filters being shift-invariant, is that each column in the matrix **L** from Equation 5.15 is a shift of the previous column. Also, the diagonals of **L** are constant, with the ℓ_k filling the kth diagonal.

When there is a finite number of filter coefficients, the filter is called a finite impulse response (FIR) filter. Another filter of importance is a causal filter. When the filter coefficients with negative indices are zero, that is, $\ell_k = 0$ for $k < 0$, we say that the filter is causal. In this thesis we consider filters which are both FIR and causal. Let N_f denote the finite number of filter coefficients with nonnegative indices so that $\ell = (\ell_0, \ell_1, \ldots, \ell_{N_f-1})$. The convolution matrix using a filter which is FIR and causal has the form

$$\mathrm{L} = \begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 0 & \ell_0 & \ell_1 & \ell_2 & \cdots & \ell_{N_f-1} & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & \ell_0 & \ell_1 & \cdots & \ell_{N_f-2} & \ell_{N_f-1} & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & \ell_0 & \cdots & \ell_{N_f-3} & \ell_{N_f-2} & \ell_{N_f-1} & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

In our case, the input vector will be a spectrum which has a finite number of elements. The number of elements is determined by the number of wavelengths for which the absorbance or reflectance of a substance has been measured. Finite length data poses a problem

near the endpoints. To understand this phenomenon consider the following example. Let $\mathbf{x} = (x_0, x_1, \ldots, x_7)^T$ be the input vector to the filter $\boldsymbol{\ell} = (\ell_0, \ell_1, \ell_2, \ell_3)$ is defined by four filter coefficients ($N_f = 4$). The output vector will be calculated by

$$s_i = \sum_{k=0}^{N_f-1} \ell_k x_{i+k}. \tag{5.16}$$

From Equation 5.16, s_0, s_1, s_2, s_3 and s_4 are calculated by

$$\begin{aligned} s_0 &= \ell_0 x_0 + \ell_1 x_1 + \ell_2 x_2 + \ell_3 x_3 \\ s_1 &= \ell_0 x_1 + \ell_1 x_2 + \ell_2 x_3 + \ell_3 x_4 \\ &\vdots \\ s_4 &= \ell_0 x_4 + \ell_1 x_5 + \ell_2 x_6 + \ell_3 x_7. \end{aligned}$$

Complications arise when s_5 is calculated. From Equation 5.16 we then have

$$s_5 = \ell_0 x_5 + \ell_1 x_6 + \ell_2 x_7 + \ell_3 x_8,$$

but x_8 is not defined. In this thesis periodic (circular) boundary conditions are applied, so that $x_0 = x_8$ and $x_1 = x_9$, or in general if our data vector has length p such that $\mathbf{x} = (x_0, x_1, \ldots, x_{p-1})^T$, then $x_i = x_{p+i}$. For details about implementing other forms of boundary conditions the reader is referred to [128, 107, 14].

Periodic boundary conditions have the effect of wrapping the filter coefficients in the convolution matrix so that

$$\mathbf{L} = \begin{pmatrix} \ell_0 & \ell_1 & \ell_2 & \ell_3 & \cdots & \cdots & \ell_{N_f-1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & & \vdots & \vdots & & & \vdots \\ \ell_2 & \ell_3 & \cdots & \cdots & \ell_{N_f-1} & 0 & 0 & \cdots & 0 & \ell_0 & \ell_1 \\ \ell_1 & \ell_2 & \cdots & \cdots & \ell_{N_f-2} & \ell_{N_f-1} & 0 & \cdots & 0 & 0 & \ell_0 \end{pmatrix}.$$

In the case above with $\mathbf{x} = (x_0, x_1, \ldots, x_7)^T$ and $\boldsymbol{\ell} = (\ell_0, \ell_1, \ell_2, \ell_3)$ then

$$\mathbf{L} = \begin{pmatrix} \ell_0 & \ell_1 & \ell_2 & \ell_3 & 0 & 0 & 0 & 0 \\ 0 & \ell_0 & \ell_1 & \ell_2 & \ell_3 & 0 & 0 & 0 \\ 0 & 0 & \ell_0 & \ell_1 & \ell_2 & \ell_3 & 0 & 0 \\ 0 & 0 & 0 & \ell_0 & \ell_1 & \ell_2 & \ell_3 & 0 \\ \ell_3 & 0 & 0 & 0 & 0 & \ell_0 & \ell_1 & \ell_2 \\ \ell_2 & \ell_3 & 0 & 0 & 0 & 0 & \ell_0 & \ell_1 \\ \ell_1 & \ell_2 & \ell_3 & 0 & 0 & 0 & 0 & \ell_0 \end{pmatrix}.$$

The discrete wavelet transform is performed by passing the discrete data vector through two filters, a low pass and a high pass filter. Together, the two filters form an analysis bank. The low pass filter is defined by the low pass filter coefficients $\ell = (\ell_0, \ell_1, \ldots, \ell_{N_f-1})$ and the high pass filter is defined by the high pass filter coefficients $\boldsymbol{h} = (h_0, h_1, \ldots, h_{N_f-1})$. The high pass filtering operations are described similarly to the low pass filtering operations. Let the output of the high pass filter be denoted by $\mathbf{w}$, then

$$w_i = \sum_{k=0}^{N_f-1} h_k x_{i+k}$$

The high pass convolution matrix has the form

$$\mathbf{H} = \begin{pmatrix} h_0 & h_1 & h_2 & h_3 & \cdots & \cdots & h_{N_f-1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & & \vdots & \vdots & & & \vdots \\ h_2 & h_3 & \cdots & \cdots & h_{N_f-1} & 0 & 0 & \cdots & 0 & h_0 & h_1 \\ h_1 & h_2 & \cdots & \cdots & h_{N_f-2} & h_{N_f-1} & 0 & \cdots & 0 & 0 & h_0 \end{pmatrix}$$

The sizes of $\mathbf{L}$ and $\mathbf{H}$ are influenced obviously by the size of the input and output vectors. Assuming the number of elements in input and output vectors is the same, then $\mathbf{L}$ and $\mathbf{H}$ must be square matrices with the number of rows and columns equal to the length of the input and output vectors.

A problem evolves from passing a spectrum through a low and a high pass filter – there is twice the amount of data, but not twice as much information. In terms of feature extraction where our goal is to reduce the data whilst retaining the majority of information, we are definitely heading in the wrong direction. Fortunately, this problem is easily overcome and the solution is simple. The filtered sequences are decimated. This means every second element in the sequence is deleted, or, stated another way, the filtered vectors are downsampled by two. The symbol $(\downarrow 2)$ will be used to indicate such a procedure. For example $(\downarrow 2)\mathbf{s} = (s_0, s_2, s_4, \ldots, s_{p-1})$ where p is an even number. The same effect can be achieved by dropping every second row in the convolution matrix. In the previous example with $\mathbf{x} = (x_0, x_1, \ldots, x_7)^T$ and $\ell = (\ell_0, \ell_1, \ell_2, \ell_3)^T$ we would then have

$$(\downarrow 2)\mathbf{L} = \begin{pmatrix} \ell_0 & \ell_1 & \ell_2 & \ell_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & \ell_0 & \ell_1 & \ell_2 & \ell_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ell_0 & \ell_1 & \ell_2 & \ell_3 \\ \ell_2 & \ell_3 & 0 & 0 & 0 & 0 & \ell_0 & \ell_1 \end{pmatrix}.$$

The low pass filter coefficients have been shifted horizontally by 2 $(=m)$ positions from the previous row.

The filtering operations which have been discussed so far are

$$\begin{aligned} \mathrm{s} &= \mathrm{Lx} \\ \mathrm{w} &= \mathrm{Hx}. \end{aligned}$$

When downsampling occurs we have

$$(\downarrow 2)\mathrm{s} = (\downarrow 2)\mathbf{L}\mathbf{x} \tag{5.17}$$

$$(\downarrow 2)\mathrm{w} = (\downarrow 2)\mathbf{H}\mathbf{x}. \tag{5.18}$$

It is convenient if we define the following notation to avoid the $(\downarrow 2)$ symbols

$$\begin{aligned} \mathbf{c} &= (\downarrow 2)\mathrm{s} \\ \mathbf{C} &= (\downarrow 2)\mathbf{L} \\ \mathbf{d} &= (\downarrow 2)\mathrm{w} \\ \mathbf{D} &= (\downarrow 2)\mathbf{H} \end{aligned}$$

then, Equations 5.17 and 5.18 become

$$\mathbf{c} = \mathbf{C}\mathbf{x} \tag{5.19}$$

and

$$\mathbf{d} = \mathbf{D}\mathbf{x}, \tag{5.20}$$

respectively. From here on we shall assume that the low pass and high pass filtering procedures incorporate the appropriate downsampling routines, so that $\mathbf{C}$ and $\mathbf{D}$ are now the low pass and high pass filter matrices.

The DWT of discrete data is obtained by iterating the low pass and high pass filtering operations on the scaling coefficients $\mathbf{c}$. Let $J \in \mathbb{Z}^+$ denote an arbitrary positive integer which indicates the highest level in the DWT. In practice the original data points in $\mathbf{x} = (x_0, x_1, \ldots, x_{p-1})^T$ are considered to be the scaling coefficients $\mathbf{c}_J$ at the highest level in the DWT. When $\mathbf{x}$ passes through the analysis bank we have two new sequences $\mathbf{c}$ and $\mathbf{d}$. It is convenient if the subscript $J-1$ is given to $\mathbf{c}$ and $\mathbf{d}$ such that

$$\begin{aligned} \mathbf{c}_{J-1} &= \mathbf{C}\,\mathbf{x} = \mathbf{C}\,\mathbf{c}_J \\ \mathbf{d}_{J-1} &= \mathbf{D}\,\mathbf{x} = \mathbf{C}\,\mathbf{c}_J\,. \end{aligned}$$

The elements in $\mathbf{c}_{J-1} = (c_{J-1,0}, c_{J-1,1}, \ldots, c_{J-1,\frac{p}{2}-1})$ and $\mathbf{d}_{J-1} = (d_{J-1,0}, d_{J-1,1}, \ldots, d_{J-1,\frac{p}{2}-1})$ now have two subscripts. The first subscript is simply the level of the filtering procedure and the second subscript is the element number in the vector. The number of elements in $\mathbf{c}_{J-1}$ and $\mathbf{d}_{J-1}$ is $p/2$ which is half that from the previous level. The next level of sequences is obtained by filtering the smoothed data sequence $\mathbf{c}_{J-1}$ as follows

$$\mathbf{c}_{J-2} = \mathbf{C}_{J-1}\,\mathbf{c}_{J-1} \tag{5.21}$$

$$\mathbf{d}_{J-2} = \mathbf{D}_{J-1}\,\mathbf{c}_{J-1}. \tag{5.22}$$

Notice that subscripts have been assigned to the matrices $\mathbf{C}$ and $\mathbf{D}$. This is necessary since whilst $\mathbf{C}$ and $\mathbf{D}$ will follow the same basic 'wraparound' pattern from one iteration to the next, their dimensions will change in accordance with the dimensions of $\mathbf{c}$ and $\mathbf{d}$. In general the DWT from any level j to $j-1$ can be defined by

$$\mathbf{c}_{j-1} = \mathbf{C}_j\,\mathbf{c}_j$$

$$\mathbf{d}_{j-1} = \mathbf{D}_j\,\mathbf{c}_j$$

for $j = J, J-1, \ldots, J - \max_{lev} + 1$ where $\max_{lev}$ is the maximum number of levels in the DWT.

The summation equations,

$$c_{j-1,i} = \sum_{k=0}^{N_f-1} \ell_k c_{j,2i+k} \tag{5.23}$$

$$d_{j-1,i} = \sum_{k=0}^{N_f-1} h_k c_{j,2i+k}. \tag{5.24}$$

for which the DWT is based can also be used quite useful in the construction of $\mathbf{C}_j$ and $\mathbf{D}_j$. Periodic boundary conditions as discussed earlier in this section continue to be applied so that

$$c_{j,k} = c_{j,2^j+k}$$

$$d_{j,k} = d_{j,2^j+k}.$$

Note the similarities between the DWT of discrete data using Equations 5.23 and 5.24 with the DWT of continuous functions using the recursion formulae in Equations 5.10 and 5.11. Both the discrete wavelet transforms (of continuous and discrete data) make use of the equivalent pyramidal algorithm.

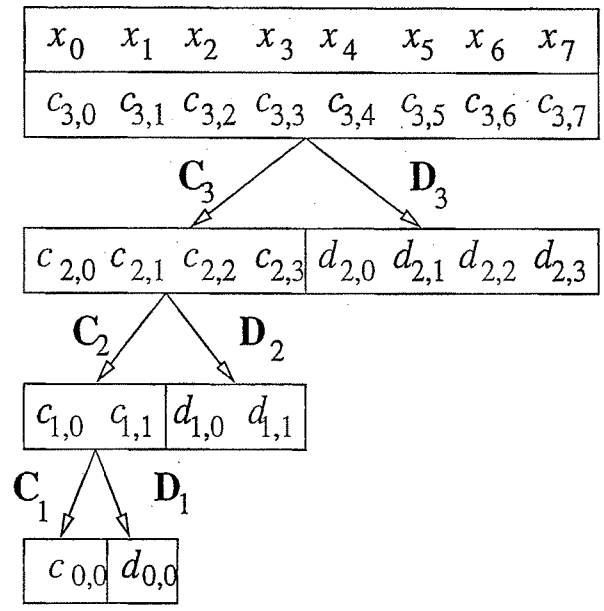


Figure 5.3: **Pictorial representation of a 2 band DWT for a signal which has been sampled 8 times.**

Figure 5.3 provides a pictorial description as to how the wavelet coefficients (and scaling coefficients) are calculated for some discretely sampled signal $\mathbf{x} = (x_0, x_1, \ldots x_7)^T$. Initially, all the data $\mathbf{c}_3$ are passed through the low and high pass filters to give the scaling $\mathbf{c}_2$ and wavelet $\mathbf{d}_2$ coefficients at the next lower level. As one progresses down the tree, the number of elements in each of the bands is reduced by half. The refiltering of the scaling procedure occurs for 3 cycles, that is, the number of levels n_{levels}, in this example is $n_{\text{levels}} = \max_{lev} = 3$. Whilst J could be some arbitrary integer we prefer to to set $J = \text{ceiling}(\log p / \log m)$.

For future reference, band(j, τ) will indicate the τth band $\tau \in (0, m-1)$ at the jth level of the DWT. The band at the top of the tree is band(3,0). At the next lower level, the bands will be labelled as band(2,0) and band(2,1) and so forth until the lowest level where the bands are denoted by band(0,0) and band(0,1) (see Figure 5.4). Note that for the two band DWT at a given j, the scaling coefficients will be contained in band$(j, 0)$, and the wavelet coefficients will be stored in band$(j, 1)$.

Figure 5.5 shows the effect of performing the DWT on an artificially generated spectrum. The spectrum contains 256 points and consists of a sine curve with a period of two, sampled over $-\pi$ to π, a block pulse, and another sine curve which has a period of 5 over the same

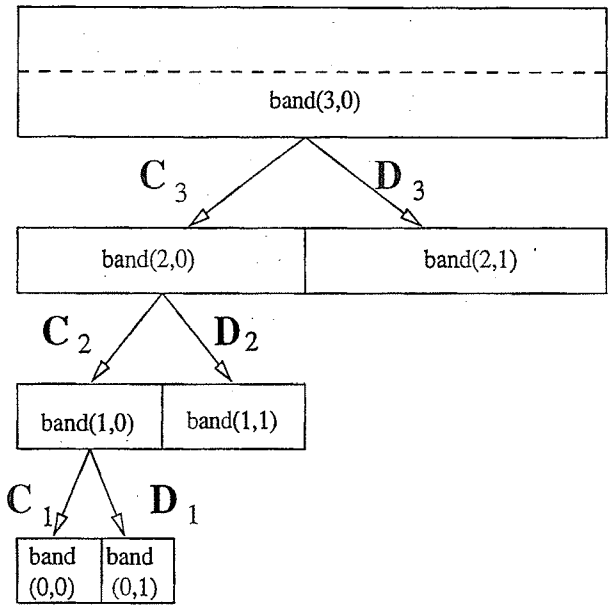


Figure 5.4: **Labelling of the bands in the DWT.**

interval $-\pi$ to π. The DWT is shown for the first six levels (j=8,7,6,5,4,3), when there are only 8 coefficients in the bands. The original spectrum in band(8,0) undergoes a low pass filtering process (which includes subsampling) to give the scaling coefficients in band(7,0). The original spectrum also undergoes the high pass filtering (and subsampling) process to give the wavelet coefficients which lie in band(7,1).

Next, the scaling coefficients in band(7,0) are then passed through the low pass and high pass filters to give the scaling coefficients in band(6,0) and the wavelet coefficients in band(6,1). The same procedure continues with the scaling coefficients from band(6,0) being filtered to give the scaling coefficients in band(5,0) and the wavelet coefficients in band(5,1). This process could continue for 8 (=$\max_{lev}$) in which case there would be one scaling coefficient and one wavelet coefficient. For display purposes only the first six levels are shown.

As one moves down the tree, the filtered signal in the scaling bands become smoother and smoother. The low pass filtering process can be likened to a smoothing procedure followed by decimation (i.e. subsampling). The wavelet bands highlight the information which has not been captured by the scaling bands.

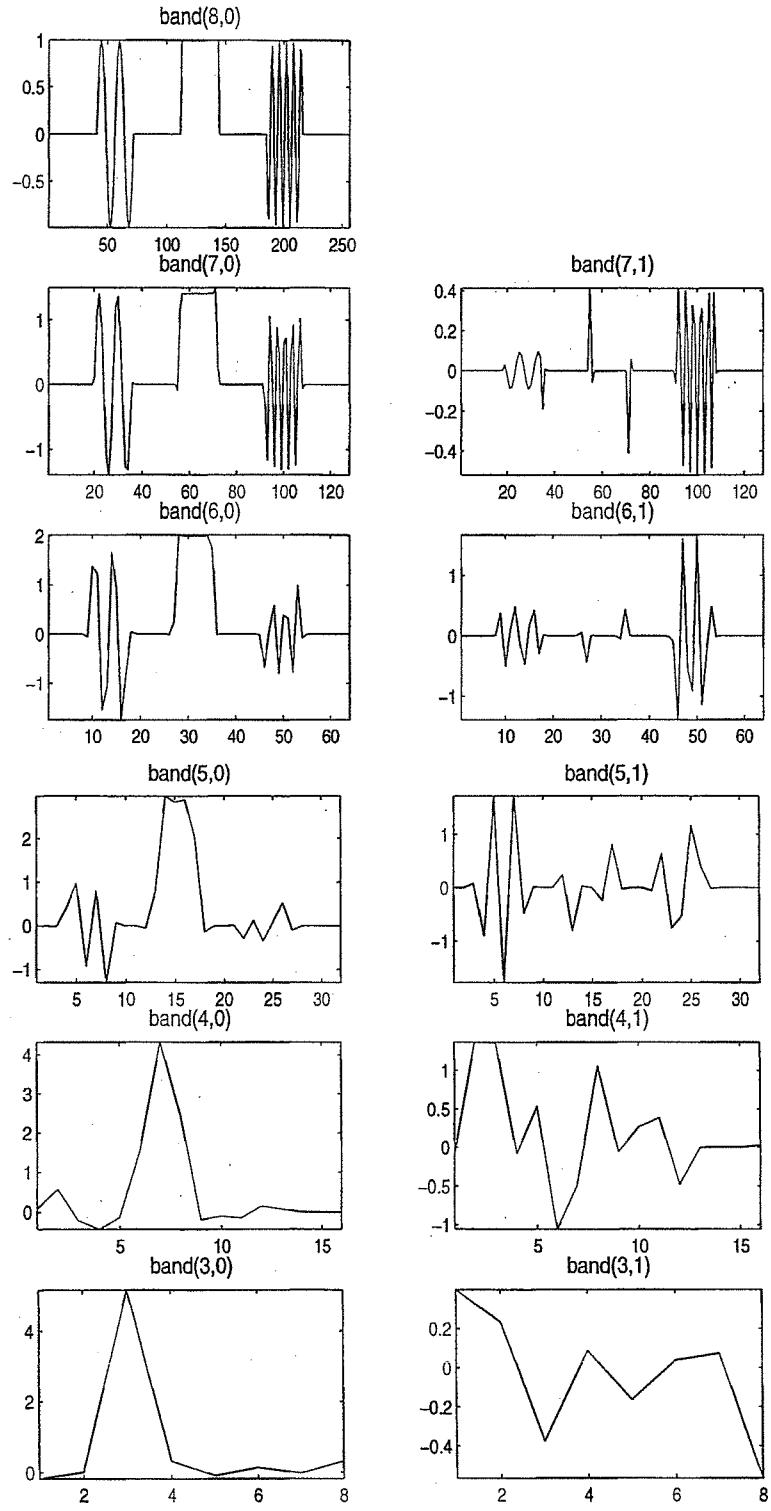


Figure 5.5: **2-band DWT performed on a generated spectrum to level three.**

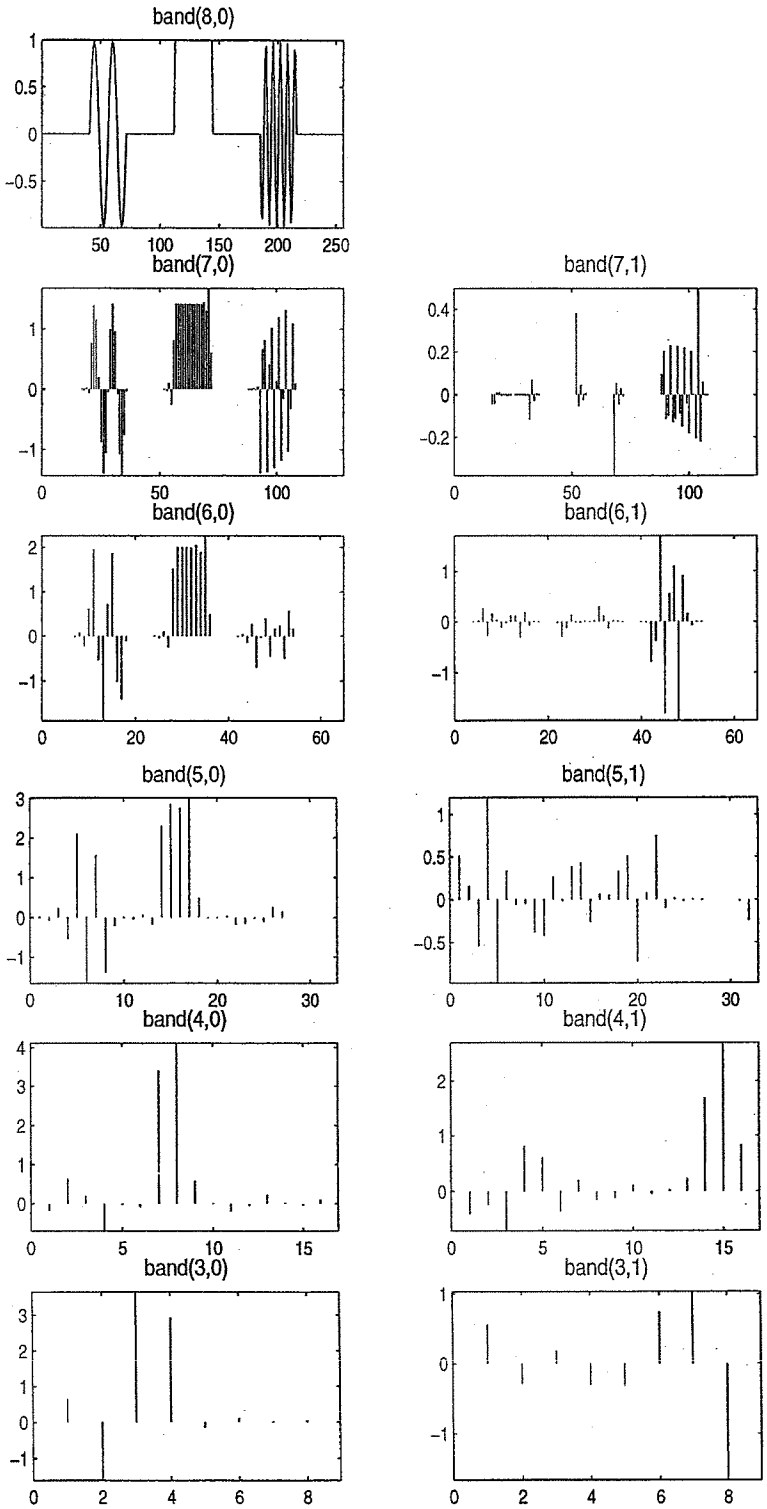


Figure 5.6: **Another presentation for a 2-band DWT performed on the generated spectrum to level three.**

The components with the highest frequency are the first to be removed from the scaling coefficients and captured by the wavelet coefficients. Consider for example the scaling coefficients in band(5,0) and band(4,0). Traces of the sine curve (which had a period of 5) are almost undetected in the scaling coefficients, whilst the remains of the sine curve $\sin(2t)$ are slightly more distinct. However, the most noticeable feature is the remains due to the block pulse.

In Figure 5.5 the wavelet and scaling coefficients were plotted against their index, which represents the element number of the coefficients in the respective **c** and **d** vectors. The plots appear somewhat continuous since the points in the plot have been joined. Figure 5.6 shows another way in which the scaling and wavelet coefficients can be displayed. Line segments proportional to the value of the scaling and wavelet coefficients are plotted at there respective index.

Now that some insight has been given to the DWT performed on a generated spectra, we present an example of the 2-band DWT applied to a spectrum which is similar to those analysed in this thesis. Figure 5.7 shows the effect of performing the DWT on a single spectrum for the levels $j = 8, 7, 6, 5, 4, 3$. Again, the first six levels have been chosen for display purposes only. The original spectrum which has been sampled 2^8 times lies in band(8,0). It can be seen that the wavelet coefficients at higher levels extract information about the smaller peaks, while the wavelet coefficients at lower levels in the tree, extract information pertaining to the larger, more significant peaks of the original spectra.

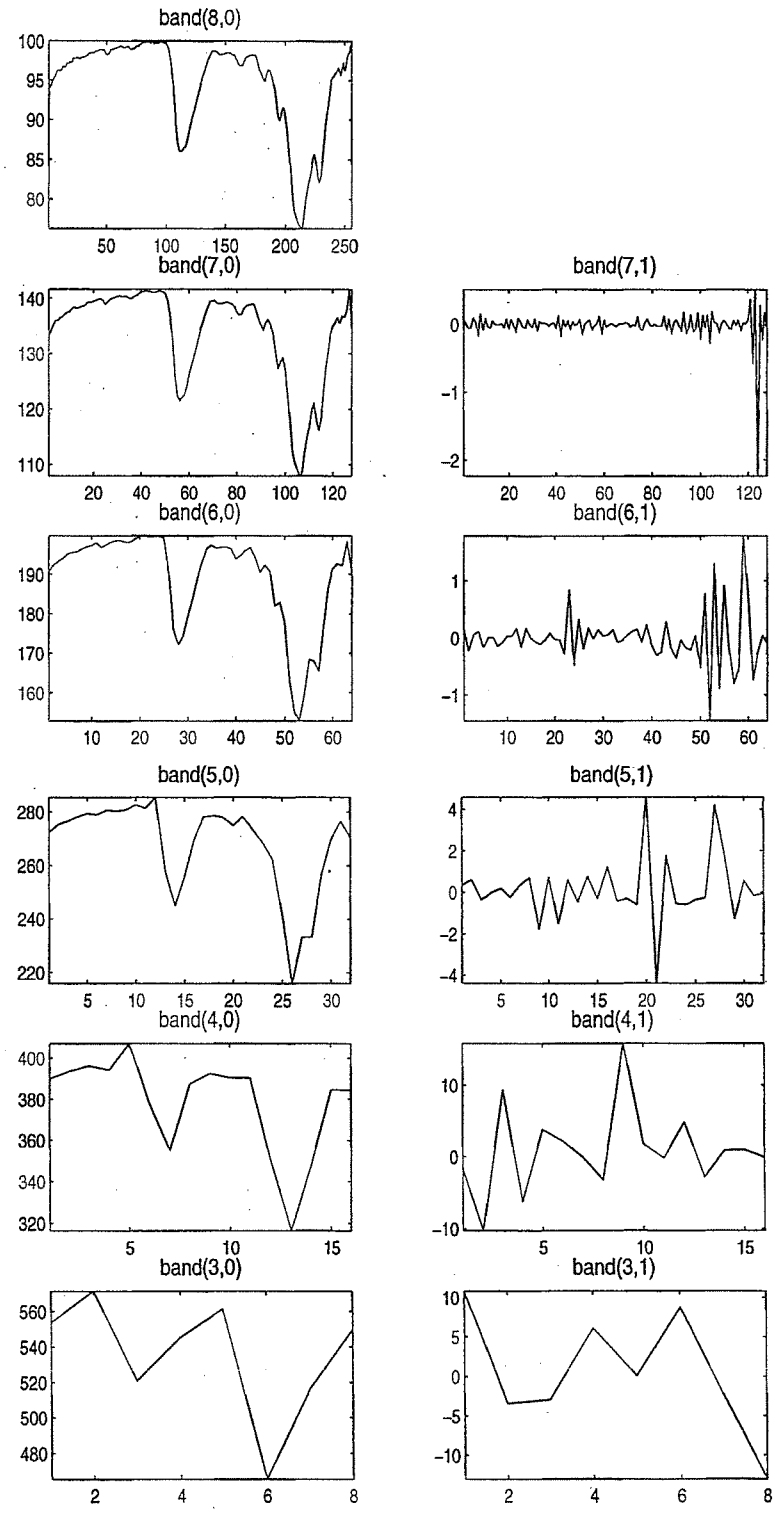


Figure 5.7: **Two-band DWT for a spectrum to six levels.**

5.10 The m−band Discrete Wavelet Transform of Discrete Data

Similar recursion formulae for calculating the scaling and wavelet coefficients can be derived for the $m-$band DWT of discrete data as those derived for the DWT of continuous function using higher multiplicity wavelets. Recall that when higher multiplicity wavelets were introduced in Section 5.8 there was one scaling function defined by one set of low pass filter coefficients, and $m-1$ wavelet functions which were defined by $m-1$ sets of high pass filter coefficients. The DWT with higher multiplicity wavelets on continuous data corresponds to performing the DWT on discrete data using a filter system which contains one low pass filter and $m-1$ high pass filters. This is referred to as a $m-$band DWT [127] of discrete data. For the m-band DWT, the downsampling rate is by a factor of m. This corresponds to shifting the filter coefficients in each row of the filter matrices by m. This is explained further in the example presented next.

A 3-band DWT for the spectrum $\mathbf{x} = (x_0, x_1, \ldots, x_8)$ is shown in Figure 5.8. There is one low pass and two high pass filters producing one set of scaling (or smoothed) coefficients and two sets of wavelet (or detailed) coefficients. As before, to go from one level to the next, only the scaling coefficients are filtered and, the number of coefficients in each band is reduced by one third when moving from one level to the next. For this example, $n_{\text{levels}} = \max_{lev} = 2$.

Following the same notation as introduced earlier, band(j, τ) will be referred to as the τth band $\tau \in \{0, 1, \ldots, m-1\}$ at the jth level $j \in \{J, J-1, \ldots, J - \max_{\text{lev}} + 1$ of the DWT. The band at the top of the tree is band(2,0). At the next level the bands from left to right are referred to as band(1,0), band(1,1) and band(1,2). Similarly, the bands in the last level of the DWT are band(0,0), band(0,1) and band(0,2).

In previous sections, the DWT, has been described by using a single convolution matrix for the low pass filtering operation, and a single convolution matrix for the high pass filtering operation. Now that we have several high pass filters it is necessary to introduce a convolution matrix for each high pass filter. For the case $m = 3$ and say $N_f = 6$ the filter coefficient matrices which decomposed the original data at level 3 to the next lower

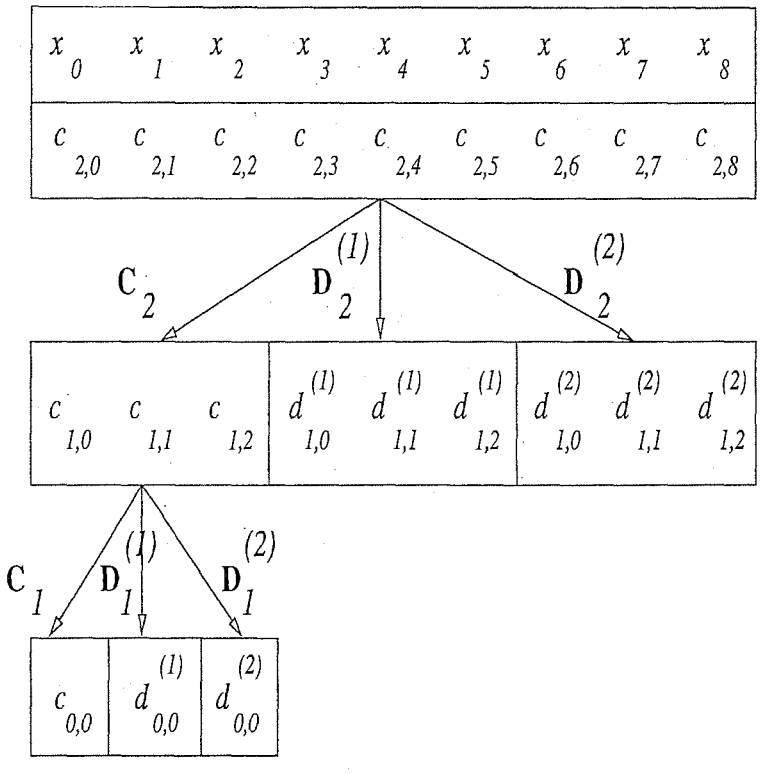


Figure 5.8: **A 3-band discrete wavelet transform.**

level 2, would be represented as follows

$$\mathbf{C}_2 = \begin{pmatrix} \ell_0 & \ell_1 & \ell_2 & \ell_3 & \ell_4 & \ell_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ell_0 & \ell_1 & \ell_2 & \ell_3 & \ell_4 & \ell_5 \\ \ell_3 & \ell_4 & \ell_5 & 0 & 0 & 0 & \ell_0 & \ell_1 & \ell_2 \end{pmatrix}$$

$$\mathbf{D}_2^{(1)} = \begin{pmatrix} h_0^{(1)} & h_1^{(1)} & h_2^{(1)} & h_3^{(1)} & h_4^{(1)} & h_5^{(1)} & 0 & 0 & 0 \\ 0 & 0 & 0 & h_0^{(1)} & h_1^{(1)} & h_2^{(1)} & h_3^{(1)} & h_4^{(1)} & h_5^{(1)} \\ h_3^{(1)} & h_4^{(1)} & h_5^{(1)} & 0 & 0 & 0 & h_0^{(1)} & h_1^{(1)} & h_2^{(1)} \end{pmatrix}$$

$$\mathbf{D}_2^{(2)} = \begin{pmatrix} h_0^{(2)} & h_1^{(2)} & h_2^{(2)} & h_3^{(2)} & h_4^{(2)} & h_5^{(2)} & 0 & 0 & 0 \\ 0 & 0 & 0 & h_0^{(2)} & h_1^{(2)} & h_2^{(2)} & h_3^{(2)} & h_4^{(2)} & h_5^{(2)} \\ h_3^{(2)} & h_4^{(2)} & h_5^{(2)} & 0 & 0 & 0 & h_0^{(2)} & h_1^{(2)} & h_2^{(2)} \end{pmatrix}$$

and the scaling and wavelet coefficients at level one in each of the bands would be calculated by

$$\begin{aligned} \mathbf{c}_1 &= \mathbf{C}_2\, \mathbf{c}_2 \\ \mathbf{d}_1^{(1)} &= \mathbf{D}_2^{(1)} \mathbf{c}_2 \\ \mathbf{d}_1^{(2)} &= \mathbf{D}_2^{(2)} \mathbf{c}_2 \end{aligned}$$

In general, the m-band DWT from some level j to the next lower level $j-1$ is performed

using

$$\begin{aligned} \mathbf{c}_{j-1} &= \mathbf{C}_j\,\mathbf{c}_j \\ \mathbf{d}_{j-1}^{(z)} &= \mathbf{D}_j^{(z)}\mathbf{c}_j \qquad z = 1,\ldots,m-1\,. \end{aligned}$$

In summation notation one has

$$c_{j-1,i} = \sum_{k=0}^{N_f-1} \ell_k c_{j,mi+k} \tag{5.25}$$

$$d_{j-1,i}^{(z)} = \sum_{k=0}^{N_f-1} h_k^{(z)} c_{j,mi+k} \qquad z = 1,\ldots,m-1\,. \tag{5.26}$$

$$\tag{5.27}$$

The periodic boundary conditions have

$$c_{j,k} = c_{j,m^j+k}$$

$$d_{j,k}^{(z)} = d_{j,m^j+k}^{(z)}.$$

These operations can be considered equivalent to the discrete wavelet transform of a continuous signal using higher multiplicity wavelets.

5.11 The m-Band Discrete Wavelet Transform of a Discrete Data Set

Our applications involve performing the m-band DWT ($m \geq 2$) for each object vector in a data set and then using the wavelet (or scaling) coefficients as features for some multivariate modelling method. The m-band DWT has previously been described for a single data vector, but it is more convenient to redefine this using a slight change of notation. Let $\mathbf{x}^{[j]}(\tau)$ be a column vector containing the coefficients in band(j,τ) of the DWT, so that for a given j, the scaling coefficients will be stored in $\mathbf{x}^{[j]}(0)$ and $\mathbf{x}^{[j]}(\tau)$ will be a vector of wavelet coefficients for $\tau \in \{1,\ldots,m-1\}$. The DWT from level j to level $j-1$ is then described by the matrix operations

$$\begin{aligned} \mathbf{x}^{[j-1]}(0) &= \mathbf{C}_j\mathbf{x}^{[j]}(0) \\ \mathbf{x}^{[j-1]}(z) &= \mathbf{D}_j^{(z)}\mathbf{x}^{[j]}(0) \qquad z = 1,\ldots,m-1\,. \end{aligned}$$

The DWT from level j to level $j-1$ for each spectrum is then described by

$$\begin{aligned} \mathbf{X}^{[j-1]}(0) &= \mathbf{C}_j \mathbf{X}^{[j]}(0) \\ \mathbf{X}^{[j-1]}(z) &= \mathbf{D}_j^{(z)} \mathbf{X}^{[j]}(0) \qquad z = 1, \ldots . m-1 \, . \end{aligned}$$

where $\mathbf{X}^{[j]}(\tau)$, is the matrix containing the coefficients for the objects which would lie in band(j,τ). Or more specifically, if $\mathbf{x}_i^{[j]}(\tau)$ denotes the coefficients in band(j,τ) obtained for object $\mathbf{x}_i$ to level j then, this vector will form the ith column in $\mathbf{X}^{[j]}(\tau)$. The original data matrix would be represented by $\mathbf{X}^{[J]}(0)$.

It is interesting to note that when the DWT is performed on an entire data set, the scaling coefficients tend to be more correlated than the wavelet coefficients, particularly at higher levels. We already know that spectral data suffer from being highly correlated, and since the scaling coefficients are similar to smoothed versions of the original spectra, then the scaling coefficients are likely to inherent the same high correlation structure. Table 5.1 was constructed to provide the reader with some idea about the correlation structure of the data in the scaling and wavelet bands of a 2-band DWT for a spectral data set which contained 100 spectra and had $p = 512$. The columns in Table 5.1 are indicative of

- Level: the level of the discrete wavelet transform. (The original data would be at level 9).
- Number> 0.7: the number of correlation coefficients whose magnitude is greater than 0.7.
- Mean: the mean of the absolute value of the correlation coefficients.
- Variance: the variance of the absolute value of the correlation coefficients.

The number, mean and variance calculations are compared for the scaling and wavelet coefficients at the various levels of a 2-band DWT. This information is also displayed graphically using boxplots in Figure 5.9. The middle line indicates the positioning of the median, and the width of the box is proportional to the number of observations. Since the number of scaling and wavelet coefficients is reduced from one level to the next, then so do the number of correlation coefficients which can be calculated.

	Scaling Coefficients			Wavelet Coefficients		
level	number>0.7	mean	variance	number>0.7	mean	variance
8	1781	0.39	0.09	5	0.15	0.00
7	427	0.39	0.09	5	0.15	0.00
6	98	0.38	0.09	9	0.21	0.07
5	21	0.34	0.08	13	0.33	0.17
4	2	0.30	0.06	10	0.50	0.34
3	0	0.35	0.05	1	0.51	0.36

Table 5.1: **Summary statistics for the correlation coefficients of the scaling and wavelet coefficients of a spectral data set.**

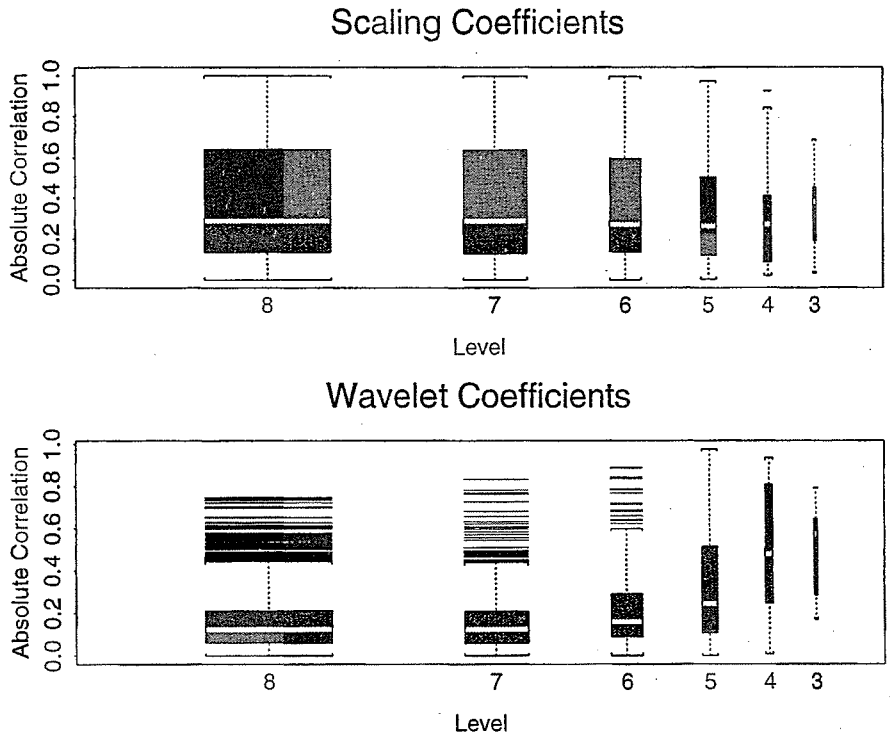


Figure 5.9: **Boxplots obtained from the correlation coefficients discussed for Table 5.1.**

5.12 Filter Coefficient Conditions

We have shown that it is possible to obtain the discrete wavelet transform of both continuous functions and discrete data points without having to construct the scaling or wavelet

functions. We only need to work with the filter coefficients. One may begin to wonder where the filter coefficients actually come from. Basically, wavelets with special characteristics such as orthogonality, can be determined by placing restrictions on the filter coefficients.

Let A denote the matrix of filter coefficients with the first row containing the low pass filter coefficients and the remaining $m-1$ rows the sets of high pass filter coefficients. If N_f is the number of filter coefficients contained in each filter, then A will be a $m \times N_f$ matrix. A can be partitioned into $m \times m$ sub-matrices as follows

$$A = (A_0\ A_1\ \cdots\ A_q).$$

Here, q is a non-negative integer such that $q = (N_f/m) - 1$. If for example, there were three filters ($m = 3$), with each filter containing six filter coefficients ($N_f = 6$), hence $q = 6/3 - 1 = 1$ then

$$A = \begin{pmatrix} \ell_0 & \ell_1 & \ell_2 & \ell_3 & \ell_4 & \ell_5 \\ h_0^{(1)} & h_1^{(1)} & h_2^{(1)} & h_3^{(1)} & h_4^{(1)} & h_5^{(1)} \\ h_0^{(2)} & h_1^{(2)} & h_2^{(2)} & h_3^{(2)} & h_4^{(2)} & h_5^{(2)} \end{pmatrix}$$

could be expressed as $A = (A_0\ A_1)$ with

$$A_0 = \begin{pmatrix} \ell_0 & \ell_1 & \ell_2 \\ h_0^{(1)} & h_1^{(1)} & h_2^{(1)} \\ h_0^{(2)} & h_1^{(2)} & h_2^{(2)} \end{pmatrix}$$

and

$$A_1 = \begin{pmatrix} \ell_3 & \ell_4 & \ell_5 \\ h_3^{(1)} & h_4^{(1)} & h_5^{(1)} \\ h_3^{(2)} & h_4^{(2)} & h_5^{(2)} \end{pmatrix}$$

The restrictions which are imposed on the filter coefficients so that a MRA and orthogonal wavelet basis exist are summarized as follows [78]

1. Orthogonality

$$\sum_k A_k A_{k+i}^T = \delta_{0i} I,$$

where $\delta_{0i} = 1$ if $i = 0$, and zero otherwise, I is the identity matrix.

2. The basic regularity condition

$$\sum_k \ell_k = \sqrt{m}.$$

3. The Lawton matrix

$$M_{ij} = \sum_k \ell_k \ell_{k+j-mi}$$

must have 1 as a simple eigenvalue. If more sophisticated wavelet and scaling functions are required, then more constraints need to be placed on the filter coefficients.

In practice it is common to choose a set of filter coefficients from literature such as the Daubechies or Coiflet filter coefficients, see for example [24]. Chapter 6 considers an approach for designing the wavelet matrix $\boldsymbol{A}$ with the goal of optimizing some multivariate modelling criteria.

5.13 Boundary Related Issues

In the examples presented so far, the dimensionality of the data has been set at $p = m^{\max_{lev}}$. It is not necessary that the number of variables be some integer power of m. In the case of periodic boundary conditions one requires that $p/m^{n_{\text{levels}}}$ be equal to an integer, where n_{levels} is the number of levels in the DWT as defined earlier. For instance, a 2-band DWT could be performed on data vector with length equal to 20. In this case the maximum number of levels in the DWT would be $\max_{\text{lev}} = 2$. This is the largest integer for which $p/m_{\text{lev}}^{\max} = 20/2_{\text{lev}}^{\max}$ is also an integer. For other boundary conditions such as zero padding and symmetric extension this assumption can be relaxed, but in some cases there is a penalty to pay. If for example symmetric end reflection is applied to data whose dimensionality is not divisible by $m^{n_{\text{levels}}}$, then exact reconstruction is only possible for biorthogonal wavelets [14]. The Splus wavelets user's manual [14] provides a concise summary of the advantages and disadvantages which should be considered when implementing a boundary method. As default settings they have implemented the periodic boundary treatment method for data which has p divisible by $m^{n_{\text{levels}}}$ where n_{levels} is pre-specified by the user. When biorthogonal wavelets are implemented and p is not divisible by $m^{n_{\text{levels}}}$ then the symmetric reflection boundary condition is applied. When orthogonal wavelets for the same scenario are used, then zero padding is applied. For more details about boundary treatments which can be applied the reader is referred to [14, 105, 128].

Another issue which arises is when there are more filter coefficients than wavelet coefficients. This will usually result at a lower level in the DWT transform. One has to ask if it is reasonable to have more filter coefficients than data points. As a general rule you

may wish to define $\max_{\text{lev}}$ to be the largest integer such that $p/m_{\text{lev}}^{\max}$ is an integer greater than or equal to N_f.

5.14 The Wavelet Packet Transform of Discrete Data

So far we have only considered filtering the scaling coefficients, but it seems perfectly viable to filter the wavelet coefficients. The wavelet packet transform (WPT) is obtained by filtering both the scaling and wavelet coefficients. In this section the discussion on the wavelet packet transform assumes the $m = 2$ case. Although it is not necessary, this discussion on WPT can be simplified if one assumes that $p = 2^J$.

The WPT has a tree like structure, where each band in the transform produces two new children bands at the next lower level. The tree like structure occurs because now the detailed (or wavelet) coefficients are filtered through a low pass and a high pass filter to obtain the next lower level of the WPT. This is done in the same way that the smoothed (or scaling coefficients) are filtered. Figure 5.10 presents the structure of a wavelet packet transform for some discretely sampled signal $\mathbf{x} = (x_0, x_1, \ldots, x_{2^J-1})^T = \mathbf{x}^{[J]}(0)$. Here the notation ${}^o\mathbf{x}^{[j]}(\tau)$ is used to represent the wavelet packet coefficients which occur at the jth level in the τth band of the decomposition. The DWT is simply the left most branches of the WPT.

We now describe how the filtering operations depicted in Figure 5.10 are obtained mathematically. For some $\mathbf{x} = (x_0, x_1, \ldots, x_{2^J-1})^T = \mathbf{x}^{[J]}(0)$, the $(J-1)$st level of the WPT would be obtained as for the DWT, that is the data is passed through a low pass and a high pass filter so that

$$\begin{aligned} {}^o\mathbf{x}^{[J-1]}(0) &= \mathbf{C}_J\, \mathbf{x}^{[J]}(0) \\ {}^o\mathbf{x}^{[J-1]}(1) &= \mathbf{D}_J\, \mathbf{x}^{[J]}(0) \end{aligned}$$

For the WPT, the number of bands doubles from one level to the next (lower) level, since each of the bands in the previous level is passed through a low pass and a high pass filter. At the next level, there will be four bands of wavelet packet coefficients which are obtained by

$${}^o\mathbf{x}^{[J-2]}(0) = \mathbf{C}_{J-1}\, {}^o\mathbf{x}^{[J-1]}(0)$$

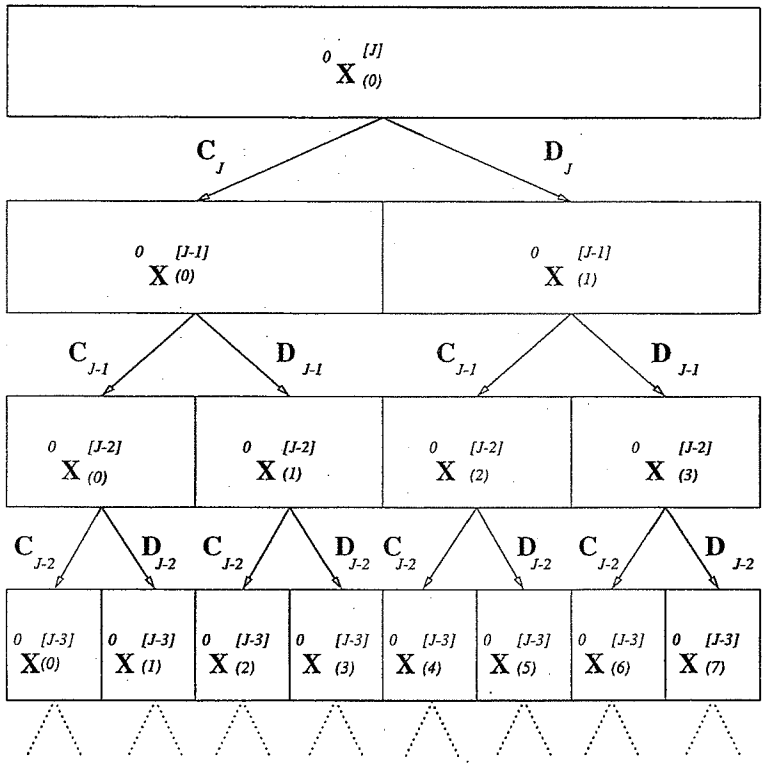


Figure 5.10: **Wavelet packet transform with $m = 2$.**

$$
\begin{aligned}
{}^{o}\mathrm{x}^{[J-2]}(1) &= \mathbf{D}_{J-1}\ {}^{o}\mathrm{x}^{[J-1]}(0) \\
{}^{o}\mathrm{x}^{[J-2]}(2) &= \mathbf{C}_{J-1}\ {}^{o}\mathrm{x}^{[J-1]}(1) \\
{}^{o}\mathrm{x}^{[J-2]}(3) &= \mathbf{D}_{J-1}\ {}^{o}\mathrm{x}^{[J-1]}(1).
\end{aligned}
$$

Continuing to the next level, one then has

$$
\begin{aligned}
{}^{o}\mathrm{x}^{[J-3]}(0) &= \mathbf{C}_{J-2}\ {}^{o}\mathrm{x}^{[J-2]}(0) \\
{}^{o}\mathrm{x}^{[J-3]}(1) &= \mathbf{D}_{J-2}\ {}^{o}\mathrm{x}^{[J-2]}(0) \\
{}^{o}\mathrm{x}^{[J-3]}(2) &= \mathbf{C}_{J-2}\ {}^{o}\mathrm{x}^{[J-2]}(1) \\
{}^{o}\mathrm{x}^{[J-3]}(3) &= \mathbf{D}_{J-2}\ {}^{o}\mathrm{x}^{[J-2]}(1) \\
{}^{o}\mathrm{x}^{[J-3]}(4) &= \mathbf{C}_{J-2}\ {}^{o}\mathrm{x}^{[J-2]}(2) \\
{}^{o}\mathrm{x}^{[J-3]}(5) &= \mathbf{D}_{J-2}\ {}^{o}\mathrm{x}^{[J-2]}(2) \\
{}^{o}\mathrm{x}^{[J-3]}(6) &= \mathbf{C}_{J-2}\ {}^{o}\mathrm{x}^{[J-2]}(3) \\
{}^{o}\mathrm{x}^{[J-3]}(7) &= \mathbf{D}_{J-2}\ {}^{o}\mathrm{x}^{[J-2]}(3)
\end{aligned}
$$

The same procedure may continue until there is one wavelet packet coefficient in each of

the bands. As for the DWT, there can be a maximum of J levels in the WPT, the main difference is that the WPT has 2^{J-j} bands at each level $j \in J, J-1, \ldots, 0$.

When both the scaling and wavelet coefficients are filtered there is a surplus of information stored in the wavelet packet tree. An advantage of this redundant information is that it provides greater freedom in choosing an orthogonal basis. Coifman *et al.* [20, 143] introduces a routine called the best basis algorithm which endeavours to find a basis in the WPT which optimizes some criterion.

5.14.1 The Best Basis Algorithm

The best basis algorithm seeks a basis in the WPT which optimizes some criterion function. Thus, the best basis algorithm is a task specific algorithm in that the particular basis is dependent upon the role for which it will be used. For example, a basis chosen for compressing data may be quite different to a basis that might be used for classifying or calibrating data, since different criterion functions would be optimized. The wavelet packet coefficients which are resultant of the best basis, may then be used for some specific task such as compression or classification for instance.

The first step in obtaining the wavelet packet coefficients from the best basis is to produce the wavelet packet decomposition tree to some level j_o. A criterion measure for each of the wavelet packet coefficients in each node (or band) in the wavelet packet decomposition is calculated and is denoted by $\mathcal{J}({}^{o}\mathbf{x}^{[j]}(\tau))$ for $j = J, \ldots, j_o$. One starts at level j_o in the tree and works up, gradually deleting the bands of coefficients in the tree which do not produce sufficiently good criterion measures. This can be formalised. Initially, the criterion measure for each of the bands of coefficients at level $j_o + 1$ are compared with the criterion measures for the bands of the coefficients in the descendants at level j_o. Here descendant nodes are used to categorize any nodes which lie beneath a node at a higher level in the tree. The node which the descendant nodes lie under is called a parent node. If the criterion measure of the parent node is superior to that of the descendant nodes, then the descendant nodes are deleted. If the descendant nodes produce a superior criterion measure, then the descendant nodes are kept and the parent node is deleted. This procedure continues all the way to the top of the tree and the coefficients in the best basis will lie in the bands which were not deleted in the elimination process.

Figure 5.11 summarizes the procedure described above, i.e. how to find the wavelet packet

Obtaining the Wavelet Packet Coefficients From the Best Basis Algorithm

1. Perform WPT for $\mathbf{x} = (x_0, \ldots, x_{2^J-1})^T$ to level j_o.
2. $\mathrm{BB}(j_o, \tau) = \mathrm{band}(j_o, \tau)$ for $\tau = 0, \ldots, 2^{J-j_o} - 1$
3. FOR $j = j_o - 1, \ldots, J$
4. FOR $\tau = 0, \ldots, 2^{J-j} - 1$
5. IF $\mathcal{J}(\mathrm{band}(j,\tau)) \leq \mathcal{J}(\mathrm{BB}(j-1, 2\tau) \cup \mathrm{BB}(j-1, 2\tau+1))$
6. $\mathrm{BB}(j,\tau) = \mathrm{band}(j,\tau)$
7. ELSE $\mathrm{BB}(j,\tau) = \mathrm{BB}(j-1, 2\tau) \cup \mathrm{BB}(j-1, 2\tau+1)$
8. END
9. END

Figure 5.11: **Best basis algorithm.**

coefficients from the best basis algorithm. Step 1 performs the WPT to some prespecified level j_o as described previously. Step 2 then initializes a current best basis or best set of bands. Initially, the best set of bands (BB) is simply all the bands at level j_o in the WPT. Steps 3 to 9 then begins to compare the cost measure of the parent nodes against the current best of bands which are descendants of the parent node being examined.

Consider finding the best basis for some signal $\mathbf{x} = (x_0, x_1, \ldots, x_7)^T$. Once the wavelet packet transform has been calculated, the next step of the best basis algorithm is to calculate the criterion measurement for each of the nodes in the wavelet packet transform. This is done for some task specific criterion. The criterion measurements for each of the nodes is shown in Figure 5.12, so that $\mathcal{J}(\mathrm{band}(1,3) = 21)$. The best basis is also highlighted in Figure 5.12 for some criterion function which is to be minimized. For this example, $j_0 = 0$ since the WPT transform is performed to the lowest level. We now describe how the best set of bands is formed by working our way up the tree, comparing descendant and parent nodes.

When $j = 1$:

BB(1,0)= { band(1,0) } since $6 < 5 + 4$,

BB(1,1)= { band(0,3), band(0,4) } since $21 > 7 + 11$

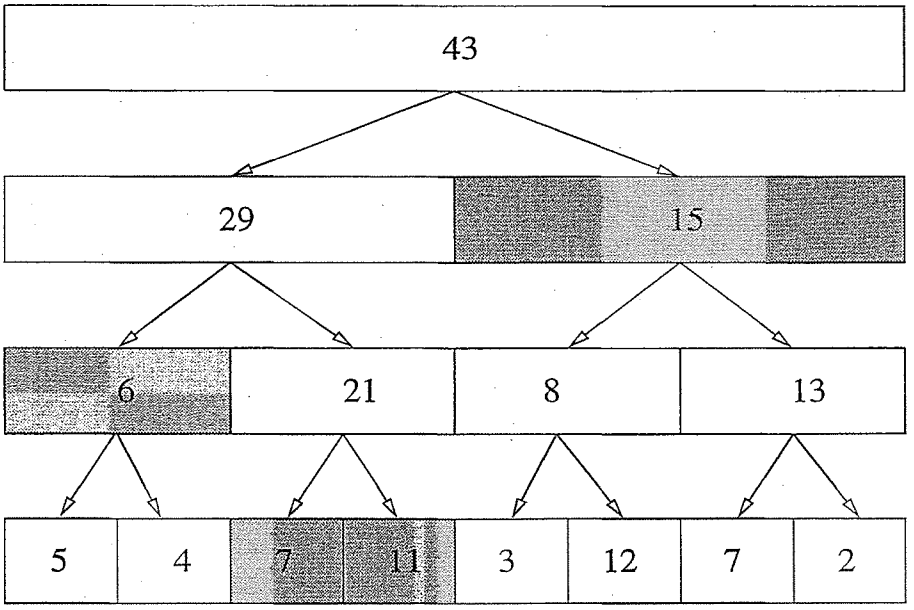


Figure 5.12: **Best basis.**

BB(1,2)= { band(1,3) } since $8 < 3 + 12$

BB(1,3)= { band(0,6), band(0,7) } since $13 > 7 + 2$.

When $j = 2$:

BB(2,0) = { band(1,0), band(0,2), band(0,3) } since $29 > 6 + 7 + 11$

BB(2,1) = { band(2,1) } since $15 < 8 + 7 + 2$.

When $j = 3$:

BB(3,0)= { band(1,0), band(0,2), band(0,3), band(2,1) } since $43 > 6 + 7 + 11 + 15$.

Saito and Coifman [118] use the best basis algorithm to determine a set of wavelet packet coefficients which are used as input to Fisher's linear discriminant analysis. This procedure is referred to as the local discriminant basis algorithm.

5.14.2 The Local Discriminant Basis Algorithm

The local discriminant bases algorithm of Saito and Coifman [118] extends the principles of the best basis algorithm [20] to allow for the classification of digitized data. There are several steps involved for selecting the wavelet packet coefficients which are to be used as input to the particular classification procedure.

For each object $\mathbf{x}_i$, the wavelet packet decomposition is performed to some level j_o. Before the best basis algorithm is applied, Saito and Coifman [118] calculate what they

refer to as an 'energy map'. This is done for each class $r = 1, 2, \ldots, R$. The energy maps have the same structure as the wavelet packet transform, hence the same indices will be used to locate items within the energy map (or tree). If $\mathbf{e}_{(r)}^{[j]}(\tau)$ denotes the energy coefficients in band(j, τ) of the energy map for class r, then,

$$\mathbf{e}_{(r)}^{[j]}(\tau) = \frac{\text{diag}\left(\sum_{i=1}^{n_r} {}^{o}\mathbf{x}_{i(r)}^{[j]}(\tau) \quad \left({}^{o}\mathbf{x}_{i(r)}^{[j]}(\tau)\right)^{T}\right)}{\text{const}}.$$

This represents the sum of squares of the coefficients which occur in the same position of the wavelet packet tree divided by a normalization constant. The energy maps were obtained from the data objects which belong to class r. The notation ${}^{o}\mathbf{x}_{i(r)}^{[j]}(\tau)$ are the wavelet packet coefficients band(j, τ) of the WPT produced from the object vector $\mathbf{x}_{i(r)}$.

Once the energy maps have been constructed, one can then begin to find the wavelet packet coefficients which correspond to the best basis. Saito and Coifman describe three criterion functions which can be used to find the best basis. These criterion functions are based on entropy and can be used to represent how differently vectors from different classes are distributed (see also Section 6.3.1). The criterion assigns a discriminatory measure to

$$\mho(j, \tau) = \mathcal{J}\left(\mathbf{e}_{(1)}^{[j]}(\tau), \ldots, \mathbf{e}_{(R)}^{[j]}(\tau)\right)$$

to each node or band in the wavelet packet transform. The wavelet packet coefficients which correspond to the best basis for discrimination give the optimum measure of $\mho$ across the entire tree.

Note that in many cases it is necessary to choose a subset of the wavelet packet coefficients, since the number of wavelet packet coefficients corresponding to the best basis is still equal to the dimensionality of the original data vector. Saito and Coifman mention that one way of selecting a subset might be to select the wavelet packet coefficients (from the best bands) which have the largest ratio of the between-groups variance to the within-groups variance as described in Section 4.1.1. Alternatively, one could select the wavelet packet coefficients based on the entropy criteria.

Chapter 6

Adaptive Wavelets

6.1 Introduction

There exists an abundant variety of wavelets which are defined by their respective filter coefficients. These are readily available for the situation when $m = 2$, and include for example the Daubechies wavelets, Coiflets, Symlets and the Meyer and Haar wavelets. The fundamental problem to overcome is deciding which set (or family) of filter coefficients will produce the best results for a particular application. In practice, several families of filter coefficients may be trialled, and the family which produces the most desirable results is used. It can be advantageous however, to design your own task specific filter coefficients rather than using a predefined set.

In this chapter, it is demonstrated how the filter coefficients can be designed to suit almost any general application. The goal is to design the wavelet matrix A which optimizes some specified modelling criterion relevant to a given multivariate prediction model, such as regression or discriminant analysis. Instead of optimizing over each of the $m \times N_f$ elements in $\boldsymbol{A}$, we make use of the factorized form [137] of a wavelet matrix and the conditions placed therein to reduce the number of parameters to be optimized. Since the filter coefficients gradually adapt to the application at hand, the procedure for designing the task specific filter coefficients is referred to as the adaptive wavelet algorithm (AWA). The adaptive wavelet algorithm forms part of an integrated feature extraction procedure since the features are repeatedly updated so they conform better to some multivariate statistical procedure.

Previous applications involving the optimization of wavelets include the work performed by Telfer *et. al.* [133] and Szu *et. al.* [130]. Telfer *et. al.* [133] consider optimizing the shift and dilation parameters of the discretization of a chosen wavelet transform, while Szu *et. al.* [130] sought the optimal linear combination of predefined wavelet bases for the classification of speech signals. In both papers the wavelet features are updated by adaptively computing the wavelet parameters and shape. This is a form of integrated feature extraction which also makes use of neural networks. Sweldens [129] also considers a lifting scheme for constructing biorthogonal second generation wavelets. Our method is made distinct because the wavelet is designed from its humble beginnings. It also allows for the general m-band wavelet transform to be utilized, as well as the more common 2-band wavelet transform.

Since the number of coefficients in the DWT is equal to the number of wavelengths in the original spectra, it is necessary to select a subset of wavelet coefficients. In our implementation, a single band of coefficients at some level in the DWT is selected. The band of coefficients produced for each spectrum are then supplied to the statistical procedure. The modelling criterion for optimizing the wavelet matrix is also based on the same coefficients.

We now consider in more detail the factorized form of a wavelet matrix, and show that A can be constructed from some set of normalized vectors, denoted by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q$, and $\boldsymbol{v}$.

6.2 Factorization of Wavelet Matrices

Recall from Section 5.11, that the wavelet matrix A can be partitioned into $m \times m$ submatrices as follows $A = (A_0, A_1, \ldots, A_q)$. Provided that the orthogonality condition: $\sum_k A_k A_{k+i}^T = \delta_{0i} I$ is satisfied, the wavelet matrix can also be written in the factorized form [137]

$$A = Q \square F_1 \square \cdots \square F_q. \tag{6.1}$$

The symbol $\square$ denotes the "polynomial product" which is defined by

$$(B_0\ B_1\ \ldots\ B_{p-1}) \square (C_0\ C_1\ \ldots\ C_{s-1}) = (G_0\ G_1\ \ldots\ G_{p+s-2})$$

with

$$G_i = \sum_k B_k C_{i-k}.$$

The factors

$$F_i = (R_i \; I - R_i) \tag{6.2}$$

where R_i is a projection matrix and $Q = \sum_i A_i$ is an orthogonal matrix.

If for example, $m = 3$ and $q = 2$ then $A = (A_0 \; A_1 \; A_2)$ with each A_j having dimension 4×4 thus, A has size $m \times [m(q+1)] = 3 \times 9$. Assuming the orthogonality condition is satisfied then

$$\begin{aligned} A &= Q \square F_1 \square F_2 \\ &= Q \square (R_1 \; I - R_1) \square (R_2 \; I - R_2) \\ &= [Q R_1 R_2 \;\; Q(R_1 - 2R_1R_2 + R_2) \;\; Q(I - R_1)(I - R_2)]. \end{aligned}$$

Essentially, we strive for representations of Q and each projection matrix R_i (for $i = 1, \ldots, q$). First consider the representation of Q.

The regularity condition $\sum_k \ell_k = \sqrt{m}$, places a constraint on the first row of Q. The regularity condition is equivalent to setting the first row of Q to $1/\sqrt{m} \; 1_m^T$ where 1_m denotes a $m \times 1$ column vector of ones. The remaining $m-1$ rows are calculated ensuring the orthogonality of Q is maintained. This is satisfied if the last $m-1$ rows are calculated by $(I - 2vv^T)T \odot D$ where v is a normalized vector, T is an upper triangular matrix with diagonal elements $T_{ii} = i - m$ and off-diagonal elements equal to 1. The symbol $\odot$ indicates a form of element by element scalar multiplication across two matrices such that $B \odot C = G \rightarrow B_{ij}C_{ij} = G_{ij}$. This scalar product of T with some matrix D normalizes the rows of T. The $m \times m$ orthogonal matrix Q is partitioned as follows,

$$Q = \begin{pmatrix} 1/\sqrt{m} \; 1_m^T \\ \\ (I - 2vv^T)T \odot D \end{pmatrix} \tag{6.3}$$

Now consider the projection matrices. A symmetric projection matrix of rank ϱ can be written $R = UU^T$ where $U_{m \times \varrho}$ is a matrix with orthonormal columns. For the wavelet matrix to be non-redundant, the ranks of the projection matrices must form a monotonically increasing sequence [137], that is the $\text{rank}(R_1) \leq \text{rank}(R_2) \leq \cdots \leq \text{rank}(R_q)$. For simplicity, we restrict the ranks of each projector matrix to be 1, and so,

$$R_i = u_i u_i^T \tag{6.4}$$

where $u_i^T u_i = 1$.

The following example illustrates how $\boldsymbol{A}$ with $m = 3$ and $q = 2$ can be constructed. The example begins by defining the column vector $\boldsymbol{v}$ of length $m - 1$ and two columns vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ both of length m. Let

$$\begin{aligned} v &= (-0.7918, -0.6107)^T \\ u_1 &= (-0.3873, -0.9097, 0.1497)^T \\ u_2 &= (-0.9062, 0.1674, 0.3884)^T \end{aligned}$$

First, consider calculating the symmetric projectors $\boldsymbol{R}_1 = \boldsymbol{u}_1\boldsymbol{u}_1^T$ and $\boldsymbol{R}_2 = \boldsymbol{u}_2\boldsymbol{u}_2^T$.

$$\boldsymbol{R}_1 = \begin{pmatrix} 0.1500 & 0.3523 & -0.0580 \\ 0.3523 & 0.8276 & -0.1362 \\ -0.0580 & -0.1362 & 0.0224 \end{pmatrix} \text{ and } \boldsymbol{R}_2 = \begin{pmatrix} 0.8212 & -0.1517 & -0.3520 \\ -0.1517 & 0.0280 & 0.0650 \\ -0.3520 & 0.0650 & 0.1509 \end{pmatrix}$$

Now consider calculating $\boldsymbol{Q}$. The first row of $\boldsymbol{Q}$ is $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$, and the remaining two rows are calculated by $(\boldsymbol{I} - 2\boldsymbol{v}\boldsymbol{v}^T)(\boldsymbol{T} \odot \boldsymbol{D})$ where

$$\begin{aligned} \boldsymbol{T} \odot \boldsymbol{D} &= \begin{pmatrix} -2 & 1 & 1 \\ 0 & -1 & 1 \end{pmatrix} \odot \begin{pmatrix} 1/\sqrt{(6)} & 1/\sqrt{(6)} & 1/\sqrt{(6)} \\ 1/\sqrt{(2)} & 1/\sqrt{(2)} & 1/\sqrt{(2)} \end{pmatrix} \\ &= \begin{pmatrix} -0.8165 & 0.4802 & 0.4802 \\ 0 & -0.7071 & 0.7071 \end{pmatrix} \end{aligned}$$

$$\boldsymbol{I} - 2\boldsymbol{v}\boldsymbol{v}^T = \begin{pmatrix} -0.2539 & -0.9671 \\ -0.9671 & 0.2541 \end{pmatrix}$$

which together give

$$\boldsymbol{Q} = \begin{pmatrix} 0.5774 & 0.5774 & 0.5774 \\ 0.2073 & 0.5802 & -0.7875 \\ 0.7896 & -0.5745 & -0.2151 \end{pmatrix}$$

Now consider forming the wavelet matrix $\boldsymbol{A}$. Using the factorized form of the wavelet matrix one has

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{Q}\square\boldsymbol{F}_1\square\boldsymbol{F}_2 \\ &= \boldsymbol{Q}\square(\boldsymbol{R}_1 \ \boldsymbol{I} - \boldsymbol{R}_1)\square(\boldsymbol{R}_2 \ \boldsymbol{I} - \boldsymbol{R}_2) \\ &= [\boldsymbol{Q}\boldsymbol{R}_1\boldsymbol{R}_2 \ \ \boldsymbol{Q}(\boldsymbol{R}_1 - 2\boldsymbol{R}_1\boldsymbol{R}_2 + \boldsymbol{R}_2) \ \ \boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{R}_1)(\boldsymbol{I} - \boldsymbol{R}_2)]. \end{aligned}$$

then substituting for $\boldsymbol{Q}$, $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ one arrives at the following result for $\boldsymbol{A}$.

$$\boldsymbol{A} = \begin{pmatrix} 0.1542 & -0.0285 & -0.0661 & 0.1316 & 0.6257 & -0.0456 & 0.2917 & -0.0198 & 0.6891 \\ 0.1690 & -0.0312 & -0.0724 & 0.3027 & 0.6566 & -0.1179 & -0.2643 & -0.0451 & -0.5972 \\ -0.0430 & 0.0079 & 0.0184 & 0.8258 & -0.3336 & -0.3569 & 0.0069 & -0.2488 & 0.1234 \end{pmatrix}$$

or

$$A_0 = \begin{pmatrix} 0.1542 & -0.0285 & -0.0661 \\ 0.1690 & -0.0312 & -0.0724 \\ -0.0430 & 0.0079 & 0.0184 \end{pmatrix}$$

$$A_1 = \begin{pmatrix} 0.1316 & 0.6257 & -0.0456 \\ 0.3027 & 0.6566 & -0.1179 \\ 0.8258 & -0.3336 & -0.3569 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 0.2917 & -0.0198 & 0.6891 \\ -0.2643 & -0.0451 & -0.5972 \\ 0.0069 & -0.2488 & 0.1234 \end{pmatrix}$$

We have now discussed how A can be constructed from the normalized vectors $u_1, \ldots, u_q$ and v. Initially, $u_1, \ldots, u_q$ and v are randomly assigned elements from the uniform distribution. The optimization routine then proceeds to update the elements of these vectors so that some modelling criterion can be optimized. We describe the different modelling criteria for discriminant and regression analysis in Section 6.3.1 and 6.3.2, respectively.

6.3 Criteria Measures for Optimization

The adaptive wavelet algorithm can be used for a variety of situations, and its goal is reflected by the particular criterion which is to be optimized. In this thesis, we apply the filter coefficients produced from the adaptive wavelet algorithm for discriminant analysis and regression analysis. It was stated earlier, that the dimensionality is reduced by selecting some band(j_o, τ_o) of wavelet coefficients from the discrete wavelet transform. It then follows that the criterion function $\mathcal{J}$ will be based on the same coefficients i.e. $\boldsymbol{X}^{[j_o]}(\tau_o)$. Some suitable criterion functions which are to optimized for the various statistical procedures are discussed next.

6.3.1 Discriminant Criterion Functions

If the filter coefficients are to be used for discriminatory purposes, then the criterion function (which is referred to as a discriminant criterion function) should strive to reflect differences among classes. In this section three suitable discriminant criterion functions are described. These discriminant criterion functions are Wilk's lambda ($\mathcal{J}_\Lambda$), entropy ($\mathcal{J}_E$), and the cross-validated quadratic probability measure ($\mathcal{J}_{\text{cvqpm}}$).

Wilks Lambda

The Wilks' Λ criterion can be used to test the significance of the differences between group centroids [132]. A smaller value for Λ is preferred since this indicates a larger significance.

Wilks' Λ is the ratio of the determinant of the within covariance matrix to the determinant of the total covariance matrix and is defined to be

$$\begin{aligned}\Lambda &= \frac{|\, S_W \,|}{|\, \mathrm{S}_T \,|} \\ &= \frac{|\, S_W \,|}{|\, S_B + S_W \,|}\end{aligned}$$

where the total covariance matrix $\mathrm{S}_T = S_B + S_W$ is the sum of the between (S_B) and within (S_W) covariance matrix.

Entropy

Saito and Coifman [118] discuss a cross entropy measure which can be used to measure how differently vectors are distributed. Let $\zeta_{(1)}$ and $\zeta_{(2)}$ be vectors from classes 1 and 2 respectively. If the elements in $\zeta_{(1)}$ and $\zeta_{(2)}$ are nonnegative and sum to unity, then cross entropy is defined by

$$E_{\text{cross}}(\zeta_{(1)}, \zeta_{(2)}) = \sum_{i=1}^{p} \zeta_{i(1)} \log \frac{\zeta_{i(1)}}{\zeta_{i(2)}} \tag{6.5}$$

where $p = \text{length}(\zeta_{(1)})$ =length($\zeta_{(2)}$), i.e. dimensionality of vectors. Equation 6.5 is not symmetric, that is the measure of discrepancy for $E_{\text{cross}}(\zeta_{(1)}, \zeta_{(2)})$, will be different to that for $E_{\text{cross}}(\zeta_{(2)}, \zeta_{(1)})$. For our purposes we prefer to use a symmetric criterion which is defined in [118] as

$$E_{\text{sym}}(\zeta_{(1)}, \zeta_{(2)}) = E_{\text{cross}}(\zeta_{(1)}, \zeta_{(2)}) + E_{\text{cross}}(\zeta_{(2)}, \zeta_{(1)}).$$

Measuring the distinctness of several vectors from different classes, involves calculating E_{sym} for each combination of vectors. Call this entropy measure the total entropy E_{tot}. For example, the total symmetric entropy for $\zeta_{(1)}, \zeta_{(2)}$ and $\zeta_{(3)}$ is calculated as follows

$$E_{\text{tot}}(\zeta_{(1)}, \zeta_{(2)}, \zeta_{(3)}) = E_{\text{sym}}(\zeta_{(1)}, \zeta_{(2)}) + E_{\text{sym}}(\zeta_{(1)}, \zeta_{(3)}) + E_{\text{sym}}(\zeta_{(2)}, \zeta_{(3)}).$$

It is necessary to construct a single vector which in some way is representative of the classes, this could for instance be a mean vector. In Saito and Coifman [118], the representative vector from each class is an energy vector. More specifically, define the class energy vector of the wavelet coefficients from band(j, τ) as

$$\mathrm{e}_{(r)}^{[j]}(\tau) = \frac{\text{diag}\left(\mathrm{X}_{(r)}^{[j]}(\tau)\right)\left(\mathrm{X}_{(r)}^{[j]}(\tau)\right)^T}{\text{const}} \qquad r = 1, \ldots, R$$

and if the wavelet packet coefficients are being used then

$$\mathrm{e}_{(r)}^{[j]}(\tau) = \frac{\mathrm{diag}\left({}^{\circ}\mathrm{X}_{(r)}^{[j]}(\tau)\right)\left({}^{\circ}\mathrm{X}_{(r)}^{[j]}(\tau)\right)^{T}}{\mathrm{const}} \qquad r = 1, \ldots, R.$$

The denominator is a normalization constant. The numerator is simply the sum of squares of the wavelet coefficients from either the DWT or WPT which occur in the same position of the wavelet trees, where the DWT or WPT has been performed for objects belonging to the same class. The discriminatory criterion function is then

$$\begin{aligned} \mathcal{J}_E\left(\mathrm{e}_{(r)}^{[j]}(\tau)\right) &= E_{\mathrm{tot}}\left(\mathrm{e}_{(1)}^{[j]}(\tau), \ldots, \mathrm{e}_{(R)}^{[j]}(\tau)\right) \\ &= \sum_{l}\sum_{r:\, r\neq l} E_{\mathrm{sym}}(\mathrm{e}_{(l)}^{[j]}(\tau), \mathrm{e}_{(r)}^{[j]}(\tau)) \end{aligned}$$

Cross-Validated Quadratic Probability Measure (CVQPM)

The cross-validated quadratic probability measure (CVQPM) assesses the trustworthiness of the class predictions made by the discriminant model. The CVQPM ranges from 0 to 1. Ideally, larger values of the CVQPM are preferred, since this implies the classes can be differentiated with a higher degree of certainty. The CVQPM was previously discussed in greater detail in Section 2.8. The CVQPM criterion function based on a band of coefficients $\mathbf{X}^{[j]}(\tau)$ would be defined as follows

$$\mathcal{J}_{\mathrm{CVQPM}}\left(\boldsymbol{X}^{[j]}(\tau)\right) = \frac{1}{n}\sum_{i=1}^{n} a_Q(\mathrm{x}_i^{[j]}(\tau), -i).$$

where

$$a_Q(\mathrm{x}_i^{[j]}(\tau), -i) = \frac{1}{2} + P_{-i}\left(r \mid \mathrm{x}_{i(r)}^{[j]}(\tau)\right) - \frac{1}{2}\sum_{r=1}^{R} P_{-i}\left(r \mid \mathrm{x}_i^{[j]}(\tau)\right)^2.$$

6.3.2 Regression Criterion Functions

A suitable criterion function for regression analysis should reflect how well the response values are predicted. In the adaptive wavelet algorithm, the criterion function considered for regression is based on the PRESS statistic and is then converted to a cross-validated R-squared measure as discussed in Section 3.6.

Cross-Validated R-Squared

The cross-validated R-squared criterion function is defined as

$$\mathcal{J}_{\text{CVRSQ}}\left(\mathbf{X}^{[j]}(\tau)\right) = 1 - \text{PRESS/TSS}$$

where the TSS and the PRESS statistic are calculated by

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

and

$$\text{PRESS} = \sum_{i=1}^{n}(y_i - \hat{y}_{-i})^2,$$

respectively. The actual regression model used for predicting the response is

$$\hat{\mathbf{y}} = \left(\mathbf{X}^{[j]}(\tau)\right)^T \mathbf{b}.$$

6.4 The Adaptive Wavelet Algorithm

The algorithm shown in Figure 6.1 summarizes the adaptive wavelet algorithm. Step 1 of the algorithm sets values for the parameters m, q, j_o and τ_o and Step 2 initializes $\boldsymbol{v}$ and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q$. Steps 3-6 go about constructing the filter coefficient matrix $\boldsymbol{A}$, so that the m-band DWT can be performed based on the filter coefficients in $\boldsymbol{A}$. This is done in Step 7 to level j_o. The coefficients $\mathbf{X}^{[j_o]}(\tau_o)$ are then extracted in Step 8, and the multivariate modelling criterion $\mathcal{J}\left(\mathbf{X}^{[j_o]}(\tau_o)\right)$ is calculated for the extracted data. Step 9 assesses if the stopping criterion of the algorithm has been reached. The stopping criterion are discussed further at the end of this section. If the stopping criterion has not been reached, then the parameters $\boldsymbol{v}$ and $\{\boldsymbol{u}_i\}_{i=1}^{q}$ are updated and the algorithm proceeds to Step 3. If some stopping criterion has been reached, then the algorithm proceeds to Step 10 where the Lawton matrix condition is verified. Provided Conditions 1 and 2 of Section 5.12 hold, then the Lawton matrix condition will not be satisfied for exceptional degenerate cases, thus the Lawton matrix is verified after the adaptive wavelet has been found. Finally, the multivariate statistical procedure can be performed using the coefficients $\mathbf{X}^{[j_o]}(\tau_o)$. The optimizer used in the adaptive wavelet algorithm is the unconstrained MATLAB optimizer [4], for which the default algorithm is the quasi-Newton method which alsc incorporates a mixed quadratic and cubic line search procedure.

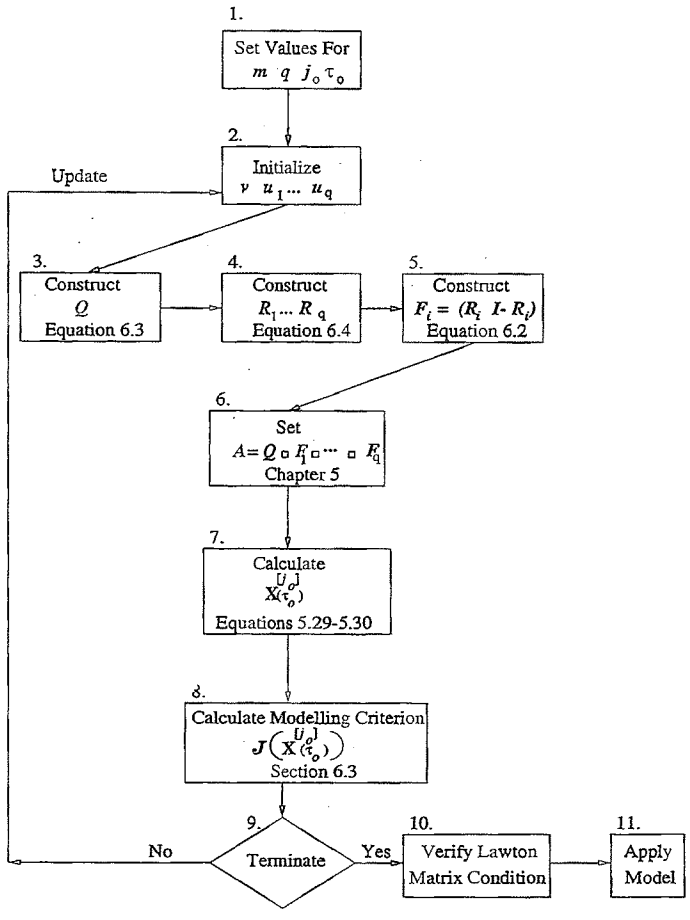


Figure 6.1: **The adaptive wavelet algorithm.**

Before applying the adaptive wavelet algorithm, values for m, q, j_o and τ_o need to be specified. There is no empirical rule for determining these parameters. In fact, the only way to know which values will be the best is to try all of them. To reduce the labour of this intensive task, some heuristics for choosing appropriate parameter values can be suggested.

First consider some heuristics for choosing the values m and q. The value of m determines the number of bands in the DWT and the downsampling factor, so m is chosen such that $p/m^{(J-j_o+1)}$ is an integer value. Since m combines with q to determine the number of the filter coefficients ($N_f = m(q+1)$) another constraint is placed on m so that N_f does not become too large. Similarly, a constraint is placed on q for the same reason. It is preferred that the number of filter coefficients be less than 25. The analyses which follow in proceeding chapters typically use 12 or 16 filter coefficients, since experimentation has

revealed this to be adequate.

Now, consider selecting values for j_o and τ_o. These parameters simultaneously determine the band(j_o, τ_o) and hence the coefficients $\boldsymbol{X}^{[j_o]}(\tau_o)$ for which optimization of the discriminant criterion is based. The coefficients $\boldsymbol{X}^{[j_o]}(\tau_o)$ are later used as inputs to the multivariate statistical method.

The value for j_o determines the level of the DWT that the spectra are to be decomposed. A value for j_o should be chosen such that $p/m^{(J-j_o+1)}$ which is the number of coefficients in band(j_o, τ_o), is suitable (not too large) for classification. Each of the appropriate values of j_o should be tested. To perform this task, a value for τ_o is also required. To ensure the best j_o and τ combination, each of the appropriate values of j_o should be individually tested with each value of τ_0. To reduce this computational burden, we have chosen to select τ_o as the band which gives the largest $\mathcal{J}\left(\boldsymbol{X}^{[j_o]}(\tau_o)\right)$ at initialization. It is recommended that if one suspects the basic shape of the data will be useful for classification, then optimization over the scaling band may prove worthwhile.

The discussion so far has not eluded to the various criterion which can be used for deciding when the adaptive wavelet algorithm should cease updating the parameters $v, u_1, \ldots, u_q$. Based on tolerance settings which control the convergence of the algorithm, the algorithm may halt when the optimal value for the modelling criterion has been achieved or when a preset number of iterations of the optimization routine has been reached – which ever occurs sooner. Of course stopping the algorithm after a prespecified number of iterations does not ensure an optimal value will be produced, but does assist in the practical experimentation of the model.

When searching for optimal values, there is always the issue of whether or not a global or local optimal solution has been found. Unless the problem is continuous and has only one optimal point, there can be no guarantee that a global optimal value has been found. It is suggested in [4] that starting the optimization routine from different starting values may assist in overcoming this problem.

6.5 Example

To obtain a better understanding of the adaptive wavelet algorithm, we apply its concepts to a spectral data set. The goal is to assign the spectra to one of several predefined

categories. The adaptive wavelet is then designed for a discriminant analysis task. In this example the classifier used is Bayesian linear discriminant analysis. The training spectral data [118] contains 20 spectra in each of five classes and the test set also contains 20 spectra per class. The dimensionality p of the data (i.e. number of variables) is 512. The five classes represent different kinds of minerals, and this data set is subsequently referred to as the mineralogical data set and is discussed in greater detail in Section 7.2.2.

In this example, the parameters m, q and j_o were set at 4,3 and 3, respectively. Optimization was based on the coefficients $\mathbf{X}^{[3]}(\tau)$ which gave the maximum $\mathcal{J}\left(\mathbf{X}^{[3]}(\tau)\right)$ at initialization where $\tau \in \{0,1,2,3\}$. Three discriminant criterion functions were considered, these were $\mathcal{J}_\Lambda$, $\mathcal{J}_\mathcal{E}$ and $\mathcal{J}_{\text{cvqpm}}$. The results for each of the criterion functions are displayed in Tables 6.1. Here the classification rates of the individual bands at initialization and at completion of the algorithm are shown. Note that the same starting parameters for $\boldsymbol{v}$, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ have been used for the implementation involving the different modelling criteria, hence the same results occur at initialization for each of the criterion functions $\mathcal{J}_\Lambda$, $\mathcal{J}_\mathcal{E}$ and $\mathcal{J}_{\text{cvqpm}}$. The asterisk indicates which band optimization was based upon.

	τ	0	1	2	3	$\mathcal{J}$
Initialization	Train	97	96	97	97	$\mathcal{J}_\Lambda$,$\mathcal{J}_\mathcal{E}$, $\mathcal{J}_{\text{cvqpm}}$
	Test	90	90	91	88	
Termination	Train	98	96	95	100*	$\mathcal{J}_\Lambda$
	Test	91	89	88	90*	
Termination	Train	97	94	94*	97	$\mathcal{J}_\mathcal{E}$
	Test	86	89	90*	87	
Termination	Train	100*	98	96	95	$\mathcal{J}_{\text{cvqpm}}$
	Test	96*	92	89	87	

Table 6.1: **The percentage of correctly classified spectra, using the coefficients $\{\mathbf{X}^{[3]}(\tau)\}$ for $\tau = 0,\ldots,3$ at initialization and at termination of the adaptive wavelet algorithm. The discriminant criterion functions were Wilk's Lambda, symmetric entropy and the CVQPM.**

For the Wilk's Lambda criterion, optimization was based on band(3,3), while the entropy criterion optimized over band(3,2). The CVQPM criterion optimized over the scaling band(3,0). Some features which we might expect from the adaptive wavelet algorithm, is that at termination, the band which optimization was based would outperform the other bands, at least in terms of the percentage of correctly classified training objects. This is

the case with the CVQPM and Λ criterion, but is not so, for the symmetric entropy criterion. In defence however, band(3,2) for the symmetric entropy does produce the largest percentage of correctly classified objects for the testing data, and has not overfitted as significantly to the training data as perhaps the Λ criterion function. Overall, for the results presented in Table 6.1, the CVQPM seems to be performing most adequately. It is the only criterion function which has improved the test classification rate from those obtained at initialization. One reason for the success of the CVQPM, maybe due to the fact that optimization and hence classification is based on scaling coefficients. Since one can observe from Figure 7.2 which shows some sample spectra of the mineral data, that perhaps information about the basic shape of the data might be potentially useful. For this reason, the optimization routine using the Wilk's Lambda, and symmetric criterion functions was repeated, this time forcing optimization over the scaling band. These results are summarized in Table 6.2, where for ease of comparison, we have reproduced the same results from Table 6.1 for the percentage of correctly classified objects for $\mathcal{J}_{\text{cvqpm}}$.

	$\tau =$	0	1	2	3	$\mathcal{J}$
Initialization	Train	97	96	97	97	$\mathcal{J}_\Lambda$,$\mathcal{J}_E$, $\mathcal{J}_{\text{cvqpm}}$
	Test	90	90	91	88	
Termination	Train	100*	95	96	96	$\mathcal{J}_\Lambda$
	Test	91*	89	86	90	
Termination	Train	96*	94	85	91	$\mathcal{J}_E$
	Test	92*	90	76	87	
Termination	Train	100*	98	96	95	$\mathcal{J}_{\text{cvqpm}}$
	Test	96*	92	89	87	

Table 6.2: **The percentage of correctly classified spectra, using the coefficients $\{\mathbf{X}^{[3]}(\tau)\}$ for $\tau = 0, \ldots, 3$ at initialization and at termination of the adaptive wavelet algorithm. Optimization was based on $\{\mathbf{X}^{[3]}(0)\}$ and the discriminant criterion functions were Wilk's Lambda, symmetric entropy and the CVQPM.**

Optimization over the scaling band did improve the results slightly for the Wilk's Lambda and symmetric entropy criterion, but these criterion functions were not able to improve upon the results previously obtained with the CVQPM criterion function.

Chapter 7

Classification Applications

7.1 Overview

In this chapter, different strategies are investigated for classifying spectral data. A strategy may refer to the particular classifier utilized such as Bayesian linear discriminant analysis, or a feature extraction technique. A strategy may even refer to the combination of feature extraction techniques with a particular classifier.

As an initial step to a discriminant analysis, one would generally experiment with the original variables, and then perhaps try other kinds of features. In this chapter we initially supply the original data to the classifiers and then investigate the performance of the discriminant techniques using wavelet coefficients as features. We use standard wavelet filter coefficients from the Daubechies family and filter coefficients which are derived from the adaptive wavelet algorithm (AWA).

It should be mentioned that the goal of this chapter is not necessarily to find the best discriminant model. Rather, we would like to investigate the effect of wavelet coefficients when used as features for discriminant techniques, as opposed to the original variables. The application of the AWA involves the use of an integrated feature extraction method. Whilst much emphasis will be placed on numerical measures which reflect the assignment accuracy of the discriminant strategies, we will also qualitatively assess if wavelet coefficients can help us understand more about the group structure of the data as well as regions which may contain useful discriminatory information.

Data Set		Class 1	Class 2	Class 3	Class 4	Class 5	Total
Seagrass	Train	55	55	55	-	-	165
	Test	34	34	34	-	-	102
Mineral	Train	20	20	20	20	20	100
	Test	20	20	20	20	20	100
Paraxylene	Train	25	25	25	-	-	75
	Test	25	25	25	-	-	75
Butanol	Train	21	27	-	-	-	48
	Test	21	26	-	-	-	47

Table 7.1: **Description of the spectral data sets used for classification.**

7.2 The Data Sets

Four spectral data sets will be used for investigating the various classification procedures. Each data set initially contains 512 variables (i.e. $p = 512$). The data sets will be referred to as the seagrass (s), mineral (m), paraxylene (p) and butanol (b) data. The number of training and testing spectra in the group categories is listed in Table 7.1 for each set of data. A further description about the data is now presented.

7.2.1 Seagrass Data

The training seagrass data set contains 165 digitized spectra, for which $\log 1/\text{reflectance}$ was measured for the 512 wavelengths $400, 404, \ldots, 2444$ nm. The data consists of three classes of seagrass species — Halophila ovalis (class 1), a mixture of Halodule uninervis and Halodule pinifolia (class 2) and Halophila spinulosa (class 3). The training data comprises of 55 spectra in each group and the testing data has 34 spectra in each class. Figure 7.1 shows five sample spectra from each of the classes. This data is particularly relevant to environmental scientists investigating the eating habits of dugongs, a whale like mammal, whose diet constitutes a substantial proportion of seagrasses. The same data is also important for taxonomic purposes. The seagrass data was provided by Lem Aragones and Dr Bill Foley, from the Department of Zoology at James Cook University.

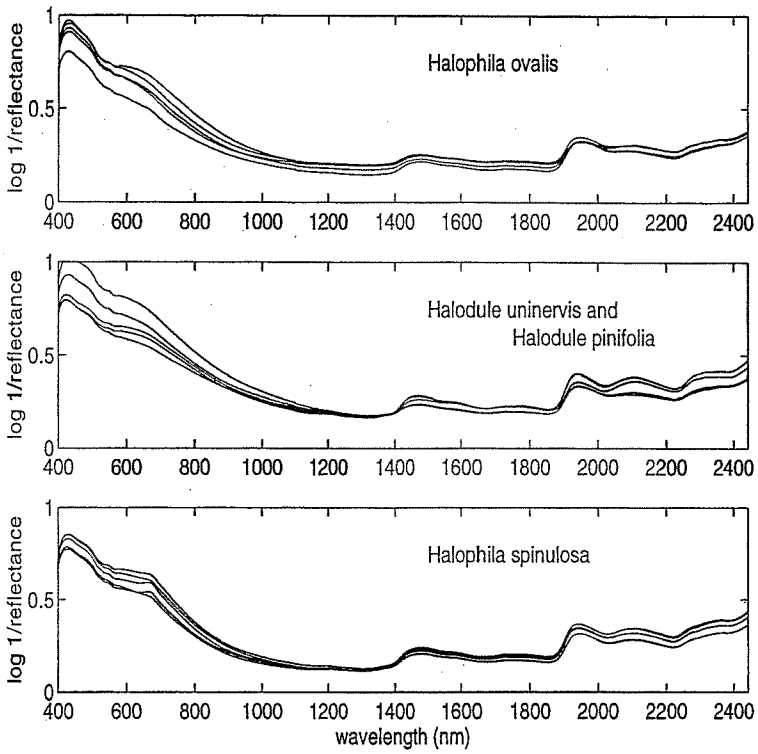


Figure 7.1: **Five sample spectra from the seagrass data.**

7.2.2 Mineral Data

The mineral data was provided by Dr Danny Aswen, from the Department of Earth Sciences at James Cook University. The mineral data set which has undergone the hull quotient transformation as described in Section 4.2.1, contains 100 digitized spectra, for which absorbance was measured at the 512 wavelengths 1478,1480,...,2500 nm. The data consists of five mineralogical groups — amphilolites (class 1), calsilicates (class 2), granite (class 3), mica (class 4) and soil (class 5). The training and testing data comprise of 20 spectra in each of the classes. Figure 7.2 shows five sample spectra from each of the classes. With the exception of the soil spectra the rock data exhibit some within variation particularly at the peaks of the spectra. Whilst it could be worthwhile to seek some transformation such as the SNV transformation which may assist in dampening the variation, we elected to leave the data in the hull quotient format only, and compare the discriminant techniques on this data as it is presented.

The development of automated classification models for the discrimination of miner-

alogical spectra is important to geologists for obvious practical and economic reasons. Experienced geologists may be able to distinguish among various minerals by observing the position and shapes of certain peaks at different wavelengths. The presence of noise and lack of experience can however, distort ones judgement. In these situations, an automatic discriminant model could be of great value to a geologist.

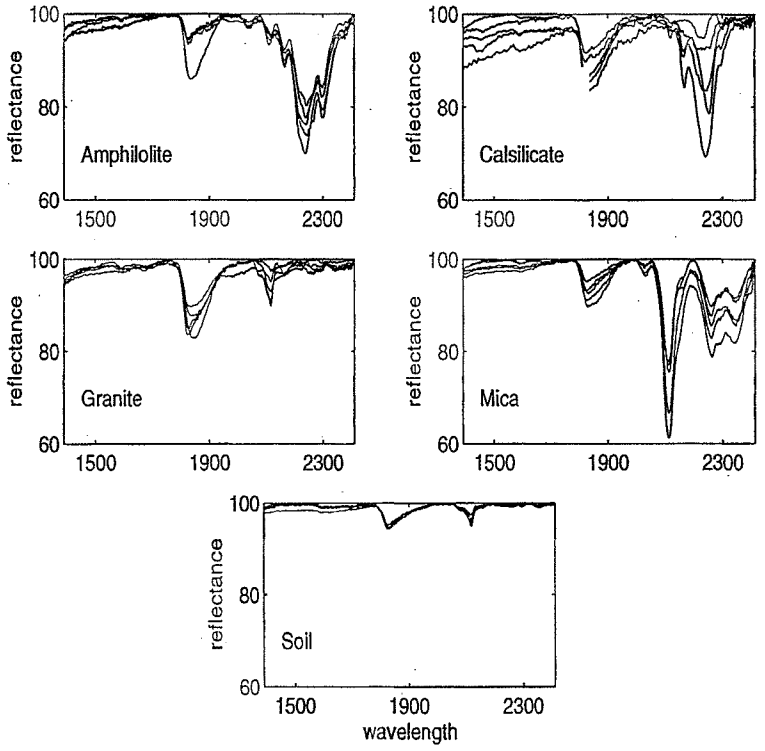


Figure 7.2: **Five sample spectra from the mineral data.**

7.2.3 Paraxylene Data

The paraxylene data was kindly provided by Professor Massart at the Pharmaceutical Institute, The Free University, Brussels. The data was produced by Dr Wim Penninckx at the same institute. The training paraxylene data set contains 75 digitized spectra, for which absorbance was measured at the 512 wavelengths 1289,1291,...,2311 nm. The data consists of three groups. Pure paraxylene (class 1), paraxylene plus 10% orthoxylene (class 2) and paraxylene plus 20% orthoxylene (class 3). The training and testing data comprise of 25 spectra in each of the classes. Figure 7.3 shows five sample spectra from each of the classes.

This data set is important for quality control procedures in pharmaceutical science. When drugs are being devised, it is possible for impurities to form in the substance. Production rates of such drugs can be increased if there are relatively quick, nondestructive techniques which can be implemented for detecting levels of impurities which have formed in substances.

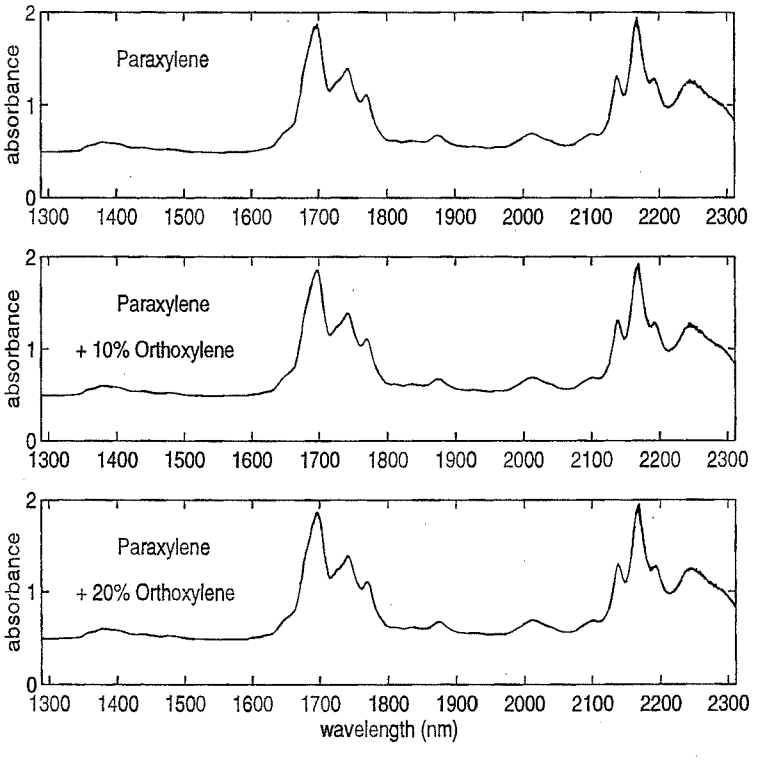


Figure 7.3: **Five sample spectra from the paraxylene data.**

7.2.4 Butanol Data

The butanol data was accessed from Professor Massart and Wu Wen at the Pharmaceutical Institute, The Free University, Brussels. The training butanol data set contains 48 digitized spectra, for which absorbance was measured at 512 wavelengths in the range of 1200 nm to 2400 nm. The data consists of two groups. Pure butanol (class 1) and butanol containing different concentrations of water (class 2). Class 1 in the training set contains 21 spectra and class 2 in the training set contains 27 spectra. Class 1 in the test set contains 21 spectra and class 2 in the test data has 26 spectra. As for the paraxylene data, this data set also relates to the detection of impurities. Figure 7.4 shows five sample

spectra from each of the classes. The exact wavelength number for each absorbance value is unavailable. For this reason the horizontal axis is labelled with wavelength indices.

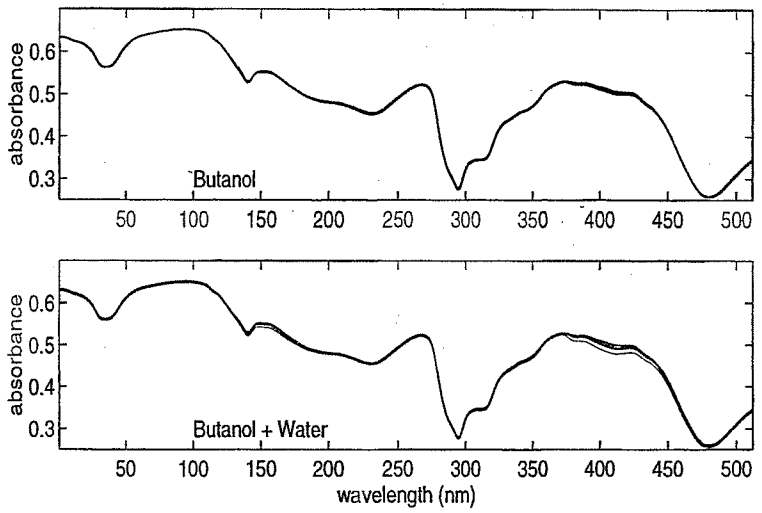


Figure 7.4: **Five sample spectra from the butanol data.**

7.3 Discriminant Analysis Based on the Original Variables

In this section, the original variables are the features which are inputted to the discriminant techniques BLDA, BQDA, FDA, PDA and RDA. For PDA and RDA, no feature selection (i.e. dimension reduction) was performed. The set of grid values representing the combination of (a, b) pairs trialled are listed below. Recall that $a \in [0, 1]$, controls the degree to which the pooled covariance matrix should be used, and $b \in [0, 1]$ determines the degree to which $\mathbf{S}_{(r)}(a)$ is shrunken toward a multiplier of the identity matrix in the RDA model.

		a				
		0.00	0.25	0.50	0.75	1.00
	0.00	(0.00,0.00)	(0.00,0.25)	(0.00,0.50)	(0.00,0.75)	(0.00,1.00)
	0.25	(0.25,0.00)	(0.25,0.25)	(0.25,0.50)	(0.25,0.75)	(0.25,1.00)
b	0.50	(0.50,0.00)	(0.50,0.25)	(0.50,0.50)	(0.50,0.75)	(0.50,1.00)
	0.75	(0.75,0.00)	(0.75,0.25)	(0.75,0.50)	(0.75,0.75)	(0.75,1.00)
	1	(1.00,0.00)	(1.00,0.25)	(1.00,0.50)	(1.00,0.75)	(1.00,1.00)

The default settings (described in the code of Hastie [60]) were used for PDA and FDA, where the regression model used in FDA was BRUTO which accommodates a variable

selection routine. The variables selected for BLDA and BQDA were based on a forward stepwise selection strategy and will be now referred to as SBLDA and SBQDA, respectively. Throughout this section, bold type setting will be used to identify the highest classification rates calculated from the testing data for the various models in a particular table.

Before presenting the results for FDA, PDA and RDA based on the original data, we would first like to explain how the stepwise procedures were implemented. Three forward stepwise strategies which will be referred to as CF1, CF2 and CF3 were applied to each of the data sets, and are described in greater detail below.

- CF1: The CF1 procedure starts with an empty subset and at each step adds the variable (or wavelength) which produces the largest increase in the correct classification rate (CCR). Since the CCR is a discrete measure, there may be instances when several variables give the same significant increase in the CCR. Should such a situation arise, then the variable (from the set of tied variables) which gives the largest quadratic probability measure (QPM) will enter the model.

- CF2: The CF2 procedure starts with an empty subset and at each step adds the variable (or wavelength) which produces the largest increase in the quadratic probability measure. Since the QPM is a continuous measure, the event of a tie is unlikely to occur. In the event of a tie you could randomly select the variable, but for convenience we chose to use the variable which had the smallest wavelength, since this is automatically done in the Matlab programming language.

- CF3: The CF3 procedure starts with an empty subset and at each step adds the variable (or wavelength) which produces the largest increase in the cross-validated quadratic probability measure. The same tie-breaking mechanism as CF2 is implemented.

The same stopping rule was used for each of the stepwise strategies CF1, CF2 and CF3. The procedures cease to enter variables into the model when one of the following stopping criterion is reached:

- The change in the correct classification rate is less than $1/n$ where n is the number of samples in the data set. That is, from one iteration to the next of the stepwise

routine, the inclusion of another variable does not improve the correct classification rate by more than $1/n$.

- The correct classification rate reaches 100%. At each iteration, both stopping criteria are checked, and if one of the stopping criteria has been met, the stepwise procedure will not enter any more variables.

Data		SBLDA			SBQDA		
		CF1	CF2	CF3	CF1	CF2	CF3
Seagrass	Train	99.39	100	99.39	100	100	100
	Test	**100**	100	100	97.06	97.06	**97.06**
	dimension	3	8	6	6	6	6
Mineral	Train	99	100	100	100	100	100
	Test	86	87	**88**	92	90	**93**
	dimension	5	5	5	3	3	3
Paraxylene	Train	98.67	100	100	100	100	100
	Test	78.67	**89.33**	87.33	**80**	68	78.67
	dimension	9	7	7	6	6	7
Butanol	Train	87.50	87.50	87.5	100	93.75	89.58
	Test	**78.22**	72.39	72.39	**86.60**	68.09	76.60
	dimension	3	3	3	7	3	4

Table 7.2: **Correct classification rates (%) for the stepwise procedures.**

Table 7.2 shows the correct classification rates the stepwise procedures. The numbers which appear in bold face identify the highest classification rates based calculated from the testing data for each for the stepwise procedures CF1, CF2 and CF3. The correct classification rates have been separately highlighted for SBLDA and SBQDA. In the event that two strategies produce the same (highest) testing classification rate, the forward method which utilizes the least number of variables is highlighted. If the strategies then have the same number of variables, the particular method highlighted will have the highest testing quadratic probability measure. Also shown is the resulting dimension or number of variables in the stepwise modes. Table 7.3 shows the variables in the order which they were selected by the stepwise models. For the seagrass data it is reasonable for one to be skeptical about the selection of variables 1 and 3 by SBQDA-CF2. Concern arises since these variables (or wavelenghts) lie close to the ends of the spectra, and also because, SBQDA-CF1 and SBQDA-CF3 did not select these same wavelengths. Likewise,

	Seagrass			Mineral			Paraxylene			Butanol		
SBLDA	CF1	CF2	CF3	CF1	CF2	CF3	CF1	CF2	CF3	CF1	CF2	CF3
	153	149	148	458	458	458	236	417	417	377	405	405
	476	6	6	265	266	266	497	380	380	470	402	402
	416	30	30	467	410	411	227	464	464	127	263	263
		8	9	314	444	282	63	414	414			
		124	102	264	281	445	143	187	187			
		69	66				135	113	198			
		400					214	161	472			
		506					512					
							259					
	CF1	CF2	CF3	CF1	CF2	CF3	CF1	CF2	CF3	CF1	CF2	CF3
SBQDA	141	148	148	458	458	458	471	417	417	145	405	406
	69	232	232	359	357	356	234	380	380	39	402	417
	231	70	71	199	424	351	470	226	226	405	420	402
	182	1	181				411	198	198	423		276
	71	3	91				413	464	363	38		405
	392	221	122				432	227	355	402		419
										420		

Table 7.3: **Original variables selected by SBLDA and SBQDA.**

some concern may arise from variable 6 being selected by SBLDA-CF2 and SBLDA-CF3. However, since two stepwise methods selected this variable at an early stage in the stepwise routine, i.e. in the first three steps, there is perhaps less cause for concern.

We now compare the performance of each of the classification methods SBLDA, SBQDA, FDA, PDA, and RDA. The correct classification rates and quadratic probability measures for the training and testing data are displayed in Table 7.4 and Table 7.5, respectively. The best results based on the performance of the test sets have been typed in bold face. Figure 7.5 was produced to facilitate interpretation of Tables 7.4 and 7.5. Only the classification rates and quadratic probability measures based on the testing data have been displayed in this figure.

The seagrass data tends to have better classification results than the remaining data sets. The mineral data are the next easily classified. The butanol and paraxylene data seem to be more difficult to assign the spectra into their appropriate classes.

Data		SBLDA	SBQDA	FDA	PDA	RDA
Seagrass	Train	100	100	98.18	96.97	99.39
	Test	**100** $_{CF1}$	97.06 $_{CF3}$	99.02	95.10	99.02
Mineral	Train	100	100	100	100	100
	Test	88 $_{CF3}$	93 $_{CF3}$	95	**100**	95
Paraxylene	Train	100	100	100	86.67	100
	Test	89.33 $_{CF2}$	80.00 $_{CF1}$	86.67	81.33	**100**
Butanol	Train	87.5	100	75	43.68	87.50
	Test	78.22 $_{CF1}$	86.60 $_{CF1}$	70.21	43.75	**87.23**

Table 7.4: **Correct classification rates (%)**

Data		SBLDA	SBQDA	FDA	PDA	RDA
Seagrass	Train	0.990	1.000	0.987	0.978	0.994
	Test	**0.997**	0.973	0.990	0.968	0.986
Mineral	Train	0.997	0.997	1.000	1.000	1.000
	Test	0.904	0.942	0.990	**1.000**	0.956
Paraxylene	Train	0.997	0.984	1.000	0.706	1.000
	Test	0.908	0.837	0.876	0.699	**1.000**
Butanol	Train	0.906	0.994	0.826	0.767	0.888
	Test	0.845	0.779	0.828	0.765	**0.881**

Table 7.5: **Quadratic probability measures**

In terms of the actual discriminant methods, no method performs the best for all of the data sets, although RDA performs quite well overall. PDA produces the highest test CCR for one data set – the mineral data. For the butanol data, PDA performs quite poorly. The performance measures for the low dimensional classifiers is much more diverse.

Analysis of the quadratic probability measures reflect a similar outcome as that of the correct classification rates. One interesting feature to note however, is that, for the seagrass and mineral data, FDA and RDA produce the same test classification rate, but in both instances the QPM for FDA is higher than that for RDA. This indicates that perhaps the class assignments made by FDA have been made with greater certainty than the class assignments for RDA. Another point of interest arises from the seemingly optimistic QPM value for the application of PDA to the butanol data. The correct classification rates are quite low, yet the QPM measures whilst smaller compared to the other QPM measures for butanol, may still seem a little high. It is a phenomenon that the QPM can have a

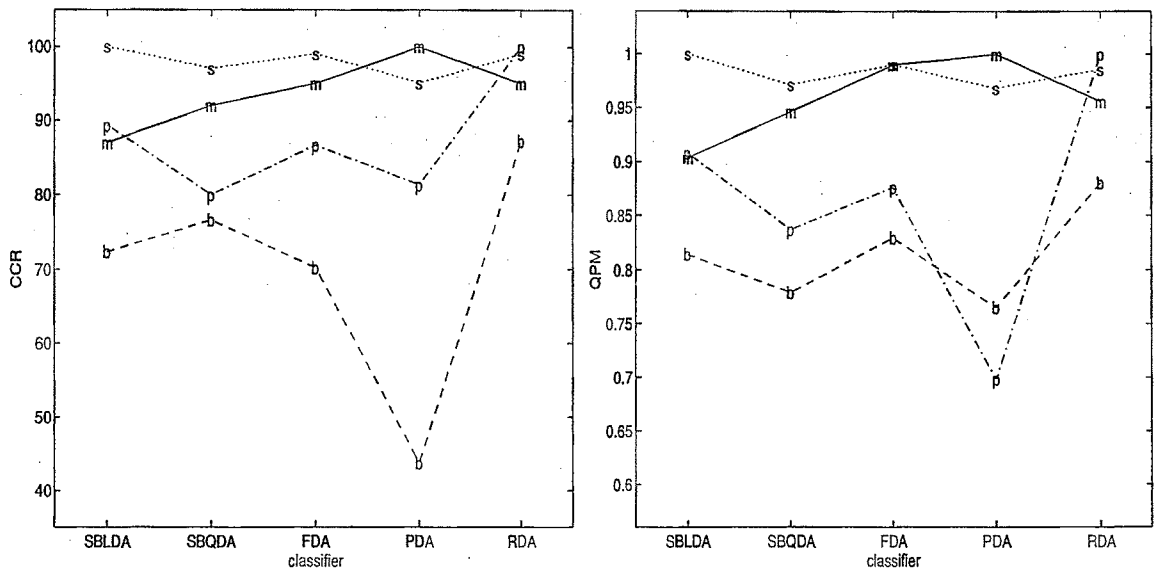


Figure 7.5: **Correct classification rates (CCR) and quadratic probability measures (QPM) for the seagrass (s), mineral (m), paraxylene (p) and butanol (b) data.**

tendency to produce overly optimistic values [1]. Another issue arises for PDA, with the paraxylene data. Now the QPM measures are quite low especially when compared to the QPM measures for SBLDA, which produced similar test classification rates to PDA.

Not shown in Table 7.4 are the grid values which produced the results for RDA. The setting (1.00,0.25) was used for the seagrass, mineral and butanol data. For these data, this indicates that a pooled covariance matrix is preferred to one that is not pooled. Conversely, the combination (0.25,0.25) used for the paraxylene data which weighs more heavily the individual class covariance matrices as opposed to the pooled class covariance matrix.

The next section explores the use of wavelet coefficients as features for discriminant analysis.

7.4 Discriminant Analysis Based on Wavelet Coefficients

In this section we investigate the use of wavelet (and scaling) coefficients as features for classification. Before embarking on the feature extraction procedure we explore the effects of the DWT when applied to the spectral data sets described in Section 7.2. Here, the DWT is applied using filter coefficients from the Daubechies family.

After examining the wavelet and scaling coefficients (and their backtransformations) of our data, feature selection techniques will be applied to the coefficients of the DWT where the filter coefficients are again, from the Daubechies family. Filter coefficients from the adaptive wavelet algorithm (AWA) will also be be used for calculating the coefficients from the DWT.

7.4.1 Exploring the DWT

One item of interest when using wavelet based features for classification, is whether the wavelet coefficients or the scaling coefficients should be used. Sometimes, a combination of the two may also prove to be worthwhile. To help us better understand what the wavelet and scaling coefficients represent, Figures 7.6–7.9 have been produced. For reasons outlined in Section 5.11, it is worthy to remember that the scaling coefficients, particularly from a higher level in the DWT, exhibit strong collinearity.

Figures 7.6–7.9 show two components – (i) the scaling and wavelet coefficients from DWT and (ii) the reconstructed spectra produced for the respective bands of coefficients in the DWT. In each of the figures, the wavelet transformation has been performed on a sampled spectrum from each group category. The sampled spectra used are the same as those in Figures 7.1– 7.4 and are overlayed in the plots. The DWT has been performed using the Daubechies filter with $N_f = 16$ to level 3, which is when 8 coefficients remain in the scaling and wavelet bands. The scaling coefficients for each of the levels (8 through to 3) are shown in the first column. The next column shows the reconstructed spectra produced by backtransforming the scaling coefficients (the wavelet coefficients at the same level have been set to zero). Column 3 shows the wavelet coefficients for each of the spectra for the levels 8 through to 3. The final column, displays the reconstructed spectra produced by backtransforming the wavelet coefficients (the scaling coefficients at the same level have been set to zero).

Consider the fourth row of plots in Figure 7.6. The fourth row corresponds to level 5 of the DWT. The coefficients in band(5,0) are the scaling coefficients which have been plotted against their index. The reconstructed spectra in the next column were obtained by thresholding the wavelet coefficients in band(5,1) to zero and then performing the inverse DWT on the scaling coefficients and the thresholded wavelet coefficients. The

original (unthresholded) wavelet coefficients at level 5 in the DWT are shown in the third column of row 5 in Figure 7.6. The final column of the same row shows the reconstructed spectra which results when the inverse DWT is performed on the wavelet coefficients from band(5,1) and on the thresholded scaling coefficients from band(5,0) which have been set to zero. The reconstructed spectra, illustrate in an approximate sense, the spectra which would be obtained, when the coefficients are linearly combined with their respective basis functions.

Coefficients potentially useful for classification should display some (between class) variability for the sampled class spectra. For the seagrass data this is visible for the scaling coefficients at most levels and the wavelet coefficients at lower levels in the DWT. Likewise, for the mineral data, it would appear that both the scaling and wavelet coefficients may provide useful information to the classification procedure. The variation with the paraxylene and butanol data are very slight. Some minor differences in the scaling and wavelet coefficients can however be detected for the butanol data. It is important to remember when inspecting these figures, that only a single spectrum from each class has been used in the construction of the plots, and that spectra from the same class can exhibit some slight within-class variability.

The aim of this section was to allow the reader to visualize what the various scaling and wavelet coefficients from the different levels represent. The next section applies various wavelet based feature selection strategie

7.4.2 Banded Discriminant Analysis

In this section we consider two banded approaches. The first which we refer to as BBLDA is banded Bayesian linear discriminant analysis (BBLDA), and the second which we refer to as BBQDA is banded Bayesian quadratic discriminant analysis. Both banded procedures use all of the coefficients from the same band in the wavelet transform, as input to the particular discriminant method, i.e. BLDA or BQDA. The discriminant analysis is then based on some set of coefficients $\mathbf{X}^{[j]}(\tau)$ at some level j, belonging to some band τ. The number of coefficients in band(j,τ) should be small when compared to the sample size so that an ill- or poorly-posed situation is avoided. The banded approach is a very simple procedure for feature selection of the wavelet coefficients. Previously, Bos [8] has used a

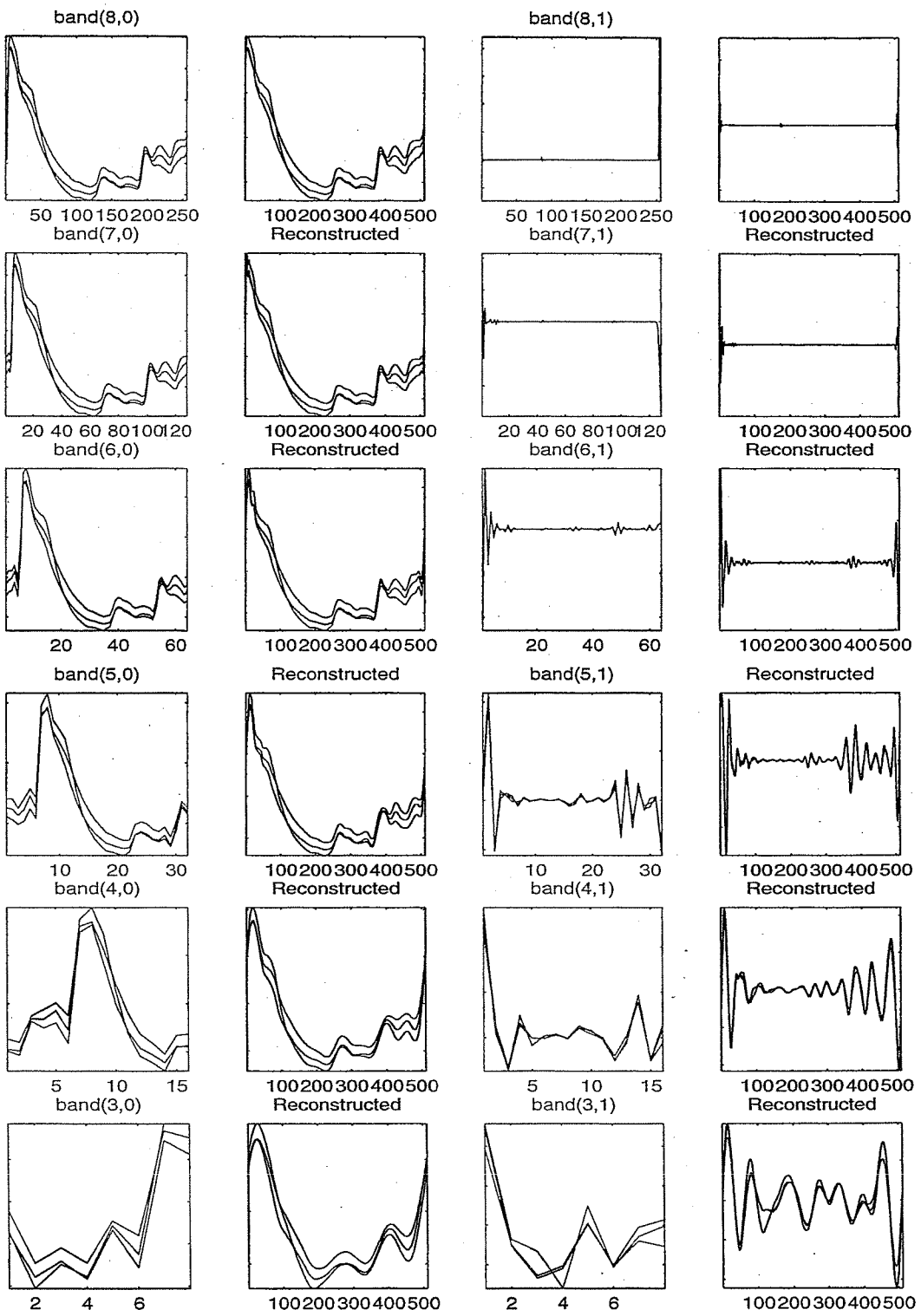


Figure 7.6: The DWT and inverse DWT performed on the seagrass data.

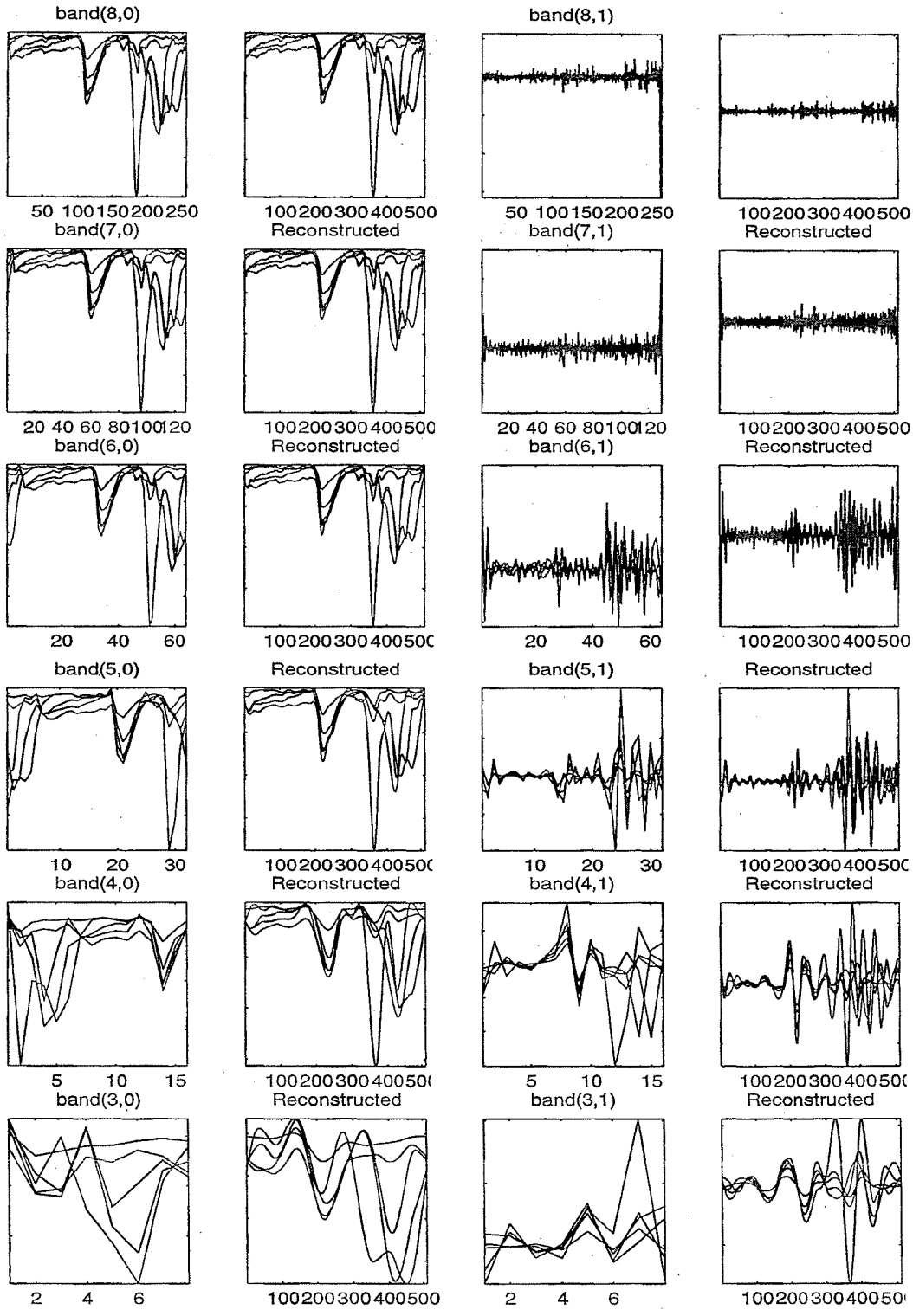


Figure 7.7: The DWT and inverse DWT performed on the mineral data

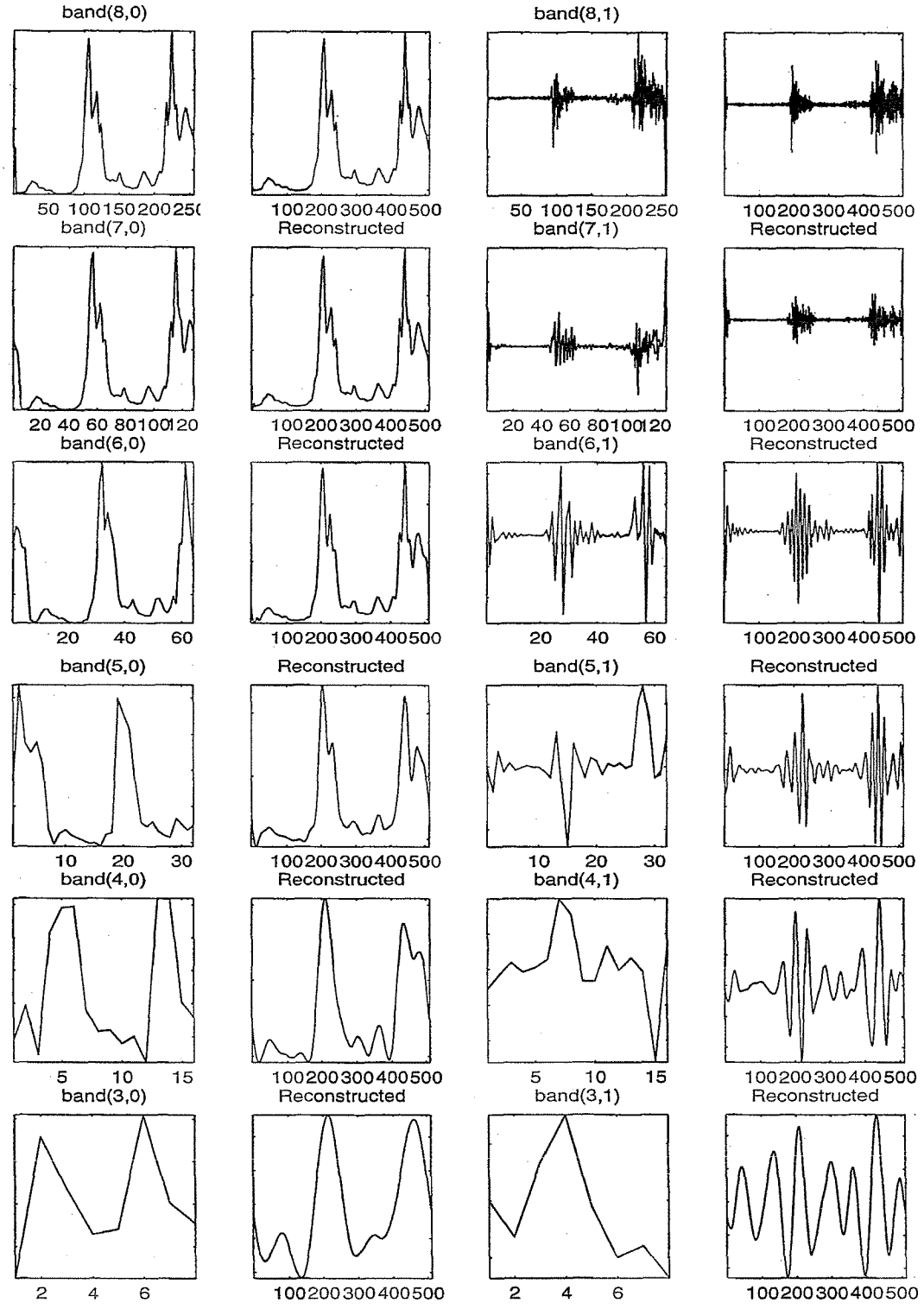


Figure 7.8: The **DWT and inverse DWT performed on the paraxylene** data.

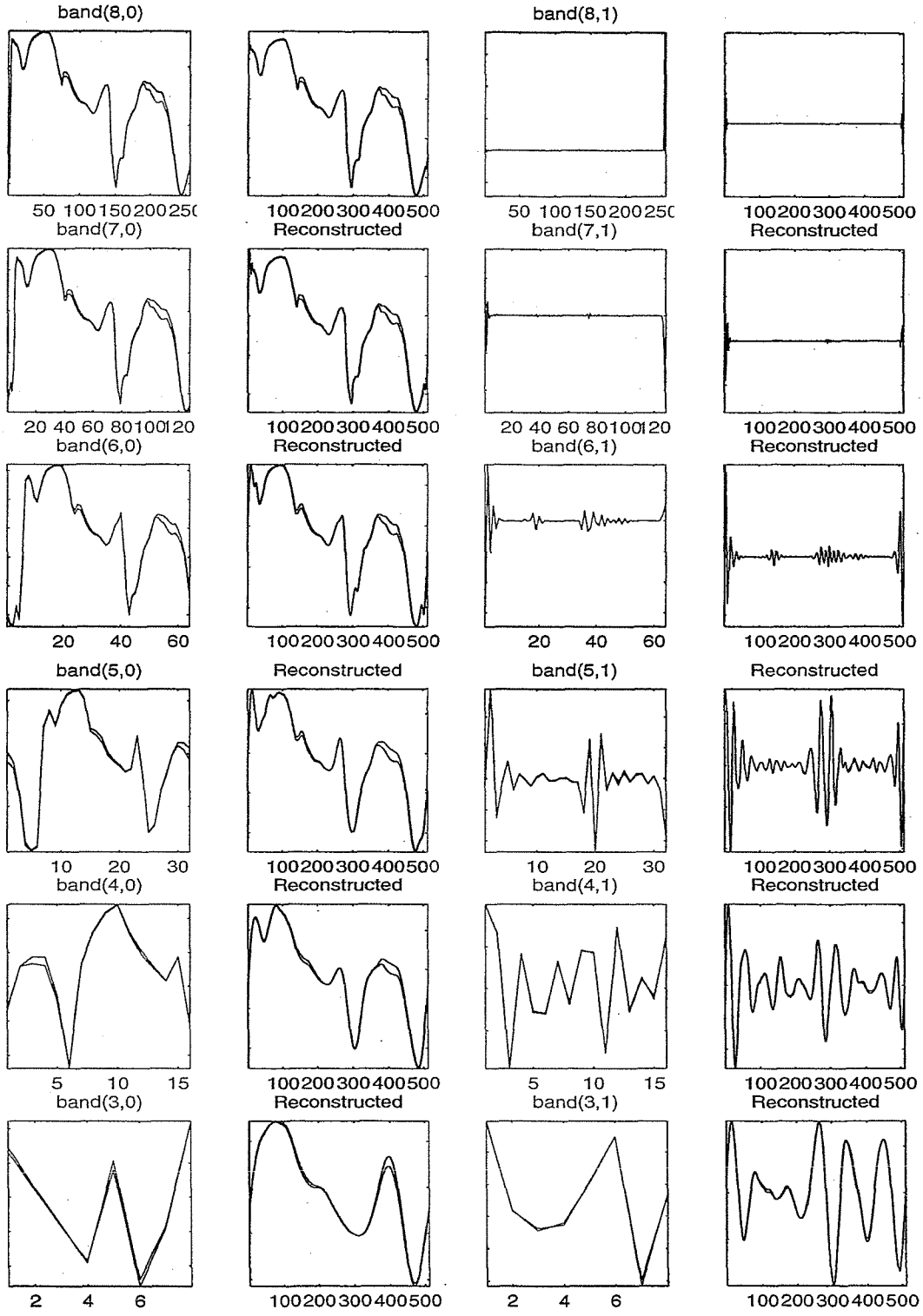


Figure 7.9: **The DWT and inverse DWT performed on the butanol data.**

similar approach, except that the bands of coefficients were supplied to neural networks.

In the banded procedure for this section the scaling $X^{[3]}(0)$ and wavelet $X^{[3]}(1)$ coefficients from level 3, and the scaling $X^{[4]}(0)$ and wavelet $X^{[4]}(1)$ coefficients from level 4 have been used for classification. At level 3, there are 8 coefficients in each of the bands, while level 4 has 16 coefficients in each of the bands. The classification results for BBLDA and BBQDA when applied to the wavelet coefficients produced for each of the data sets of Section 7.2 are summarised in Tables 7.6 and 7.7, respectively.

Data		$X^{[3]}(0)$	$X^{[3]}(1)$	$X^{[4]}(0)$	$X^{[4]}(1)$
Seagrass	Train	98.79	99.39	100	100
	Test	100	98.04	**100**	99.02
Mineral	Train	97	95	97	98
	Test	87	90	94	**98**
Paraxylene	Train	62.67	68.00	81.33	80.00
	Test	50.67	58.67	56.00	**61.33**
Butanol	Train	85.42	87.50	93.75	87.50
	Test	82.98	82.98	76.60	**87.23**

Table 7.6: **Classification results for BBLDA.**

We first comment on the results for BBLDA. The figures typed in boldface have the highest (test) classification rate for each of the data. If the same test classification rate appears for two or more bands, then the figure typed in bold will have the highest (test) quadratic probability measure. For the seagrass data each of the scaling bands have outperformed the wavelet bands, while for the mineral, paraxylene and butanol data wavelet bands have produced better classification results than the respective scaling bands. For the butanol data the performance between band(3,0) and band(3,1) is relatively marginal however.

For BBQDA, numerical instabilities arose for the mineral data when 16 coefficients were supplied to the classifier. This can be attributed to the fact that for BQDA, the class sample size should be large compared to the dimensionality. For the mineral data there are 20 objects per class which is only marginally larger than 16, hence it was not possible to produce accurate results for this setting. When only 8 wavelet coefficients were used however there is a 6 per cent improvement in using BQDA as opposed to BLDA. This is seen for both the scaling (band(3,0)) and wavelet bands (band(3,1)). There is

Data		$X^{[3]}(0)$	$X^{[3]}(1)$	$X^{[4]}(0)$	$X^{[4]}(1)$
Seagrass	Train	100	100	100	100
	Test	100	99.02	100	**100**
Mineral	Train	100	100	-	-
	Test	93	**96**	-	-
Paraxylene	Train	88	86.67	100	100
	Test	66.67	72.00	**81.33**	76.00
Butanol	Train	89.58	87.50	100	100
	Test	**74.47**	72.34	63.83	57.45

Table 7.7: **Classification results for BBQDA.**

also an improvement in results for the paraxylene data as well. For the seagrass data the results are comparable to those obtained by BBLDA. The results for the butanol data using BBQDA are not as favourable as those obtained using BBLDA.

7.4.3 Stepwise Feature Extraction from the DWT

The DWT will be performed to level 3 using a Daubechies wavelet defined by 16 filter coefficients. The total set of features consists of the scaling coefficients at level 3, and the wavelet coefficients at level 3 up to and including the wavelet coefficients at level 8. These coefficients constitute the commonly used Mallat's right hand pyramidal tree.

In this section the stepwise methods SWBLDA and SWBQDA are applied to the wavelet and scaling coefficients produced from the seagrass, mineral, paraxylene and butanol data. Each of the forward stepwise strategies CF1, CF2 and CF3 are applied. The classification results of the forward stepwise strategies are summarized in Table 7.8.

The boldface type identifies the stepwise procedure producing the highest CCR. If two or more strategies produce the same "highest" CCR, then the number marked in bold will have fewer variables. Should both the strategies have the same number of variables, then the method giving the largest QPM for the testing data will be highlighted. This procedure is much the same as that performed on the original data, and is done separately for SWBLDA and SWBQDA. The CF3 procedure tends to be outperforming the CF1 and CF2 strategies.

Also of interest is the coefficients which have been selected by the stepwise procedures. Table 7.9 shows the indices of the coefficients from the DWT which have been selected

Data		SWBLDA			SWBQDA		
		CF1	CF2	CF3	CF1	CF2	CF3
Seagrass	Train	99.39	100	99.39	100	100	100
	Test	95.10	97.06	**98.04**	97.06	97.06	**97.06**
	dimension	3	4	3	4	4	4
Mineral	Train	99	100	100	100	100	100
	Test	**97**	93	93	89	**92**	90
	dimension	6	5	6	4	3	4
Paraxylene	Train	100	100	98.67	100	100	97.33
	Test	69.33	81.33	**81.33**	77.33	78.67	**82.67**
	dimension	7	7	6	6	5	6
Butanol	Train	100	85.42	85.42	100	100	91.67
	Test	72.34	85.11	**85.11**	68.09	65.57	**74.47**
	dimension	6	6	5	5	5	4

Table 7.8: **Correct classification rates for SWBLDA and SWBQDA.**

	Seagrass			Mineral			Paraxylene			Butanol		
SWBLDA	CF1	CF2	CF3	CF1	CF2	CF3	CF1	CF2	CF3	CF1	CF2	CF3
	62	205	205	51	51	51	299	299	299	217	217	217
	201	201	201	20	2	2	423	423	381	465	155	155
	56	419	2	12	6	7	476	282	476	318	257	318
		265		66	340	456	324	75	368	427	471	465
				501	459	116	460	344	389	309	260	313
				2		358	123	486	170	257	334	
							457	409				
SWBQDA	CF1	CF2	CF3	CF1	CF2	CF3	CF1	CF2	CF3	CF1	CF2	CF3
	204	204	204	51	51	51	299	299	299	217	217	217
	34	34	34	40	3	3	491	491	192	465	155	77
	202	202	202	374	54	164	168	197	168	118	257	411
	14	14	171	482		23	243	17	191	130	133	334
							368	257	450	467	433	
							197		45			

Table 7.9: **Coefficients selected by the forward schemes for SWBLDA and SWBQDA.**

by each of the forward stepwise schemes, for each of the sets of data. The data has been stored as follows

band(3,0)	band(3,1)	band(4,1)	band(5,1)	band(6,1)	band(7,1)	band(8,1)
1:8	9:16	17:32	33:64	65:128	129:256	257:512

so that the first 8 coefficients are from the scaling band at level 3, while the next 8 coefficients are the wavelet coefficients at level 3. The next 16 coefficients are wavelet coefficients from band(4,1) and so on. The only set of scaling coefficients which formed part of the feature set were those contained in band(3,0). Consider for example SWBLDA-CF3 applied to the seagrass data. This technique selected coefficients with index labels of 205, 201 and 2. The indices 205 and 201 refer to the position of the coefficients in the DWT. Using the table above, we can see that these coefficients are from band(7,1), while the coefficient with an index of 2, is the second scaling coefficient in band(3,0). Note that instead of using indices we could have used the two subscipts (j, k) to identify their positions in the wavelet tree. Instead we chose to use a single number so that one can quickly compare the indices which were selected.

There is a some variation between the coefficients which have been selected for SWBLDA by the forward selection schemes CF1, CF2 and CF3, although by examination of Table 7.9 one can see, that the coefficients generally pertain to similar regions of the DWT. A similar observation can be made for SWBQDA.

Figure 7.10 was produced to help provide some idea where the coefficients in Table 7.9 lie in relation to the bands of the DWT. This was done for each of the data sets, but for SWBLDA using one selection scheme – CF1, CF2 or CF3. The coefficients selected from the CF3 forward strategy have been shown for the seagrass, paraxylene and butanol data, while the coefficients displayed for the mineral data were produced using the CF1 strategy.

For the paraxylene and butanol data where discrimination appears to be somewhat challenging, selection of the wavelet coefficients pertaining to a higher level in the DWT is more predominant. For the mineral data, SWBLDA (CF1) has selected a range of coefficients from the DWT. With the exception of band(7,1) a coefficient has been selected from each of the bands constituting the DWT. This indicates that a range of high and low frequency information is utilized by the stepwise discriminant techniques. There are only three features which have been selected by the stepwise method SWBLDA (CF3) for the

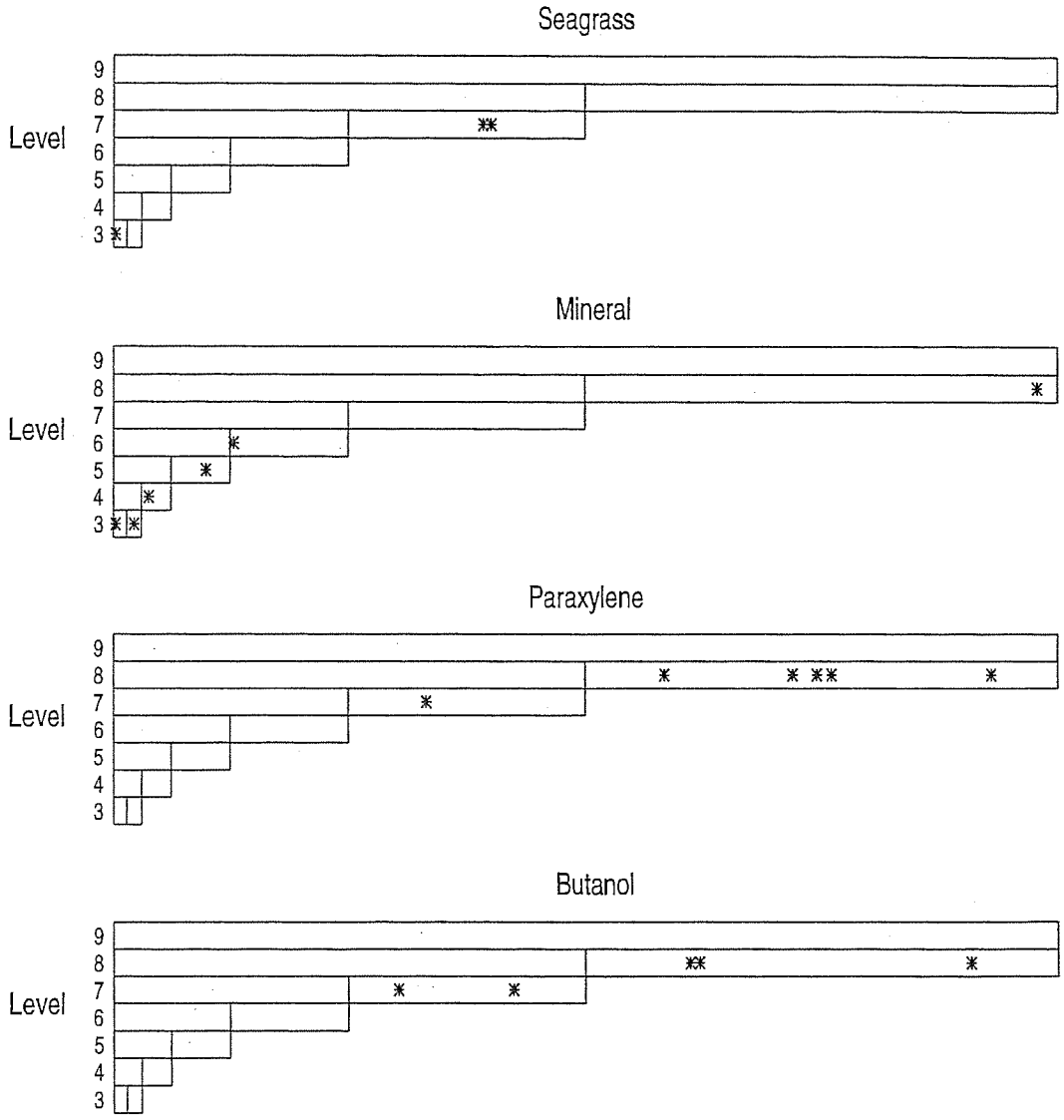


Figure 7.10: **Coefficients selected from the DWT by SWBLDA.**

seagrass data, where a coefficient from the scaling band and two wavelet coefficients both from band(7,1).

Next we present the results which were obtained from the LDB algorithm.

7.4.4 Local Discriminant Bases

In this section a feature extraction method is applied to wavelet packet coefficients. This procedure is referred to as the local discriminant bases (LDB) algorithm and was previously discussed in Section 5.14.2.

The wavelet packet transform was calculated to level 3 using the Daubechies filter coefficients with $N_f = 16$. Once the wavelet packet decomposition has been formed, it is necessary to determine the best basis from the energy maps. The criterion which we have used to form the best basis is the symmetric entropy criterion. Figure 7.11 marks the best basis selected by the LDB algorithm for each of the sets of data. For the paraxylene data, the selected best basis is the original data.

Once the best basis has been found, it is then necessary to select a subset of wavelet packet coefficients from the best basis. It was decided to select the 16 wavelet packet coefficients based on the same discriminant measure which produced the best basis, i.e. symmetric entropy.

The asterisks in Figure 7.11 show the positions of the 16 wavelet packet coefficients which were selected from the best basis and supplied to the classifier BLDA. There is a tendency for the wavelet packet coefficients to be selected from band(4,0) and band(4,1) of the WPT. These bands will contain the same coefficients as those in band(4,0) and band(4,1) from the DWT. The wavelet packet coefficients are quite clustered, which is a likely consequence of selecting the coefficients by a univariate strategy, that is without consideration given to previously selected features.

The sixteen wavelet packet coefficients with the largest symmetric entropy measures were then supplied to the classifier BLDA in a top-down approach, that is, the first $1, 2, \ldots, 16$ coefficients were used for classification. That is, initially a single wavelet packet coefficient (with the largest discriminant measure) is supplied to the classifier. Then, the two wavelet packet coefficients with the largest discriminant measures are supplied to the classifier. This procedure continues until all 16 coefficients have formed part of the BLDA model.

Table 7.10 gives the classification rates for the training and testing data for each of the discriminant data sets where the first $1, \ldots, 16$ wavelet packet coefficients have been

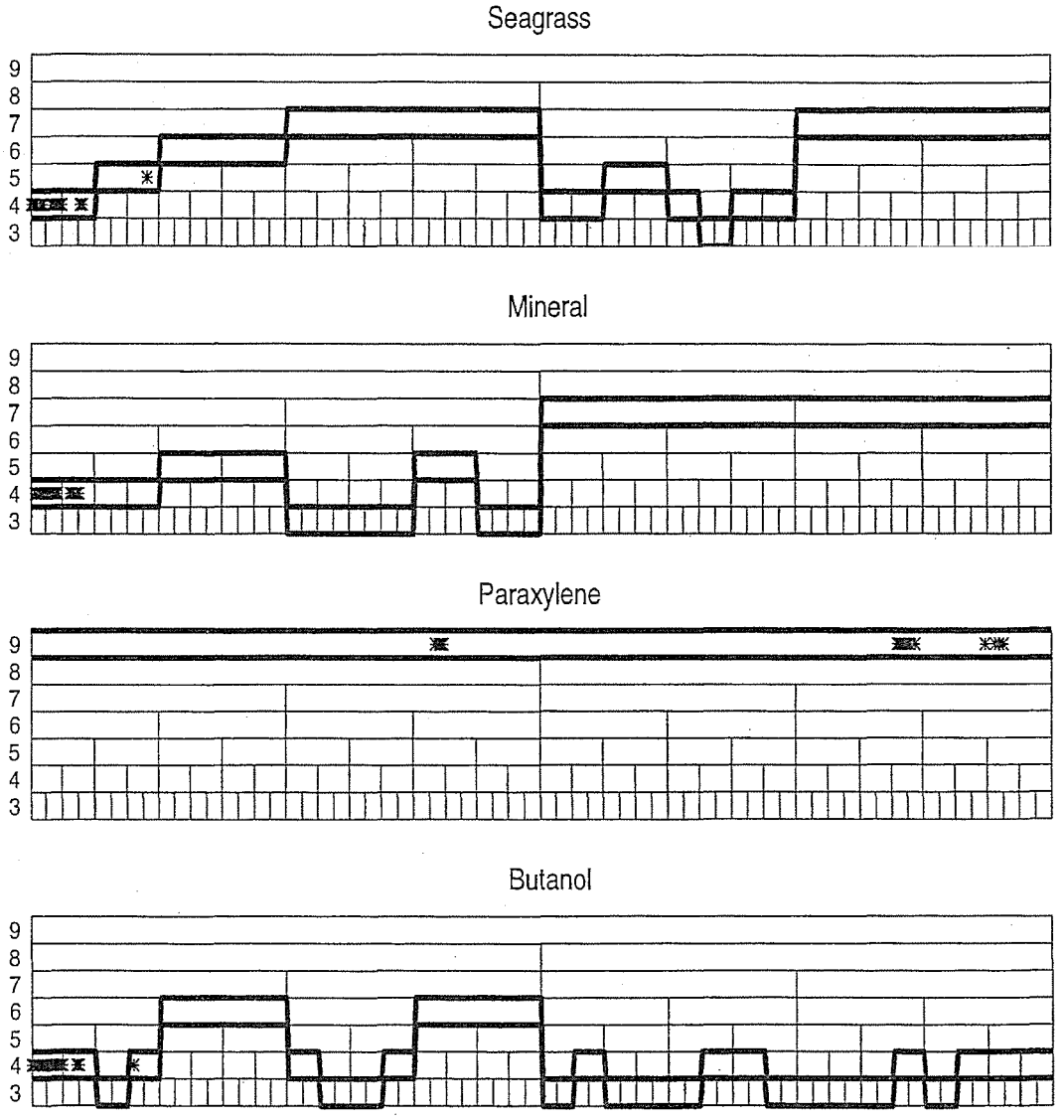


Figure 7.11: **Selected wavelet coefficients (asterisks) from the best bases.**

selected from the best basis, and supplied to BLDA. The numbers highlighted in bold type have the largest (test) CCR and the fewest terms in the discriminant model.

This application of the LDB approach has followed closely that outlined in [118], the main difference is that we have used BLDA as opposed to FLDA. There are some issues which arise from the LDB algorithm. One question is if it is indeed worthwhile to select the final set of coefficients for classification from the best basis, as opposed to searching through the entire wavelet packet transform. Walczak *et. al* [141] have compared performance of feature selection from the LDB and the full WPT using a univariate feature selection

Number	Seagrass		Mineral		Paraxylene		Butanol	
of WPC	Train	Test	Train	Test	Train	Test	Train	Test
1	74.55	76.47	67	72	42.67	41.33	72.92	74.47
2	73.33	69.61	87	89	48.00	65.33	75.00	76.60
3	86.67	86.27	88	90	54.67	53.33	79.17	80.85
4	93.94	94.12	93	87	68.00	62.67	81.25	80.85
5	90.91	89.22	93	87	69.33	57.33	83.33	**87.23**
6	93.33	91.18	94	85	72.00	57.33	83.33	80.85
7	98.18	97.06	94	88	78.67	77.33	83.33	80.85
8	100.00	**100.00**	96	88	78.67	**78.67**	83.33	85.11
9	100.00	100.00	96	88	78.67	77.33	85.42	85.11
10	100.00	100.00	96	88	84.00	73.33	83.33	82.98
11	99.39	99.02	97	**93**	84.00	72.00	83.33	80.85
12	99.39	99.02	97	93	84.00	77.33	85.42	82.98
13	100.00	100.00	98	91	88.00	76.00	85.42	82.98
14	100.00	100.00	98	92	89.33	77.33	83.33	78.72
15	100.00	100.00	98	92	89.33	77.33	87.50	76.60
16	100.00	100.00	99	93	92.00	73.33	87.50	76.60

Table 7.10: **Classification performance of the LDB algorithm.**

method based on Fisher's criterion (see Section 4.1.1). For their data, they concluded that no gain was bought about by the LDB algorithm. One advantage for the LDB algorithm without taking into consideration the time to calculate the best basis, is that it will be computationally quicker to select coefficients from the best basis as opposed to selecting them from the larger set of wavelet packet coefficients. The other advantage is that it does help to reduce the inter-dependencies that exist between the coefficients in the parent nodes with the coefficients in the children nodes. Although, a feature selection procedure which looked at combinations of features, such as stepwise procedure, would also take into consideration the inter-dependencies between the coefficients.

7.4.5 Adaptive Wavelet Algorithm

In the previous sections, the DWT has been performed using the filter coefficients from the Daubechies family. There are many filter coefficients which we could have chosen, but the Daubechies filter coefficients were chosen since they tend to be documented quite frequently. There is no reason however, that another set of coefficients could not have

been used. Of course, the problem which we face is deciding whether or not we might obtain better results using other filter coefficients. In this section we design our own task specific filter coefficients using the adaptive wavelet algorithm of Chapter 6. The idea behind the adaptive wavelet algorithm is to avoid the decision of which wavelet basis we should select and design our own wavelets to suit the current task at hand, which in this case is discriminant analysis.

The adaptive wavelet algorithm is applied using several settings of the m, q and j_o parameters. The particular (m, q, j_o) triplets used were (4,3,2), (4,2,2) , (8,1,1), (2,5,3), (2,5,4), (2,7,3), and (2,7,4). These settings were chosen because (i) they provide suitable ratios of the dimensionality of the wavelet bands to the sample size and, (ii) so that the number of filter coefficients is $N_f = 12$ and $N_f = 16$. Section 6.4 describes some heuristics for choosing values for these parameters as well as τ.

The discriminant criterion function implemented by the adaptive wavelet algorithm is the CVQPM criterion function. A form of banded selection is performed, whereby the criterion function is calculated from a band of coefficients $\mathcal{J}_{\text{CVQPM}}\left(\mathbf{X}^{[j]}(\tau)\right)$. The same coefficients are later supplied to the classifier.

The value τ is chosen as the band which gave the highest CVQPM value at initialization for a particular (m, q, j_o) triplet. The coefficients in band(j, τ) are then supplied to the classifier. In some cases the algorithm chose to optimize over a scaling band. This would occur if the discriminant criterion for a scaling band was higher than that for the wavelet bands (at initialization). We have discussed earlier that the scaling coefficients may prove to be useful when the basic shape or low frequency event contains discriminatory information. If a scaling band (i.e. $\tau = 0$) were selected for a particular setting, then for the same (m, q, j_o) settings it was decided to repeat the experiment and optimize over the wavelet band having the largest discriminant measure at initialization.

Some stopping rules were applied to the optimization routine. The optimization routine halted if 2000 iterations of the optimization routine had been performed or sooner if an optimal value was obtained. For the seagrass data we found it was necessary to have only 500 iterations, since the discriminant measure was already quite high in the early stages of the AWA. Whilst having a preset number of iterations does not allow for the best optimal value to be found, from an applied point of view it is more practical. In

our experimentations we generally found that the classification rates did not improve very much, if at all, after 1500 iterations.

The results of the adaptive wavelet algorithm are presented in Table 7.11, also shown is the number of filter coefficients (N_f), used in computing the DWT and the number of coefficients (N_{coef}) in each of the bands for the respective (m, q, j_o) settings. For each data set the highest CCR based on the testing data, obtained with the least number of coefficients is typed in boldface. The adaptive wavelet algorithm performs quite well for each of the settings for the seagrass data and eventually, the setting (4,3,2) produced the best results using fewer coefficients. Quite good results are also obtained for the mineral data with the setting (2,7,4) for band(4,1). For this setting optimization was initially based on the scaling coefficients, but when optimization for the (2,7,4) setting was performed on the wavelet coefficients the results were further improved. For the paraxylene data the classification performance was generally improved when optimization was based on the wavelet coefficients. This was not necessarily the case for the butanol data, where for the settings (2,5,4) and (2,7,4) classification based on the scaling coefficients improved the test CCR by more than 10% when compared to their respective wavelet bands.

Seagrass	m	q	j_o	N_f	N_{coef}	τ	Train	Test
	4	3	2	16	8	1	100	**100**
	4	2	2	12	8	0	100	99.02
						1	99.39	97.06
	8	1	2	16	8	1	99.39	99.02
	2	5	6	12	8	0	100	99.02
						1	99.39	99.02
	2	5	5	12	16	0	100	100
						1	100	100
	2	7	6	16	8	1	99.39	99.02
	2	7	5	16	16	1	100	100
Mineral	m	q	j_o	N_f	N_{coef}	τ	Train	Test
	4	3	2	16	8	0	100	96
						1	100	92
	4	2	2	12	8	0	97	89
						1	98	89
	8	1	1	16	8	4	96	90
	2	5	3	12	8	1	98	90
	2	5	4	12	16	1	99	95
	2	7	3	16	8	1	100	95
	2	7	4	16	16	0	100	93
						1	99	**99**
Paraxylene	m	q	j_o	N_f	N_{coef}	τ	Train	Test
	4	3	2	16	8	2	94.67	76.00
	4	2	2	12	8	2	88.00	68.00
	8	1	1	16	8	2	97.33	74.67
	2	5	3	12	8	0	84.00	58.67
						1	85.33	66.67
	2	5	4	12	16	0	70.67	50.67
						1	94.67	**86.67**
	2	7	3	16	8	0	78.67	61.33
						1	84.00	74.67
	2	7	4	16	16	1	96.00	81.33
Butanol	m	q	j_o	N_f	N_{coef}	τ	Train	Test
	4	3	2	16	8	3	97.92	61.70
	4	2	2	12	8	1	97.92	57.45
	8	1	1	16	8	2	97.92	74.47
	2	5	3	12	8	1	93.75	**87.23**
	2	5	4	12	16	0	93.75	82.98
						1	95.83	70.21
	2	7	3	16	8	1	93.75	65.96
	2	7	4	16	16	0	97.92	85.11
						1	97.92	74.47

Table 7.11: **Classification results for the adaptive wavelet algorithm.**

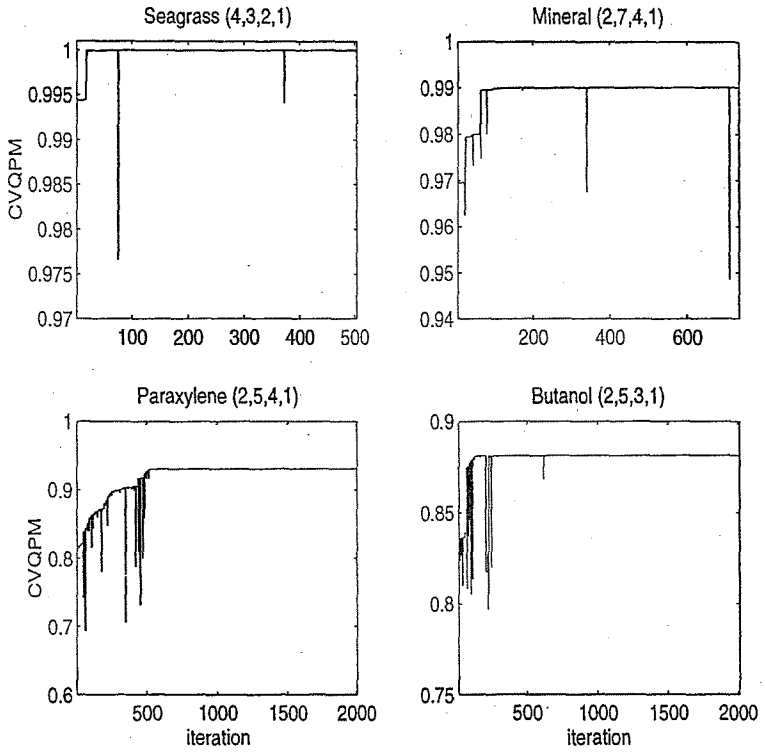


Figure 7.12: **Discriminant measure versus iteration for the adaptive wavelet algorithm.**

To demonstrate convergence of the adaptive wavelet algorithm in at least a local sense, the values of the criterion measure $\mathcal{J}_{\text{CVQPM}}\left(\mathrm{X}^{[j]}(\tau)\right)$ have been plotted against the iteration number of the optimization routine as shown in Figure 7.12. This was done for the setting $(m, q, j_o, \tau) = (4, 3, 2, 1)$ for the seagrass data, the (2,7,4,1) setting for the mineral data, the (2,5,4,1) setting for the paraxylene data and the (2,5,3,1) setting for the butanol data. The CVQPM values were initially very high. This is especially the case for the seagrass data, which is why a maximum of 500 iterations were used in the optimization routine. For the mineral data the optimization routine halted after approximately 700 iterations. When comparing the output for the paraxylene and butanol data against that of the seagrass and mineral data, one can see that more work was required by the optimizer to improve the CVQPM measures for the paraxylene and butanol data and both data sets used 2000 iterations of the optimization routine. The optimization routine will make several evaluations of the discriminant criterion function before choosing a search direction. Some of the trialled discriminant criterion values will be quite small and this

contributes to the sharp drops in Figure 7.12.

In the next experiment we decided to optimize over a scaling band, and a wavelet band, since sometimes information is needed about the low and high frequency events. The aim was to use at least a 3 band DWT and optimize over a scaling and wavelet band, with the aim of 'pushing' the information which is not useful for discrimination into the remaining band(s). We opted to use the $m = 4$ band DWT for this experiment With the $m = 4$ band DWT, it was anticipated that the noise would be pushed into the two remaining bands. The adaptive wavelet algorithm was applied to the (m, q, j_o)=(4,3,2) setting. Now that optimization involves a scaling and wavelet band (at the same level), we need only to select the wavelet band at initialization. The discriminant measure was formulated based on the coefficients from the scaling and wavelet band which produced the highest measurement at initialization. For the mineral data, the wavelet coefficients $\mathbf{X}^{[2]}(1)$ were combined with the scaling coefficients $\mathbf{X}^{[2]}(0)$ from the 4 band DWT. For the paraxylene data, the wavelet coefficients $\mathbf{X}^{[2]}(2)$ were combined with the scaling coefficients and for the butanol data, the wavelet coefficients$\mathbf{X}^{[2]}(3)$ were used together with the scaling coefficients.

Data	τ	Train	Test
Mineral	$\mathbf{X}^{[2]}(0)$ and $\mathbf{X}^{[2]}(1)$	100	96.00
Paraxylene	$\mathbf{X}^{[2]}(0)$ and $\mathbf{X}^{[2]}(2)$	92.00	69.33
Butanol	$\mathbf{X}^{[2]}(0)$ and $\mathbf{X}^{[2]}(3)$	100	61.70

Table 7.12: **Classification results for the adaptive wavelet algorithm where optimization was over a scaling and wavelet band for the (4,3,2) setting.**

The results from this combined approach are shown in Table 7.12 for the mineral, paraxylene, and butanol data. The seagrass data was not applied, since the adaptive wavelet algorithm on a single band already performs quite adequately for this data. Despite the attempt of using low and high frequency information, the combined approach was less effective than using the single band approach. This may simply be due to the data not performing as well for the $m = 4$ scenario.

7.4.6 Summary of the Wavelet Feature Extraction Strategies

In the previous sections we have investigated different feature extraction procedures which involve the application of wavelet (and scaling) coefficients for classification. We have considered, the banded feature selection procedures BBLDA and BBQDA, the stepwise methods SWLBDA and SWBQDA, and the LDB and AWA algorithms.

In this section we summarize the results of these strategies and compare how these methods performed on our discriminant data sets. The correct classification rates for these procedures are listed in Table 7.13. In the previous sections we elected not to present details about the quadratic probability measures so as to avoid presenting too many details. In this section we present the quadratic probability measures for the corresponding models whose correct classification rates are listed in Table 7.13. The quadratic probability measures are in Table 7.14. The classification rates and quadratic probability measures (for the testing data only) are also displayed in Figure 7.13.

Data		BBLDA	BBQDA	SWBLDA	SWBQDA	LDB	AWA
Seagrass	Train	100	100	99.39	100	100	100
	Test	**100**	**100**	98.04_{CF3}	97.06_{CF3}	**100**	**100**
	Dimension	16	16	3	4	8	8
Mineral	Train	98	100	99	100	97	99
	Test	98	96	97_{CF1}	92_{CF2}	93	**99**
	Dimension	16	8	6	3	11	16
Paraxylene	Train	80.00	100	98.67	97.33	78.67	94.67
	Test	61.33	81.33	81.33_{CF3}	82.67_{CF3}	78.67	**86.67**
	Dimension	16	16	6	6	8	16
Butanol	Train	87.50	89.58	85.42	91.67	83.33	93.75
	Test	**87.23**	74.47	85.11_{CF3}	74.47_{CF3}	**87.23**	**87.23**
	Dimension	16	8	5	4	5	8

Table 7.13: **Correct classification rates for the wavelet based feature extraction strategies.**

The tabulated results for BBLDA correspond to the largest test CCR for each of the data sets, similarly for BBQDA. This information was taken from Tables 7.6 and 7.7, respectively. The results for SWBLDA and SWBQDA were extracted from Table 7.8. For SWBLDA the results for the CF3 strategy were tabulated for the seagrass, paraxylene and butanol data., and the results for the CF1 strategy were tabulated for the mineral data.

Data		BBLDA	BBQDA	SWBLDA	SWBQDA	LDB	AWA
Seagrass	Train	1.000	1	0.994	1	1	1
	Test	**1.000**	**1.000**	0.985	0.973	**1.000**	**1.000**
	Dimension	16	16	3	4	8	8
Mineral	Train	0.987	0.997	0.991	0.993	0.978	0.99
	Test	0.980	0.960	0.973	0.941	0.941	**0.982**
	Dimension	16	8	3	3	11	16
Paraxylene	Train	0.866	1	0.982	0.973	0.834	0.954
	Test	0.751	0.819	0.866	0.853	0.838	**0.876**
	Dimension	16	16	6	6	8	16
Butanol	Train	0.918	0.927	0.869	0.957	0.88	0.935
	Test	0.905	0.866	0.879	0.789	0.883	**0.890**
	Dimension	16	8	5	6	5	8

Table 7.14: **Quadratic probability measures for the wavelet based feature extraction strategies.**

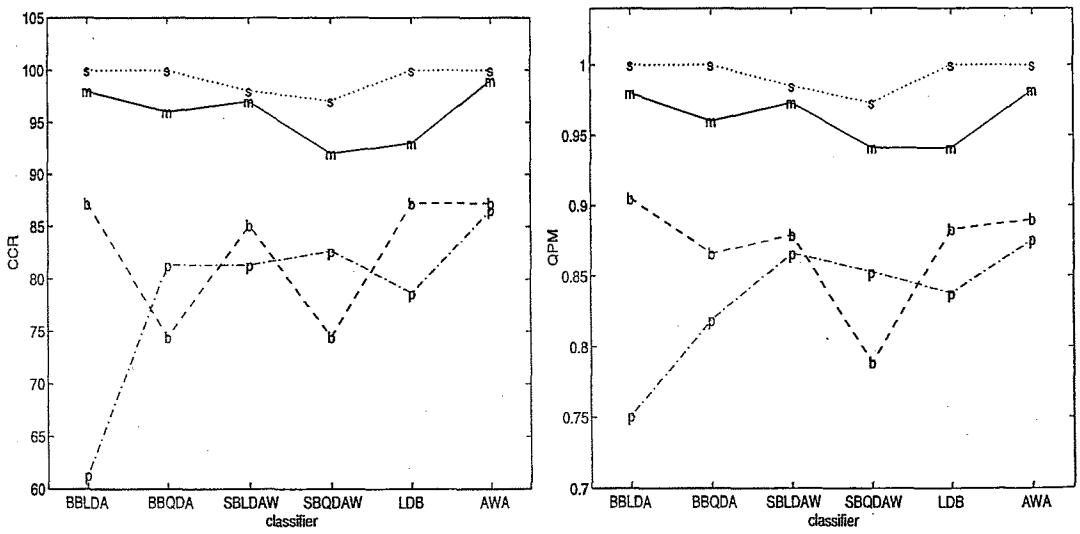


Figure 7.13: **Correct classification rates (CCR) and quadratic probability measures (QPM) for the wavelet based methods applied to the seagrass (s), mineral (m), paraxylene (p) and butanol (b) data.**

For SWBQDA the results for the CF3 strategy are presented for the seagrass, paraxylene and butanol data, and the results for the CF2 scheme are listed for the mineral data. The results for the LDB model were taken from Table 7.10 and the results for the AWA were taken from Table 7.11.

Whilst the results presented in this section are based on the largest testing correct clas-

sification rates, in practice it will not be known in advance which particular model will give the best test CCR, and such solutions require further investigation. Since we are fortunate enough to have independent test sets available, we have decided to compare the results based on the performance of the testing data but are aware that further investigation is required to determine, based on some procedure involving the training data only, which model will have the best test classification rates. Since our desire is to compare the performance of the best possible models, we feel that the approach adopted herein is satisfactory.

In Tables 7.13 and 7.14 the largest CCR and QPM based on the testing data have been typed in boldface. Despite the somewhat limited feature extraction procedure implemented by the AWA as opposed to the more flexible feature extraction procedures which can be implemented when a predefined family of filter coefficients is used for calculating the DWT, the results for the adaptive wavelet algorithm are consistently favourable across each of the data sets and, in terms of both the classification rates and quadratic probability measures.

The SWBLDA routines based on the Daubechies scaling and wavelet coefficients also performs quite well. The LDB algorithm performs well for the seagrass and butanol data, and the BBLDA performs well for the seagrass and mineral data. The only data set which seemed to perform well under SWBQDA was the paraxylene data.

When examining Figure 7.13 it is interesting to note if the profiles for the CCR, are followed closely by the profiles for the QPM (for each data set individually). The profiles for each data set are mostly the same with the exception occurring for the QPM for SWBQDA on the paraxylene data. The ranking then, of our models with respect to the classification rate based estimates are similar to the ranking of the models with respect to their probability based estimates.

7.5 Which Classification Strategy?

This chapter has presented several classification strategies which can be applied to spectral data. Of course such strategies extend to similar forms of data. Three main approaches have been presented:

1. Classification using all of the original data and a high dimensional classifier,
2. selecting features from the original data and applying a low dimensional classifier, and
3. selecting wavelet coefficients as features and applying a low dimensional classifier.

In the first approach, PDA and RDA were applied to the original data without any prior feature selection. The second approach applied SBLDA, SBQDA and FDA to the original data. The third approach which supplied the wavelet coefficients to the classifiers included BBLDA, BBQDA, SWBLDA, SWBQDA, the LDB and AWA algorithms.

This section investigates which of the these approaches might be more suitable for the classification of spectral data in general, and particularly to the seagrass, mineral, paraxylene and butanol data. This will be done in two stages. Firstly, the correct classification rates and quadratic probability measures will be examined for each of the classification procedures. This performance based assessment is presented in Section 7.5.1. We are also interested in what a particular classification strategy can tell us about our data. This qualitative assessment will be given in Section 7.5.2.

7.5.1 Performance Based Measures

In this section a summary of the previous classification results obtained for the original data and the coefficients from the wavelet transforms is given. Figure 7.14 displays the correct classification rates for each of the classification strategies, again the information displayed is calculated from the testing data. The results for the high dimensional methods PDA and RDA appear at the top of the graphs for each of the data sets. Following this the results for the low dimensional classifiers SBLDA, SBQDA and FDA are shown. For these methods the features have been selected from the original data. The last six classification strategies have extracted features from the wavelet coefficients. The methods shown are BBLDA, BBQDA, SWBLDA, SWBQDA, LDB and AWA. To enable easier interpretation of Figure 7.14, crosses have been used to indicate the results of the high dimensional methods likewise, circles have been used to indicate the low dimensional classification performed on the original data, and the asterisks are indicators for the low dimensional classification methods based on the coefficients from the DWT. To further

enhance interpretation of Figure 7.14, the line types have also changed accordingly with the marker indicators.

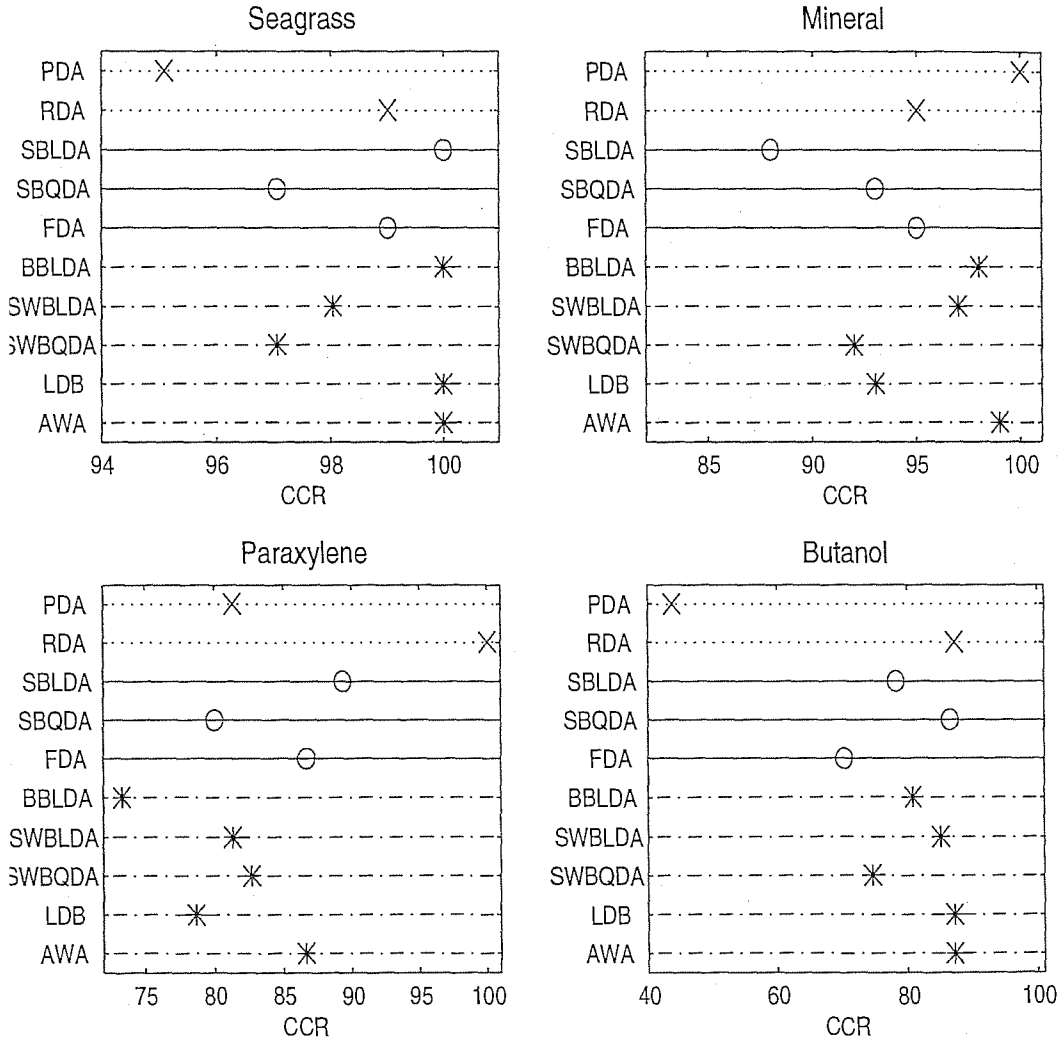


Figure 7.14: **Correct classification rates for each of the discriminant strategies.**

There are three main items which can be noted from the classification summary.

- RDA tends to outperform the other high dimensional classifier PDA, with the exception being for the mineral data where PDA outperforms RDA. If one compares RDA against the low dimensional methods which entail some form of feature extraction (or selection), then one can clearly notice that RDA also consistently produces high classification results across each of the sets of data. For the mineral data however, RDA is outperformed by three of the wavelet based approaches (and PDA).

- For the classification methods which are based on a subset of the original variables, there does not appear to be any one approach which consistently performs well.
- Of the classification methods which are based on a selection of the wavelet coefficients, the adaptive wavelet algorithm consistently produces high test classification rates. The AWA also tends to produce higher classification rates than the classification strategies which involve feature selection on the original variables. The only exception occurs for the paraxylene data where SBLDA on the original data assigns 89.33% of the observations for the test data to their correct class. The AWA also compares favourably with the high dimensional classification methods.
- There is no method which clearly outperforms the other methods, although the AWA does consistently classify a large proportion of the testing objects into their appropriate class categories.

In deciding which discriminant approach should be applied to the classification of spectral data it is important to identify the main goals of the discriminant procedure, that is the kind of information which is required. From the results presented in this section, if one is purely interested in assigning objects to their appropriate classes, then RDA and the AWA tend to consistently perform well. Sometimes, information other than the percentage of correctly classified objects is required. For instance, a common question relating to the discriminant analysis of spectral data is often posed – "which features are important for classification?" Some of the classification strategies which have been discussed in this chapter can assist in answering such questions, while other methods are really only useful for assigning spectra to their particular classes. The next section examines any additional information apart from correct classification rates that can be provided by each of the discriminant approaches.

7.5.2 Qualitative Assessment

This chapter has focused predominantly on the classification of spectral data. Discriminant analysis can involve more than just assigning an object into a particular class. In terms of description there are two main items which are often of interest in discriminant procedures. One item of importance is to determine which parts of the spectra are most useful for

discriminating among the various classes. Another item of interest is understanding how the different groups relate to one another as a whole. This can most easily be visualized with the aid of discriminant plots.

Some of the discriminant strategies discussed previously will be able to assist in uncovering such information while others will not. We proceed to investigate the qualitative information which can be obtained from the lower dimensional methods.

For deducing which variables may be important for classification, one typically relies on the features extraction strategies to determine such information. For example, BLDA may provide little knowledge as to which variables are important, but this classifier combined with a stepwise procedure can help to identify which variables contain discriminatory information.

If feature selection is being performed on the original variables, then it is possible to deduce the variables which are likely to contain discriminatory information by simply observing the variables which have been selected. When features other than the original variables are used then interpretation of important wavelengths becomes more involved. For instance, if we are to use wavelet coefficients, then it becomes slightly more difficult to say if a particular wavelength is important or not. What one can deduce from wavelet (and scaling) coefficients is (i) the kind of information which is useful e.g. the high frequency components or the low frequency components, and (ii) which regions of our original spectra are useful for classification. We now investigate the different approaches which can be used to highlight information about the discriminatory regions of our spectra. We consider each approach separately to highlight the capabilies of each strategy individually.

Low Dimensional Classification Using the Original Data

As previously mentioned when feature selection is based on the original data it is interesting to examine the actual features, i.e. wavelengths which have been selected. Figure 7.15 superimposes the wavelengths which have been selected by the stepwise routines SBLDA and SBQDA. The spectra shown in Figure 7.15 are the same sampled class spectra used previously in this chapter. Although, SBLDA and SBQDA will often select wavelengths pertaining to the same region of the spectra, there exists some variability with respect to the variables selected. This is not surprising since the variables selected by SBLDA will

classify using linear boundaries, and SBQDA will classify using quadratic decision boundaries. Recall from Section 7.3 that the classification strategies used for each of SBLDA and SBQDA is as follows:

Data	SBLDA	SBQDA
Seagrass	CF1	CF3
Mineral	CF3	CF3
Paraxylene	CF2	CF1
Butanol	CF1	CF1

Also shown in Figure 7.15 is the variables selected by FDA.

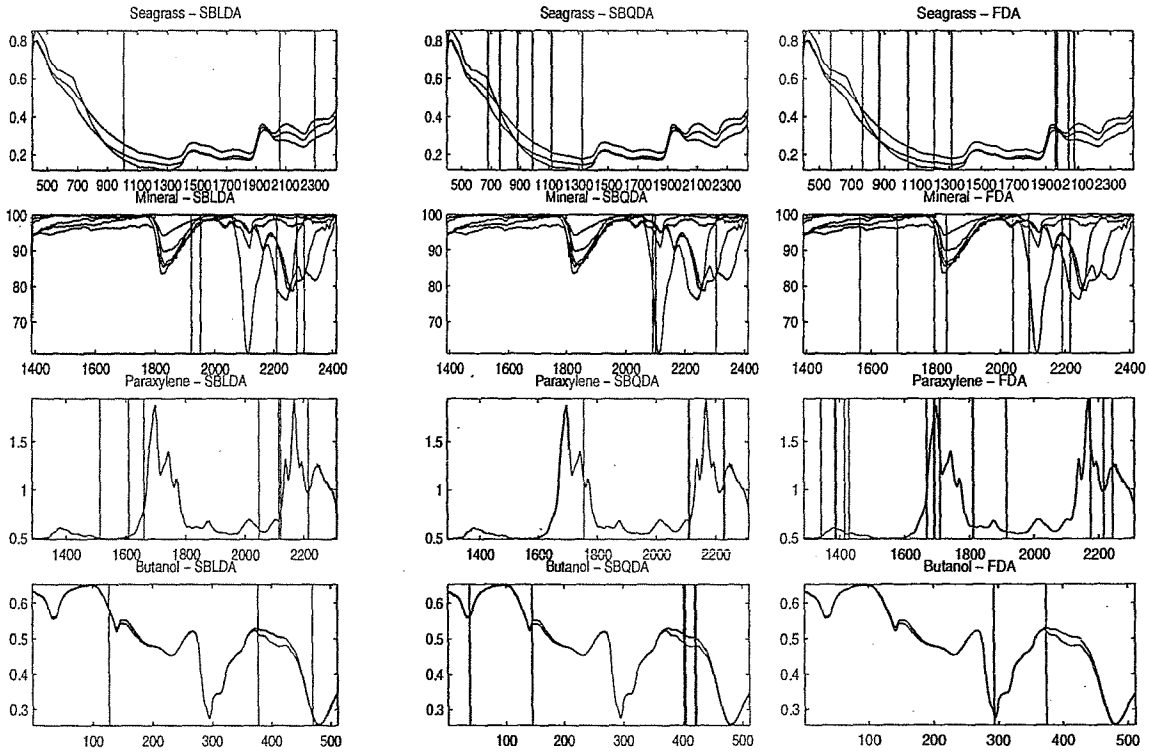


Figure 7.15: **Wavelengths selected by SBLDA, SBQDA and FDA.**

For the seagrass data in the approximate range of 500 nm to 1300 nm, FDA has selected variables similar to those selected from SBQDA. It has also selected variables nearer to the peak occurring around the 1900-2100 nm range. SBLDA selects variables from a similar region. For the mineral data, there seems to be two main areas of interest – those around the peak at 1800 nm (FDA, SBLDA) and in the range of 2100 – 2300 nm (SBLDA,

SBQDA, FDA). The wavelengths selected by SBLDA and SBQDA for the paraxylene data tend to concentrate around the region of approximately 2100 to slightly over 2200 nm, while FDA selects wavelengths pertaining to the peaks occurring in the vicinity of 1400 nm, 1600 – 1800 nm and 2150 – 2250 nm. It seems that for the butanol data, SBQDA has concentrated a selection of wavelengths pertaining to the index near 400. Both FDA and SBLDA do not concentrate heavily on this region and each select only a single wavelength from this region. Other areas of interest for the butanol data include the trough near index 130 (SBLDA, SBQDA) and the minima at index 300 (FDA).

With the exception of the butanol data, FDA had the tendency to select more variables than either SBLDA or SBQDA. There seems to be some slight differences to the particular wavelengths selected by SBLDA, SBQDA and FDA, although the variables selected by each of the methods often pertain to similar regions of the spectral data.

With FDA it also possible to obtain information about the segregation of the group categories. Since FDA stems from a Fisher-based method, it is possible to obtain discriminant plots. The discriminant plots for the FDA models are displayed in Figure 7.16 where the points in the plots are based on the testing data. The numerals represent their respective group categories. The seagrass and paraxylene data each have two discriminant variables, the butanol data has a single discriminant variable, the mineral data has four discriminant variables but we have only displayed the first three discriminant variables.

The discriminant plot of the seagrass data forms a v-shape. The discriminant plot for mineral data has reasonably separated each of the mineral groups. Whilst the discriminant plot for the paraxylene data show some separation of the three groups, there is obviously a great deal of spread in the plot, one can also observe the overlap in the classes for the butanol data.

We now proceed to discuss the qualitative information which can be determined from the classification strategies based on the wavelet (and scaling) coefficients. We discuss separately the output from the banded procedures BBLDA and BBQDA, the stepwise methods SBLDA and SBQDA, and the LDB and AWA algorithms.

Banded Discriminant Analysis

The banded approach is rather limited in what it can identify as important from the original variables. This is because the way in which we select the bands is based purely on

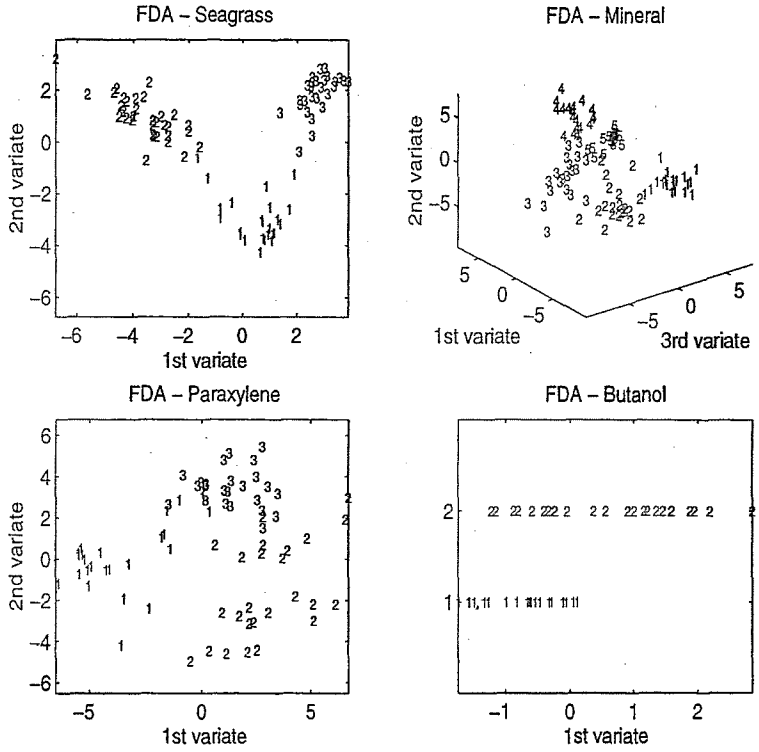


Figure 7.16: **Discriminant plots produced by FDA.**

the number of coefficients in the bands. It is necessary that the number of coefficients in the bands will not make the situation ill- or poorly-posed. What we can of course compare is if the scaling or wavelet bands produce more desired results, and one can then gain some information as to whether the low frequency or high frequency information may be more useful for classification. The coefficients from the bands which were supplied to BBLDA and BBQDA are shown in Figures 7.6-7.9.

Stepwise Discriminant Analysis

As for the stepwise methods based on the original data, it is interesting to determine which variables, or wavelet (and scaling) coefficients in this case, have been selected. We now present a figure which identifies the wavelet and scaling coefficients selected by the stepwise methods SWBLDA and SWBQDA when the following forward stepwise searches were implemented:

Data	SWBLDA	SWBQDA
Seagrass	CF3	CF3
Mineral	CF1	CF2
Paraxylene	CF3	CF3
Butanol	CF3	CF3

For each of the discriminant data sets, Figure 7.17 shows the wavelet coefficients at levels $j = 8$ to $j = 3$, and the scaling coefficients at level $j = 3$ of the discrete wavelet transform for the same sampled class spectra presented earlier. The coefficients selected by the stepwise procedures are then superimposed on the figures. The dotted lines shows the coefficients selected by SWBLDA and the dashed lines identifies the coefficients selected by SWBQDA. Sometimes both methods selected the same coefficients, so to make this more visible asterisks have also been plotted for the coefficients selected by SWBLDA, and circles at the coefficients selected by SWBQDA.

The original sampled spectra are shown in the top row of Figure 7.17 where the horizontal axis is labelled in nanometers (this information is not available for the butanol data). The reason for plotting the original data is to relate the regions of the original data with the selected wavelet and scaling coefficients. In an approximate sense, if vertical lines are extended from the asterisks and circles to the original spectra, then where the lines meet the original spectra will indicate the approximate region which is represented by the coefficients. This region is wider for the coefficients selected at lower levels of the DWT, and narrower for the coefficients selected at higher levels in the DWT.

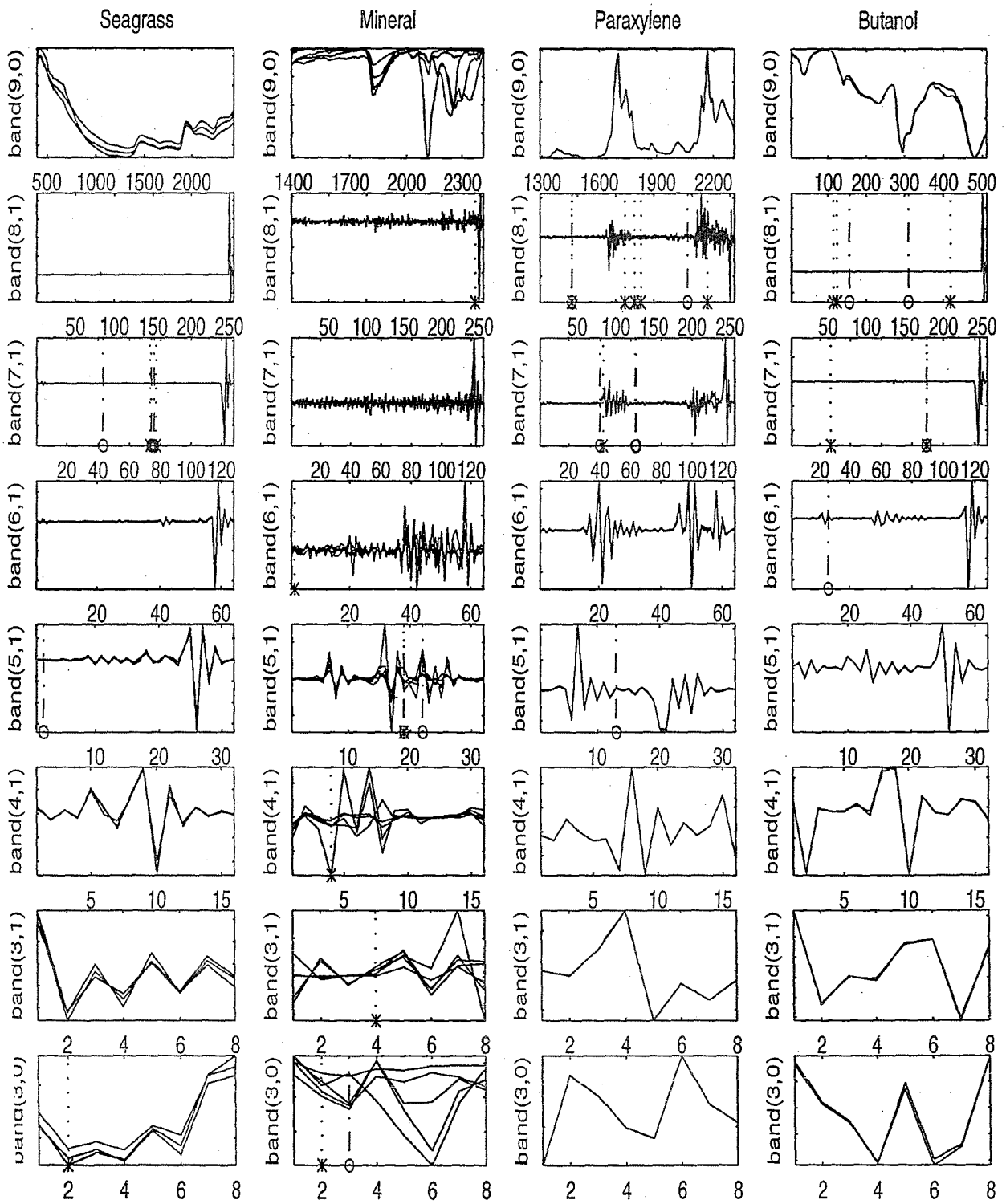


Figure 7.17: Coefficients from the DWT which were selected by SWBLDA (asterisk) and SWBQDA (circle).

Another interesting procedure which can be performed, is to reconstruct the spectra by backtransforming the coefficients which were selected by the stepwise procedures. The reconstructed spectra can however be a little difficult to interpret. For instance, the reconstructed spectra produced from the coefficients selected by SWBLDA for the seagrass and mineral data are similar. Their similarity can be attributed to the presence of a scaling term. If one examines the magnitudes of the reconstructed wavelet terms one can see the wavelet terms are quite small, and will have a minor effect on the reconstruction process, if a scaling term is present. When no scaling terms are present, the reconstruction procedure produces spectra which reflect the high frequency components of the spectra. The reconstructed spectra highlight the information represented by the selected coefficients. In an approximate sense, the classifiers will utilize the information from the regions of the reconstructed spectra that are not zero. For instance, SWBQDA performed on the paraxylene data tends to utilize information around the 1900 nm region (between the two major peaks) while SWBLDA utilizes the information nearer the two peaks at the regions 1700 nm and 2200 nm.

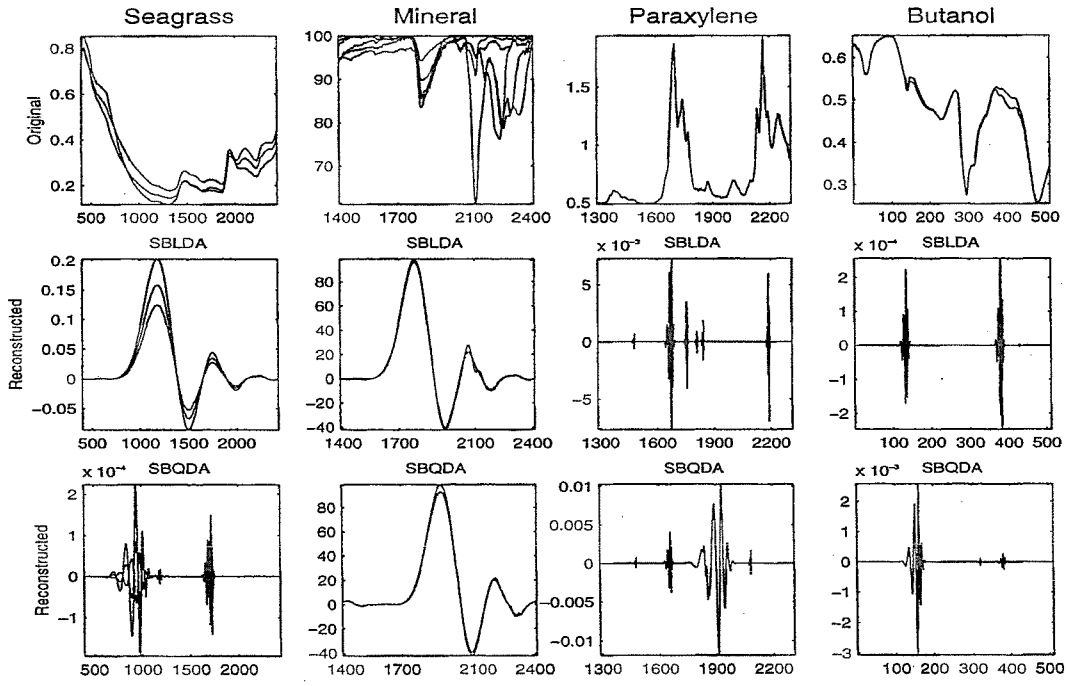


Figure 7.18: **Reconstructed spectra produced from the coefficients selected by SWBLDA and SWBQDA.**

The LDB Algorithm

Another wavelet feature extraction procedure applied is the LDB algorithm. The LDB algorithm utilizes a wavelet packet transform. Since the coefficients from the wavelet packet transform are obtained by passing the data in each of the bands through a low pass and a high pass filter, it can be more challenging to interpret the coefficients from the bands in the wavelet packet transform. With the exception of the coefficients selected from the paraxylene data (which were the original variables) the coefficients selected from the best basis produced from the LDB algorithm were mostly from the left hand tree (i.e. the DWT). Similar methods for displaying and interpreting the coefficients from the best basis can still be performed. That is the coefficients can be plotted against their index, and the nonzero coefficients can be backtransformed to produce the reconstructed spectra. To avoid reiteration we have elected not to produce such plots. Since most of the coefficients are from the DWT, coefficients from the LDB algorithm can be visualized by the use of Figures 7.11 and Figures 7.6-7.9.

The AWA Algorithm

We now proceed to the interpretation of the adaptive wavelet algorithm. The AWA performs a kind of banded selection process whereby coefficients pertaining to the band of the DWT are supplied to the classifier. The high pass (or low pass) filter coefficients are also constructed based on the classification of these 'banded coefficients'. We have previously mentioned that the banded approach is rather limited in what it can identify as important from the original variables, although it is interesting to identify if the scaling or wavelet bands produce more results that are more desirable.

Figure 7.19 shows the adaptive wavelet coefficients which produced the highlighted results in Table 7.11. Also shown are the reconstructed spectra, which were obtained by setting the coefficients in the remaining bands to zero, and then backtransforming the thresholded data.

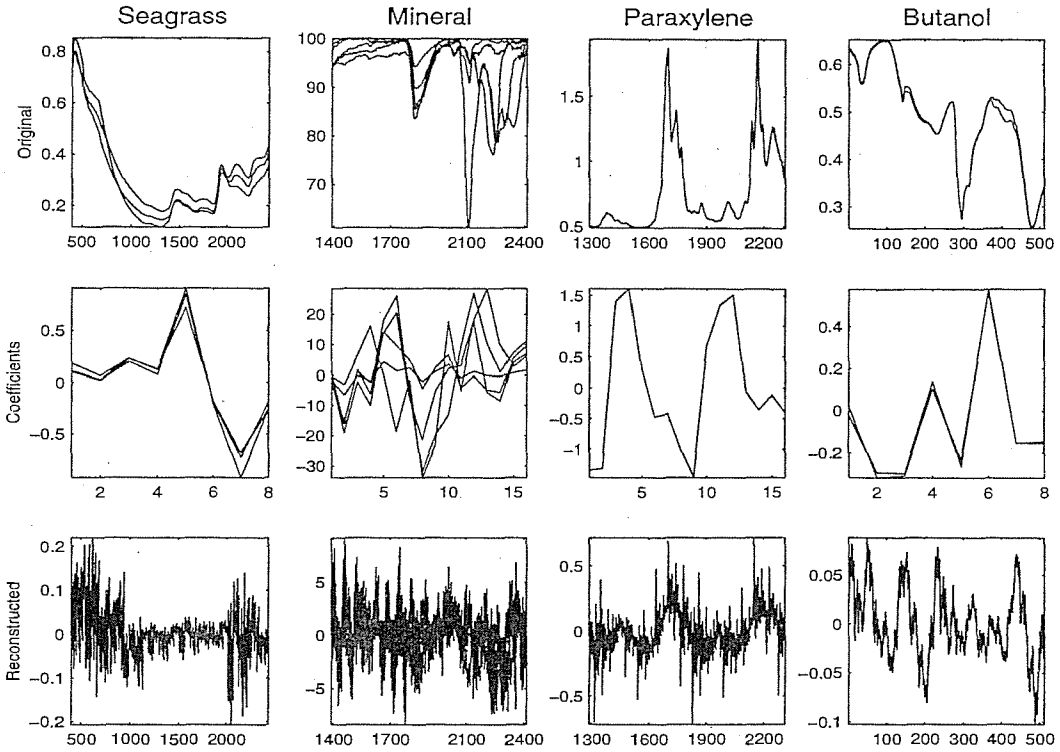


Figure 7.19: **The wavelet coefficients and reconstructed spectra produced from the AWA.**

The reconstructed spectra are very difficult to interpret for the AWA. When the reconstructed spectra are plotted in colour, the reconstructed spectra for the sampled class spectra from the mineral data are more distinguishable than presented here. The same can not be said however for the reconstructed seagrass, paraxylene and butanol data.

Here, one can look for the positions of the wavelet coefficients where they differ the most for the individual class spectra, and then see where these differences relate (approximately) to the original data. For instance with the butanol data, there appears to be some visible difference in the 2nd, 3rd and 4th coefficients produced from the sampled spectra. This indicates that useful discriminatory information might be in the approximate indices of 50 – 250. Again, only a vague interpretation can be provided, because only the coefficients from a single sampled spectra from each of the classes is shown, and there is likely to be some variability of the spectra within each of the classes. So whilst some separation is evident for the spectra which we have selected, there is no guarantee that this same separability can be visualized if other spectra were selected.

Discriminant plots were obtained for the adaptive wavelet coefficients which produced the highlighted results in Table 7.11. These are displayed in Figure 7.20. Although the classifier used in the AWA was BLDA, it was decided to supply the coefficients available upon termination of the AWA to FLDA, so we could visualize the spatial separation between the classes. The discriminant plots are produced from the testing data. There is a good deal of separation for the seagrass data, while for the paraxylene there is some overlap between the objects of class 1 and 3. The distinctness of the paraxylene data appears more evident for the AWA discriminant plots than in the FDA plots of Figure 7.16. Also, by comparison with Figure 7.16 we can see that Figure 7.20 achieves slightly more separation for the butanol data.

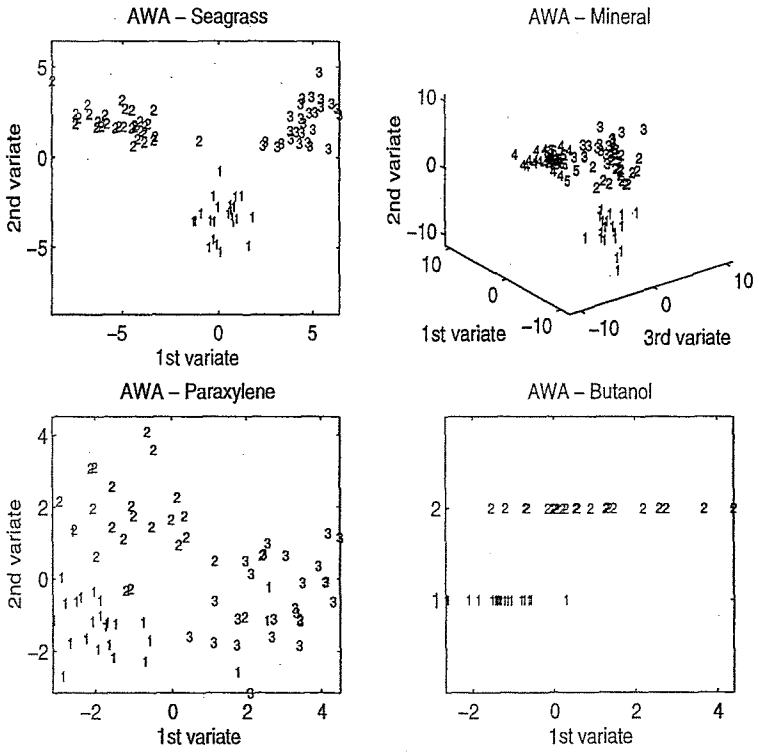


Figure 7.20: **Discriminant plots produced by from the coefficients produced by the AWA.**

High Dimensional Classification Methods

Consider now the high dimensional classifiers, PDA and RDA. Since RDA stems from Bayesian classification theory, the main information which can be extracted from a RDA model is how accurately it can assign objects into the respective classes. Also, by exam-

ining the parameters (a, b) chosen for the RDA model one can determine how much the pooled covariance matrix has been utilized as opposed to the class covariance matrix, and thus whether the RDA model is closer to a BLDA model or a BQDA model.

With PDA it also difficult to obtain information about which variables may contribute significantly to the discrimination of the various groups, but since PDA stems from a Fisher-based method, it is possible to obtain discriminant plots. The discriminant plots for the PDA models are displayed in Figure 7.21. The points in the space represent the testing objects from their respective classes.

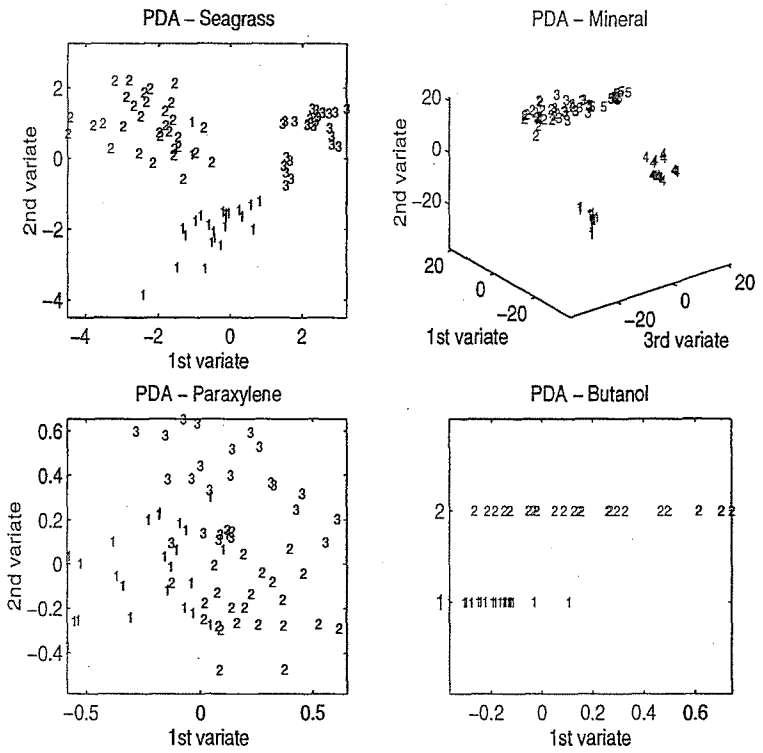


Figure 7.21: **Discriminant plots produced by PDA.**

For the mineral data, more groups appear to be easily recognized than those obtained from FDA when applied to the same data. Another interesting features which can be observed from the PDA discriminant plots is that for the seagrass data, the objects from class 3 appear in subclusters. The PDA discriminant plots for the paraxylene data is more scattered than that produced for FDA and AWA.

7.6 Summary

In this chapter we have investigated several discriminant approaches which can be applied to spectral data sets. The discriminant approaches which we have considered are:

1. classification using all of the original data and a high dimensional classifier,
2. applying a low dimensional classifier to a selection of the original variables, and
3. applying a low dimensional classifier to a selection of the coefficients from the DWT.

In practice, the particular discriminant method which is to be implemented, will ultimately depend on the goal of the discriminant analysis. The goals of discriminant analysis are twofold – to assign objects into a predefined group category, and to understand more about our data. This may include determining which regions are more important for discriminating, or how the groups are related, for instance one group maybe much easier to distinguish from the remaining classes.

We have mentioned that if assignment accuracy is the foremost goal of the discriminant procedure, then it could be worthwhile to apply RDA and the adaptive wavelet algorithm. In terms of extracting and interpretting information, the stepwise methods using either the original data or wavelet coefficients are relatively easy to understand. FDA is also relatively easy to interpret and has the added advantage of producing discriminant plots.

PDA, RDA, BBLDA, and the LDB and adaptive wavelet algorithms seemed to be less trivial to investigate which regions contain discriminatory information. PDA however was able to provide some information about the group structure of the data with the aid of discriminant plots. Discriminant plots were only available for FDA and PDA since these methods were based on Fishers linear discriminant analysis. It should be noted however, that in instances where BLDA was the classifier, discriminant plots may have been produced if the same data which was supplied to BLDA, was also supplied to Fishers linear discriminant analysis. This was done in Figure 7.20 for the adaptive wavelet coefficients.

Another item which should be addressed when considering the kind of discriminant procedure to be implemented is the computational expense which is inherited by the various methods. Although a comprehensive analysis of the computational complexities for the classification strategies was not undertaken, we would like to comment about our expe-

rience with run times of the various procedures. It is our experience that RDA and the AW algorithm tended to be more computationally expensive than the remaining methods. Perhaps this extra expense, contributed to the models performing quite well overall. These two methods involved some form of optimization routine. For RDA, this entailed finding the optimal (a, b) pairs, and for the adaptive wavelet algorithm this involved the development of a wavelet basis. The stepwise procedure which used the CF3 forward stepwise search was also computationally burdensome, even when fast updating formulae were implemented. If the computational time is an issue, then perhaps FDA, or one of the stepwise methods using a CF1 or CF2 approach could be applied. The DWT is relatively quick to calculate (faster than the fast Fourier transform), so the stepwise methods could be applied to the wavelet coefficients with minimal fuss (if a standard wavelet basis is used).

The next chapter applies regression methods to spectral data sets and follows a similar format to this chapter.

Chapter 8

Regression Applications

8.1 Overview

This chapter investigates various regression strategies when applied to spectral data of relevance to the agricultural industry. As for the previous chapter, the word strategy may refer to a regression method, a feature extraction technique, or a combination of the two.

Following an introduction to the data sets which are analyzed in this chapter, some regression methods which are commonly applied for the regression of spectral data will be investigated. This includes the application of partial least squares regression and two stepwise strategies. The first stepwise strategy simply enters the original variables into the multiple linear regression model, while the second is a stepwise procedure applied to the principal components and is referred to as stepwise principal component regression (SPCR). Regression analysis using features from the DWT is also investigated. We apply the DWT using the defined filter coefficients from the Daubechies family, as well as derived filter coefficients produced from the adaptive wavelet algorithm.

The structure and goals of this chapter will follow much the same format as that for the previous chapter on classification applications. The goals of this chapter are not necessarily to find the very best regression model, but to investigate the various regression approaches applicable to spectral data, and to assess quantitatively and qualitatively the advantages of each.

8.2 The Data Sets

Two data sets and three responses were used for evaluating the performance of the various regression procedures. These data sets will be referred to as the sugar and wheat data. A summary of each of these data sets are presented in Table 8.1. Here the number of spectra in each training and test set is displayed, as well as the response(s) which are to be modelled by each spectral data set. Further details about the sugar and wheat data are provided in Sections 8.2.1 and 8.2.2, respectively. The dimensionality of both data sets is $p = 512$.

Data Set	Train	Test	Responses
Sugar	100	89	brix (b), fibre (f)
Wheat	60	40	protein (p)

Table 8.1: **Description of the spectral data sets used for regression.**

8.2.1 Sugar Data

The sugar data was supplied by Dr Nils Burding at the Bureau of Sugar Experiment Station in Gordonvale. The training sugar data contains 100 digitized spectra for which log 1/reflectance was measured at the 512 wavelengths 916,918,...,1938 nm. The test set contains 89 spectra. Figure 8.1 shows five sample spectra from the sugar training data which were used to model the responses, brix and fibre. At 1100 nm there is a distortion which arises from a change in instrumentation. One detector is used to measure the radiation reflected for wavelengths less than 1100 nm and another detector is used to measure the radiation reflected for wavelengths greater than 1100 nm (inclusively). The change in receptors gives rise to the jump.

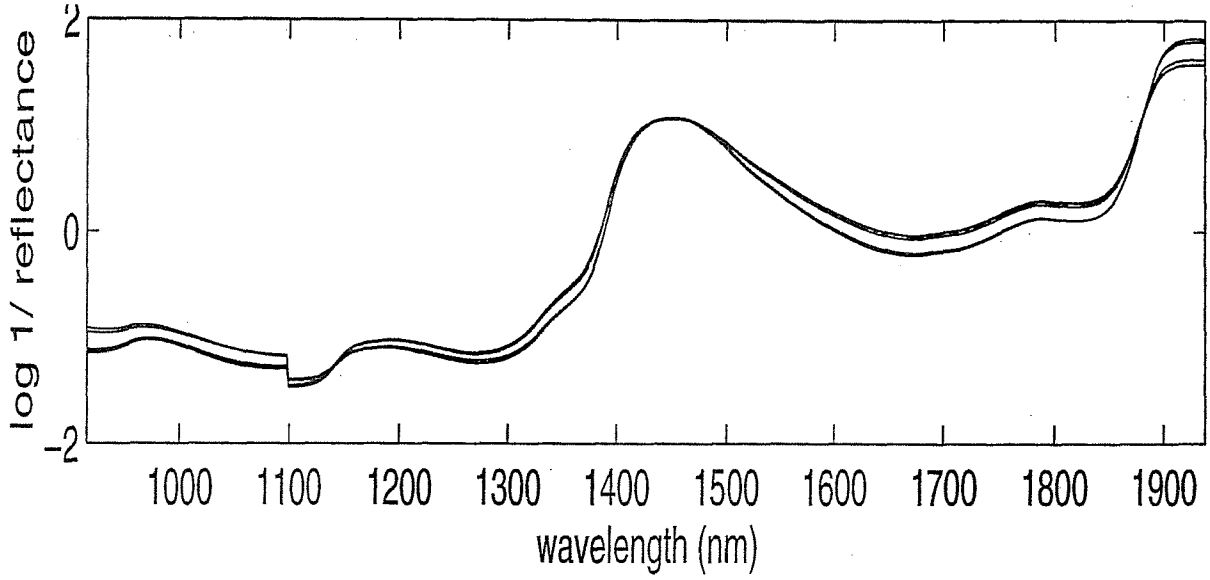


Figure 8.1: **Five sample spectra from the sugar data.**

8.2.2 Wheat Data

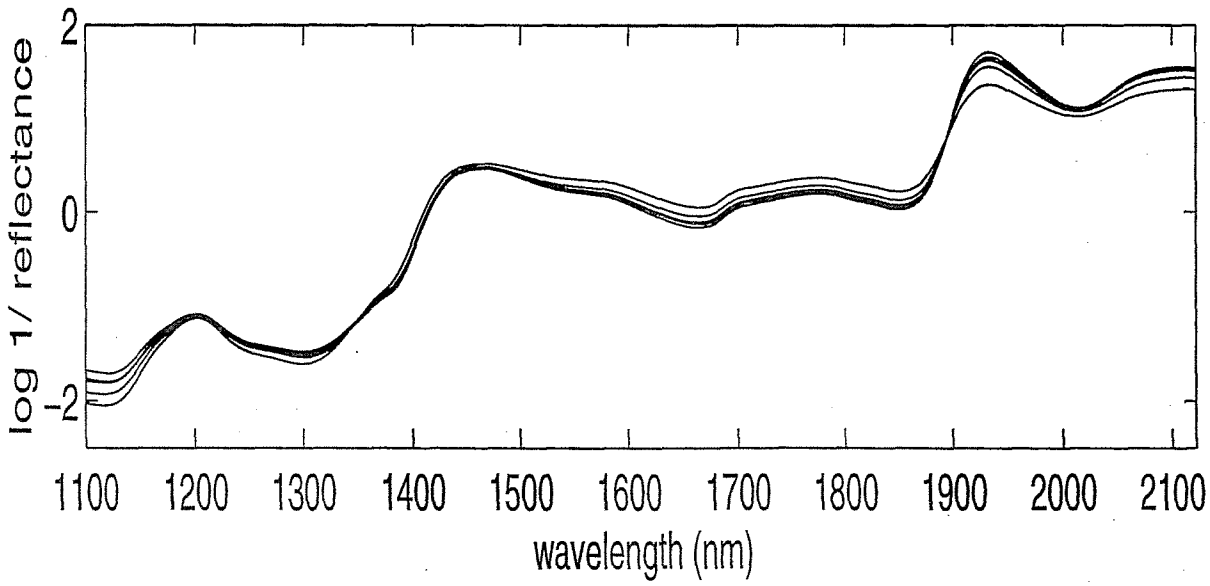


Figure 8.2: **Five sample spectra from the wheat data.**

The wheat data set was accessed from Professor Philip K. Hopke and has previously been discussed in literature, see for example [77]. The training wheat data contains 60 spectra for which log 1/reflectance was measured at the 512 wavelengths 1100,1102,...,2122 nm.

The test set contains 40 spectra. Figure 8.2 shows five sample spectra from the wheat training data which were used to model protein content.

8.3 Common Approaches for the Regression of Spectral Data

This section considers the performance of several regression methods. Multiple linear regression (MLR) is applied using the original variables. Since MLR is considered to be a low dimensional regression method, it is first necessary to reduce the dimensionality of the data. This is done by selecting the original variables by a stepwise routine. Principal component regression (PCR) is also applied where again, a stepwise routine is used to reduce the dimensionality of the problem by selecting a smaller set of the principal components. The stepwise routine implemented for SPCR is similar to that used for selecting the original variables for MLR, the main difference for SPCR, is that the stepwise routine is now selecting principal components as opposed to the original variables. This is essentially MLR using the principal components as features. Much of the literature refers to this technique as 'principal component regression' and this thesis adopts the same terminology. Partial least squares (PLS) regression is also applied. The regression strategies in this section have each been applied to centered data. That is, the independent variables and response variables have all been mean centered (see Section 4.2.1).

Table 8.2 shows the R-squared scores based on the training and test set for each of the regression strategies applied to all three responses. The figures typed in bold, highlight the largest R^2_{test} score for each of the regression procedures. Some clarification is now given for the column headings.

- SMLR-S1: stepwise MLR where stopping rule 'S1' is applied. The variable which gives the largest increase in the R^2_{train} enters the model. At each iteration all the variables in the model are tested for removal. Variables are removed if their t-statistic is less than 0.674. Since variables are removed from the model when their t-statistic is less than 0.674, this value is also referred to as the t-to-remove statistic. The procedure stops when no more variables can be retained in the model, or, when there are 16 variables in the model, which ever occurs sooner. In Table 8.2 all the models for SMLR-S1 had 16 variables.

- SMLR-S2: stepwise MLR where stopping rule 'S2' is applied. As for SMLR-S1, the t-to-remove statistic for a variable remains at 0.674, but the procedure stops when the change in R^2_{train} from one iteration to the next is less than 0.005 or when 16 variables have been entered into the model, which ever results sooner. Stopping rule 'S2' tended to produce simpler models which contained fewer variables, at the expense of a slight decrease in performance.

  Comment: the t-to-remove statistic of 0.674 may seem very low. Traditionally default values in statistical packages are set much higher. For the spectral data sets in this chapter, setting higher values for the t-to-remove statistic, tended to halt the algorithm with only 3-5 terms in the model, consequently the prediction performance of the model for both the training and test set was very poor. Setting the t-to-remove statistic at 0.674 reduced this effect and produced more superior results since it allowed for more contributing terms to be included in the model.

- SPCR: stepwise principal component regression. At each step, the principal component which produces the largest increase in R^2_{train} enters the model. At each iteration, all the principal components in the model were tested for removal. Principal components with a t-statistic less than 2.71 were removed from the model. The SPCR routine stops, when no more variables can be retained in the model, or, when 16 principal components are in the model, which ever occurs sooner. Note that we only consider selecting from the first 50 principal components.

  Comment: It was considered appropriate to have a higher t-to-remove value for SPCR, than for SMLR-S1 and SMLR-S2. When the t-to-remove value for SMLR (0.674) was used for the selection of principal components, the model tended to involve more terms than necessary. Consequently, it was decided to use a larger t-to-remove statistic of 2.71 for SPCR since this resulted in fewer terms being included in the model which produced equal or better measures of prediction.

- PLS-S3: performs partial least squares regression where the number of components is chosen by method 'S3'. That is, the number of PLS components to be retained in the model corresponds to the minimum PRESS statistic for the first 16 components only. For example, if for the first 16 components, the PRESS statistic was minimum

when 12 components were used, then the PLS model would have 12 components.

PLS-S4 : performs partial least squares regression where the number of components is chosen by method 'S4'. This method simply chooses the number of PLS components which corresponds to the highest R^2_{test}. As for PLS-S3, the maximum number of components which can be in a PLS model is 16. This is not a traditional approach for choosing the number of PLS components, since this approach would not be possible if an independent test set was unavailable. PLS-S4 was simply included in the table to show the very best results that could be obtained by PLS.

Data		SMLR-S1	SMLR-S2	SPCR	PLS-S3	PLS-S4
Brix	Train	0.981	0.964	0.966	0.976	0.977
	Test	0.963	0.960	0.953	0.971	**0.972**
	dimension	16	11	12	14	15
Fibre	Train	0.908	0.885	0.876	0.898	0.898
	Test	**0.820**	0.800	0.796	0.805	0.805
	dimension	16	12	11	15	15
Protein	Train	0.991	0.962	0.954	0.983	0.966
	Test	0.808	0.792	0.799	0.800	**0.832**
	dimension	16	6	13	16	14

Table 8.2: **Training and test R-squared values.**

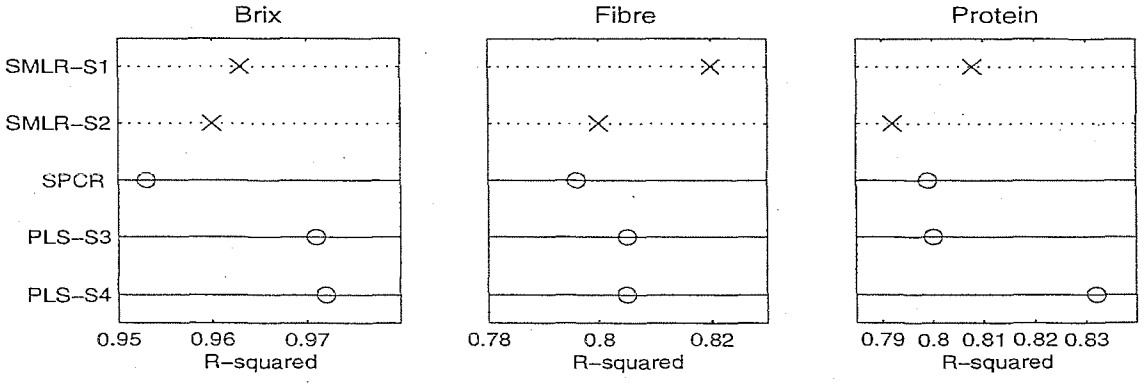


Figure 8.3: **Test r-squared values corresponding to the brix, fibre and protein responses.**

To facilitate interpretation of Table 8.2, Figure 8.3 was produced. In this figure, the R^2_{test} values obtained using SMLR-S1, SMLR-S2, SPCR, PLS-S3 and PLS-S4 have been plotted. The crosses indicate the results for both SMLR methods and the circles indicate the results for SPCR and the two PLS procedures. In the previous chapter, plots were produced which overlayed the results for each of the data sets. With the regression results for brix being much higher than those for fibre and protein, such an overlay of plots made interpretation less precise. Hence, this chapter presents the plots displaying the regression results for each response separately.

Consider first the stepwise MLR methods. Here, SMLR-S1 tends to produce higher R^2_{test} values than SMLR-S2. This is a likely consequence of SMLR-S1 having more terms than SMLR-S2. The simplified SMLR-S2 model usually outperforms SPCR, with the model for the protein response being the exception. Here, SPCR outperforms SMLR-S2, but not SMLR-S1. Both procedures for partial least squares consistently produce a higher R^2_{test} value than that obtained by SPCR. For the modelling of the protein response however, SPCR compares favourably with PLS-S3, but not with PLS-S4. Overall, the SMLR and PLS models are performing quite adequately.

Table 8.3 shows the variables which were included in each SMLR model as well as the principal components which were selected by SPCR.

Some of the variables selected by SMLR-S2 are not necessarily selected by SMLR-S1. This result occurs for the brix response because SMLR-S1 has allowed more terms to enter the model. It is possible that a variable which was part of the SMLR-S1 model at an early stage of the algorithm, maybe removed at a later stage. Conversely, the SMLR-S2 routine has the tendency to stop earlier, thereby retaining terms that the SMLR-S1 routine may remove. The wavelengths selected in Table 8.3 are examined further in Section 8.5.2.

It can be worthwhile to pursue other feature extraction strategies which may help to improve the performance of MLR. In the next section we investigate if the performance of MLR can be improved when the features are coefficients from the DWT.

8.4 Regression Analysis Using Features From the DWT

This section considers different procedures for selecting coefficients (both wavelet and scaling) from the DWT which are then supplied to MLR. The feature extraction methods

which are applied include banded multiple linear regression (BMLR). This procedure is similar to BBLDA and BBQDA, but instead of supplying coefficients to a classifier, BMLR supplies the coefficients to MLR. Another feature extraction method which involves the use of the DWT, is stepwise MLR. Here, the variables selected are the scaling and wavelet coefficients from the DWT. This technique will be referred to as SMLRW. A stepwise procedure which involves the selection of principal components will also be investigated. Here, PCA is first performed on a selection of the coefficients from the DWT. This technique will be referred to as SPCRW. For BMLR, SMLRW and SPCRW the DWT is performed using a Daubechies filter defined by 16 coefficients. For SMLRW and SPCRW, the DWT is performed to level 3, for BMLR, the scaling and wavelet coefficients from level 3 and 4 are extracted. The adaptive wavelet algorithm will also be applied using a criterion function of relevance to regression. Prior to implementing the wavelet feature extraction methods,

Brix			Fibre			Protein		
SMLR-S1	SMLR-S2	SPCR	SMLR-S1	SMLR-S2	SPCR	SMLR-S1	SMLR-S2	SPCR
1880	1880	2	1114	1114	1	1178	1178	1
1136	1136	8	1410	1410	2	1254	1254	4
1324	1324	6	1128	1128	19	1300	1300	3
1506	1506	10	1324	1324	4	1280	1280	13
1840	1840	3	1098	1098	9	1878	1878	6
1758	1758	1	1276	1276	18	1934	1934	16
1674	1674	12	1080	1080	13	1306		14
1028	1028	9	1066	1066	8	1932		2
1448	1448	14	1082	1082	20	2078		9
1630	1630	11	1456	1456	17	2092		5
1878		13	1884	1884	29	2098		21
1678		16	1516	1516		2108		11
1822			1428			1930		20
1778			1398			1162		
1748			1388			1172		
1638			1274			1314		

Table 8.3: **Wavelengths selected by the SMLR routines, and the principal components selected by SPCR.**

we first examine the coefficients from the DWT obtained for the sugar and wheat spectral data sets.

8.4.1 Exploring the DWT

Figures 8.4 and 8.5 were constructed by applying the DWT to a single spectrum from the sugar and wheat data. At the time the DWT was performed, these spectrum were in their original (uncentered) form. The scaling and wavelet coefficients from each of the bands in the DWT are plotted against their index. The reconstructed spectra are shown in the second and fourth columns. Here the inverse DWT has been applied to the bands of scaling and wavelet coefficients, respectively. One distinguishing feature in Figure 8.4 is that the wavelet coefficients have detected the change of instrumentation which occur at the 1100 nm mark of the original spectra. For a more detailed description about the construction of Figures 8.4 and 8.5, the reader is referred to Section 7.4.1 where similar figures were constructed for the classification of spectral data sets.

8.4.2 Banded Multiple Linear Regression (BMLR)

Banded multiple linear regression (BMLR) uses all of the coefficients from the same band in the wavelet transform, as input to the regression technique, MLR. The particular bands of coefficients which are supplied to MLR are the scaling $\mathbf{X}^{[3]}(0)$ and wavelet $\mathbf{X}^{[3]}(1)$ coefficients from level 3, and the scaling $\mathbf{X}^{[4]}(0)$ and wavelet $\mathbf{X}^{[4]}(1)$ coefficients from level 4. Here, the wavelet transform was produced using the Daubechies wavelet with 16 filter coefficients ($N_f = 16$). The DWT was performed on the original (uncentered data), but the coefficients and response variables were centered, prior to them entering the MLR model. The R^2_{train} and R^2_{test} values for each of the responses are displayed in Table 8.4 where the largest R^2_{test} for each of the responses has been highlighted. Due to numerical instabilities, it was not possible to obtain regression results for the protein model when the scaling coefficients from band(4,0) were supplied to MLR. This problem arises from the condition number of the matrix $\left(\mathbf{X}^{[4]}(0)^T\mathbf{X}^{[4]}(0)\right)$ having a large condition number (3.133e+17). Care should also be taken when interpreting the results for the scaling coefficients from the wheat data in band(3,0). Here, the coefficients $\left(\mathbf{X}^{[3]}(0)^T\mathbf{X}^{[3]}(0)\right)$ also had a relatively large condition number (2.0937e+16). (The condition number of $\left(\mathbf{X}^{[3]}(1)^T\mathbf{X}^{[3]}(1)\right)$ was 5.5276e+03).

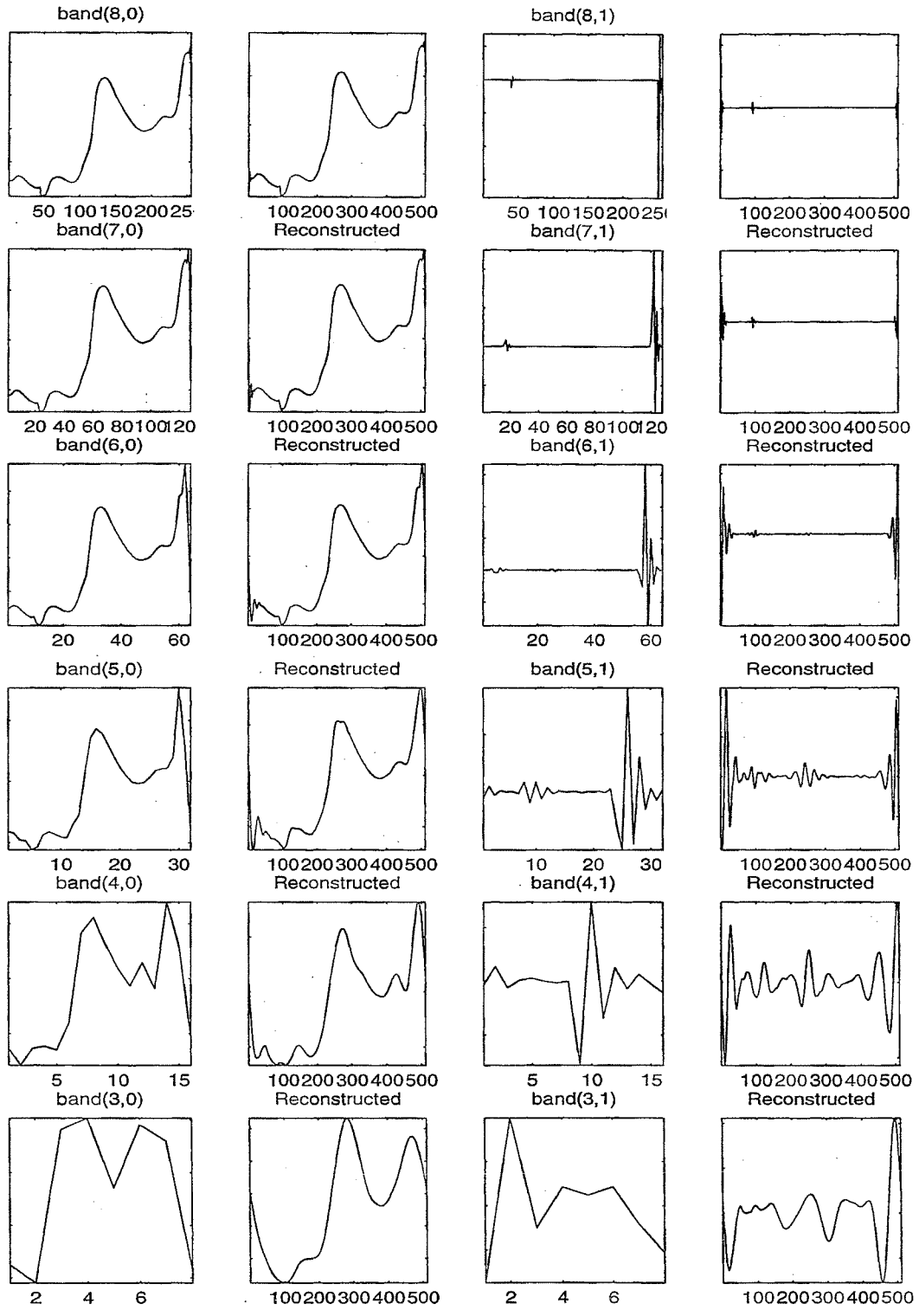


Figure 8.4: The DWT and inverse DWT performed on the sugar data.

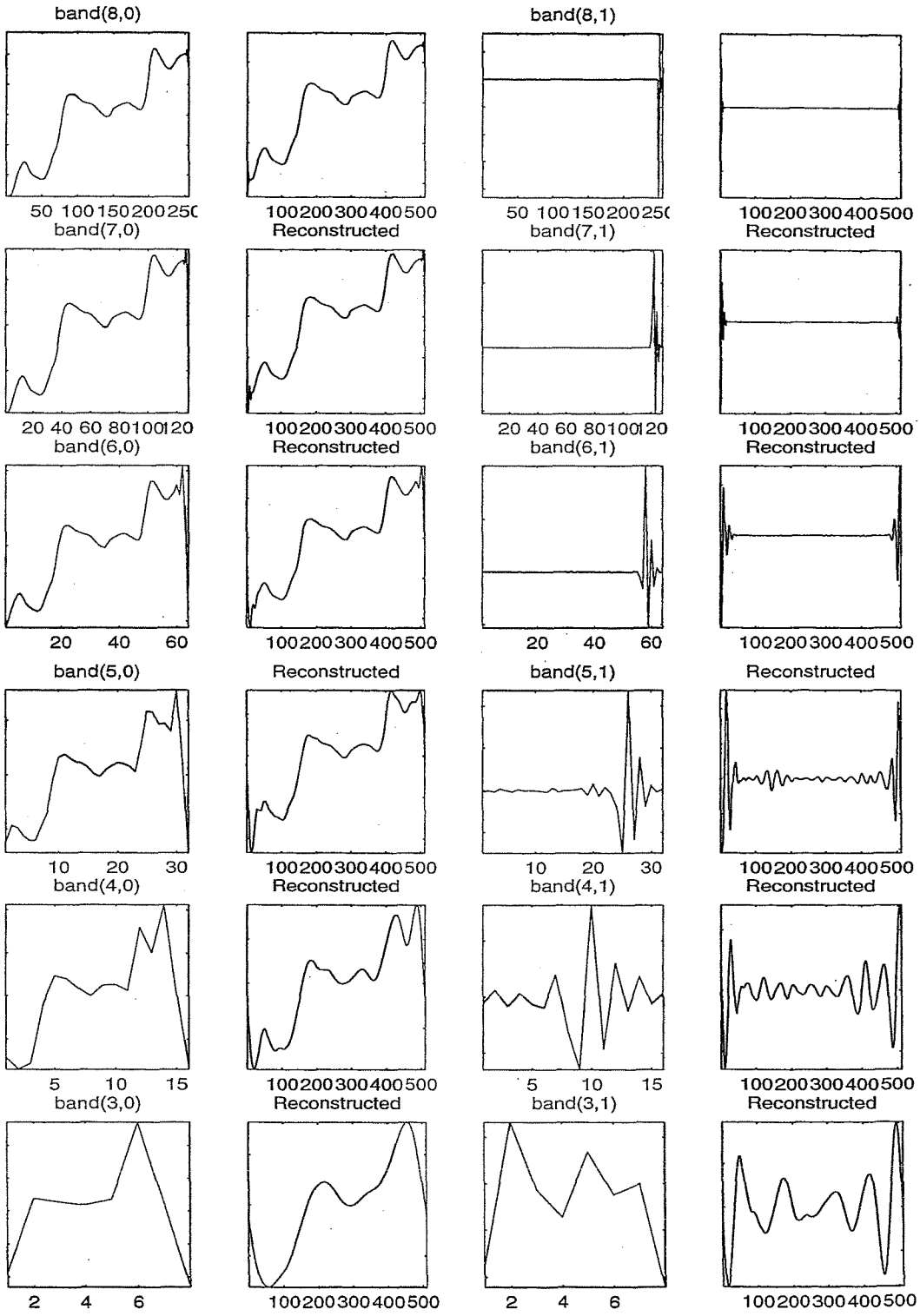


Figure 8.5: **The DWT and inverse DWT performed on the wheat data.**

Data		$X^{[4]}(0)$	$X^{[4]}(1)$	$X^{[3]}(0)$	$X^{[3]}(1)$
Brix	Train	0.975	0.961	0.740	0.525
	Test	**0.973**	0.949	0.753	0.530
Fibre	Train	0.781	0.797	0.647	0.707
	Test	0.692	**0.723**	0.533	0.569
Protein	Train	-	0.952	0.763	0.795
	Test	-	**0.704**	0.263	0.108

Table 8.4: **Classification results for banded BLDA.**

8.4.3 Stepwise Feature Extraction

In this section, two stepwise strategies are investigated. The first will involve stepwise selection of wavelet coefficients from the DWT, and the second will involve stepwise selection of principal components, where the principal components have been calculated from a selection of wavelet and scaling coefficients.

Stepwise Feature Extraction from the DWT (SMLRW)

SMLRW applies a stepwise strategy which selects coefficients from the DWT. The DWT is applied to the original (uncentered) spectral data sets to level 3 using a Daubechies wavelet defined by 16 filter coefficients. The total set of features consists of the scaling coefficients at level 3, and the wavelet coefficients at level 3 up to and including the wavelet coefficients at level 8. The wavelet and scaling coefficients, along with the response variables are mean centered. SMLRW will be applied using two stopping rules. These are the same stopping rules previously implemented by SMLR (see Section 8.3) except, the t-to-remove value for SMLRW is now set at 2.71. The two SMLRW applications will be referred to as SMRLW-S1 and SMLRW-S2, the results for which are displayed in Table 8.5. Both SMLRW-S1 and SMLRW-S2 produce the same results, since the same model for each technique was produced. SMLRW produces reasonable results for the training response values but performs much less adequately when predicting the test response values, particularly for fibre and protein.

Table 8.6 shows the indices of the coefficients which have been selected from the DWT. Figure 8.6 identifies where the selected coefficients lie in relation to the DWT. For each

Data		SMLRW-S1	SMLRW-S2
Brix	Train	0.886	0.886
	Test	0.767	0.767
	dimension	6	6
Fibre	Train	0.765	0.765
	Test	0.451	0.451
	dimension	6	6
Protein	Train	0.958	0.958
	Test	0.500	0.500
	dimension	6	6

Table 8.5: **R-squared values for SMLRW-S1 and SMLRW-S2.**

Brix		Fibre		Protein	
SMLRW-S1	SMLRW-S2	SMLRW-S1	SMLRW-S2	SMLRW-S1	SMLRW-S2
23	23	25	25	43	43
50	50	111	111	44	44
18	18	107	107	65	65
293	293	214	214	135	135
211	211	241	241	212	212
153	153	449	449	286	286

Table 8.6: **Coefficients selected from the DWT by SMLRW-S1 and SMLRW-S2.**

of the models no coefficients were selected from level 3 in the DWT. Another common occurrence is that each of the response models contained 6 terms. After this, no more variables (coefficients) could be entered into the model which had a t value greater than the specified t-to-remove value of 2.71.

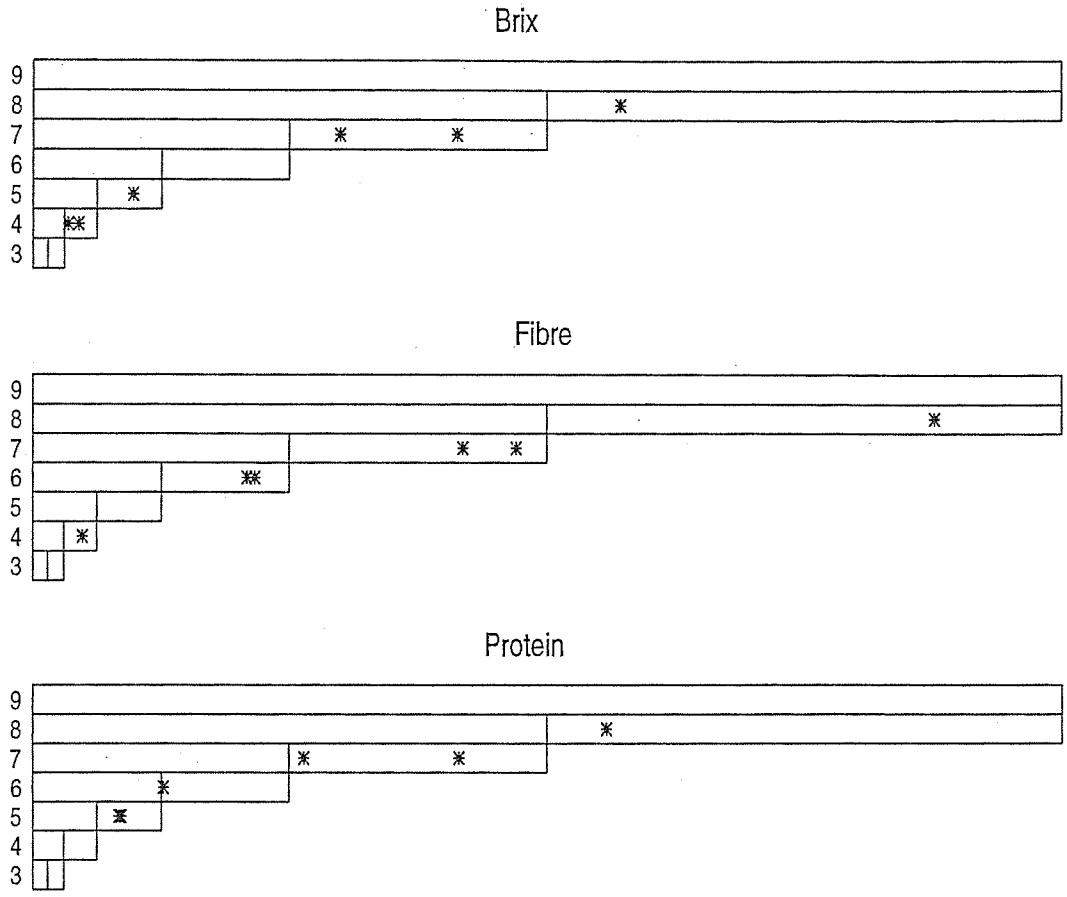


Figure 8.6: **Coefficients selected from the DWT by SMLRW.**

Stepwise Feature Extraction from the Principal Components (SPCRW)

This section applies a stepwise procedure which involves the selection of principal components where the principal components are formed from a reduced subset of coefficients from the DWT. This technique will be referred to as SPCRW. Principal component analysis is performed on the coefficients from the DWT which have an absolute correlation coefficient of more than 0.5 with the response. Stepwise PCR is then performed using methods 'S1' and 'S2' as described for SMLRW (also with the same t-to-remove value of 2.71).

The SPCRW procedure is summarized as follows:

1. Perform the DWT on the original (uncentered) data to level 3 of the transform.

2. Measure the correlation between the coefficients indexed for each $j \in \{3, \ldots, 8\}$

$k \in \{0, \ldots, 2^j - 1\}$ pair.

3. Perform PCA on the the coefficients which have an absolute correlation of more than 0.5 with the response.

4. Select the principal components using a stepwise routine. Selection is from the first 50 principal components only.

The results of the SPCRW routines are presented in Table 8.7, where again both stopping rules produce the same model. For each response variable SPCRW stops at the tenth iteration with 10 principal components in each of the models. Table 8.8 shows the principal components which enter the models.

Data		SPCRW-S1	SPCRW-S2
Brix	Train	0.936	0.936
	Test	0.891	0.891
	dimension	10	10
Fibre	Train	0.825	0.825
	Test	0.664	0.664
	dimension	10	10
Protein	Train	0.934	0.934
	Test	0.656	0.656
	dimension	10	10

Table 8.7: **R-squared values for SPCRW-S1 and SPCRW-S2.**

8.4.4 Adaptive Wavelet Algorithm

We also apply the adaptive wavelet algorithm (AWA) to the regression spectral data sets. The AWA is applied with similar settings as those used in Section 7.4.5 of the previous chapter. The (m, q, j_o) settings for which the AWA is applied, are again (4,3,2), (4,2,2), (8,1,1), (2,5,3), (2,5,4), (2,7,3), and (2,7,4). The most conceivable difference between the AWA when applied for regression (as opposed to classification) is the criterion function which is implemented. Here, the cross-validated R-squared criterion which is based on the

Brix		Fibre		Protein	
SPCRW-S1	SPCRW-S2	SPCRW-S1	SPCRW-S2	SPCRW-S1	SPCRW-S2
1	1	1	1	1	1
4	4	2	2	3	3
5	5	4	4	6	6
8	8	23	23	4	4
6	6	21	21	16	16
3	3	9	9	13	13
2	2	8	8	2	2
7	7	22	22	17	17
11	11	19	19	12	12
18	18	32	32	10	10

Table 8.8: **Components selected from the DWT by SPCRW-S1 and SPCRW-S2.**

PRESS statistic is the regression criterion which is implemented by the AWA. A similar banded selection strategy used for classification is used for regression. Here, each band at some level j_o in the DWT, the band (i.e. τ) which produces the largest regression criterion measure ($\mathcal{J}_{\text{CVRSQ}}\left(\mathbf{X}^{[j]}(\tau)\right)$) forms the basis of the optimization routine. The same coefficients are later supplied to MLR. If the algorithm chose to optimize over a scaling band (i.e. $\tau = 0$), then for the same (m, q, j_o) settings the experiment was repeated, where optimization was over the wavelet band producing the largest CVRSQ measure at initialization. The optimization routine halted if 2000 iterations of the optimization routine had been performed or sooner if an optimal value was obtained.

The results of the adaptive wavelet algorithm for each setting are presented in Table 8.9. For each response, the largest R^2_{test} is typed in boldface. For the brix response the (2,5,4) setting produced the same results (to 3 decimal places) for both the scaling and wavelet bands, indicating that low frequency and high frequency components perform well for this setting. When the fibre response was modelled using the AWA, the best setting in terms of the R^2_{test} measure was (2,5,4). The best results for the wheat data were also obtained with the (2,5,4) setting where optimization was over a wavelet band.

Brix	m	q	j_o	N_f	N_{coef}	τ	Train	Test
	4	3	2	16	8	1	0.955	0.949
	4	2	2	12	8	3	0.977	0.967
	8	1	1	16	8	3	0.972	0.968
	2	5	3	12	8	1	0.927	0.930
						2	0.971	0.969
	2	5	4	12	16	0	0.975	**0.971**
						1	0.975	**0.971**
	2	7	3	16	8	1	0.950	0.946
	2	7	4	16	16	1	0.976	0.968
Fibre	m	q	j_o	N_f	N_{coef}	τ	Train	Test
	4	3	2	16	8	2	0.791	0.676
	4	2	2	12	8	0	0.721	0.636
						2	0.855	0.799
	8	1	1	16	8	5	0.872	0.801
	2	5	3	12	8	0	0.718	0.638
						1	0.731	0.603
	2	5	4	12	16	1	0.869	**0.814**
	2	7	3	16	8	0	0.703	0.612
						1	0.777	0.641
	2	7	4	16	16	0	0.863	0.794
						1	0.868	0.737
Protein	m	q	j_o	N_f	N_{coef}	τ	Train	Test
	4	3	2	16	8	0	0.677	0.260
						3	0.959	0.671
	4	2	2	12	8	2	0.937	0.781
	8	1	1	16	8	6	0.970	0.797
	2	5	3	12	8	1	0.781	0.369
	2	5	4	12	16	1	0.975	**0.825**
	2	7	3	16	8	1	0.838	0.365
	2	7	4	16	16	1	0.974	0.790

Table 8.9: **Regression results for the adaptive wavelet algorithm.**

Figure 8.7 plots $\mathcal{J}_{\text{CVRSQ}}\left(\mathbf{X}^{[j]}(\tau)\right)$ against the iterations of the optimization routine. This was done for the $(2, 5, 5, 2)$ setting for the brix response, the (2,7,5,2) setting for the fibre response and the (2,5,5,2) setting for the protein response. The CVRSQ values at initialization (and completion) were lower for fibre, than those for the brix and protein responses. The same can be said when comparing the R^2_{test} measures for fibre, implying that the AWA performed more adequately for the brix and protein responses, as was the trend with most of the regression applications.

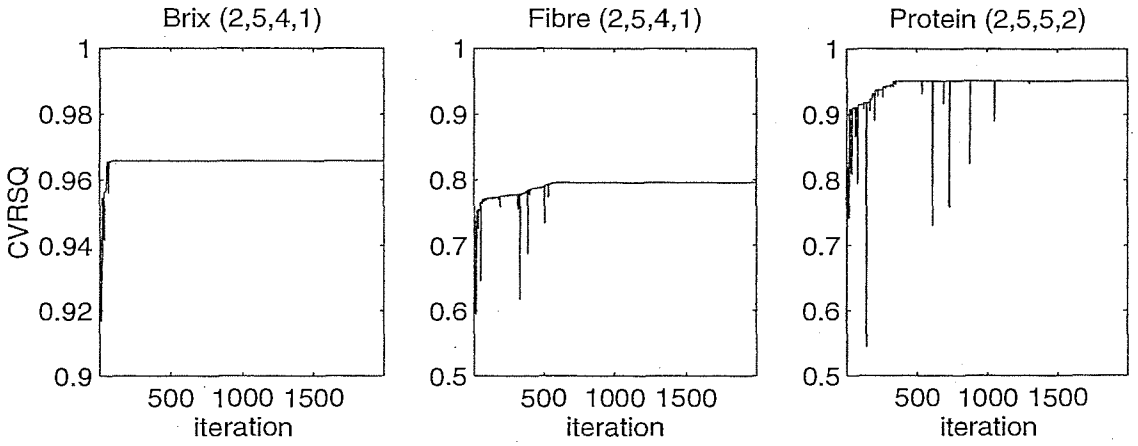


Figure 8.7: **Regression criterion measure versus iteration for the adaptive wavelet algorithm.**

One noticeable feature for the protein response in Table 8.9 is the extremely low R^2_{test} values occurring for the (4,3,2,0), (2,5,3,1), and (2,7,3,1) settings. This is a consequence of the extremely high condition numbers for the matrices $\mathbf{X}^{[j]}(\tau)^T\mathbf{X}^{[j]}(\tau)$ for each of these (m, q, j_o, τ) settings. The condition numbers of the matrices for the (4,3,2,0), (2,5,3,1), and (2,7,3,1) settings are 2.9260e+16, 4.9224e+16 and 1.7691e+16, respectively.

8.4.5 Summary of Wavelet Based Feature Extraction Strategies

In this section we summarize the results which were obtained by the wavelet based regression approaches. This is done in Table 8.10. The first application of a regression procedure involving the DWT was the banded multiple linear regression procedure (BMLR). The results which have been tabulated for BMLR are when the coefficients from band(4,0) were used for modelling brix, and when the coefficients from band(4,1) were used for the fibre

and protein responses. Also shown in Table 8.10 are the results from each of the stepwise procedures – SMLRW and SPCRW. Recall from Section 8.4.3 that each of the stopping procedures which were applied in conjunction with SMLRW and SPCRW, produced the same model. The results which have been presented for the AWA correspond to the highlighted values in Table 8.9.

Data		BMLR	SMLRW	SPCRW	AWA
Brix	Train	0.975	0.886	0.936	0.975
	Test	**0.973**	0.767	0.891	0.971
	dimension	16	6	10	16
Fibre	Train	0.797	0.765	0.825	0.870
	Test	0.723	0.451	0.664	**0.814**
	dimension	16	16	10	16
Protein	Train	0.952	0.958	0.934	0.975
	Test	0.704	0.500	0.656	**0.825**
	dimension	16	6	10	16

Table 8.10: **Training and test r-squared values for the wavelet based regression approaches.**

Based on the R^2_{test} measures, most of the regression procedures produce reasonable results when modelling the brix response, particularly BMLR and the AWA. For the modelling of the fibre and protein responses, the AWA appears to outperform the other wavelet based regression methods, in terms of the R^2_{test} value. This is more clearly seen in Figure 8.8 which displays the R^2_{test} values from Table 8.10 for each response separately.

8.5 Which Regression Strategy?

This chapter has presented several regression strategies for calibrating spectral data. Some traditional approaches have been explored as well as some modern feature extraction techniques. In this section we wish to elaborate further on the regression procedures applied thus far. This will be done in two parts for each of the response models. The first part will involve an assessment of some performance based measures. The second part

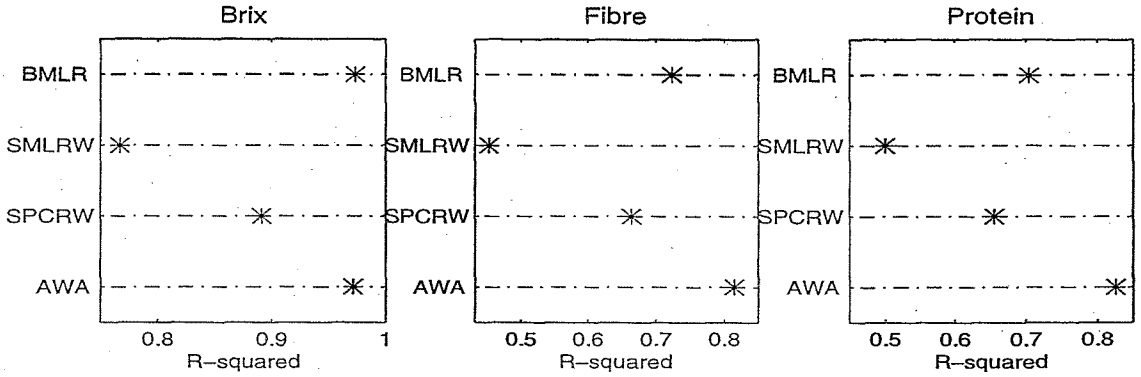


Figure 8.8: **Test r-squared values for the wavelet based regression methods.**

will explore the additional descriptive information which can be obtained about our data from each of the regression strategies. Particular attention is focused on which regions of the spectra maybe informative for regression purposes.

8.5.1 Performance Based Measures

In this section we summarize the R^2_{test} measures for each of the regression strategies presented in this chapter. We will also examine the p-values associated with each of the models and their coefficients. Plots of the residuals and fitted versus observed values will also be examined.

R-squared Values (Test)

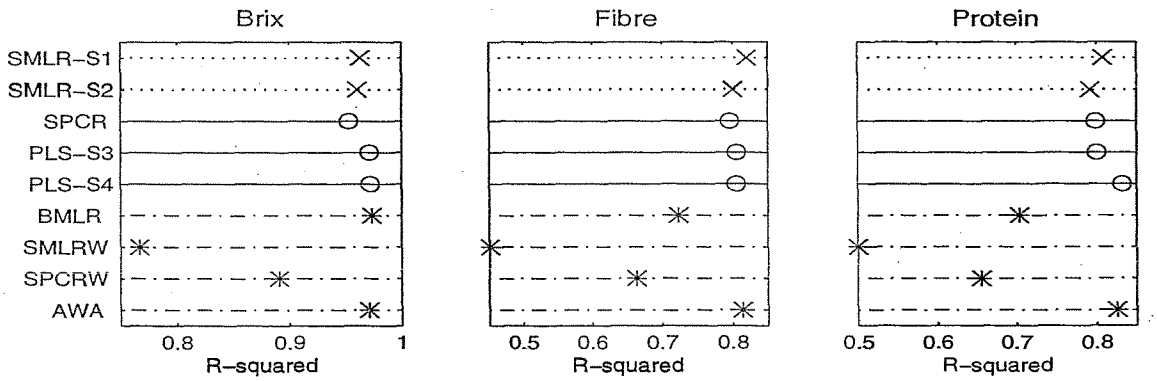


Figure 8.9: **Test r-squared values each of the regression strategies.**

Figure 8.9 displays the R^2_{test} values for each regression strategy. The wavelet based stepwise procedures, SMLRW and SPCRW produce reasonably low R^2_{test} values for each

of the response models. To allow for easier interpretation of the remaining methods Figure 8.10 has been produced without the R^2_{test} values displayed for these techniques. The

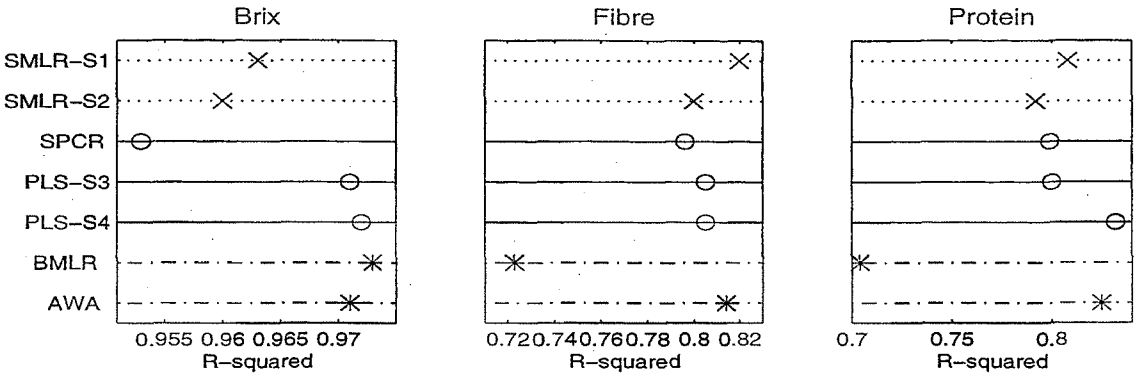


Figure 8.10: **Test r-squared values each of the regression strategies (SMLRW and SPCRW not shown).**

following information can be conveyed:

- There is no method which consistently produces the highest R^2_{test} measures across all response models.
- For the brix response, the wavelet methods BMLR and AWA, perform well as does the PLS-S4 procedure.
- The stepwise techniques (SMLR-S1) for fibre seem to perform the most adequate (in terms of the test measure) and again, the PLS method produces the next highest R^2_{test} score.
- PLS-S4 gives the highest R^2_{test} measure closely followed by the AWA for the protein models.
- The PLS approaches consistently perform well as does the AWA method.

Model Assessment

An extensive summary of regression diagnostics for most of the models discussed throughout this chapter are presented in Appendix A. All regression models with the exception of PLS-S4 have been presented, since these results will resemble closely those for PLS-S3. The statistics which are presented for the overall model include the residual standard error,

the F-statistic and the corresponding p-value. The statistics which are presented for the individual terms in the models include the regression coefficients (coef), the standard error of the coefficients (std.err), the t-statistic (t.stat) and corresponding p-value (p.value) for each of the coefficients.

All of the regression models are shown to be significant, but the same can not be said for all of the regression coefficients in the model. There are some regression coefficients which are not considered to be significantly different from zero, if we use a significance level of 0.05. Table 8.11 summarizes the p-values for each of the terms for the models shown in Appendix A. The notation 'Vi' means the ith term (or variable) in the model. Note that both AWA models for brix have been shown. The first column is for the model when optimization was over the scaling band, and the second column is for the model when optimization was over the wavelet band. The p-values in Table 8.11 which are larger than 0.05 have been shaded.

Table 8.11 presents the p-values separately for the non-wavelet (top) and wavelet (bottom) based regression methods. For the non-wavelet based methods, the SMLR-S1 contains several insignificant terms for Brix, two insignificant terms for fibre and one insignificant term for protein. One of the effects of changing the stopping rule for SMLR-S1 to SMLR-S2, is a reduction in the number of insignificant terms. For SMLR-S2 there is only one insignificant term at the 0.05 level, this is for the brix response.

For the wavelet based strategies, insignificant terms are apparent for the procedures which involve some banded selection process. Many of the coefficients for BMLR are not significant. This phenomenon can be attributed to the problems which we have previously discussed about band selection. That is, we are selecting the band not because we know it contains useful information for regression, but rather because the number of coefficients in the band is convenient for regression purposes. For the adaptive wavelet algorithm, this problem is also present.

When there are insignificant terms in the model, a common procedure is to refit the model in absence of these variables. This is performed in an attempt to obtain a simpler, yet just as effective model (or more effective in some cases) for explaining the sample variability. This can be seen as a second-stage procedure to the regression analysis. In this thesis we are primarily concerned not with finding the very best model, but investigating

	SMLR-S1			SMLR-S2			SCPR			PLS		
TERM	BRIX	FIBRE	PROT	BRIX	FIBRE	PROT	BRIX	FIBRE	PROT	BRIX	FIBRE	PROT
V1	0.000	0.000	0.145	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
V2	0.001	0.035	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
V3	0.000	0.000	0.002	0.113	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
V4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
V5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
V6	0.068	0.000	0.015	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000
V7	0.032	0.000	0.002	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000
V8	0.072	0.000	0.001	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000
V9	0.000	0.216	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000
V10	0.244	0.007	0.000	0.000	0.006		0.000	0.000	0.000	0.000	0.000	0.000
V11	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.001	0.000	0.000	0.000
V12	0.309	0.039	0.000		0.001		0.000		0.001	0.000	0.000	0.000
V13	0.000	0.008	0.001						0.002	0.000	0.000	0.000
V14	0.001	0.000	0.000							0.033	0.000	0.000
V15	0.009	0.009	0.003								0.039	0.000
V16	0.015	0.126	0.012									0.000

	BMLR			SMLRW			SPCRW			AWA			
TERM	BRIX	FIBRE	PROT	BRIX	FIBRE	PROT	BRIX	FIBRE	PROT	BRIX	BRIX	FIBRE	PROT
V1	0.763	0.642	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.824	0.000	0.000	0.729
V2	0.045	0.045	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.752	0.001
V3	0.000	0.006	0.498	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.353	0.777
V4	0.010	0.123	0.054	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.003	0.137
V5	0.000	0.450	0.210	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000
V6	0.000	0.000	0.391	0.001	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.590
V7	0.000	0.001	0.011				0.000	0.000	0.000	0.000	0.000	0.488	0.000
V8	0.000	0.110	0.499				0.000	0.001	0.000	0.056	0.000	0.001	0.000
V9	0.007	0.722	0.055				0.002	0.001	0.000	0.000	0.000	0.010	0.141
V10	0.000	0.813	0.247				0.007	0.004	0.006	0.000	0.000	0.221	0.000
V11	0.000	0.593	0.081							0.000	0.146	0.847	0.000
V12	0.000	0.000	0.071							0.000	0.109	0.304	0.105
V13	0.698	0.003	0.369							0.015	0.068	0.404	0.000
V14	0.000	0.106	0.589							0.152	0.136	0.013	0.000
V15	0.937	0.738	0.619							0.049	0.078	0.000	0.000
V16	0.082	0.495	0.087							0.359	0.049	0.000	0.000

Table 8.11: **Summary of p-values for the regression models.**

the performance of the various regression strategies on a first-stage basis, that is without modifying the regression model. Whilst this thesis does not consider the removal of terms and refitting the various models (at a second-stage level), we wish to make the reader aware, that if it is necessary to find the 'best' regression model, then such a procedure is worthy of investigation.

Residual and Fitted Value Plots

A residual analysis is another procedure which can be useful for determining which changes might be useful to make to a model. If the residuals are plotted against the independent variables, or the vector of fitted (i.e. estimated/predicted) response values, then some clues can be given about the suitability of the model, or if additional terms should be incorporated into a model. For example if some second degree curvature is present when the residuals are plotted against some variable, then it might be worth adding a quadratic term for that variable to the model. For a more detailed review of residual analysis, the reader is referred to [29, 106]. We do not actually plot the residuals against each of the variables in the model, since with so many models, this would be quite cumbersome. Here we will examine:

1. Plots of the residuals versus the fitted values.

2. Histograms of the residuals.

3. Plots of the fitted values versus the observed values.

Figures 8.11–8.13 show the residuals versus the fitted values for each of the responses (brix, fibre and protein), separately. A plot which has observations scattered about the horizontal line passing through the origin (that is the line $y = 0$) is preferred to one that has any structure. To help visualize if any structure is present in the data, a line has been superimposed on the residual plots. This line represents a smoothed version of the residuals and is called a smooth. If there is little structure in the residuals, then the smooth should resemble closely the horizontal line $y = 0$. The residual plots for the brix response hint at some slight curvature, but in general, there is little evidence of any structure among the residuals.

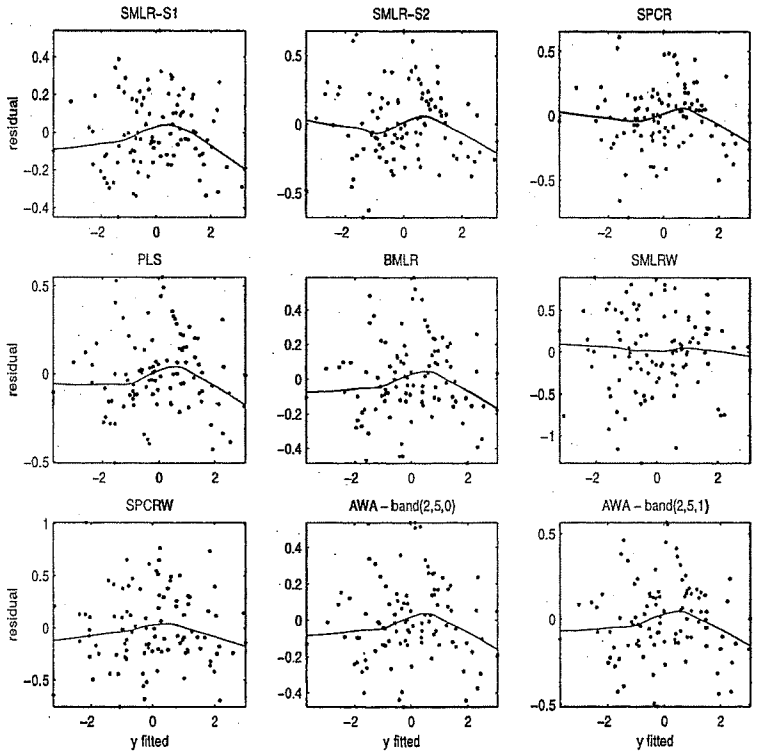


Figure 8.11: **Residuals versus fitted values for the brix response models.**

Figures 8.14–8.16 shows histograms of the residuals. Ideally, the residuals should follow a normal distribution with a mean of zero and constant variance. Generally the histograms produced for the brix response in Figure 8.14, appears to be normal, with the exception of the residuals from the SMLRW model, which has many values spread relatively constant in the tails of the distribution, and some clustered around the zero point. The residuals for both the SMLRW and SPCRW models exhibit are larger degree of variability, than the remaining models. The histograms shown in Figure 8.15 for the residuals of fibre are less symmetric than those in Figure 8.14, though some bell-shaped appearance is present in most plots. Histograms in Figure 8.16 for protein portray a similar outcome to the histograms in Figure 8.14, with the residuals from the SMLRW and SPCRW models having a wider range of residuals than the remaining methods.

The next selection of plots show the fitted values versus the obtained values. Although plotting the fitted values against the actual values is not technically considered to be part of a residual analysis, it is interesting to observe how close the predicted values are to the actual values. The closer the objects conform to a straight line the better the predicted

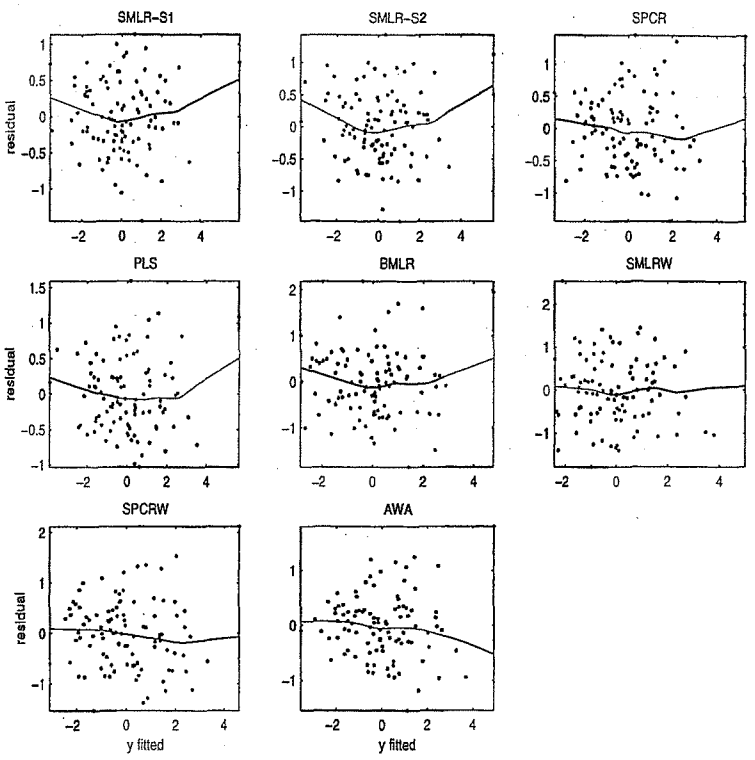


Figure 8.12: **Residuals versus fitted values for the fibre response models.**

alues. Figures 8.17–8.19 plot the fitted training response values against the actual training esponse values. One can observe again, that the scattering for the SMLRW and SPCRW nodels appears to be greater than those for the remaining models.

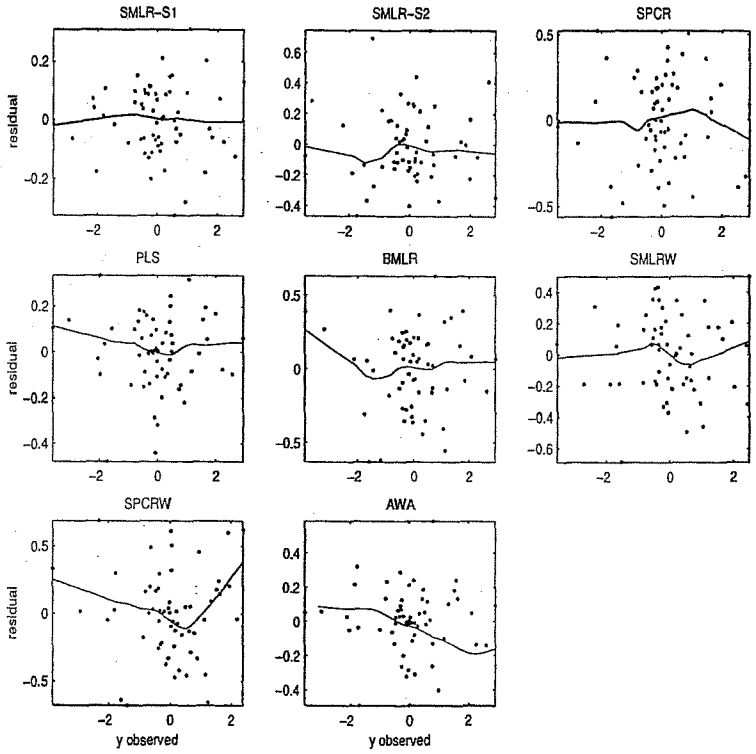


Figure 8.13: **Residuals versus fitted values for the protein response models.**

8.5.2 Qualitative Assessment

This section investigates the additional information that can be obtained from our data by using each of the regression strategies which have been discussed throughout this chapter. Of particular interest, are the regions or features which are important for regression. Of course there are no fixed guidelines for determining such information, and indeed the different regression strategies may even utilize different information. It is possible however, that some regression methods may suggest that similar regions or features are more useful than others. From this some subjective conclusions can be drawn regarding which wavelengths, regions or particular features maybe more important than others.

The additional information which can be obtained from SMLR, SPCR and PLS are discussed. We also consider the qualitative information which can be obtained by the wavelet based regression strategies. The regression methods which have been applied in conjunction with the DWT are BMLR - banded MLR, SMLRW - stepwise MLR on the coefficients from the DWT, where the DWT consisted of the scaling and wavelet coefficients

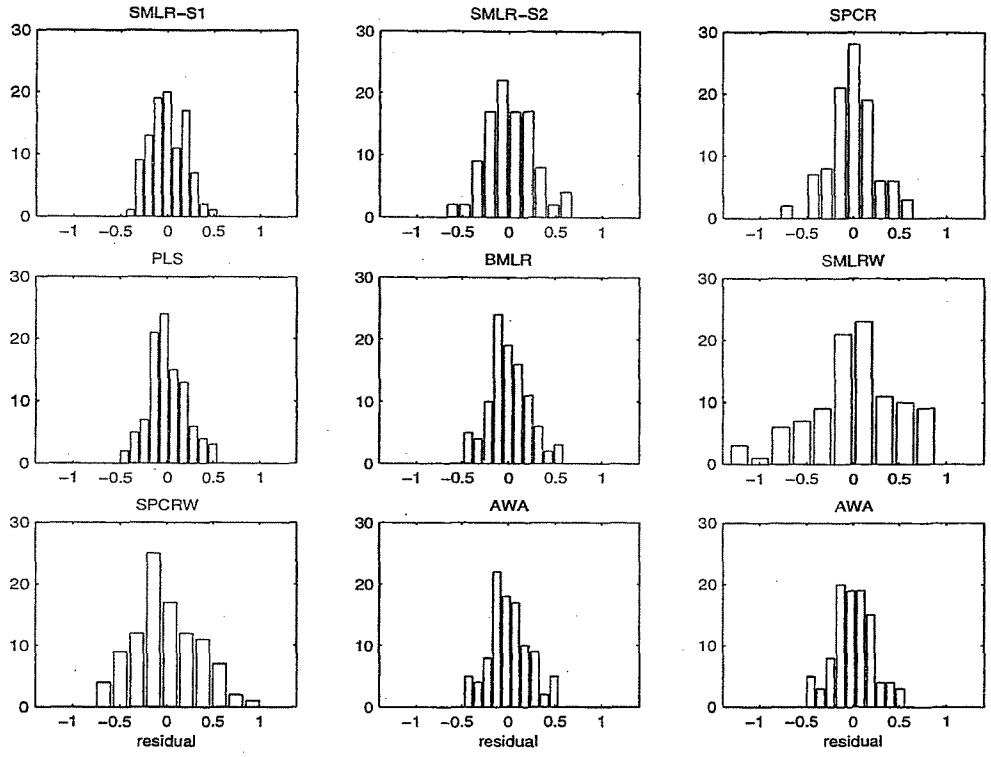


Figure 8.14: **Histograms of the residuals from the brix response models.**

at level 3, the wavelet coefficients at level 4 up to and including the wavelet coefficients at evel 8, SPCRW - SPCR where PCA has been applied to the coefficients from the DWT and the final wavelet based regression method which was applied is the AWA - adaptive wavelet algorithm. We consider each of the wavelet based strategies separately, and note that much of the same techniques as those discussed in in Section 7.5.2 in the chapter on classification applications have been used here for extracting information of relevance to regression analysis.

SMLR

We consider first the stepwise methods applied to the original data. Here, it is of interest to simply observe which variables have been selected by the SMLR models for each of the data sets. In Section 8.3 two SMLR models were investigated SMLR-S1 and SMLR-S2. The variables selected by SMLR-S1 and SMLR-S2 are displayed in Figure 8.20. The wavelengths selected for the modelling of brix are spread widely across the spectrum while areas for the fibre and protein responses are more concentrated at particular features. For

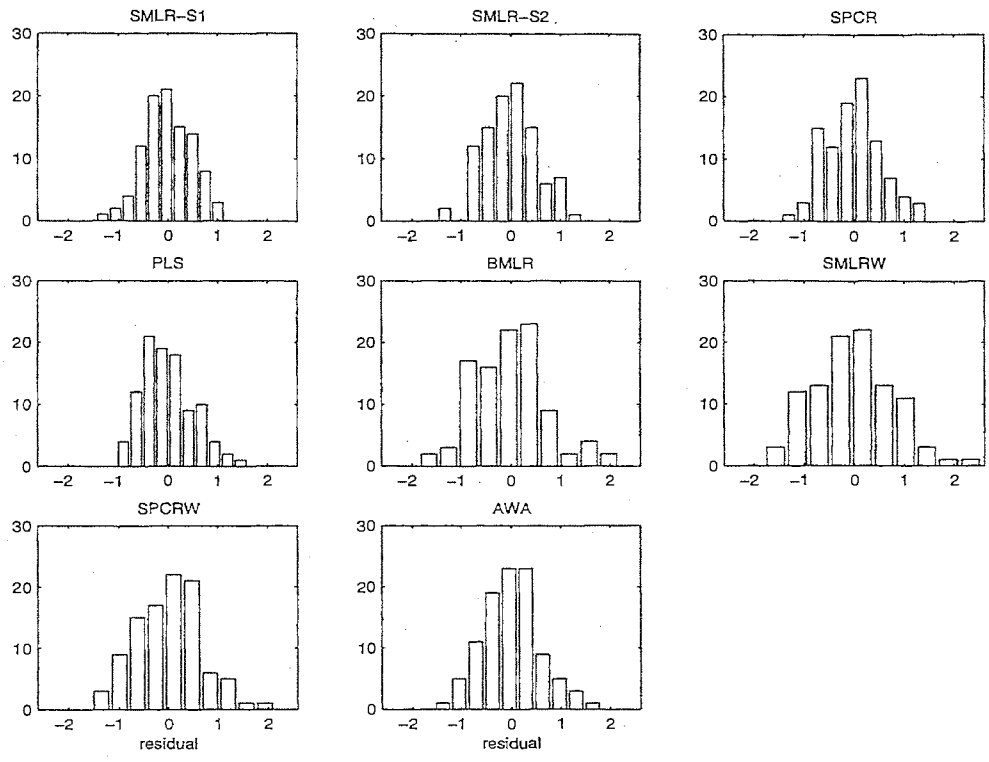


Figure 8.15: **Histograms of the residuals from the fibre response models.**

the modelling of fibre several wavelengths prior to and following the sharp discontinuity at 1100 nm have been selected as well as the region near the peak at 1400 nm. Unlike the wavelengths selected for the brix response, the fibre response does not utilize information from the 1600-1800 nm region. The concentrated regions for the protein stepwise models occur in the trough around the 1300 nm mark,

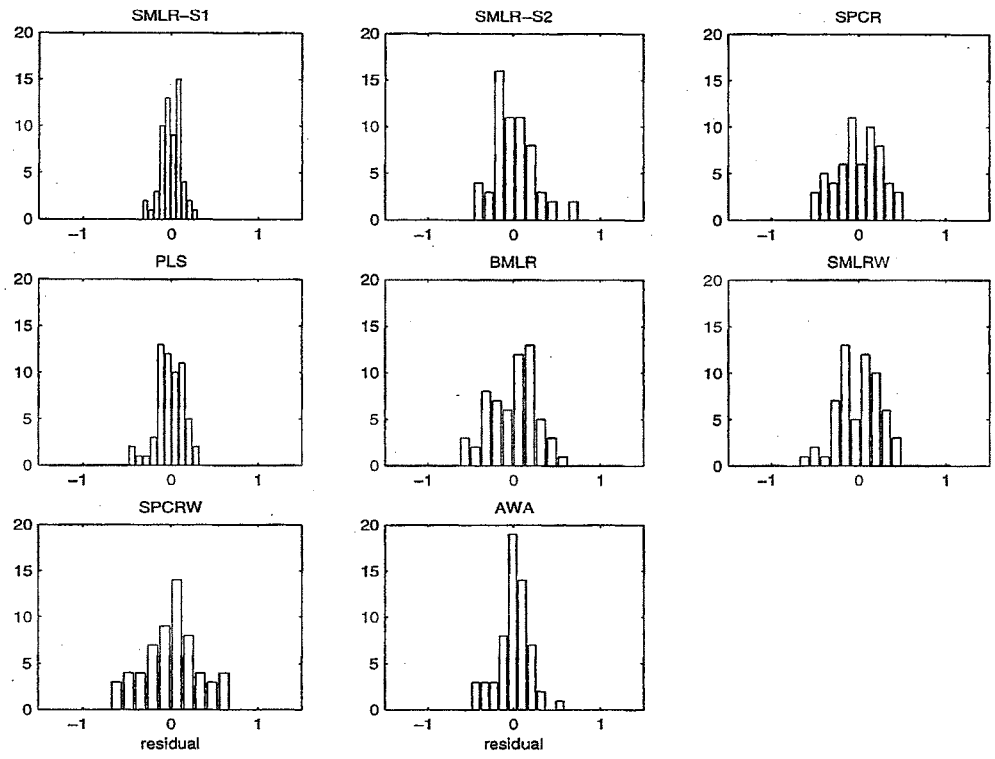


Figure 8.16: **Histograms of the residuals from the protein response models.**

SPCR

Stepwise PCR was also applied where PCA was performed on the original data, and the principal components were then selected by a stepwise routine. Since PCA involves linearly combining the original variables it is slightly more difficult to directly link which variables or wavelengths are important. Of course if we only had two principal components, we could simply examine biplots to note which variables had high loadings, but with several significant principal components in the model, this becomes less of an option. What could be done however is to establish which variables have a strong dependence on the principal components, and then if it is known which principal components are important for regression, we can deduce which variables form an important role in the SPCR model.

To achieve this, an average absolute correlation measure has been calculated. This involves calculating the correlation between each variable with the each of the selected principal components. If $\hat{\rho}_{ij}$ denotes the correlation between the ith variable and the jth

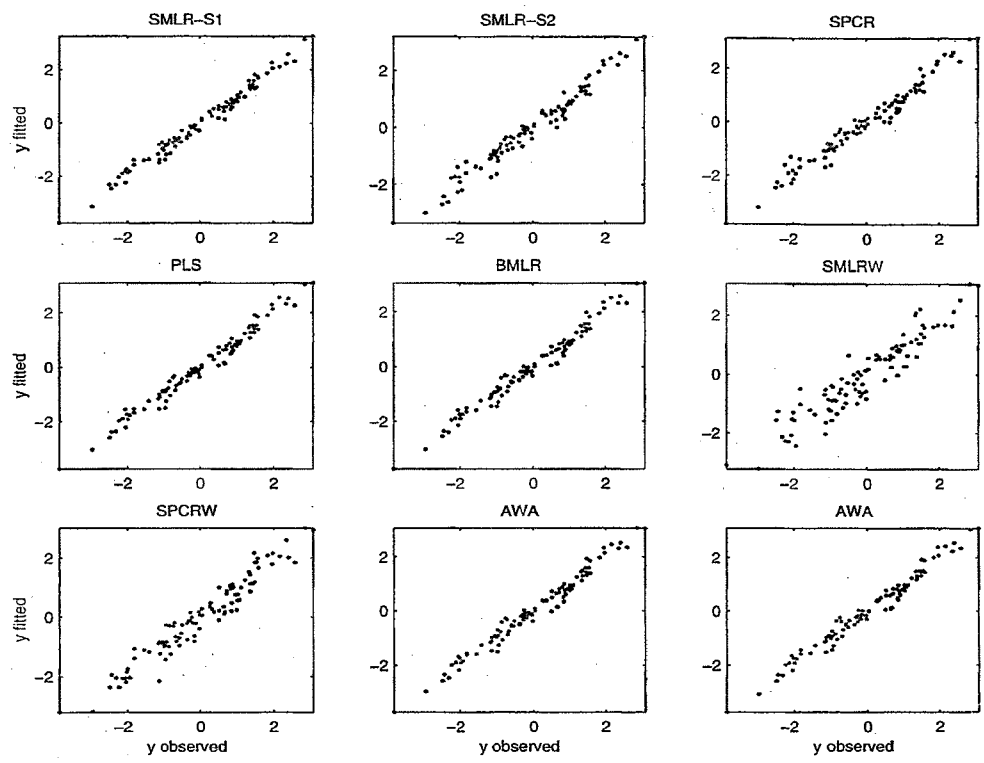


Figure 8.17: **Plots of the residuals versus the fitted values for each of the models for brix.**

principal component, then the average absolute correlation (AAC) is determined by,

$$\text{AAC} = \text{average}(|\hat{\rho}_i|) = \frac{\sum_j |\hat{\rho}_{ij}|}{\text{dimensionality}}$$

where the dimensionality is the number of selected principal components. Figure 8.21, plots the average absolute correlation measure against each of the 512 wavelengths indices for each of the responses. Also shown is the absolute correlation measures (AC= $|\hat{\rho}_i|$) against each wavelength, for each principal component selected by the stepwise routine.

If one considers only the AAC and assumes that a large AAC measure infers a more significant attribute, then essentially all we can conclude by examining the middle row of plots in Figure 8.21 is which regions are not important. These occur where the local minima appear. If one considers the absolute correlation (AC) measure separately for each of the selected principal components, then one can see that sometimes a variable may produce a large AC value for one component but a small AC value for another component. This makes the interpretation of important variables or regions of our spectra quite difficult to assess.

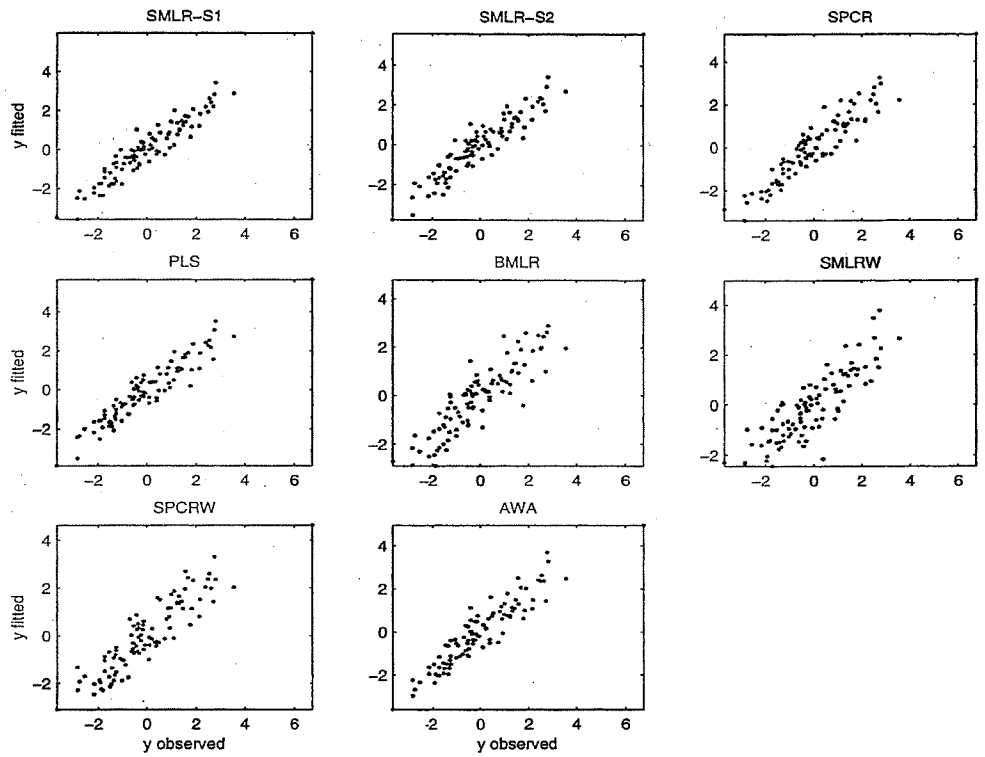


Figure 8.18: **Plots of the residuals versus the fitted values for each of the models for fibre.**

PLS

Before discussing how one might deduce which variables are significant from a PLS model we first note that in Section 8.3 two PLS models were considered, PLS-S3 and PLS-S4. In this section we search for the significant variables from the PLS-S4 model. If one makes the assumption that a large regression coefficient implies a more significant variable, then one can examine the size of the regression coefficients from the PLS model. If the data are not standardized then a large coefficient may indicate an important variable but may also reflect a variable which has a small magnitude and large variance [50]. If the data are standardized, then this problem relating to the interpretation of important variables can be somewhat lessened. Here, we standardize our original data, and then calculate the regression coefficients for each of the responses, using the same number of PLS components that were used in PLS-S4, that is 15 for the brix and fibre responses, and 14 for the protein response. Figure 8.22 plots the absolute values of the regression coefficients obtained from PLS-S4, for each of the responses.

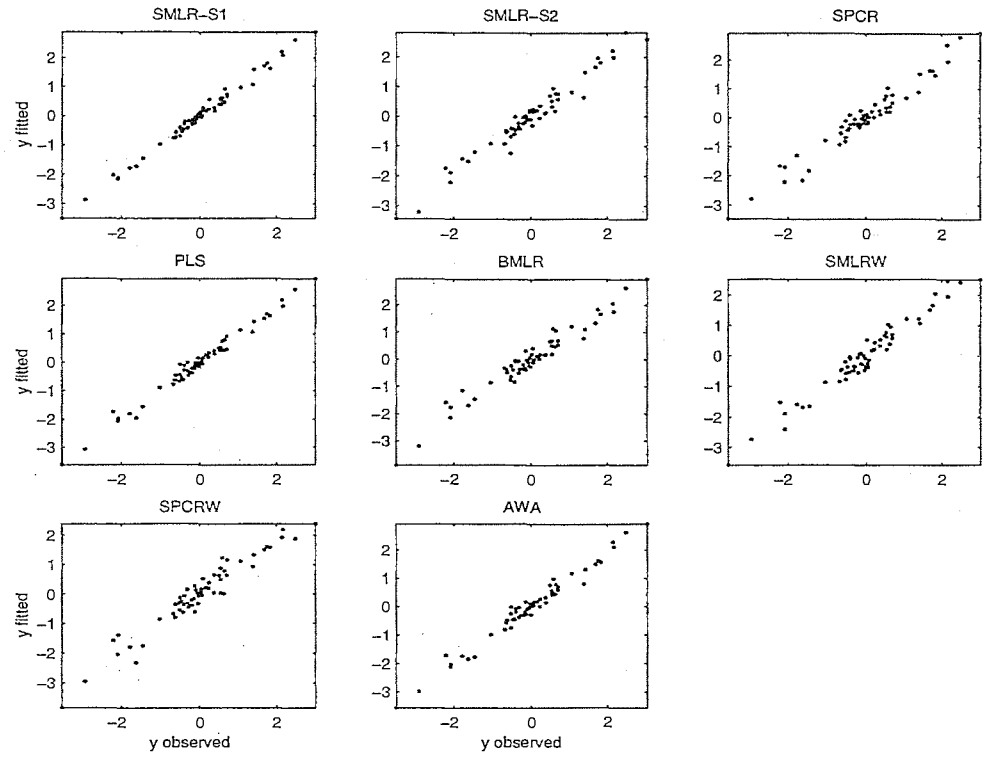


Figure 8.19: **Plots of the residuals versus the fitted values for each of the models for protein.**

The PLS model for the regression of brix has larger coefficients at approximately 1400 nm and between 1600-1800 nm. The larger (in magnitude) PLS coefficients for the modelling of fibre occur around the 1100 nm mark, around the 1400 nm position, the 1600 nm position and in the vicinity of 1850 nm. These regions are similar to those utilized by SMLR for predicting fibre. The exception however occurs for the 1600 nm mark which appears important for the PLS model, but not for the SMLR model. The larger PLS coefficients for the protein model lie predominantly between the 1200-1300 nm range.

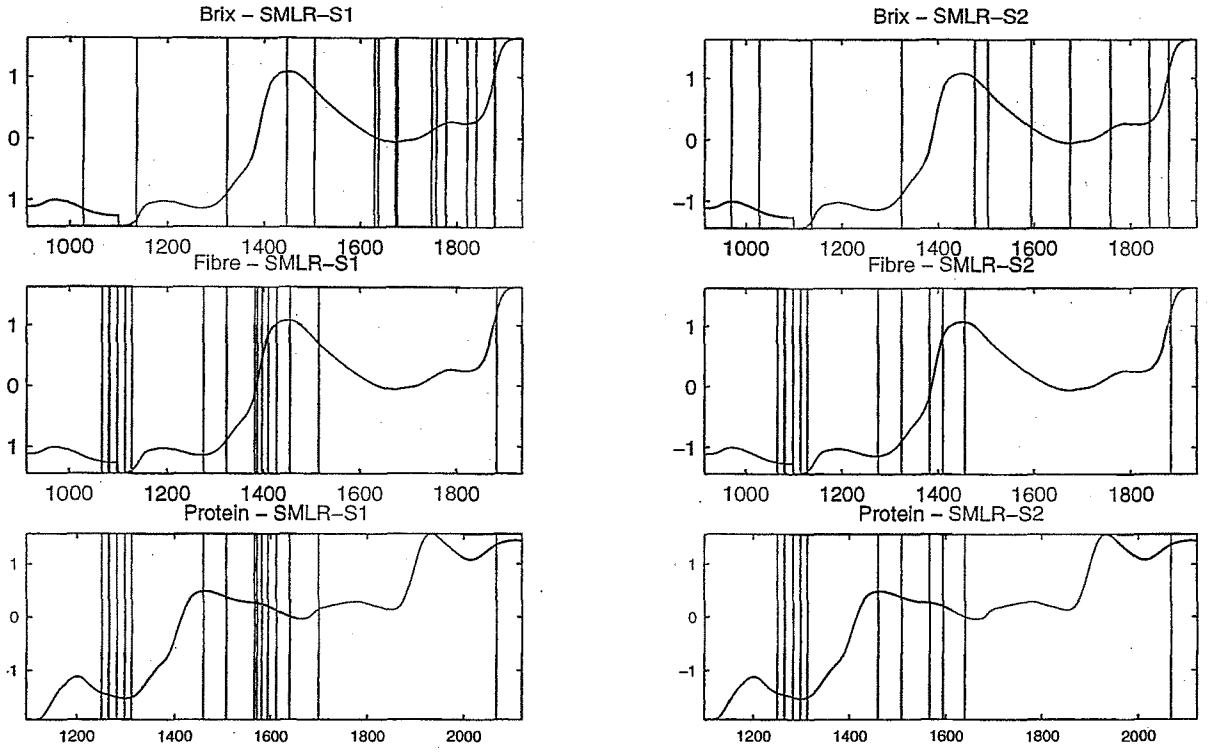


Figure 8.20: Wavelengths selected by SMLR-S1 and SMLR-S2.

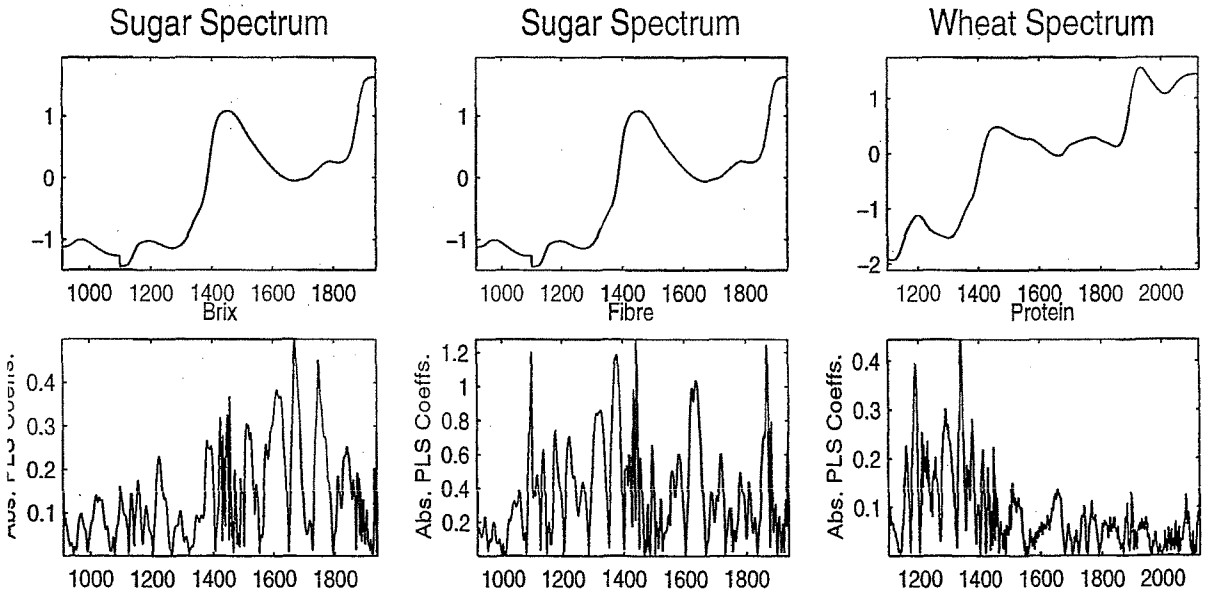


Figure 8.22: Regression coefficients obtained from PLS, when the data has been standardized.

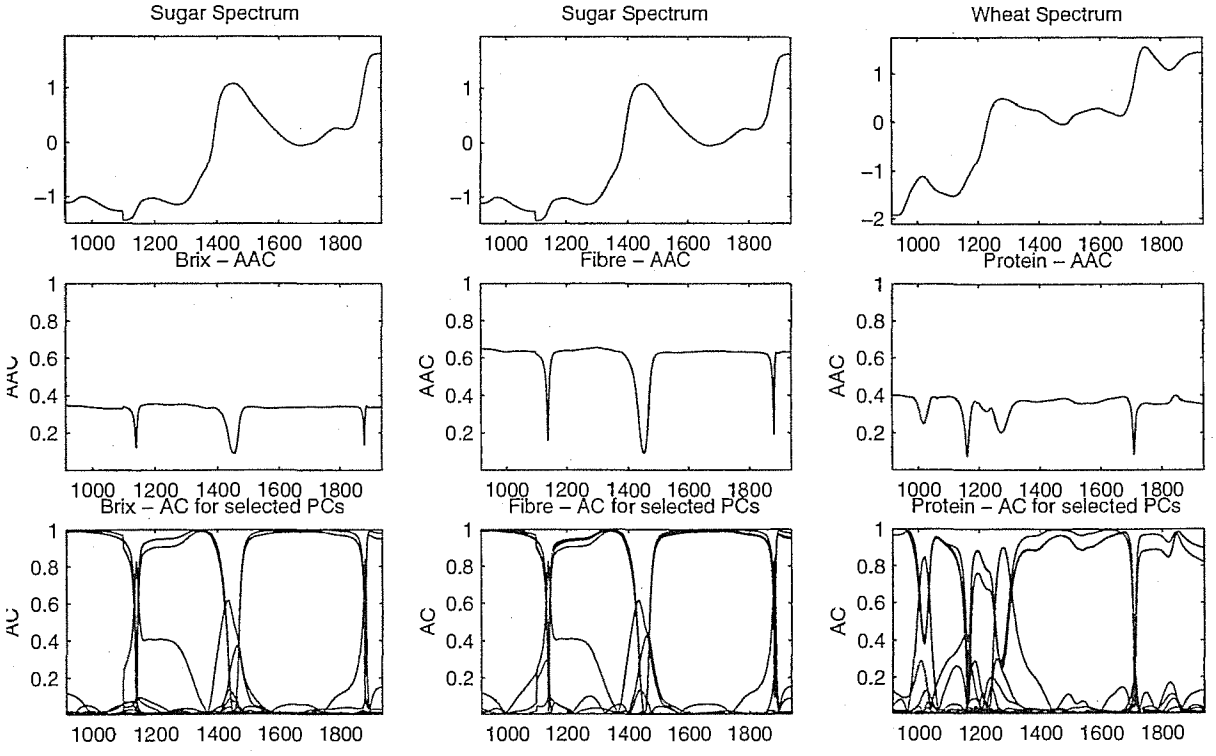


Figure 8.21: **Absolute correlations between each wavelength and the principal components selected by SPCR.**

BMLR

For BMLR, one can observe the coefficients from the bands which were supplied to MLR. These can be seen in Figures 8.4–8.5 for the sugar and wheat data, respectively. For reasons discussed in Section 7.5.2, it is difficult to draw any conclusions about the important regions of the spectra from BMLR. This is because, we have preselected the bands of coefficients without any consideration of the data, but rather, so that the number of coefficients in the bands allows for a well-posed regression problem.

SMLRW

The results for SMLRW-S1 and SMLRW-S2 are identical since the different stopping rules produced the same models, we will subsequently refer to either SMLRW-S1 or SMLRW-S2 as SMLRW. Figure 8.23 (page 206) shows the coefficients from the DWT which have been selected by SMLRW for each response. The original sampled spectra are shown in the top row of Figure 8.23. Recall from Chapter 7, that if vertical lines are extended from

the asterisks to the original spectra, then where the lines meet the original spectra will indicate in an approximate sense, the region which is represented by the coefficients.

The main areas of focus for the SMLRW-brix model appears to be prior to the 1100 nm jump and around the central peak of 1400 nm. For the SMLRW-fibre model, information prior to the peak at 1400 nm and between 1600-1800 nm appears to be represented by the wavelet coefficients which were supplied to MLR. The coefficients for the SMLRW-protein model relate to the peaks at approximately 1200 nm, 1400 nm and the small elevation at approximately 1750 nm.

We also consider backtransforming the wavelet coefficients which were selected by the SMLRW models for each of the responses to visualize the spectra which would be obtained when the selected wavelet coefficients are linearly combined with their respective basis functions. Figure 8.24 presents these reconstructed spectra.

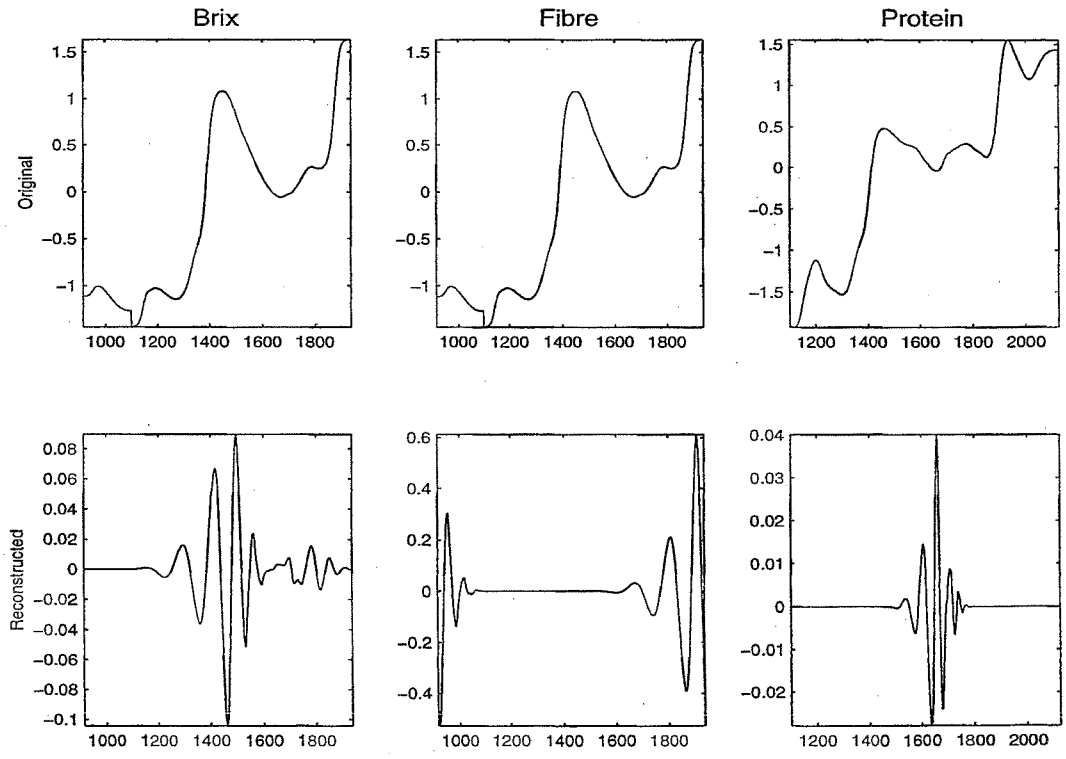


Figure 8.24: **Reconstructed spectra produced from the coefficients selected by SMRLW.**

If were to assume that the regions which are not equal to zero contain the useful information for regression, then there would appear to be some disagreement with the information

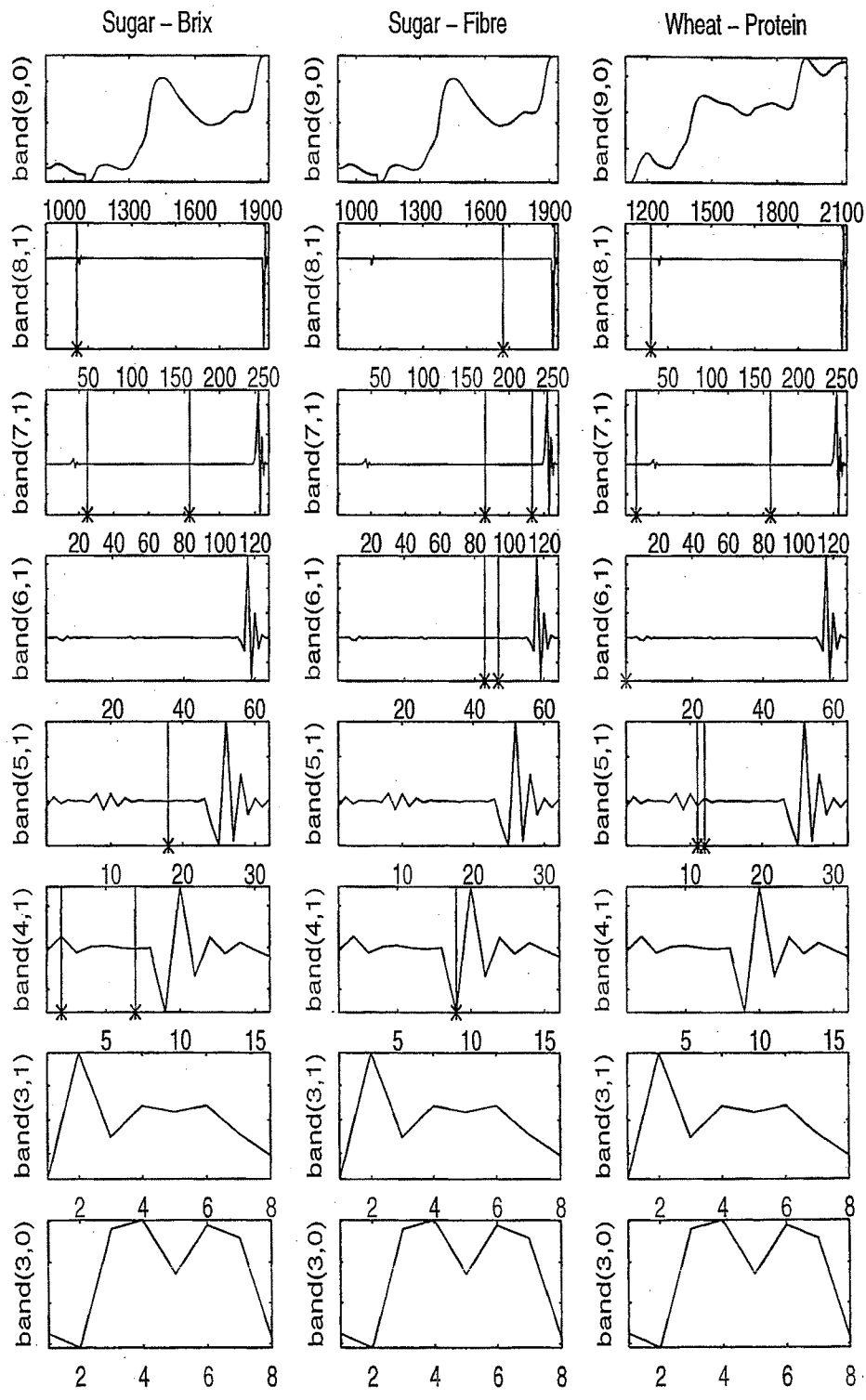


Figure 8.23: **Coefficients from the DWT which were selected by SMLRW.**

presented in Figure 8.24 (page 205) to that in Figure 8.23 (page 206). Take the protein response for example. From Figure 8.23 we conjectured that information around the peaks at approximately 1200 nm, 1400 nm and at 1750 nm contain useful information for regression. However, in Figure 8.24 there seems to be one main area which has been extracted for regression, that is around 1600 nm. One problem which arises from superimposing each of the backtransformed spectra associated with each of the wavelet coefficients, is that the magnitude of some of the backtransformed spectra will be larger than others, hence impeding the visibility of the backtransformed spectra with smaller magnitudes. This is especially the case when we produce a reconstructed spectra from coefficients from different levels in the DWT. The next series of plots shows the reconstructed spectra produced by backtransforming the coefficients from the individual bands in the DWT. That is if two coefficients were selected from the same band by SMLRW, then both coefficients will be backtransformed to produce a reconstructed spectrum. These reconstructed spectra are shown in Figure 8.25. Now, the reconstructed spectra with a smaller magnitude can be more clearly visualized, and the informative regions are now in agreement with that in Figure 8.23. Note that some of the axes do not contain any plots. These 'blank' axes have simply been shown so that direct comparison with Figure 8.23 is made easier. Also, at the expense of having less 'cluttered' plots, the scales on the vertical axis have been omitted. Generally though, the higher up the transform, the smaller the scale, since at the higher levels in the transform, the wavelet coefficients contain information about the high frequency events.

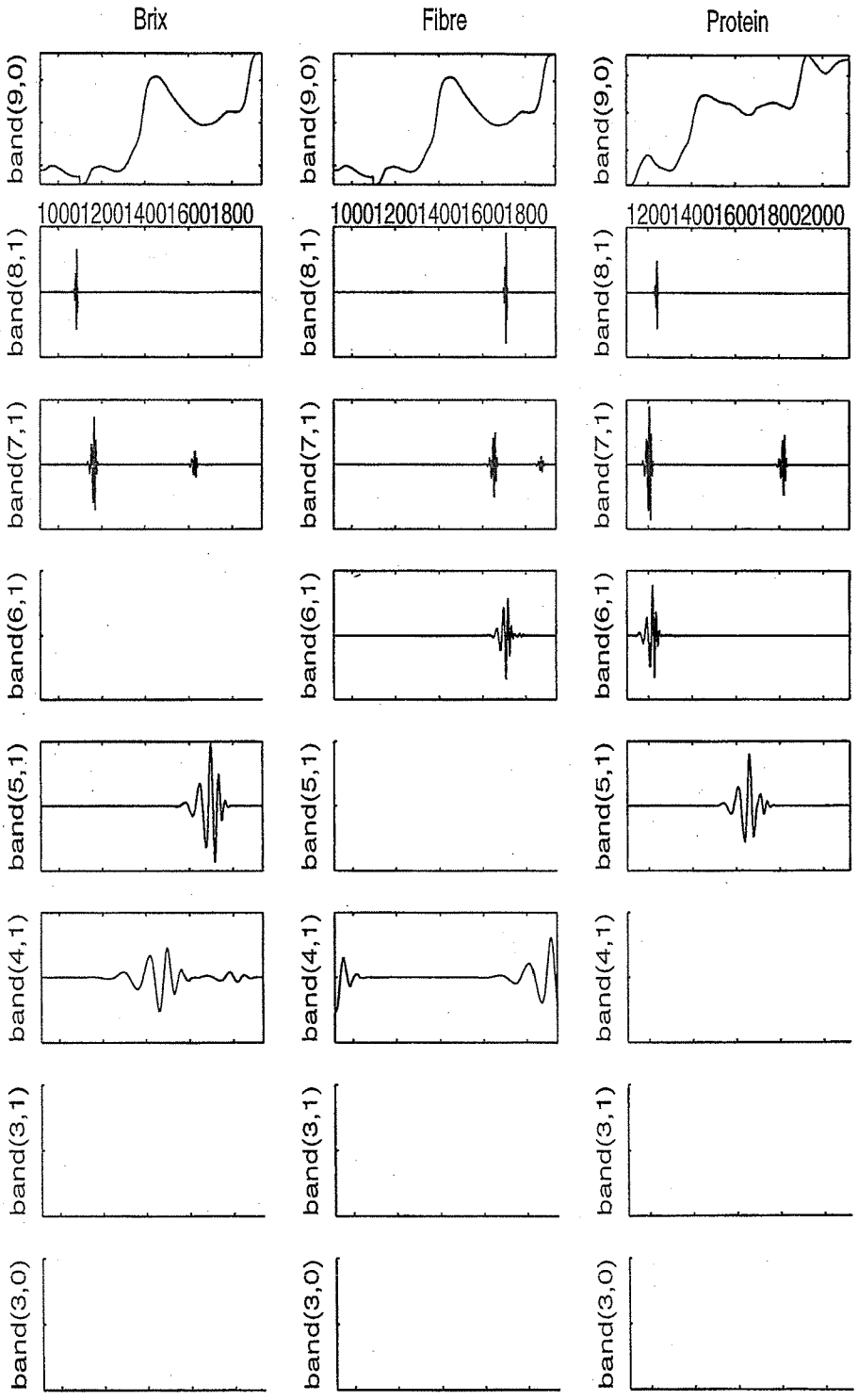


Figure 8.25: **Reconstructed spectra produced from the coefficients selected by SMRLW that pertain to the same band.**

SPCRW

The process for determining which features are utilized is not straight forward to calculate for SPCRW. We note that both SPCRW-S1 and SPCRW-S2 produced the same models so we can speak of SPCRW in general. In order to determine which features are more predominantly involved in the SPCRW models there are several steps which need to be performed. First, it is necessary to consider the wavelet coefficients for which PCA was performed. Here we selected the wavelet coefficients which had an absolute correlation of more than 0.5 with the response. Next we need to consider which of the selected coefficients are more significant to the principal components which were selected by the stepwise routine. This could be done as before by measuring the correlation between the reduced set of wavelet (and scaling) coefficients with the selected principal components. If we could determine which wavelet coefficients were important, then this information could be translated to determine which regions of the spectra are important. We believe that such an approach would be too subjective. When SPCR was applied to the original data we previously mentioned that interpretation was difficult, since some variables are more important for certain principal components and less important for other principal components. Now that an intermediate step has been included i.e. performing the DWT, the level of subjectivity is further enhanced.

AWA

In this section we investigate the information which was extracted from the original spectra by the AWA. Here we will examine the adaptive wavelet coefficients (Figure 8.26) which produced the highlighted results in Table 8.9 as well as the reconstructed spectra, which were obtained by setting the coefficients in the remaining bands to zero, and then backtransforming the thresholded data.

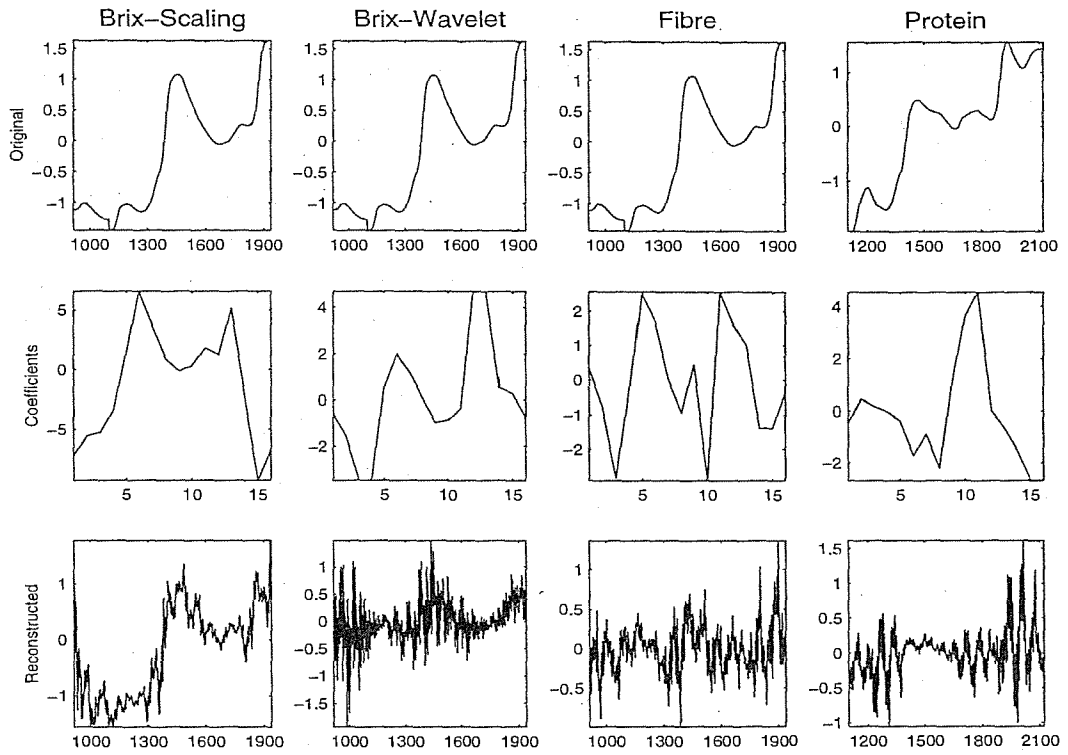


Figure 8.26: **The wavelet coefficients and reconstructed spectra produced from the AWA.**

Interpretation of the adaptive wavelet coefficients is not straight forward. One might immediately assume that a larger coefficient implies a more important variable. From the p-values in Table 8.11, we see this is not necessarily true. Take for example the brix model. When optimization was over the wavelet band the 12th and 13th coefficients are quite large, but have quite high p-values (see Table 8.11) indicating that these coefficients are not significant terms in the model. Perhaps what is more interesting is if we could examine the reconstructed spectra which would be obtained if all the adaptive wavelet coefficients which were used for regression are backtransformed. The reconstructed spectra are shown in the bottom row of plots in Figure 8.26. As was the case in the previous chapter, interpretation of the reconstructed spectra is not straight forward. One interesting feature to note however, is that when the scaling coefficients from the brix response are backtransformed, we obtain a spectrum similar to the original spectrum, which is typical of backtransformations involving the scaling terms.

One might assume that for the backtransformed wavelet coefficients, a region which

deviates from the horizontal line y= 0 contains information which is represented by the adaptive wavelet coefficients and hence contains useful information for regression. Based on this assumption, then one might conclude that the information extracted by the AWA comes from the 1000 nm and 1300-1600 nm regions for the brix response, the region prior to 1000 nm and the region after 1700 nm for the fibre response, and for the protein response, from the 1100-1300 nm and 1900-2000 nm regions.

8.6 Summary

This chapter has explored several regression procedures which can be applied to spectral data. We have considered using features which are the original data such as SMLR, and features which involve some (linear) combination of the original variables, such as SPCR and the PLS routines. Wavelet coefficients which can also be thought of as a linear combination of the original variables have also been used as features for the regression procedures BMLR, SPCRW, SMLRW and the AWA.

In terms of performance measures namely the R^2_{test} measure, the two techniques which tended to produce higher measures than the remaining methods were PLS, the AWA and SMLR-S1, while SMLRW and SPCRW tended to produce lower values. If one then considers the proportion of insignificant terms in the three models mentioned above, we can see from Table 8.11 that the SMLR-S1 and AWA models tended to have a higher proportion than the PLS models which did not have any insignificant terms at the 0.05 level. It is interesting to note that despite the poor performance of SMLRW and SPCRW, all terms in this model were significant.

If one is interested in discovering any additional information about our data such as, which regions might be potentially useful for regression, then one would conclude that of the methods presented here, such information was more easily accessible by the SMLR routines applied to the original data as well as the wavelet coefficients. By examining the PLS coefficients it was also possible to make some decisions about the regions which the PLS model paid particular attention to, though the PLS coefficients were rather 'jagged' in appearance, thereby making such judgements more difficult. Some attempt was made with the AWA algorithm for deducing similar information, but as for the PLS method, such information was not well defined, unlike the SMLR and SMLRW strategies. The SPCR

methods for the original and wavelet coefficients were also found to be quite difficult to interpret.

With the exception of the AWA, most regression methods considered in this chapter were not computationally laborious, unlike several of the discriminant procedures. One of the major factors which contributes to the relatively fast implementation of the regression procedures is the use of the PRESS statistic which does not actually require the removal of the validation object to obtain a cross-validated estimate of the regression performance.

This is the final chapter which requires the analysis of data. The next chapter summarizes the work presented throughout the thesis, and makes some concluding remarks.

Chapter 9

Concluding Remarks

9.1 Original Contribution

This thesis has investigated several strategies that can be used for the discrimination and regression of spectral data. A problem which frequently occurs when modelling spectral data, is that the dimensionality (i.e., number of variables) is usually quite large, especially when compared to the number of spectra that are available. This leads to a substantial deterioration in performance of traditionally favoured classifiers and regression methods. There are basically two approaches that can be implemented to help overcome this problem. One option is to apply a high dimensional technique which is capable of handling a large number of variables. An alternative procedure, and perhaps more commonly applied, is to first reduce the dimensionality by some feature extraction preprocessing method, and then use an appropriate low dimensional classification or regression technique.

This thesis has introduced some novel dimension reducing techniques as well as some low and high dimensional multivariate methods which have evolved quite recently (e.g. FDA and PDA). The original part of this thesis comes in the exploration of wavelet coefficients as features for the multivariate analysis of spectral data. In particular, the discrimination and regression of near infrared spectral data. The discrete wavelet transform was introduced as a method for extracting features. Wavelets were considered as features for discriminant and regression analysis because of their ability to detect local changes in a spectrum. Whilst, there have been previous applications of the discrete wavelet transform as a feature extraction procedure for classification and regression, this field remains relatively unexplored and any work performed in this area is of much interest.

This thesis has considered two main approaches where wavelets form the foundation of the feature extraction procedure. The first procedure selected wavelet coefficients from the discrete wavelet transform that was produced using standard wavelet bases. A new and innovative feature extraction scheme was also proposed, which avoids the need to preselect a wavelet basis. We demonstrated how wavelets can be designed to suit almost any general application, but we focused on designing wavelets for the classification and regression of spectral data. The wavelet gradually adapts to the application at hand, and is therefore referred to as an 'adaptive wavelet'. The adaptive wavelet methodology simultaneously reduces the dimensionality and attempts to find the wavelet which optimizes some multivariate modelling criteria. The adaptive wavelet methodology stems from the work peformed by Kautsky and Turcajová (1995) [78] who introduce a procedure for designing wavelets for removing disturbances in signals. The adaptive wavelet algorithm applied in this thesis is based on a similar algorithm to that in [78]. The main difference between the two algorithms is the criteria which are to be optimized and the particular application that the adaptive wavelets are designed for.

9.2 Summary of Results

This section provides a summary of the results obtained using the various classification and regression strategies. Since a summary was provided at the end of Chapters 7 and 8, only a brief summary will be provided here.

Each of the discriminant and regression applications can be divided into into one of three main categories.

1. High dimensional multivariate statistical methods using all of the original variables.
2. Low dimensional multivariate statistical methods which have selected from the original data.
3. Low dimensional multivariate statistical methods which have selected wavelet coefficients as features.

The multivariate methods that were employed for discriminant analysis include:

- PDA: penalized discriminant analysis
- RDA: regularized discriminant analysis
- SBLDA: stepwise Bayesian linear discriminant analysis
- SBQDA: stepwise Bayesian quadratic discriminant analysis
- FDA: flexible discriminant analysis
- BBLDA: banded Bayesian linear discriminant analysis
- BBQDA: banded Bayesian quadratic discriminant analysis
- SWBLDA: stepwise Bayesian linear discriminant analysis applied to the wavelet coefficients
- SWBQDA: stepwise Bayesian quadratic discriminant analysis applied to the wavelet coefficients
- LDB: local discriminant bases algorithm
- AWA: adaptive wavelet algorithm.

The multivariate methods which were applied for the regression of spectral data include:

- SMLR: stepwise multiple linear regression
- SPCR: stepwise principal component regression
- PLS: partial least squares regression
- BMLR: banded multiple linear regression
- SMLRW: stepwise multiple linear regression applied to wavelet coefficients
- AWA: adaptive wavelet algorithm.

Throughout the previous two chapters we have discussed the advantages and disadvantages of each of the strategies listed above. This was done by taking into consideration two main objectives:

1. how well the model performed in terms of specified performance measures, and
2. consideration was also given to the descriptive information which could be provided by each of the approaches.

In terms of the performance based measures for classification it was noted that RDA and the AWA consistently performed well, and for regression, PLS tended to dominate, although the AWA with a simpler regression model often produced comparable results.

The high dimensional methods RDA and PLS tended to give quite reasonable performance measures, but were relatively difficult to convey infomation about the spectral regions which were important for either discrimation or regression. The AWA also performed well but also seemed to be difficult to extract information about the useful features of the spectra. This is a consequence of the banded selection procedure. However, the AWA is able to provide more information about the spectra than RDA. If one is interested in a technique that produces average performance measures and is relatively easy to interpret or to extract additional information, then one might like to pursue some of the strategies involving stepwise feature selection.

In terms of the wavelet based methods, the AWA seemed to be more superior than the remaining wavelet based strategies which were investigated. This is despite the somewhat restrictive procedure of basing optimization over an entire band. The success of the adaptive wavelets can be attributed to their ability to adapt to different tasks. This is a primary advantage of using adaptive wavelets as opposed to predefined wavelets. Of course predefined wavelets are more readily available and generally do not require the use of optimization procedures.

9.3 Future Work and General Remarks About the AWA

There are several items regarding the adaptive wavelet algorithm which warrant further discussion. These items are now considered separately.

- Number of Iterations

  One can argue that using a prespecified number of iterations in the AWA (as we have done) does not necessarily allow for a optimal value to be found. This is quite true, but from a practical perspective it is more convenient. We have also noted that generally there is little improvement to the model after 2000 iterations. It is important to mention however, that this was not rigorously tested, and is simply an observation which has been made. It is possible however that with more extensive experimentation on additional real and simulated data, that a more suitable number of maximum iterations could be found.

- Local and Global Minima

  If the AWA algorithm does converge to an optimal value prior to reaching the maximum number of iterations then one can query if it is indeed a local or global minima. As we have discussed previously in Chapter 6 unless the problem is continuous and has only one optimal point, there can be no guarantee that a global optimal value has been found. One suggestion offered in [4] is that starting the optimization routine using different values for parameters at initialization may assist in overcoming this problem. Due to time constraints this was not done for every model produced by the AWA. It was however trialled for a few settings where the criterion function did converge to the same optimal value.

- Constrained Optimization Versus Unconstrained Optimization

  In the the adaptive wavelet algorithm, it was possible to avoid using constraints which ensured orthogonality. This is due to some clever algebraic factorizations of the wavelet matrix for which much credit is due to [79]. However, one constraint which we have not discussed in very much detail is that the vectors $\boldsymbol{v}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_q$ are required to have unit length. This normalization procedure occurs during the optimization routine. An alternative strategy which could be employed, is to place constraints on these vectors requiring them to be normalized.

- Choosing the best (m, q, l, τ) settings

  Selecting the (m, q, l, τ) combination involved trialing several suitable combinations of these values. Presently, it is unknown how one might be able to predetermine with any degree of certainty which setting combinations may produce more preferred results. One general observation however, is that the case $m = 2$ does seem to be producing a higher proportion of larger correct classification rates as well as higher R-squared measures. In order to determine which settings are more preferable remains to be further explored.

- Validation without an independent test set

  Each of the data sets applied in this thesis have consisted of a training and independent test set. Assessment of the various techniques often involved an analysis of the prediction accuracy for the testing data. It was previously mentioned in Section 2.8 and in Section 3.6 that if there are two few observations to allow for an independent testing and training data set, then cross-validation could be used to assess the prediction performance of the discriminant or regression method. Should this be the situation, it is necessary to mention that it would be an extremely computational exercise to implement a full cross-validation routine for the AWA. That is, it would be too time consuming to leave out one observation, build the AWA model, predict the deleted observation, and then repeat this leave-one-out procedure for each observation separately. In the absence of an independent test set, a more realistic approach would be to perform cross-validation using the wavelet produced at termination of the AWA, but it is important to mention that this would not be a full validation.

The tasks discussed above remain the work of future investigations for further exploring and enhancing the adaptive wavelet algorithm. In conclusion we wish to further highlight the potential of wavelets as methods for feature extraction and their interesting way of viewing data through different windows.

Appendix A

SMLR-S1 BRIX
Residual Standard Error = 0.2075, Multiple R-Square = 0.9812 N = 100, F-statistic = 273.412 on 16 and 84 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	-413.3916	82.8005	-4.9926	0.0000
V2	33.1535	9.1333	3.6300	0.0005
V3	-44.0129	7.2449	-6.0750	0.0000
V4	-140.0819	16.6898	-8.3933	0.0000
V5	-557.4451	76.8799	-7.2509	0.0000
V6	185.3640	100.2743	1.8486	0.0680
V7	-474.5068	217.0552	-2.1861	0.0316
V8	-14.0633	7.7300	-1.8193	0.0724
V9	81.2956	12.9722	6.2669	0.0000
V10	-169.9041	144.8098	-1.1733	0.2440
V11	485.5299	88.4825	5.4873	0.0000
V12	228.1292	222.8518	1.0237	0.3089
V13	476.8709	103.2429	4.6189	0.0000
V14	-276.0465	78.8086	-3.5027	0.0007
V15	206.1745	76.7681	2.6857	0.0087
V16	387.5836	155.2515	2.4965	0.0145

SMLR-S1 FIBRE
Residual Standard Error = 0.5372, Multiple R-Square = 0.9077 N = 100, F-statistic = 51.6229 on 16 and 84 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	-947.1048	120.1131	-7.8851	0.0000
V2	368.3170	171.4882	2.1478	0.0346
V3	898.3101	128.9032	6.9689	0.0000
V4	-981.5167	108.2112	-9.0704	0.0000
V5	-1686.1420	288.5763	-5.8430	0.0000
V6	814.6825	91.7905	8.8755	0.0000
V7	6949.1739	1464.5789	4.7448	0.0000
V8	-1224.5805	245.0395	-4.9975	0.0000
V9	229.4859	184.1823	1.2460	0.2162
V10	-4092.2697	1473.1539	-2.7779	0.0067
V11	393.2618	73.5601	5.3461	0.0000
V12	33.6203	16.0073	2.1003	0.0387
V13	-59.1393	21.8135	-2.7111	0.0081
V14	-526.9223	139.3476	-3.7814	0.0003
V15	-561.1551	208.7378	-2.6883	0.0087
V16	443.1769	287.0874	1.5437	0.1264

SMLR-S1 PROTEIN
Residual Standard Error = 0.1387, Multiple R-Square = 0.9905 N = 60, F-statistic = 287.6991 on 16 and 44 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	-103.4695	69.6912	-1.4847	0.1448
V2	-569.0466	47.6506	-11.9421	0.0000
V3	-575.0757	170.2500	-3.3778	0.0015
V4	760.8207	128.3992	5.9254	0.0000
V5	33.4229	6.1876	5.4016	0.0000
V6	140.9795	55.6816	2.5319	0.0150
V7	506.3303	149.5141	3.3865	0.0015
V8	-320.0526	88.2420	-3.6270	0.0007
V9	178.1800	30.8989	5.7665	0.0000
V10	-298.5406	55.4835	-5.3807	0.0000
V11	298.7727	59.3241	5.0363	0.0000
V12	-150.1197	36.2051	-4.1464	0.0002
V13	195.9957	54.2381	3.6136	0.0008
V14	-159.2740	37.0052	-4.3041	0.0001
V15	325.7624	102.4385	3.1801	0.0027
V16	-242.6675	92.9929	-2.6095	0.0123

SMLR-S2 BRIX
Residual Standard Error = 0.2785, Multiple R-Square = 0.9641 N = 100, F-statistic = 216.9789 on 11 and 89 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	24.6889	8.6760	2.8457	0.0055
V2	59.7694	12.3162	4.8529	0.0000
V3	14.1401	8.8403	1.5995	0.1132
V4	-47.7623	8.6337	-5.5321	0.0000
V5	-241.7219	18.4566	-13.0968	0.0000
V6	136.4132	22.3162	6.1128	0.0000
V7	144.9032	18.0175	8.0424	0.0000
V8	-167.3372	12.4628	-13.4269	0.0000
V9	251.2473	26.8518	9.3568	0.0000
V10	-114.9007	17.5833	-6.5347	0.0000
V11	-50.2653	10.9866	-4.5751	0.0000

SMLR-S2 FIBRE
Residual Standard Error = 0.5866, Multiple R-Square = 0.8847 N = 100, F-statistic = 56.2599 on 12 and 88 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	-818.3904	92.1510	-8.8810	0.0000
V2	-213.7444	45.0951	-4.7399	0.0000
V3	802.9916	93.7831	8.5622	0.0000
V4	-831.1576	106.5776	-7.7986	0.0000
V5	-1234.9175	268.0046	-4.6078	0.0000
V6	673.1537	78.2243	8.6054	0.0000
V7	6576.0327	1584.2692	4.1508	0.0001
V8	-914.5712	229.8370	-3.9792	0.0001
V9	350.6474	59.4780	5.8954	0.0000
V10	-4475.4210	1602.1541	-2.7934	0.0064
V11	104.1997	24.7005	4.2185	0.0001
V12	56.4568	15.7666	3.5808	0.0006

SMLR-S2 PROTEIN
Residual Standard Error = 0.2525, Multiple R-Square = 0.9615 N = 60, F-statistic = 224.6625 on 6 and 54 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	84.8764	8.9748	9.4572	0.0000
V2	-624.9640	34.2107	-18.2681	0.0000
V3	-362.7573	38.0906	-9.5235	0.0000
V4	873.3815	60.8132	14.3617	0.0000
V5	24.7410	4.9389	5.0094	0.0000
V6	13.6638	4.1881	3.2625	0.0019

SPCR BRIX

Residual Standard Error = 0.2726, Multiple R-Square = 0.9659 N = 100, F-statistic = 208.0096 on 12 and 88 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	4.2093	0.1358	30.9984	0.0000
V2	33.5693	1.3521	24.8276	0.0000
V3	9.7738	0.6385	15.3069	0.0000
V4	-27.3959	2.1091	-12.9894	0.0000
V5	2.8332	0.2325	12.1868	0.0000
V6	0.2003	0.0215	9.3183	0.0000
V7	-29.1912	3.2209	-9.0629	0.0000
V8	-17.8065	2.0230	-8.8021	0.0000
V9	-29.7696	4.7879	-6.2177	0.0000
V10	16.9209	2.9502	5.7356	0.0000
V11	-22.2199	4.0771	-5.4499	0.0000
V12	39.3850	8.9062	4.4222	0.0000

SPCR FIBRE

Residual Standard Error = 0.6047, Multiple R-Square = 0.8761 N = 100, F-statistic = 57.1892 on 11 and 89 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	0.8967	0.0477	18.8058	0.0000
V2	2.0134	0.3012	6.6845	0.0000
V3	-211.0782	32.1852	-6.5582	0.0000
V4	-4.4383	0.7670	-5.7868	0.0000
V5	-24.9685	4.4873	-5.5642	0.0000
V6	132.1473	23.8125	5.5495	0.0000
V7	47.1311	9.0437	5.2115	0.0000
V8	14.0427	2.9992	4.6822	0.0000
V9	135.6085	35.2008	3.8524	0.0002
V10	80.5381	20.9538	3.8436	0.0002
V11	300.4022	80.8993	3.7133	0.0004

SPCR PROTEIN
Residual Standard Error = 0.2951, Multiple R-Square = 0.9542 N = 60, F-statistic = 75.361 on 13 and 47 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	-0.6344	0.0339	-18.7189	0.0000
V2	-5.2481	0.3939	-13.3245	0.0000
V3	-2.7032	0.2363	-11.4395	0.0000
V4	-34.3029	4.0530	-8.4636	0.0000
V5	-7.3966	0.9429	-7.8445	0.0000
V6	41.3803	6.5338	6.3332	0.0000
V7	27.4129	4.4115	6.2140	0.0000
V8	0.6697	0.1131	5.9221	0.0000
V9	-7.1017	1.4912	-4.7624	0.0000
V10	2.9666	0.7775	3.8157	0.0004
V11	64.4310	17.3055	3.7232	0.0005
V12	-7.8878	2.2502	-3.5053	0.0010
V13	48.8068	15.0334	3.2466	0.0022

PLS BRIX
Residual Standard Error = 0.2287, Multiple R-Square = 0.9766 N = 100, F-statistic = 256.0153 on 14 and 86 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	0.1393	0.0023	59.8683	0.0000
V2	0.4696	0.0081	58.1210	0.0000
V3	8.1823	0.1869	43.7777	0.0000
V4	12.4374	0.3068	40.5358	0.0000
V5	19.3918	0.6230	31.1256	0.0000
V6	9.9626	0.3726	26.7346	0.0000
V7	55.2129	3.1950	17.2812	0.0000
V8	51.9918	3.2378	16.0578	0.0000
V9	53.0374	4.6422	11.4250	0.0000
V10	197.7996	20.7199	9.5464	0.0000
V11	508.6738	81.6619	6.2290	0.0000
V12	354.9034	74.9618	4.7345	0.0000
V13	555.0606	141.6994	3.9172	0.0002
V14	231.6885	107.1771	2.1617	0.0334

PLS FIBRE
Residual Standard Error = 0.5614, Multiple R-Square = 0.898 N = 100, F-statistic = 49.8674 on 15 and 85 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	0.0113	0.0004	27.3498	0.0000
V2	1.4776	0.0807	18.2986	0.0000
V3	12.3480	0.7483	16.5023	0.0000
V4	85.5599	5.5652	15.3742	0.0000
V5	20.5716	1.3794	14.9130	0.0000
V6	81.4700	5.8077	14.0280	0.0000
V7	373.0133	28.7931	12.9549	0.0000
V8	219.4928	17.8641	12.2868	0.0000
V9	186.2217	15.5385	11.9845	0.0000
V10	482.4133	41.4357	11.6425	0.0000
V11	1308.8532	130.9792	9.9928	0.0000
V12	313.4975	53.7439	5.8332	0.0000
V13	1386.8892	277.8799	4.9910	0.0000
V14	1584.8630	341.7175	4.6379	0.0000
V15	1384.0165	659.6079	2.0982	0.0389

PLS PROTEIN
Residual Standard Error = 0.1865, Multiple R-Square = 0.9829 N = 60, F-statistic = 157.8145 on 16 and 44 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	0.0372	0.0007	50.2497	0.0000
V2	1.2601	0.0314	40.1534	0.0000
V3	1.4897	0.0422	35.2642	0.0000
V4	6.9472	0.2371	29.2957	0.0000
V5	36.5421	1.4108	25.9013	0.0000
V6	122.7007	5.5415	22.1421	0.0000
V7	80.8912	4.0994	19.7325	0.0000
V8	110.8226	6.5049	17.0368	0.0000
V9	249.5727	20.3698	12.2521	0.0000
V10	435.0110	42.3651	10.2681	0.0000
V11	501.2799	51.3976	9.7530	0.0000
V12	239.3746	26.5649	9.0109	0.0000
V13	437.6089	54.7417	7.9941	0.0000
V14	749.7086	104.5505	7.1708	0.0000
V15	1040.0230	158.0652	6.5797	0.0000
V16	384.9349	95.5418	4.0290	0.0002

BMLR BRIX
Residual Standard Error = 0.2396, Multiple R-Square = 0.9749 N = 100, F-statistic = 203.7632 on 16 and 84 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	1.0592	3.5050	0.3022	0.7633
V2	-6.2497	3.0676	-2.0373	0.0448
V3	23.1697	5.1022	4.5411	0.0000
V4	-22.6871	8.5814	-2.6437	0.0098
V5	15.5482	3.9389	3.9474	0.0002
V6	-22.7034	2.4196	-9.3832	0.0000
V7	24.0416	2.4895	9.6572	0.0000
V8	-20.6519	4.4916	-4.5979	0.0000
V9	-20.3396	7.3185	-2.7792	0.0067
V10	46.5804	6.4574	7.2135	0.0000
V11	-50.8380	5.1534	-9.8650	0.0000
V12	42.4398	6.4154	6.6153	0.0000
V13	1.9550	5.0120	0.3901	0.6975
V14	-16.1397	2.4828	-6.5005	0.0000
V15	-0.2121	2.6536	-0.0799	0.9365
V16	-6.3390	3.5985	-1.7616	0.0818

BMLR FIBRE
Residual Standard Error = 0.796, Multiple R-Square = 0.7973 N = 100, F-statistic = 20.6499 on 16 and 84 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	15.7990	33.8356	0.4669	0.6418
V2	-187.2881	91.8747	-2.0385	0.0446
V3	-288.6024	102.5862	-2.8133	0.0061
V4	-163.6692	105.0642	-1.5578	0.1230
V5	-94.2978	124.2972	-0.7586	0.4502
V6	-678.7155	160.8204	-4.2203	0.0001
V7	-472.9067	133.2914	-3.5479	0.0006
V8	62.8721	38.9508	1.6141	0.1102
V9	12.6681	35.5259	0.3566	0.7223
V10	-17.6691	74.3150	-0.2378	0.8126
V11	-33.6705	62.8214	-0.5360	0.5934
V12	-156.1085	37.7196	-4.1387	0.0001
V13	-427.7977	138.5951	-3.0867	0.0027
V14	-243.5802	149.0532	-1.6342	0.1060
V15	-66.3361	197.8717	-0.3352	0.7383
V16	29.7571	43.4330	0.6851	0.4952

BMLR PROTEIN

Residual Standard Error = 0.3128, Multiple R-Square = 0.9518 N = 60, F-statistic = 54.3503 on 16 and 44 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	234.7303	86.1259	2.7254	0.0092
V2	157.9888	38.7069	4.0817	0.0002
V3	26.0952	38.2142	0.6829	0.4983
V4	89.0377	45.0397	1.9769	0.0543
V5	21.0500	16.5506	1.2719	0.2101
V6	5.1615	5.9630	0.8656	0.3914
V7	33.5142	12.5365	2.6733	0.0105
V8	12.7502	18.6861	0.6823	0.4986
V9	-26.7414	13.5606	-1.9720	0.0549
V10	-19.5341	16.6508	-1.1732	0.2470
V11	38.5200	21.5308	1.7891	0.0805
V12	58.5335	31.6643	1.8486	0.0712
V13	-20.2556	22.3299	-0.9071	0.3693
V14	4.1250	7.5704	0.5449	0.5886
V15	9.5707	19.0961	0.5012	0.6187
V16	112.6701	64.4422	1.7484	0.0874

SMLRW BRIX

Residual Standard Error = 0.4823, Multiple R-Square = 0.8861 N = 100, F-statistic = 121.9396 on 6 and 94 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	112.0421	12.6864	8.8317	0.0000
V2	-587.7874	47.5490	-12.3617	0.0000
V3	-18.7954	3.0312	-6.2007	0.0000
V4	36749.8410	10050.2465	3.6566	0.0004
V5	-3334.3371	755.7048	-4.4122	0.0000
V6	6452.2226	1921.0921	3.3586	0.0011

SMLRW FIBRE
Residual Standard Error = 0.811, Multiple R-Square = 0.7645 N = 100, F-statistic = 50.8587 on 6 and 94 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	-4.3691	0.8745	-4.9963	0.0000
V2	2592.0316	399.9554	6.4808	0.0000
V3	2233.2707	397.1550	5.6232	0.0000
V4	7700.8274	1322.0250	5.8250	0.0000
V5	2341.1482	502.1983	4.6618	0.0000
V6	-19969.2631	6635.5992	-3.0094	0.0034

SMLRW PROTEIN
Residual Standard Error = 0.2643, Multiple R-Square = 0.9578 N = 60, F-statistic = 204.2821 on 6 and 54 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	517.8027	27.4632	18.8544	0.0000
V2	-163.2285	14.2193	-11.4794	0.0000
V3	-615.1766	135.4929	-4.5403	0.0000
V4	-1542.8067	303.7776	-5.0787	0.0000
V5	1200.0862	266.8916	4.4965	0.0000
V6	-957.3443	242.8538	-3.9421	0.0002

SPCRW BRIX

Residual Standard Error = 0.3694, Multiple R-Square = 0.936 N = 100, F-statistic = 131.7058 on 10 and 90 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	6.3170	0.2925	21.6000	0.0000
V2	76.7129	5.6634	13.5454	0.0000
V3	-105.4595	8.6689	-12.1653	0.0000
V4	385.0874	32.2931	11.9248	0.0000
V5	-132.5143	13.1893	-10.0471	0.0000
V6	27.5505	2.7619	9.9752	0.0000
V7	13.4658	1.3743	9.7986	0.0000
V8	228.5009	28.8609	7.9173	0.0000
V9	304.0115	96.1552	3.1617	0.0021
V10	1594.3647	575.1809	2.7719	0.0068

SPCRW FIBRE

Residual Standard Error = 0.7143, Multiple R-Square = 0.8251 N = 100, F-statistic = 42.4544 on 10 and 90 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	0.8973	0.0563	15.9230	0.0000
V2	-2.0412	0.3669	-5.5636	0.0000
V3	4.6783	0.9291	5.0356	0.0000
V4	-393.1734	82.8737	-4.7442	0.0000
V5	301.7367	64.5316	4.6758	0.0000
V6	-24.1737	5.4611	-4.4265	0.0000
V7	-15.3318	3.6534	-4.1966	0.0001
V8	271.2654	77.0088	3.5225	0.0007
V9	181.8684	52.5139	3.4632	0.0008
V10	886.4472	300.3162	2.9517	0.0040

SPCRW PROTEIN
Residual Standard Error = 0.3437, Multiple R-Square = 0.9339 N = 60, F-statistic = 70.677 on 10 and 50 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	0.6465	0.0399	16.1893	0.0000
V2	-4.0626	0.3446	-11.7898	0.0000
V3	12.8359	1.4676	8.7461	0.0000
V4	-4.7927	0.5881	-8.1490	0.0000
V5	-138.8801	21.0014	-6.6129	0.0000
V6	-49.2715	8.5493	-5.7632	0.0000
V7	0.8674	0.1560	5.5611	0.0000
V8	-117.9042	22.8729	-5.1548	0.0000
V9	-22.0719	4.9422	-4.4660	0.0000
V10	8.0787	2.7894	2.8963	0.0056

AWA - band(2,5,0) BRIX
Residual Standard Error = 0.2381, Multiple R-Square = 0.9752 N = 100, F-statistic = 206.4704 on 16 and 84 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	0.9624	4.3145	0.2231	0.8240
V2	16.8904	3.6243	4.6603	0.0000
V3	-37.4660	8.8290	-4.2435	0.0001
V4	28.6036	8.2427	3.4702	0.0008
V5	-25.5626	2.9157	-8.7674	0.0000
V6	23.7097	2.7388	8.6571	0.0000
V7	-35.9495	3.2954	-10.9091	0.0000
V8	12.1630	6.2623	1.9423	0.0555
V9	35.4129	7.6807	4.6106	0.0000
V10	-50.2272	6.0778	-8.2641	0.0000
V11	57.8430	6.7563	8.5613	0.0000
V12	-31.1792	5.9031	-5.2819	0.0000
V13	-10.0719	4.0500	-2.4869	0.0149
V14	3.7706	2.6101	1.4447	0.1523
V15	-5.1649	2.5863	-1.9970	0.0491
V16	3.7216	4.0358	0.9221	0.3591

AWA - band(2,5,1) BRIX
Residual Standard Error = 0.2388, Multiple R-Square = 0.975 N = 100, F-statistic = 205.154 on 16 and 84 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	52.0331	12.6895	4.1005	0.0001
V2	-59.6844	21.3923	-2.7900	0.0065
V3	35.5449	12.9744	2.7396	0.0075
V4	-54.8479	6.9832	-7.8542	0.0000
V5	37.0434	11.7281	3.1585	0.0022
V6	-43.8805	11.3191	-3.8767	0.0002
V7	-54.8110	12.9727	-4.2251	0.0001
V8	71.8989	16.7222	4.2996	0.0000
V9	-113.2040	14.8415	-7.6275	0.0000
V10	79.3447	14.6540	5.4146	0.0000
V11	-13.7873	9.3965	-1.4673	0.1460
V12	-24.6244	15.2165	-1.6183	0.1094
V13	8.6526	4.6731	1.8516	0.0676
V14	-33.7096	22.3995	-1.5049	0.1361
V15	9.3147	5.2253	1.7826	0.0783
V16	-39.6199	19.7891	-2.0021	0.0485

AWA FIBRE
Residual Standard Error = 0.6407, Multiple R-Square = 0.8687 N = 100, F-statistic = 34.7199 on 16 and 84 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	287.1120	41.5706	6.9066	0.0000
V2	-14.6522	46.2303	-0.3169	0.7521
V3	-30.4353	32.5657	-0.9346	0.3527
V4	162.7707	53.7136	3.0303	0.0032
V5	-162.0510	37.4868	-4.3229	0.0000
V6	197.5076	51.6857	3.8213	0.0003
V7	-29.3467	42.1692	-0.6959	0.4884
V8	-156.3376	46.9818	-3.3276	0.0013
V9	96.7725	36.7527	2.6331	0.0101
V10	51.9793	42.1593	1.2329	0.2210
V11	-9.5745	49.5830	-0.1931	0.8473
V12	35.6617	34.5059	1.0335	0.3043
V13	35.2423	42.0055	0.8390	0.4039
V14	-115.3310	45.2348	-2.5496	0.0126
V15	329.3379	45.1346	7.2968	0.0000
V16	-435.9561	73.5939	-5.9238	0.0000

AWA PROTEIN
Residual Standard Error = 0.2241, Multiple R-Square = 0.9753 N = 60, F-statistic = 108.4756 on 16 and 44 df, p-value = 0

	coef	std.err	t.stat	p.value
V1	-2.8615	8.2078	-0.3486	0.7290
V2	52.8530	15.5264	3.4041	0.0014
V3	-5.6918	19.9444	-0.2854	0.7767
V4	31.1330	20.5265	1.5167	0.1365
V5	61.1461	14.3145	4.2716	0.0001
V6	-6.0746	11.2027	-0.5422	0.5904
V7	78.5630	10.5020	7.4808	0.0000
V8	30.0703	6.1721	4.8720	0.0000
V9	-11.3970	7.5945	-1.5007	0.1406
V10	60.8383	10.7431	5.6630	0.0000
V11	62.6580	7.7381	8.0974	0.0000
V12	-5.5753	3.3728	-1.6530	0.1054
V13	122.4131	9.9775	12.2689	0.0000
V14	-74.8368	9.5503	-7.8360	0.0000
V15	25.7174	4.8644	5.2869	0.0000
V16	93.1781	8.0314	11.6018	0.0000

Bibliography

[1] S. Aeberhard. "Pattern classification in a high dimensional setting," Honours Thesis, School of Computing Science, Mathematics and Physics, James Cook University (1991).

[2] S. Aeberhard, O. de Vel and D. Coomans. "New fast algorithms for variable selection based on classifier performance," *SIAM* (1997) accepted.

[3] S. Aeberhard, D. Coomans and O. De Vel. "Comparison analysis of statistical pattern recognition methods in high dimensional settings," *Pattern Recognition* **27** 1065-1077 (1994).

[4] G. Andrew. "Optimization toolbox user's guide," Math Works, Natick, (1994).

[5] R. Barnes, M. Dhanoa and S. Lister. "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Applied Spectroscopy* **43** 772-777 (1989).

[6] W. Belson. " Matching and prediction on the principal of biological classification," *Applied Statistics* **8** 65-75 (1959).

[7] K. Bewig, A. Clarke, C. Roberts and N. Unklesbay. "Discriminant analysis of vegetable oils by near-infrared reflectance spectroscopy," *JOACS* **71** 195-200 (1994).

[8] M. Bos and J. Vrielink. "The wavelet transform for pre-processing IR spectra in the identification of mono- and di-substituted Benzenes," *Chemometrics and Intelligent Laboratory Systems* **23** 115-122 (1994).

[9] L. Breiman and C. Stone. "Parsimonious binary classification trees," Technical Report, Technology Service Corporation, Santa Monica, CA (1978).

[10] L. Breiman. "Classification and regression trees," Wadsworth International Group, California (1984).

[11] L. Breiman and R. Ihaka. "Nonlinear discriminant analysis via scaling and ace," technical report, Department of Statistics, University of California, Berkeley (1984).

[12] L. Breiman. "Fitting additive models to regression data," *Computational Statistics and Data Analysis* **15** 13-46 (1993).

[13] S. Brier. *Monthly Weather Review* **78** 1-31 (1950).

[14] A. Bruce and H. Gao. "S+Wavelets Users Manual Version 1.0," Seattle: StatSci, a division of MathSoft, Inc. (1994).

[15] P. Brown. "Measurement, regression and calibration," Oxford University Press, England (1993).

[16] W. Brügel. "An introduction to infrared spectroscopy," John Wiley and Sons, New York (1962).

[17] J. Chambers and T. Hastie. "Statistical models in S," Chapman and Hall, New York (1993).

[18] Y. Chan. "Wavelet basics," Kluwer Academic Publishers, Boston (1995).

[19] R. Churchill. "Fourier series and boundary value problems," McGraw-Hill, New York (1941).

[20] R. Coifman and M. Wickerhauser. "Entropy-Based Algorithms for Best Basis Selection," *IEEE Transactions on Information Theory* **38** 713-718 (1992).

[21] D. Coomans and I. Broeckaert. "Potential pattern recognition in chemical and medical decision making," Research Studies Press, John Wiley and Sons, Chichester, (1986).

[22] I. Daubechies. "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics* 41 909-996 (1988).

[23] I. Daubechies. "The wavelet transform, time frequency localization and signal analysis," *IEEE Transactions on Information Theory* **36** 961-1005 (1990).

[24] I. Daubechies. "Ten lectures on wavelets," SIAM (1992).

[25] S. de Jong. "SIMPLS: an alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems* **18** 251-263 (1993).

[26] M. Denham. "Implementing partial least squares," *Statistics and Computing* **5** 191-202 (1995).

[27] D. Domine, J. Devillers, M. Chastrette and W. Karcher. "Non-linear mapping for structure-activity and structure-property modelling," *Journal of Chemometrics* **7** 227-242 (1993).

[28] G. Downey, P. Robert, D. Bertrand and P.M. Kelly. *Applied Spectroscopy* 44 150- (1990).

[29] N. Draper and H. Smith, "Applied regression analysis," John Wiley and Sons, New York (1981).

[30] B. Efron. "Bootstrap Methods: another look at the jackknife," *The Annals of Statistics* **7** 1-26 (1979).

[31] B. Efron and G. Gong. "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician* **37** 36-48 (1983).

[32] B. Efron and R. Tibshirani. "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical Science* **1** 54-77 (1986).

[33] D. Evans, C.N. Scotter, L. Day and M. Hall. "Determination of the authenticity of orange juice by discriminant analysis of near infrared spectra," *Journal of Near Infrared Spectroscopy* **1** 33-44 (1993).

[34] R. Fisher. "The use of multiple measurements in taxonomic problems," *Annals of Eugenics* **7** 179-188 (1936).

[35] T. Fearn. "Discriminant analysis," *NIR news* 4 4-5 (1993).

[36] D. Foley. "Considerations of sample and feature size," *IEEE Transactions on Information Theory* **18** 618-626 (1972).

[37] I. Frank. "Beyond linear least squares regression," *Analytical Chemistry* **6** 271-275 (1987).

[38] I. Frank and J. Friedman. "Classification: oldtimers and newcomers," *Journal of Chemometrics* **3** 463-475 (1989).

[39] I. Frank and S. Lanteri. "Classification models: discriminant analysis, SIMCA, CART," *Chemometrics and Intelligent Laboratory Systems* **5** 247-256 (1989).

[40] I. Frank. "A nonlinear PLS model," *Chemometrics and Intelligent Laboratory Systems* 8 109-119 (1990).

[41] I. Frank. "A statistical view of some chemometrics regression tools," *Technometrics* **35** (2) 109-135 (1993).

[42] M. Frazier and B. Jawerth. "A discrete transform and decompositions of distribution spaces," *Journal of Functional Analysis* **93** 34-170 (1990).

[43] J. Friedman. "A recursive partitioning decision rule for nonparametric classification," *IEEE Transaction on Computers* **26** 404-408 (1977).

[44] J. Friedman. "Regularized discriminant analysis," *Journal of the American Statistical Association* **84** 165-175 (1989).

[45] J. Friedman. "Multivariate adaptive regression splines," (with discussion) *The Annals of Statistics* **19** 1-141 (1991).

[46] K. Fukunaga and R. Hayes. "Estimation of classifier performance," *IEEE Pattern Analysis and Machine Intelligence* **11** 1087-1101 (1989).

[47] K. Fukunaga and R. Hayes. "Effects of sample size in classifier design," *IEEE Pattern Analysis and Machine Intelligence* **11** 873-885 (1989).

[48] K. Fukunaga. "Introduction to Statistical Pattern Recognition," Academic Press, Boston (1990).

[49] D. Gabor. "Theory of communication," *Journal Institute of Electrical Engineers* **93** (3) 429-457.

[50] A. Garrido Frenich, D. Jouan-Rimbaud, D. Massart, S. Kuttatharmmakul, M. Martinez Galera and J. Martinez Vidal. "Wavelength selection method for multicomponent spectrophotometric determinations using partial least squares," *Analyst* **120** 2787-2792 (1995).

[51] P. Geladi and B. Kowalski. "An example of 2-block predictive partial least-squares regression with simulated data," *Analytica chimica Acta* **185** 19-32 (1986).

[52] P. Geladi and B. Kowalski. "Partial least-squares regression: A tutorial," *Analytica Chimica Acta* **185** 1-17 (1986).

[53] C. Gilbert, S. Kokot and U. Meyer. "Application of drift spectroscopy and chemometrics for the comparison of cotton fabrics," *Applied Spectroscopy* **47** 741-748 (1993).

[54] L. Gordon and R. Ohlsen. "Consistent nonparametric regression from recursive partitioning schemes," *Journal of Multivariate Analysis* **10** 611-627 (1980).

[55] P. Gorry. "General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method," *Analytical Chemistry* **62** 570-573 (1990).

[56] A. Green and M. Craig. "Analysis of aircraft spectrometer data with logarithmic residuals" In Proceedings of the AIS Workshop, Jet Propulsion Laboratory, Pasadena, California, 85-41 (1985).

[57] M. Griep, I. Wakeling, P. Vankeerberghen and D. Massart. "Comparison of semirobust and robust partial least squares procedures," *Chemometrics and Intelligent Laboratory Systems* **29** 37-50 (1995).

[58] D. Hand. "Discrimination and Classification," Wiley, New York (1981)

[59] T. Hastie and R. Tibshirani. "Generalized Additive Models," London, Chapman and Hall, (1990).

[60] T. Hastie, R. Tibshirani and A. Buja. "Flexible discriminant analysis," *Journal of the American Statistical Association* **89** 1255-1270 (1994).

[61] T. Hastie, A. Buja and R. Tibshirani. "Penalized discriminant analysis," *Annals of Statistics* **2** 73-102 (1995).

[62] I. Helland. "On the structure of partial least squares regression," *Communications in Statistics – Simulations and Computations* **17** 139-153 (1988).

[63] P. Heller, H. Resnikoff and R. Wells (Jr). "Wavelet matrices and the representation of discrete functions." In Wavelets - a tutorial in theory and applications, C. Chui (Ed.) Academic Press, 15-50 (1992).

[64] E. Henrichon and K. Fu. "A nonparametric partitioning procedure for pattern classification," *IEEE Transactions on Computers* **18** 614-624 (1969).

[65] D. Hirst. "Error-rate estimation in multiple-group linear discriminant analysis," *Technometrics* **38** 389-399 (1996).

[66] A. Hoerl and R. Kennard. "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics* **12** 55-67 (1970).

[67] A. Hoskuldsson. "PLS regression methods," *Journal of Chemometrics* **2** 211-228 (1988).

[68] A. Jain and B. Chandrasekaran. "Dimensionality and sample size considerations in pattern recognition practice," In Handbook of statistics (Vol. 2), P. Krishnaiah and L. Kanal (Eds.) Amsterdam, 835-855 (1982).

[69] A. Jain and W. Waller. "On the optimal number of features in the classification of multivariate gaussian data," *Pattern Recognition* **10** 365-374 (1978).

[70] B. Jawerth and W. Sweldens. "An overview of wavelet based multiresolution analyses," *SIAM Review* **36** 377-412 (1994).

[71] R. Johnson and D. Wichern. "Applied multivariate statistical analysis," Prentice Hall, New Jersey (1988).

[72] I. Jolliffe. "Principal component analysis," Springer Verlag, New York (1986).

[73] P. Jonathan, W. McCarthy and A. Roberts. "Discriminant analysis with singular covariance matrices," *Journal of Chemometrics* (1995), submitted for publication.

[74] D. Jouan-Rimbaud, D. Massart, R. Leardi, and O. Denoord. "Genetic algorithms as a tool for wavelength selection in multivariate calibration" *Analytical Chemistry* **67** 4295-4301 (1995).

[75] D. Jouan-Rimbaud, B. Walczak, D. Massart, I. Last and K. Prebble. "Comparison of multivariate methods based on latent vectors and methods based on wavelength selection for the analysis of near-infrared spectroscopic data," *Analytica Chimica Acta* **304** 285-295 (1995).

[76] G. Kaiser. "A friendly guide to Wavelets," Birkhäuser, Boston, (1994).

[77] J. Kalivas. "Two reference data sets of near infrared spectra," *Chemometrics and Intelligent Laboratory Systems* submitted.

[78] J. Kautsky and R. Turcajová. "Adaptive Wavelets for Signal Analysis," In Proceedings 6th Int. Conf. Computer analysis of images and patterns, Springer Verlag, Prague, 906-911 (1995).

[79] J. Kautsky and R. Turcajová. "Pollen product factorization and construction of higher multiplicity wavelets," *Linear Algebra and it Applications* **22** 241-260 (1995).

[80] J. Kautsky. "A matrix approach to discrete wavelets." In Wavelets: Theory, Algorithms and Applications, C. Chui, L. Montefusco and L. Puccio (Eds.) 117-335 (1994).

[81] J. Kautsky. "An algebraic construction of discrete wavelet transforms," *Applications of mathematics* **3** (38) 169-193 (1993).

[82] B. Kowalski and S. Wold. "Pattern recognition in chemistry." In Handbook of statistics (Vol. 2), P. Krishnaiah and L. Kanal (Eds.) Amsterdam, 673-697 (1982).

[83] P. Lachenbruch and R. Mickey. "Estimation of error rates in discriminant analysis," *Technometrics* **10** 1-11 (1968).

[84] R. Leardi. "Genetic algorithms as a strategy for feature selection," *Journal of Chemometrics* **6** 267-281 (1992).

[85] R. Leardi. "Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection," *Journal of Chemometrics* **8** 65-79 (1994).

[86] R. Learned and A. Wilsky. "A wavelet packet approach to transient signal classification," *Applied and Computational Harmonic Analysis*, **2** 265-278 (1995).

[87] L. Lebart, A. Morineau and K. Warwick. "Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices," Wiley, New York (1984).

[88] J. Lighthill. "Introduction to Fourier analysis and generalised functions," University Press, Cambridge (1962).

[89] C. Lucasius and G. Kateman. "Understanding and using genetic algorithms," *Chemometrics and Intelligent Laboratory Systems* **19** 1-33 (1993).

[90] A. Mabbett, M. Stone and J. Washbrook. "Cross-validation selection of binary variables in differential diagnosis," *Applied Statistics* **29** 198-204 (1980).

[91] S. Mallat. "Multifrequency channel decompositions of images and wavelet models," *IEEE Transactions on Acoustics Speech and Signal Processing* **37** 2091-2110 (1989).

[92] S. Mallat. "A Theory for multi-resolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **11** 674-693, (1989).

[93] Y. Mallet, D. Coomans and O. de Vel. "Robust and non-parametric methods in multiple regression of environmental data." In Handbook of environmental chemistry (Vol. 2g), J. Einax (Ed.), Springer-Verlag, Berlin 161-208 (1995).

[94] Y. Mallet, D. Coomans, O. de Vel and J. Kautsky. "Discrimination of high dimensional data using adaptive wavelets," *Computing Science and Statistics*, **28** 389-393 (1997).

[95] Y. Mallet, O. de Vel and D. Coomans. "Wavelet technology for industrial quality control," In proceedings Intelligent processing and manufacturing of materials (1997) to appear.

[96] Y. Mallet, D. Coomans, J. Kautsky and O. de Vel. "Classification using adaptive wavelets for feature extraction," *IEEE Pattern Analysis and Machine Intelligence* (1997) in press.

[97] Y. Mallet, D. Coomans and O. de Vel. "Recent developments in discriminant analysis on high dimensional spectral data," *Chemometrics and Intelligent Laboratory Systems* **35** 157-173 (1996).

[98] R. Manne. "Analysis of two partial-least-squares algorithms for multivariate calibration," *Chemometrics and Intelligent Laboratory Systems* **2** 187-197 (1987).

[99] H. Mark and D. Tunnell. "Qualitative near-infrared reflectance analysis using mahalanobis distances," *Analytical Chemistry* **57** 1449-1456 (1985).

[100] H. Martens and T. Naes. "Multivariate calibration," Wiley, Chichester (1989).

[101] G. McLachlan. "The bias of the apparent error rate in discriminant analysis," *Biometrika* **32** 529-515 (1976).

[102] G. McLachlan. "Discriminant analysis and statistical pattern recognition," Wiley, New York (1992).

[103] W. Meisel and D. Michalopoulous. "A partitioning algorithm with application in pattern classification and the optimization of decision trees," *IEEE Transactions on Computers* **22** 93-103 (1973).

[104] Y. Meyer. "Wavelets: algorithms and applications. SIAM, Philadelphia (1993).

[105] M. Misiti. "Wavelet toolbox user's guide," Math Works, Natick, Massachutes, (1996).

[106] R. Myers. "Classical and modern regression with applications," PWS-KENT, Boston (1990).

[107] G. Nason and B. Silverman. "The discrete wavelet transform in S," *Journal of Computational and Graphical Statistics* **3** 163-191 (1994).

[108] M. Norusis/SPSS Inc. "SPSS professional statistics 6.1," SPSS Inc., Chicago (1994).

[109] M. Norusis/SPSS Inc. "SPSS for windows base system user's guide release 6.1" SPSS Inc., Chicago, (1994).

[110] D. Pavia, G. Lampman and G. Kriz. "Introduction to spectroscopy," Saunders College Publishing, Orlando (1996).

[111] PIMA II. "Operations Manual," Integrated Spectronics, Baulkham Hills (1994).

[112] W. Press, S. Teukolsky, W. Vetterling and B. Flannery. "Numerical Recipes in C: the Art of Scientific Computing," Cambridge University Press, New York (1992).

[113] S. Raudys and V. Pikelis. " On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition," *IEEE Pattern Analysis and Machine Intelligence* **2** 242-252 (1980).

[114] S. Raudys and A. Jain. "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Pattern Analysis and Machine Intelligence* **13** 252-264 (1991).

[115] W. Rayens and T. Greene, "Covariance pooling and stabilization for classification," *Computational Statististics and Data Analysis* **11** 17-42 (1991).

[116] B. Ripley. "Neural networks and related methods for classification," *Journal Royal Statistical Society B.* **56** 409-456 (1994).

[117] N. Saito. "Local feature extraction and its applications using a library of bases," PhD Thesis, Yale University (1994).

[118] N. Saito and R.R. Coifman. "Local discriminant bases," In Mathematical imaging: wavelet applications in signal and image processing II, (Vol. 2303) A. Laine and M. Unser (Eds.) (1994).

[119] G. Smith and F. Campbell. "A critique of some ridge regression methods," *Journal of the American Statistical Association* **75** 74-103 (1980).

[120] R. Snee. "Validation of regression models: methods and examples," *Technometrics* **19** 415-428 (1977).

[121] S. Sommer and R. Staudte. "Robust variable selection in regression in the presence of outliers and leverage points," *Australian Journal of Statistics* **37** 323- (1995).

[122] SAS Institute Inc. "SAS/STAT User's Guide, Release 6.03 Edition," Cary, NC: SAS Institute Inc., (1988)

[123] A. Savitzky and M. Golay. "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry* **36** 1627-1639 (1964).

[124] F. Scheinmann. "An introduction to spectroscopic methods for the identification of organic compounds,' Volume 1, Oxford, Pergamon Press (1970).

[125] G. Seber. "Multivariate observations," John Wiley and Sons, New York (1984).

[126] Statistical Sciences. "S-PLUS User's Manual, Version 3.3 for Windows," StatSci, a division of MathSoft, Inc., Seattle (1995).

[127] P. Steffen, P. Heller, R. Gopinath and C. Burrus. "Theory of m-band wavelet bases," *IEEE Transactions in Signal Processing* **41** 3497-3511 (1993).

[128] G. Strang and T. Nguyen. "Wavelets and filter banks," Wellesley-Cambridge Press, Wellesley (1996).

[129] W. Sweldens. "The lifting scheme: a construction of second generation wavelets," Preprint Department of Mathematics, University of South Carolina (1994).

[130] H. Szu, B. Telfer and S. Kadambe. "Neural network adaptive wavelets for signal representation and classification," *Optical Engineering,* **31** 1907-1916 (1992).

[131] R. Tate, D. Watson and S. Eglen. "Using wavelets for classifying human *in vivo* magnetic resonance spectra," In Wavelets and statistics, A. Antoniadis and G. Oppenheim (Eds.) (1995).

[132] M. Tatsuoka. "Multivariate analysis: techniques for educational and psychological research," Wiley, New York (1971).

[133] B. Telfer, H. Szu, G. Dobeck, J. Garcia, H. Ko, A. Dubey and N. Witherspoon. "Adaptive Wavelet Classification of Acoustic and Backscatter and Imagery," *Optical Engineering,* **33** 2192-2203 (1994).

[134] T. Li, C. Lucasisu and G. Kateman. "Optimization of calibration data with the dynamic genetic algorithm," *Analytica Chimica Acta* **268** 123-134 (1992).

[135] B. Toussaint. "Bibliography on estimation of misclassification," *IEEE Transactions on Information Theory* **20** 472-479 (1974).

[136] R. Turcajová "Commpactly supported wavelets and their generalizations: An algebraic approach," PhD Thesis, The Flinders University of South Australia (1995).

[137] R. Turcajová and J. Kautsky. "Shift products and factorizations of wavelet matrices," *Numerical Algorithms* 8 27-54 (1994).

[138] W. Venables and B. Ripley. "Modern applied statistics with S-Plus," Springer-Verlag, New York (1994).

[139] B. Vidaković and P. Muller. "Wavelets for kids: A tutorial introduction," Unpublished *ftp://ftp.isds.duke.edu/pub/brani/papers/wav4kids[A-B].ps.Z*

[140] A. Wacker and T. El-Sheikh. "Average classification accuracy over collections of Gaussian problems – common covariance matrix case," *Pattern Recognition* **17** 259-273 (1984).

[141] B. Walczak, B. van den Bogaert and D. Massart. "Application of wavelet packet transform in pattern recognition of Near-IR data," *Analytical Chemistry* **68** 1742-1747 (1996).

[142] S. Walker and H. Straw. "Spectroscopy," Volume 2, London, Chapman and Hall (1967).

[143] M. Wickerhauser. "Adapted wavelet analysis from theory to software," A. K. Peters, Ltd, (1994).

[144] S. Wold. "Pattern recognition by means of disjoint principal component models," *Pattern Recognition* 8 127-139 (1976).

[145] S. Wold. "Partial least squares," Encyclopedia of Statistical Sciences, (Vol. 6), S. Kotz and N. Johnson (Eds.), Wiley, New York, 581-591 (1985).

[146] W. Wu, Y. Mallet, B. Walczak, W. Penninckx and D.L. Massart. "Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data", *Analytica Chimica Acta* **329** 257-265 (1996).

[147] W. Wu, B. Walczak, W. Penninckx and D.L. Massart. "Feature reduction by Fourier transform in pattern recognition of NIR data," *Analytica Chimica Acta* **331** 75-83 (1996).

[148] W. Wu, B. Walczak, D. Massart, K. Prebble and I. Last. "Spectral transformation and wavelength selection in near-infrared spectra classification," *Analytica Chimica Acta* **315** 243-255 (1995).

[149] C. Yeh and C. Spiegelman. "Partial least squares classification and regression trees," *Chemometrics and Intelligent Laboratory Systems* **22** 17-23 (1994).

[150] T. Young and T. Calvert "Classification estimation and pattern recognition," Elsevier, New York (1974).