



Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Statistical Models For Food Authenticity

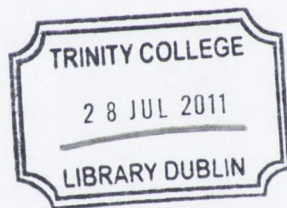
A thesis submitted to the University of Dublin, Trinity College
in partial fulfillment of the requirements for the degree of
Doctor in Philosophy

Department of Statistics, University of Dublin, Trinity College



2009

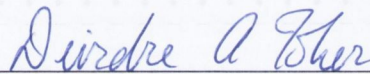
Deirdre Ann Toher



THESIS
9198

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and Deirdre Ann Toher.



Deirdre Ann Toher

Abstract

The authentication of food samples pose a particular problem for regulators. The routine testing of premium food products, most likely to be subject to manipulation for commercial gain, is only feasible if the testing method does not damage the product. Near Infrared (NIR) spectroscopy is one such method that is both fast and non-invasive. However, unlike other spectroscopic methods, peaks in the resulting NIR curves are at imprecise locations, requiring further statistical analysis if it is to be used for the classification of samples.

Three NIR datasets are examined in this thesis – two are related to the identification of adulterated samples, the third is a study on the identification of types of meats. Other commonly available, non-NIR, datasets are used for illustrative purposes.

The models developed in this thesis must be suitable for use by chemists with access only to personal computers and have reasonable computational time if they are to be adopted into practice. Some of the methods developed are refinements to existing methods such as the development of the use of information criteria for the selection of the number of parameters to use with Partial Least Squares Regression and the incorporation of a semi-supervised framework with Fisher's Linear Discriminant Analysis. A variety of dimension reduction approaches are used with model-based discriminant analysis and with classification based on a homogeneous group versus a heterogeneous group.

Throughout this thesis model assessment is on the basis of test set performance, using 50%, 25% and 10% of the observations in the datasets for training the models and 50%, 75% and 90% of the observations to test on in order to assess the robustness of the various modelling approaches to sample size.

Acknowledgements

I have been in the unusual, but fortunate position of having not one but two supervisors for the duration of my thesis. Dr Brendan Murphy has provided me with support from a statistical viewpoint, while Dr Gerard Downey provided me with the support from a food scientist's viewpoint. Brendan has supported my interest in statistical research from an undergraduate level and has given me all the support that could be wished for. Gerry has provided a fresh eye on my work, giving practical advice on what is needed for methods to be used by spectroscopists.

My research has been funded by the Walsh Fellowship scheme in Teagasc and the additional funding by Science Foundation Ireland has enabled me to travel to conferences and working groups both in Ireland and abroad.

I must thank the entire of the Statistics Department of Trinity College Dublin, especially those who contributed to the working group meetings. The feedback given at these meetings proved invaluable towards the shaping of the content of this thesis. To my fellow postgraduate students within the Department, past and present, especially those who have shared the office of room 118 – thanks for lightening the tone when required and helping me to find the appropriate words of late!

Similarly the spectroscopy group within the Ashtown Food Research Centre gave me a useful insight into the problems faced in the analysis of spectroscopy data and also into the types of problems that spectroscopy is considered a potentially valuable tool. The members of this group, among other postgraduate students in the Prepared Foods Department welcomed me into an unfamiliar setting of a food research facility, without plaguing me too much with statistics questions over coffee.

I have been privileged to have been able to attend two summer sessions of the Model-based Clustering group; in Dublin in 2007 and in Seattle in 2008. The suggestions given by members of this working group have been particularly helpful in

directing my research.

My parents have managed to survive consecutive Octobers with an offspring writing their PhD thesis. Their continuing patience and support has been duly noted and greatly appreciated; my mother in particular has spent hours proofreading a thesis with no prior knowledge of chemistry or statistics. My brother Cormac too has contributed from Germany to the proofreading process, he has also provided a useful sounding board throughout my studies.

Deirdre Ann Toher

University of Dublin, Trinity College

June 2009

Contents

Abstract	i
Acknowledgements	iii
List of Tables	ix
List of Figures	xi
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Food Authenticity Studies	2
1.3 Overview of Chapters	3
1.4 Research Contributions	4
Chapter 2 Food Authenticity Data	7
2.1 Near Infrared Spectroscopy	7
2.1.1 Structure of NIR spectrum	10
2.2 NIR Data	11
2.2.1 Honey Samples	11
2.2.2 Meat Samples	13
2.2.3 Adulterated Olive Oils	15
2.3 Other Food Authenticity Data	16
2.3.1 Geographic Origin of Olive Oil Samples	16
2.3.2 Wine	17
Chapter 3 Statistical Methodology	19
3.1 Partial least squares regression	19

3.1.1	Algorithm for PLSR	19
3.1.2	Number of Parameters	20
3.2	Soft Independent Modelling of Class Analogies	21
3.3	Likelihood Based Statistical Inference	24
3.4	Model-based Discriminant Analysis	26
3.4.1	Updating	28
3.4.2	Implementing the EM Algorithm	30
3.5	Dimension Reduction Techniques	32
3.5.1	Wavelet Analysis	32
3.5.2	Wavelength Selection Methods	34
3.6	Model Selection Techniques	36
3.6.1	Bayesian Information Criterion (BIC)	37
3.6.2	F fold cross validation	37
3.6.3	Brier's Score	38
3.7	Performance Comparison	38
3.7.1	PLSR	39
3.7.2	Model Based Methods	39
3.8	Conclusions	46
Chapter 4 Group of Interest based Classification		49
4.1	General Concept	49
4.2	Variable Selection	53
4.2.1	Variable Selection Procedure	53
4.3	Varying the Threshold	58
4.3.1	Evaluating the Threshold Directly	59
4.3.2	Behaviour of the Threshold	60
4.4	Results	60
4.4.1	NIR datasets	60
4.5	Conclusions	76
Chapter 5 Updating Fisher's Linear Discriminant Analysis		79
5.1	Fisher's LDA	79
5.1.1	Theory	80

5.1.2	Prediction	81
5.1.3	Updating	81
5.1.4	Implementation of Semi-supervised Fisher's Linear Discriminant Analysis	82
5.2	Data	86
5.2.1	Fisher's Iris Data	86
5.2.2	Wine Data	88
5.2.3	Meats Data	91
5.2.4	Honey Data	93
5.2.5	Olive Oil Data	97
5.3	Conclusions	99
Chapter 6 Generalising Fisher's Linear Discriminant Analysis		103
6.1	Further Generalisation	103
6.1.1	$\Sigma_g = \Sigma = \lambda DAD' \forall g$	104
6.1.2	$\Sigma_g = \lambda I$	104
6.1.3	$\Sigma_g = \lambda_g I$	105
6.1.4	$\Sigma_g = \lambda_g D_g A_g D'_g$	106
6.1.5	Numerical Approximations	107
6.2	Conclusions	108
Chapter 7 Conclusions and Further Work		111
7.1	Conclusions	111
7.1.1	Dimension Reduction Techniques	111
7.1.2	Identification of Adulterated Samples	113
7.1.3	Classification for Multiple Groups	113
7.2	Further Work	114
7.2.1	R package	114
7.2.2	Implementation of generalised Fisher's LDA	114
7.2.3	Variable Selection	114
7.3	Final Comments	115
Appendix A Details of Calculations		117
A.1	Rotation	117

Appendix B	Details of Numerical Approximations	119
B.1	Numerical Approximation of $\hat{\alpha}$	119
B.1.1	Gradient for α	119
B.1.2	Hessian for α	121

List of Tables

3.1	Parametrizations of the covariance matrix Σ_g	27
3.2	Classification Performance: PLSR on Honey Data	41
3.3	Classification Performance: Wavelength Selection, MBDA on Honey Data	42
3.4	Classification Performance: Wavelength Selection, MBDA on Honey Data	43
3.5	Classification Performance: Wavelength Selection, MBDA on Olive Oil Data	44
3.6	Classification Performance: Wavelength Selection, MBDA on Olive Oil Data	45
3.7	Classification Performance: PLSR on Olive Oil Data	46
4.1	Classification Performance, Honey Data: $f_o(x) = 1/V$	63
4.2	Classification Performance, Honey Data: $f_o(x) = \sum_{k=1}^K q_k f(x \mu_k, \Sigma_k)$	64
4.3	Classification Performance, Olive Oil Data: $f_o(x) = 1/V$	68
4.4	Classification Performance, Olive Oil Data: $f_o(x) = \sum_{k=1}^K q_k f(x \mu_k, \Sigma_k)$	69
4.5	Comparison of Error Rates: direct calculation of a threshold versus $\tau = (1 - p_g)/p_g V$	69
4.6	Comparison of Error Rates for τ calculated directly	72
4.7	Relationship of τ directly calculated to $1/V$: $(1 - p_g)/p_g V$ used for variable selection	73
4.8	Relationship of τ directly calculated to $1/V$: τ directly calculated used for variable selection	74
4.9	Classification Performance: Wavelength Selection using $1/V$ MBDA on Honey and Olive Oil Data	75

5.1	Classification Performance: Iris Data	88
5.2	Classification Performance: Wine Data	89
5.3	Classification Performance: Full Wine Data	89
5.4	Classification Performance: Meats Data	93
5.5	Classification Performance: Honey Data	93
5.6	Classification Performance: Olive Oil Data	98
5.7	Classification Performance: Olive Oil Data	99

List of Figures

2.1	Electromagnetic Spectrum	9
2.2	Visible Spectrum	9
2.3	NIR regions	10
2.4	Sample spectra of unmodified bacteria in MRD	11
2.5	Flow diagram of Adulteration Process	14
2.6	NIR spectra of honeys	14
2.7	NIR spectra of meats	15
2.8	NIR spectra of olive oils	16
2.9	Map of Italy	17
3.1	Volume, Shape and Orientation of Σ_g	28
3.2	Means Separated by 4, 2 and 1 Standard Deviation(s)	30
3.3	Wavelet Functions	33
3.4	Actual and Reconstructed, Thresholded Spectra	34
3.5	Wavelengths Selected using Between and Within Group Covariances .	36
4.1	Selecting the first variable	56
4.2	Selecting the second variable	56
4.3	Selecting the third variable	57
4.4	Selecting the fourth variable	58
4.5	One dimensional example of the behaviour of the threshold	59
4.6	Behaviour of τ and volume	60
4.7	Variables selected when $f_o(x) = \sum_{k=1}^K q_k f(x \mu_k, \Sigma_k)$, NIR honey data 50%/50% split	65
4.8	Variables selected when $f_o(x) = \sum_{k=1}^K q_k f(x \mu_k, \Sigma_k)$, NIR honey data 25%/75% split	66

4.9	Variables selected when $f_o(x) = \sum_{k=1}^K q_k f(x \mu_k, \Sigma_k)$, NIR honey data 10%/90% split	67
4.10	Variables selected when $f_o(x) = \sum_{k=1}^K q_k f(x \mu_k, \Sigma_k)$, NIR olive oil data 50%/50% split	70
4.11	Variables selected when $f_o(x) = \sum_{k=1}^K q_k f(x \mu_k, \Sigma_k)$, NIR olive oil data 25%/75% split	71
4.12	Relationship between τ directly calculated and $1/V$: NIR olive oil data 50%/50%	74
5.1	Projections of the Iris Dataset	87
5.2	Coefficients of the Discriminant Functions of the Iris Dataset at 10%/90% split	87
5.3	Projections of the Wine Dataset	90
5.4	Coefficients of the Discriminant Functions of the Wine Dataset at 10%/90% split	91
5.5	Projections of the Minced Meat Dataset	92
5.6	Coefficients of the Discriminant Functions of the Meats Dataset at 50%/50% split	94
5.7	Projections of the Pure and Adulterated Honey Dataset: 50%:50% split	95
5.8	Projections of the Pure and Adulterated Honey Dataset: 25%:75% split	95
5.9	Coefficients of the Discriminant Function of the Honey Dataset at 50%/50% split	97
5.10	Coefficients of the Discriminant Function of the Olive Oil Dataset at 50%/50% split	98
5.11	Projections of the Pure and Adulterated Olive Oil Dataset	100

Chapter 1

Introduction

1.1 Motivation

The main purpose of this thesis is to provide statistical methods for use by food scientists in the process of food authentication. A variety of different methods of analysing the food samples are considered, with most emphasis being placed on providing statistical techniques for Near Infrared (NIR) data.

The methods developed throughout this thesis, while targeted towards NIR data, are also designed to be suitable to more general applications: especially situations where there is high dimensional, highly correlated data.

The statistical techniques developed should remove the subjectivity from the classification process when using NIR data. Introducing a probabilistic framework for the classification process enables consistent measures of uncertainty about the individual classification decisions to be made.

Using R for all computations throughout this thesis enables the methods to be easily reproduced for a variety of computation platforms.

In order for such methods to be adopted by the chemists who use NIR, methods should be as simple as possible, with the underlying reasoning behind the modelling strategy easily comprehensible.

1.2 Food Authenticity Studies

The main aim of food authenticity studies is to detect when foods are not what they claim to be and thereby prevent economic fraud or possible damage to health. Foods that are susceptible to such fraud are those which are expensive and subject to the vagaries of weather during growth or harvesting *e.g.* coffee, various fruits, herbs and spices. Food fraud can generate significant amounts of money (*e.g.* several million US dollars) for unscrupulous traders so the risk of adulteration is real.

This type of fraud not only applies to the consumer market (generally through inaccurate labelling), but also to the industrial ingredients market, where food traceability and quality control are of increasing importance. Analytical techniques that are available to industry include “wet chemistry” techniques – invasive and destructive, but easy to interpret and other non-invasive techniques that require more interpretation of the results such as Near Infrared (NIR) spectroscopy.

Ingredient fraud can extend beyond the typical human food chain. On the 15th March 2006, the Food and Drug Administration (FDA) in the United States announced that it had discovered that some pet foods were killing cats and dogs. On further investigation, the source of contamination was found in vegetable proteins imported from China and used not only in pet food, but also in farm animal and fish feed. Although risk to people from eating the resultant foods was low, it highlighted the need for adequate controls at all points that have the potential to enter the human food chain.

An even more serious problem has emerged recently with the increased prevalence of contaminated or counterfeited drugs, resulting in ineffective, dangerous products being released onto the market. This is not a new problem – the World Health Organisation noted in a 1999 report that the problem is referred to in writings dating back to the fourth century BC; Dioscorides in the first century AD in Greece identified adulterated drugs and advised others on their detection. Newton et al. (2006) outline the scale of the problem and some of the current detection methods used in counterfeit drug detection.

The most recent major food scare emerged in China when it was reported on the 13th September by J. McDonald that a national investigation into the contamination of baby milk formula with melamine was being undertaken. The scale of the con-

tamination problem became evident in the later report by S. McDonald which noted that about 53,000 children had been sickened by the contaminated milk products.

1.3 Overview of Chapters

A brief outline of the research completed follows:

Chapter 2: Food Authenticity Data

Models developed throughout this thesis are applied to food authenticity problems. Examples of how such problems arise in practical terms are given and the motivation behind the solution of these problems are addressed. The specific data on which the methods are applied are introduced in this chapter.

Chapter 3: Statistical Methodology

Existing statistical methods are introduced, especially those that have been developed within the chemometric literature to analyse near infrared spectroscopic data. Details of the implementation of Partial Least Squares Regression, Soft Independent Modelling of Class Analogies and Model-based discriminant analysis (with and without updating procedures) are given and the issues of model selection and evaluation are addressed.

Chapter 4: Group of Interest Based Classification

In food authentication applications there is often an imbalance in the information available about samples. Comprehensive profiling of the authentic products can be undertaken. However, to accomplish the same for the unauthentic products would be almost impossible. This chapter illustrates a method whereby the authentic observations can be treated as a single, homogeneous group, while the other observations are modelled using a more flexible framework than that used in traditional discriminant problems.

Chapter 5: Updating Fisher’s Linear Discriminant Analysis

Finding a projection of the data that maximises the separability of groups enables the elimination of the need for separate dimension reduction. Fisher’s Linear Discriminant Analysis is generalised to incorporate a semi supervised perspective.

Chapter 6: Generalising Fisher’s Linear Discriminant Analysis

Fisher’s Linear Discriminant Analysis assumes that the covariance matrices are the same across groups. This is generalised using the likelihood ratio so that the restriction that all the covariance matrices are the same can be relaxed.

Chapter 7: Conclusions and Further Work

This chapter summarises the findings of the different approaches towards dimension reduction. It also compares the relative appropriateness of the different discrimination approaches for both two group and multiple group classification problems. Areas of future research leading from the work undertaken towards this thesis are also examined in this chapter.

1.4 Research Contributions

The following are the main contributions made by the research contained in this thesis:

1. The development of the use of information criteria for the automatic selection of the number of parameters to use for Partial Least Squares Regression (PLSR), enabling PLSR to be used in small sample situations where cross validation is infeasible.
2. The effectiveness of alternative dimension reduction methods to be used in association with model-based discriminant analysis have been studied.
3. Demonstrating that semi-supervised methods are highly dependent on initial model assumptions, hence are not always beneficial in classification problems.

4. Strategies for classifying observations in the presence of a single homogeneous group and a unknown number of other groups, treated as a single heterogeneous group, have been examined, incorporating variable selection techniques into the classification process.
5. Development and implementation of a semi-supervised framework for Linear Discriminant Analysis (LDA) and the generalisation of LDA in order to relax the assumption of equal covariances across groups in the semi-supervised framework.

Chapter 2

Food Authenticity Data

2.1 Near Infrared Spectroscopy

The NIR part of the electromagnetic spectrum ranges from about 700 nm to 2500 nm, lying between the visible and the infrared part of the electromagnetic spectrum as illustrated in Figure 2.1. The visible part of the spectrum is illustrated in Figure 2.2.

NIR spectroscopy is a fast, non-invasive method of examining substances. As samples do not require advance preparation, it has the potential to be used as part of an on-line quality control system. However, unlike other forms of spectroscopy, the peaks of the spectra are not well defined as this part of the spectrum is based on molecular overtones and vibrations. Thus a compound cannot be identified by locating a single narrow peak on the spectrum, rather it requires analysis of the entire range. Figure 2.3(b) illustrates the overlapping structure within just one part (2000–2498 nm) of the NIR region.

The NIR data examined in this thesis are taken at intervals of 2 nm using an NIRSystems 6500 instrument which can scan over the visible and near infrared regions, where one sensor scans from 700 to 1100 nm and another scans from 1100 to 2498 nm.

Bonds in molecules are produced by an atom sharing and/or giving electrons to another atom. Such bonds act similar to anharmonic springs, where the frequency

is the number of times the atom vibrates in a second. Using the equation

$$E_n = \left(n + \frac{1}{2}\right) \frac{h}{2\pi} \sqrt{\frac{k}{\mu}}$$

where h is Plank's constant, k is the force, μ the reduced mass, E_n is the vibrational energy for $n = (0, 1, 2, \dots)$ the energy levels can be determined. $n = 1$ represents the fundamental frequency of the molecule. $n = 2, \dots$ represent the overtone regions – where $n = m$ is associated with the $(m - 1)^{\text{th}}$ overtone region.

This spring-like behaviour is what gives rise to the fundamental frequencies (at lower frequencies / higher wavelengths) and to the overtones. Multiple overtones exist for different compounds, further hindering the direct identification of substances within the NIR region, which is mostly comprised of combinations and overtones rather than fundamental frequencies. The spring does not just oscillate in a plane, it also twists. This makes peaks harder to identify, particularly for compounds with more than one bond active in the NIR region.

With the exception of a few electronic transitions almost all of the overtones or combinations observed in the NIR region involve hydrogen. This is because hydrogen has a small mass and thus can travel further, leading to a more pronounced anharmonicity which in turn leads to greater intensity in the overtone bands. Most of the vibrations of the non-hydrogen based compounds are at lower frequencies so that only the second and higher overtones and multiple combinations fall in the NIR region. These are much weaker than the first harmonics as the intensity decreases by a factor of approximately ten as one moves up each harmonic level (or from one overtone region to the next).

Visible Spectrum

The visible part of the spectrum can be broken into the traditional colours of the rainbow: red, orange, yellow, green, blue, indigo and violet as illustrated in Figure 2.2. Thus the Near Infrared part of the spectrum is closest to the colour red on the visible spectrum.

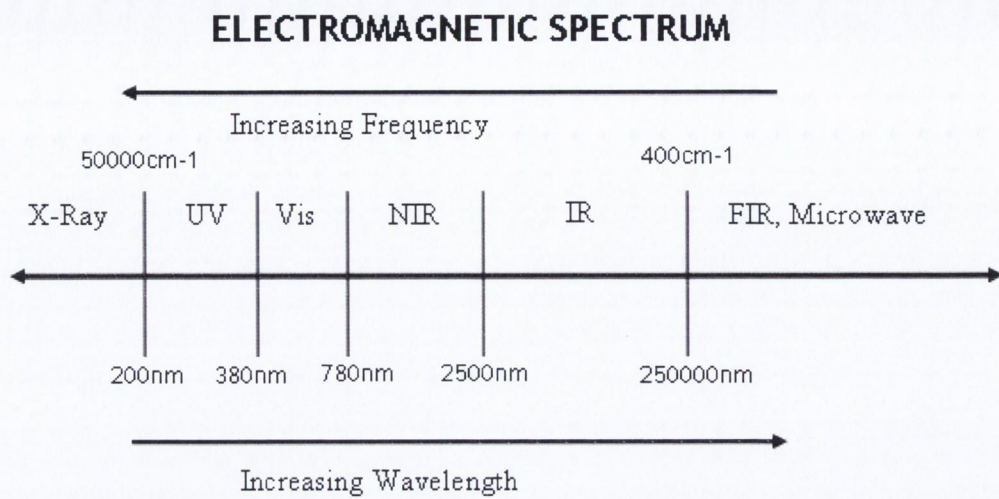


Fig. 2.1: Electromagnetic Spectrum

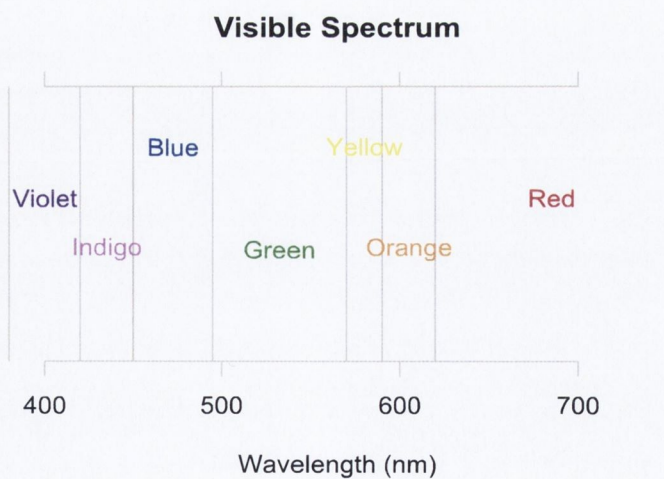


Fig. 2.2: Visible Spectrum

2.1.1 Structure of NIR spectrum

There are 4 main regions within the NIR part of the spectrum, three of which are what is known as overtone regions, while the final region is a combination region. The 3 overtone regions within the NIR spectrum exist at approximately 700–1150 nm (3rd overtone region), 1050–1650 nm (2nd overtone region) and 1470–2050 nm (1st overtone region). The combinations region (at approximately 2000–2500 nm) is illustrated in Figure 2.3(b). The theoretical peak positions illustrated by the blue bars in Figure 2.3(b) show the extent of the overlap of the different bond types in just one part of the spectrum. The overlapping of the wavelengths corresponding to different bonds is repeated in the overtone regions.

To illustrate how the same bond features in different regions of the NIR spectrum, Figures 2.4(a) and 2.4(b) illustrate the different parts of the spectrum attributed to the H₂O (water) bond.

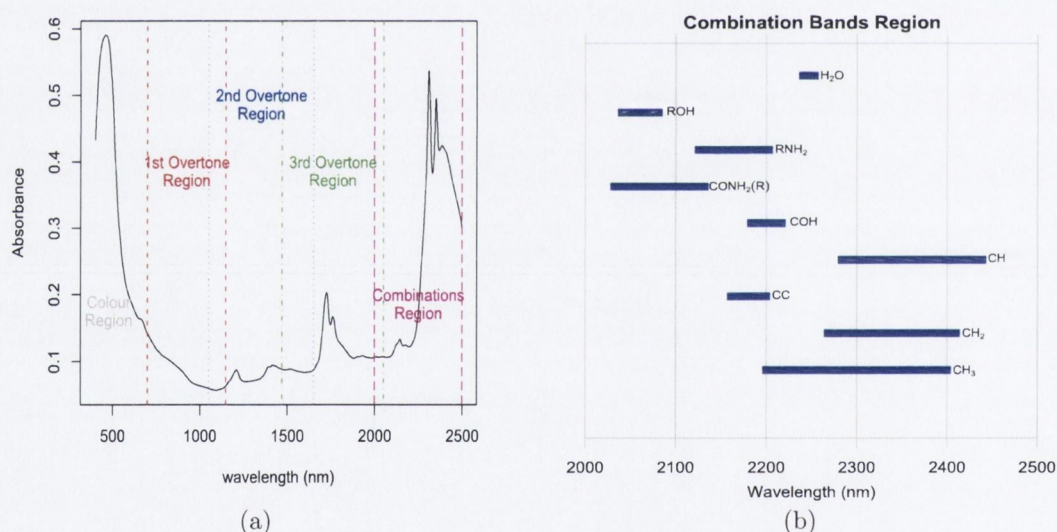


Fig. 2.3: Regions of the NIR spectrum: Figure 2.3(a) illustrates the different regions of the NIR spectrum while Figure 2.3(b) illustrates the different areas of interest within the Combination Bands Region

Figure 2.4(a) shows a sample spectra comprising almost totally of water. Superimposing the theoretical positions of the water peaks in Figure 2.4(b) illustrates the difficulty in identifying substances that comprise of more than one compound. However, this difficulty in identifying substances is somewhat counteracted by the

ability of NIR to penetrate further into a sample than other methods. It can also even measure through glass or certain types of packaging making the potential for use on industrial scale quality control evident, as products can be tested without causing damage to either the product or its packaging.

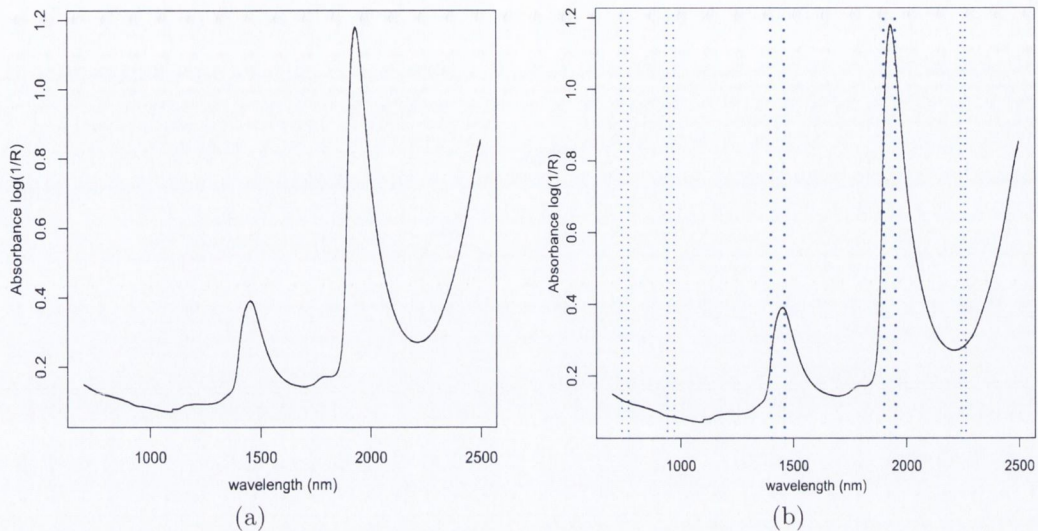


Fig. 2.4: Sample spectra of unmodified bacteria in maximum recovery diluent (MRD – mainly distilled water). Figure 2.4(a) is the spectra from 700–2498 nm; Figure 2.4(b) illustrates the theoretical locations of water peaks in the same region.

2.2 NIR Data

2.2.1 Honey Samples

Honey is defined by the EU Commission (2002) as “the natural, sweet product produced by *Apis mellifera* bees from the nectar of plants or from secretions of living plants, which bees collect, transform by combining with specific substances of their own, deposit, dehydrate, store and leave in honeycombs to ripen and mature”. As it is a relatively expensive product to produce and extremely variable in nature, honey is prone to adulteration for economic gain. Instances of honey adulteration have been recorded since Roman times when concentrated grape juice was sometimes added, although nowadays industrial syrups are more likely to be used as honey extenders.

Honey samples (157 samples) were obtained directly from bee-keepers throughout the island of Ireland. Samples were from the years 2000 and 2001; they were stored unrefrigerated from time of production and were not filtered after receipt in the laboratory. Honeys were then incubated at 40°C overnight to dissolve any crystalline material, manually stirred to ensure homogeneity and adjusted to a standard solids content (70° Brix) before spectral collection. This should help to avoid spectral complications from naturally-occurring variations in sugar concentration.

Collecting and extending the honey and recording the spectra was done at time points several months apart; the first study involved extending some of the authentic samples of honey with fructose:glucose mixtures, the second study involved extending some of the remaining authentic samples with fully-inverted beet syrup and high fructose corn syrup. All adulterant solutions were also produced at 70° Brix. Brix standardisation of honeys and adulterant solutions meant that any adulteration detected would not be simply on the basis of gross added solids.

The two studies were combined for analysis in order to reflect a more accurate picture of reality. In an ongoing practical food testing scenario, it is unlikely that all samples would be taken and processed within a very short period of time – samples would arrive for testing intermittently over an extended time period.

The fructose:glucose mixtures were produced by dissolving fructose and glucose (Analar grade; Merck) in distilled water in the following ratios:- 0.7:1, 1.2:1 and 2.3:1 w/w. Twenty-five of the pure honeys were subsampled and then adulterated with each of the three fructose:glucose adulterant solutions at three levels *i.e.* 7, 14 and 21% w/w thus producing 225 adulterated honeys.

The other adulterant solutions were generated by diluting commercially-sourced fully-inverted beet syrup (50:50 fructose:glucose; Irish Sugar, Carlow, Ireland) and high fructose corn syrup (45% fructose and 55% glucose) with distilled water. Eight authentic honeys were chosen at random to be subsampled, then were adulterated with beet invert syrup at levels of 7, 10, 14, 21, 30, 50 and 70% w/w; high fructose corn syrup was added to ten different, randomly-selected honeys (again subsampled) at 10, 30, 50 and 70% w/w. This produced 56 BI-adulterated and 40 HFCS-adulterated samples.

This adulteration scheme, as shown in Figure 2.5 was used as it represents the

most difficult classification scenario: where the aim is to differentiate between samples before and after adulteration with adulterants that have been formulated to replicate the natural composition of honey.

Both the pure and adulterated samples come from the same original source, as such there is a degree of dependence between the samples which, when considering the extent of the natural variability of honey, adds to the difficulty in differentiating between the groups. As the honey can be extremely variable, the difference introduced by the adulteration scheme may in fact be less than the natural variability within the original samples. The resultant spectra (from 1100-2498 nm) are shown in Figure 2.6.

2.2.2 Meat Samples

The spectra from a total of 231 homogenised meat samples were measured from 400-2498 nm at intervals of 2 nm. These spectra encompass both the visible and near infrared part of the electromagnetic spectrum. The samples were of raw, homogenised (minced) meat, with a total of 32 beef, 55 chicken, 34 lamb, 55 pork and 55 turkey samples. The resultant spectra are illustrated in Figure 2.7, with beef samples in black, chicken in red, lamb in green, pork in blue and turkey as the cyan coloured lines.

The meats were purchased over a period of 10-12 weeks in the form of breast meat (chicken and turkey), pork loin chops, round steak (beef) and lamb side loin chops. The samples were refrigerated overnight then prepared in order to produce the greatest quantity of lean meat in each sample by removing skin, bone, fatty and connective tissue. Excess surface moisture was removed by patting the meat samples dry before the samples were individually minced. The samples were then refrigerated again before being scanned later on the same day. The full preparation process is explained more fully by McElhinney et al. (1999).

Cross contamination or misrepresentation, intentional or otherwise, of meat products is of interest to the consumer for both religious and safety reasons. For industrial settings, the correct identification of meat products is important, especially during food scares.

The spread of the sample spectra in the 400-780 nm range in Figure 2.7 represents

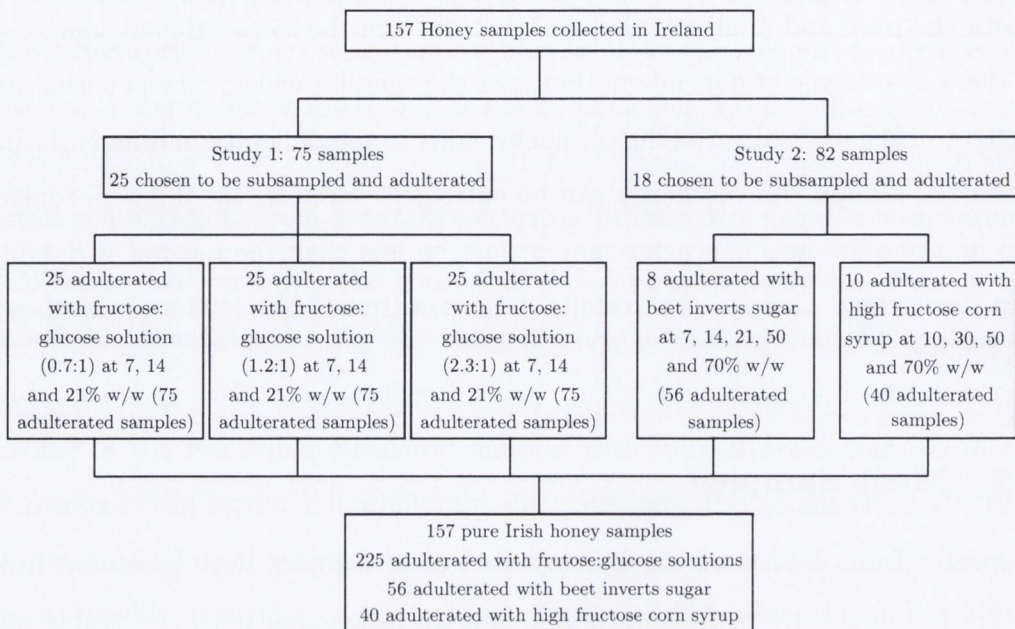


Fig. 2.5: Flow diagram of Adulteration Process

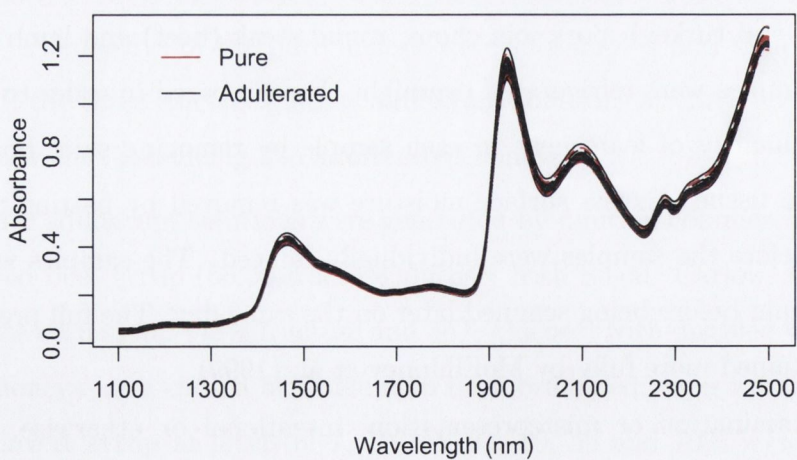


Fig. 2.6: NIR spectra of pure and adulterated honey samples

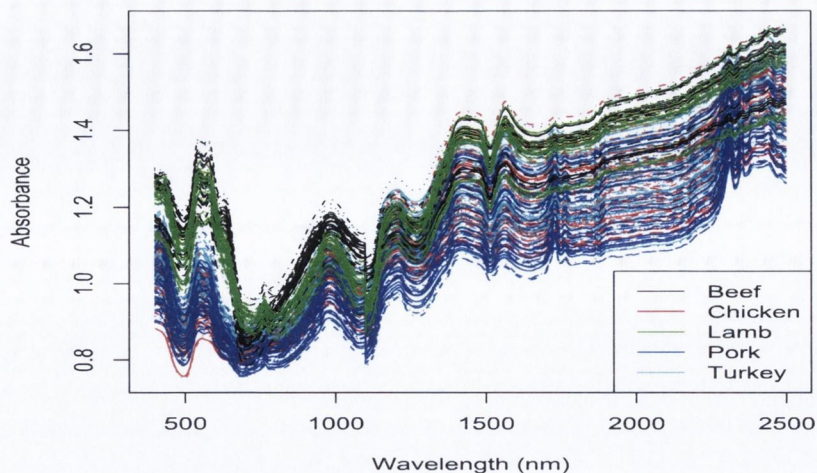


Fig. 2.7: NIR spectra of raw, homogenised meat samples

the difference in the colour of the samples, while further along the spectrum the difference originates from the different chemical composition of the types of meats. It is notable that in these samples the white and red meats are clearly separable and that the main difference in the chicken and turkey samples comes from the higher water concentrations of the chicken samples. As with most applications of NIR technology to food samples, the natural variability of the food samples is apparent in the resulting spectra.

2.2.3 Adulterated Olive Oils

This data set comprise of 46 pure extra virgin olive oil samples each of which has been subsampled into 3 samples within a laboratory setting as described by Downey et al. (2002). One subsample is left as is, another is adulterated with 1% (w/w) sunflower oil and the final subsample is adulterated with 5% (w/w) sunflower oil. Thus there are 138 spectroscopic scans in this study. The black lines (obscured) in Figure 2.8 are pure olive oil samples while the red lines are the olive oil samples that have been extended with sunflower oil.

Again, both the pure and adulterated samples come from the same original source, as such there is a degree of dependence between the samples which, when considering the extent of the natural variability of olive oils, adds to the difficulty in differentiating between the groups.

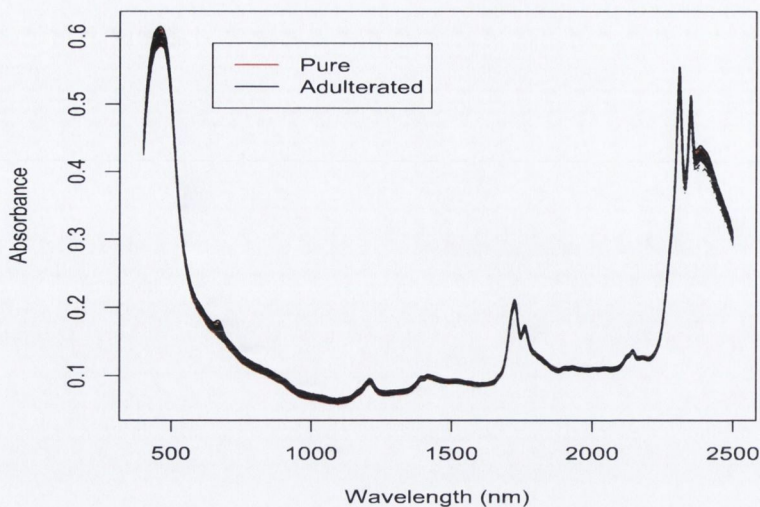


Fig. 2.8: NIR spectra of pure and adulterated olive oils

Extra virgin olive oil is a premium product, priced accordingly. If manufacturers are able to obscure the adulteration of the olive oils with much cheaper sunflower oils, the market could be compromised with a lower quality product. This, in turn, would damage not only the reputation of the company involved in the fraud, but also of the entire extra virgin branding of olive oils.

2.3 Other Food Authenticity Data

2.3.1 Geographic Origin of Olive Oil Samples

Forina and Tiscornia (1982) aimed to classify a total of 572 olive oil samples according to geographic origin from various regions in Italy. Also included in this paper were a smaller number of samples from Portugal, Israel, Lebanon, Crete and Syria; however, discriminating between the regions of Italy is of interest here, so only the Italian olive oil samples are studied. This is mainly because the Portuguese samples omit one of the variables (eicosenoic acid percentage) and were taken over different years while the other countries each had very few samples. For each of the Italian samples the percentages of eight different fatty acids were measured. These acids were *palmitic*, *palmitoleic*, *stearic*, *oleic*, *linoleic*, *linolenic*, *arachidic* and *eicosenoic* acids.

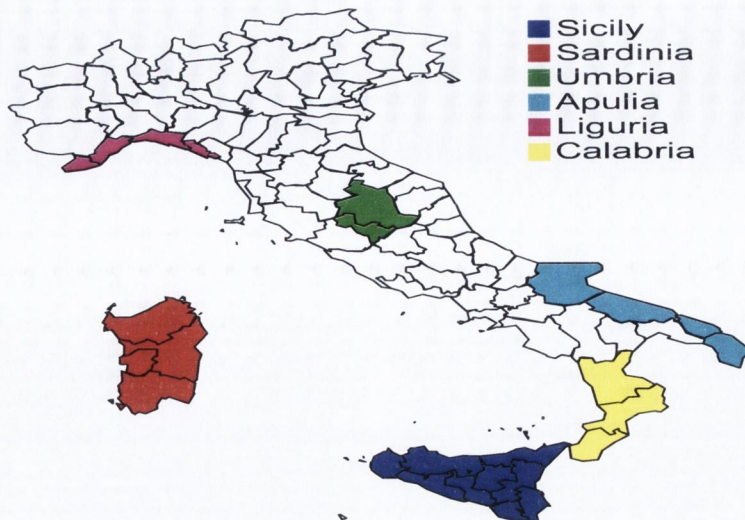


Fig. 2.9: Map of Italy with the regions of interest highlighted

The regional breakdown of the data is as follows: North Apulia (25 samples), Calabria (56 samples), South Apulia (206 samples), Sicily (36 samples), Inland Sardinia (65 samples), Coastal Sardinia (33 samples), East Liguria (50 samples), West Liguria (51 samples) and 51 samples from Umbria.

Geographic origin is another factor that increases the value of olive oil. This extends further than just on a country level – some regions within Italy produce olive oil that can be sold at a greater price than others. Misrepresenting the region of origin is thus not really a health and safety issue, rather a regulatory issue in order to prevent fraud and maintain the premium value of particular regions.

2.3.2 Wine

Forina et al. (1986) collected and analysed wine samples from the Piedmont region of Italy, which is to the north of Liguria in Figure 2.9. The common, incomplete data set using only the information about *alcohol*, *malic acid*, *ash content*, *alcalinity*, *magnesium content*, *total phenols*, *flavanoids*, *nonflavanoid phenols*, *proanthocyanins*, *intensity*, *hue*, *OD280/OD315 of phenols* and *proline* is available in the *gclus* (Hurley, 2004) package of R (R Development Core Team, 2007).

For the purposes of identifying wines solely into *Barolo*, *Grignolino* and *Barbera* the study was not well designed – the 59 Barolo wine samples come from the years

1971, 1973–4; the 71 Grignolino wines are from the years 1970–6 while the 48 Barbera wines come from the years 1974, 1976, 1978–9. Thus one cannot be certain that discrimination has not been made on the basis of year of production rather than solely by type, as intended. However, there is a significant difference in the price of these wine varieties, hence the motivation for the incorrect labelling.

With accurate information on one variable unobtainable, the rest of the variables associated with this wine data are: *sugar-free extract, fixed acidity, tartaric acid, uronic acids, pH, potassium, calcium, phosphate, chloride, OD280/OD315 of flavanoids, glycerol, 2-3-butanediol, total nitrogen and methanol*. Analysis is undertaken on both the “full” and “incomplete” wine datasets.

Chapter 3

Statistical Methodology

3.1 Partial least squares regression

Partial least squares discriminant analysis is commonly used in food authentication studies based on spectroscopic data. This method uses partial least squares regression with a binary outcome variable for two-group classification problems and seeks to optimise both the variance explained and correlation with the response variable (Hastie et al., 2001, p66-p68). Downey et al. (2003) found it to outperform other chemometric methods commonly used in the study of near infrared transfectance spectra such as Soft Independent Modelling of Class Analogies (SIMCA), which is described in Section 3.2. It has the advantage in that it can utilise highly-correlated variables for classification purposes.

Partial least squares regression (PLSR) was developed by Wold (1966*a,b*) and is based on the assumption of a linear relationship between the observed variables (*e.g.* the spectroscopy measurements) and the outcome variable (*e.g.* pure or adulterated). It is similar to principal components regression (PCR). Stone and Brooks (1990) formally explain the connection between PLSR and PCR.

3.1.1 Algorithm for PLSR

As outlined by (Hastie et al., 2001, p66-p68) each variable \mathbf{x}_j is standardized to have 0 mean and variance of 1. The response variable is y and p is the number of variables.

The original formulation of the PLS algorithm was restated in vector notation

by Frank and Friedman (1993). However an easier to follow version of the algorithm was presented by Helland (1990) as follows:

1. Define starting values for the \mathbf{x} residuals, \mathbf{e}_m , and the y residuals, f_m :

- (a) $\mathbf{e}_0 = \mathbf{x} - \mu_x$

- (b) $f_0 = y - \mu_y$

For $m = 1, 2, \dots$

2. The scores t are linear combinations of the \mathbf{x} residuals from the last step weighted by the covariances with y residuals to make the scores more closely related to y :

- (a) $\mathbf{w}_m = \text{Cov}(\mathbf{e}_{m-1}, f_{m-1})$ $\mathbf{w}_1, \mathbf{w}_2, \dots$ are orthogonal

- (b) $t_m = \mathbf{e}'_{m-1} \mathbf{w}_m$

3. Determine the \mathbf{x} loadings, \mathbf{l}_m , and y loading, q_m , using least squares:

- (a) $\mathbf{l}_m = \text{Cov}(\mathbf{e}_{m-1}, t_m) / \text{Var}(t_m)$

- (b) $q_m = \text{Cov}(f_{m-1}, t_m) / \text{Var}(t_m)$

4. Find the new residuals

- (a) $\mathbf{e}_m = \mathbf{e}_{m-1} - \mathbf{l}_m t_m$

- (b) $f_m = f_{m-1} - q_{m-1} t_{m-1}$

It is further noted that the sequence of PLS coefficients for $m = 1, \dots, p$ represent the conjugate gradient sequence for computing least squares solutions.

3.1.2 Number of Parameters

PLSR uses m relevant loadings/components in the model. However, deciding on m is not trivial as discussed by Helland (2001). Even when m is known, the number of parameters in the model is open for debate. Van Der Voet (1999) illustrated the problem in calculating the degrees of freedom of a model using PLSR. Calculating the number of parameters in a model is especially relevant when using a complexity

penalty as part of the model selection criterion. For the purposes of this thesis, the number of parameters in the population model was assumed to be

$$\frac{p(p+1)}{2} + m + 1.$$

So that if $m = 0$ no correlation between \mathbf{X} and \mathbf{y} exists and $m = p$ is the full least squares model; this agrees with Helland (2001).

3.2 Soft Independent Modelling of Class Analogies

Soft Independent Modelling of Class Analogies (SIMCA) was developed by Wold (1976). The underlying concept is to model each class separately using Principal Components (PCs). A different number of components may be selected for each class. In order to classify a new observation one of two approaches is then taken:

1. Find the Mahalanobis distance of the new observation to each of the existing classes and place the observation into the “closest” class. Using this method, the probabilities of an observation belonging to each of the groups can be easily calculated.
2. Create a set around each class developed in the training set. If the new observation falls within these sets, it is then classified as belonging to this class. This method can result in observations being classified into multiple classes, a single class or no classes at all and is quite a common approach used in the chemometrics literature.

More specifically: considering discriminant analysis to use the function

$$d_{g^*}(x) = \min_g [(x - \bar{x}_g)' \Sigma_g^{-1} (x - \bar{x}_g) + \log |\Sigma_g| - 2 \log \pi_g]$$

where \bar{x}_g is the mean (vector) of the variables associated with group g , Σ_g is the covariance matrix corresponding to group g and π_g is the (prior) probability of an observation belonging to group g . As Σ_g is a positive semi definite matrix, it has the spectral decomposition:

$$\Sigma_g = \sum_{j=1}^p \lambda_{jg} e_{jg} e'_{jg}$$

where λ_{jg} is the j^{th} eigenvalue (in descending order) of Σ_g and e_{jg} is the corresponding eigenvector.

If Σ_g is positive definite all of the λ 's will be greater than 0 thus Σ_g^{-1} can be written as

$$\Sigma_g^{-1} = \sum_{j=1}^p \frac{1}{\lambda_{jg}} e_{jg} e'_{jg}.$$

SIMCA instead uses

$$d_{g^*}(x) = \min_k [(x - \bar{x}_g)' \Sigma_g^{-1}(C_g)(x - \bar{x}_g)]$$

where

$$\Sigma_g^{-1}(C_g) = \frac{\sum_{j=C_g+1}^p e_{jg} e'_{jg}}{\sum_{j=C_g+1}^p \lambda_{jg}}$$

For $g = 1, \dots, G$ the number of principal components, C_g , is estimated using F fold cross validation as outlined by (Wold, 1976, Section 2.2.1) The cross validation algorithm to implement this minimization procedure is:

- For each group $g = 1, \dots, G$:
- **Step 1:** Create the relevant submatrix X_g containing observations in the training data that belong to group g .
- **Step 2:** Cross validation (within each group g): For $f = 1, \dots, F$
- **Step 3:** Withhold observations in group g that belong to fold f to create X_g^- , a matrix with p variables and n_g^- observations. The n_g^+ withheld observations are denoted as X_g^+
- **Step 4:** Find the variable means of X_g^- : \bar{X}_g^- .
- **Step 5:** Find the singular value decomposition $(X_g^- - \bar{X}_g^-) = \underbrace{U_g}_{n_g^- \times a} \underbrace{S_g}_{a \times a} \underbrace{V_g'}_{a \times p}$ where $a = n_g^- - 1$ if $n_g^- \leq p$ or $a = p$ if $n_g^- > p$.
- **Step 6:** U_g is the normalised score matrix, V_g the loading matrix and S_k is a diagonal matrix containing the singular values ordered so that $\lambda_{1g} > \lambda_{2g} > \dots > \lambda_{ag}$.

- **Step 7:** For $r = 1, \dots, a$ find $T_g^+ = (X_g^+ - \bar{X}_g^-) \underbrace{V_g}_{p \times r}$ (using the first r columns of V_g) and hence find the value predicted object $\hat{X}_g^+ = \bar{X}_g^- + \underbrace{T_g}_{n_g^+ \times r} \underbrace{V_g'}_{r \times p}$. Find the associated residuals $E_g = X_g^+ - \hat{X}_g^+$.
- **Step 8:** For each r find the sum of squares of the residuals contained in each of the matrices E_g 's, Δ_r^f .
- **Step 9:** If $f < F$, let $f = f + 1$ and return to step 3.
- **Step 10:** Choose the r that minimises $D_r = \sum_{f=1}^F \Delta_r^f$ to determine the number of components to include for group g , call this C_g .
- **Step 11:** If $g < G$, let $g = g + 1$ and return to step 2.
- **Step 12:** Now, for each group g the number of components C_g to include in the model had been decided.

In methods where cross validation is designed to maximise classification performance it has a tendency to over-estimate out of sample classification performance. The selection criterion for the number of principal components to use for each group does not lend itself easily to use with Information Criteria (such as BIC). Thus models developed using the SIMCA method are not easily compared to other commonly used methods for near infrared spectroscopic data.

SIMCA effectively partitions the subspace. In the primary subspace (first C_g eigenvalues and eigenvectors) it assumes that the eigenvalues of each group are infinitely large so that $1/\lambda_{jg} = 0$. In the complement of this space (of dimension $p - C_g$), the eigenvalues are estimated by:

$$\lambda_{jg} = \frac{(n_g - 1)}{(n_g - C_g - 1)} \sum_{j=C_g+1}^p \lambda_{jg}$$

Although designed for large p small n problems one of the disadvantages of the SIMCA method is that it requires enough observations belonging to every group to be included in the training data in order to use cross validation to select the number of principal components for each group. Therefore it is not suitable as a method where there are a large number of groups relative to the number of observations in the training data.

When using a set to determine group membership rather than the relative probabilities of belonging to each group, many observations may not be classified at all. While placing an observation into multiple classes indicates uncertainty, but still partially informs on the decision process, failing to place an observation into any group does not advance the knowledge about the sample. Assigning group membership based on relative probabilities avoids this problem, but this also loses the uniqueness of SIMCA as a method. The tendency of SIMCA to classify observations as outliers, thus not belonging to any group is examined for NIR spectra by De Maesschalck et al. (1999).

Frank and Friedman (1989) proposes an amendment to this scheme – Discriminant Analysis with Shrunk COvariances (DASCO), however this is not widely used in the chemometrics literature, thus not used as one of the reference methods within this thesis.

3.3 Likelihood Based Statistical Inference

Likelihood based statistical inference is based on the premise that everything that can be learned about the parameters from the data is contained in the likelihood function.

Suppose there is a sample x_1, x_2, \dots, x_n where each observation is independently generated from a \mathbb{P}_θ distribution. The density of the j^{th} observation can be written as $f(x_j|\theta)$.

The (joint) density of the whole sample is:

$$f(\mathbf{x}_n|\theta) = f(x_1, x_2, \dots, x_n|\theta) = \prod_{j=1}^n f(x_j|\theta)$$

The joint density is called the likelihood function. The joint density of the data, without the condition of independence is a likelihood function, written as:

$$L(\theta|\mathbf{x}_n) = f(\mathbf{x}_n|\theta)$$

with unknown parameters θ . f is a function describing the generating process of the data, with parameters θ .

Maximum Likelihood Estimate (MLE)

Let $L(\theta|\mathbf{x}_n)$ be the likelihood function (defined for $\theta \in \Theta$). A maximum likelihood estimate is any value $\hat{\theta} \in \Theta$ for which $L(\theta|\mathbf{x}_n) \leq L(\hat{\theta}|\mathbf{x}_n)$ for all $\theta \in \Theta$.

However, it is often easier to maximise the log-likelihood function $l(\theta|\mathbf{x}_n) := \log L(\theta|\mathbf{x}_n)$. As log is an increasing function so $x > y \Leftrightarrow \log x > \log y$ for $x, y > 0$. Thus an alternative expression for the MLE is $l(\hat{\theta}|\mathbf{x}_n) \geq l(\theta|\mathbf{x}_n) \forall \theta \in \Theta$. Suppose that the log-likelihood function $l(\theta|\mathbf{x}_n)$ is a smooth function of θ . To maximise the likelihood, differentiate the log-likelihood and solve for θ .

EM Algorithm

The Expectation Maximization (EM) algorithm (Dempster et al., 1977) was developed as an iterative approach of calculating maximum likelihood estimates when the observed data could be considered to be incomplete. This incompleteness can be introduced in order to simplify other calculations, or the unknown labels in a classification problem can be considered as the missing data.

Given a joint distribution $f(x, z|\theta)$ where x are observed variables, z are unobserved variables and θ are parameters. The goal of the EM algorithm is to maximise the likelihood function $f(x|\theta)$ with respect to θ .

Let $t = 0$. Select a starting value for θ , $\theta^{(t)}$.

Repeat

1. **Estep**: Evaluate $f(z|x, \theta^{(t)})$.
2. **Mstep**: Find $\theta^{(t+1)} = \max_{\theta} Q(\theta|\theta^{(t)})$,
where $Q(\theta|\theta^{(t)}) = \int_z f(z|x, \theta^{(t)}) \log f(x, z|\theta) dz$.
3. $t = t + 1$.

Until convergence.

The EM algorithm is relatively stable and simple to use as the Q function is typically much simpler to maximise than $f(x|\theta)$. However, it is not guaranteed to reach a global maximum and is only linearly convergent.

3.4 Model-based Discriminant Analysis

Model-based discriminant analysis enables a better understanding of the generating process that discriminates between the different groups. It focuses on parameter estimation, finding a set of parameters that describe the source(s) of separation between groups.

In model-based discriminant analysis (also known as eigenvalue discriminant analysis) (Bensmail and Celeux, 1996), the model is fitted to data \mathbf{w}_n where $n = 1, 2, \dots, N$ and labels \mathbf{l}_n where $l_{ng} = 1$ if observation n belongs to group g and 0 otherwise.

The resulting likelihood function is

$$\mathcal{L}_{\text{disc}}(p_1, p_2, \dots, p_G; \theta_1, \theta_2, \dots, \theta_G | \mathbf{w}, \mathbf{l}) = \prod_{n=1}^N \prod_{g=1}^G [p_g f(\mathbf{w}_n | \theta_g)]^{l_{ng}}. \quad (3.1)$$

The log of the likelihood function (3.1) is maximized yielding parameter estimates $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_G$ and $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_G$. For stability, equal probabilities, $\hat{p}_1 = \dots = \hat{p}_G = 1/G$, are sometimes assumed.

The posterior probability of group membership for an observation \mathbf{y} whose label is unknown can be estimated as

$$\mathbb{P}(\text{Group } g | \mathbf{y}) \approx \frac{\hat{p}_g f(\mathbf{y} | \hat{\theta}_g)}{\sum_{q=1}^G \hat{p}_q f(\mathbf{y} | \hat{\theta}_q)}. \quad (3.2)$$

Assuming the density f to be a multivariate Gaussian density, ϕ , with mean μ_g and covariance matrix Σ_g

$$\phi_g(\mathbf{w}_n | \mu_g, \Sigma_g) \equiv \frac{\exp\{-\frac{1}{2}(\mathbf{w}_n - \mu_g)^T \Sigma_g^{-1} (\mathbf{w}_n - \mu_g)\}}{\sqrt{\det(2\pi \Sigma_g)}}.$$

The multivariate Gaussian densities imply that the groups are centred at the means μ_g with shape, orientation and volume of the scatter of observations within the group depending on the covariance matrices Σ_g .

The parameters p and μ can then be estimated by:

$$\begin{aligned} \hat{p}_g^{(k+1)} &\leftarrow \frac{\sum_{n=1}^N l_{ng}}{N} && \text{if estimated} \\ \hat{\mu}_g^{(k+1)} &\leftarrow \frac{\sum_{n=1}^N l_{ng} \mathbf{w}_n}{\sum_{n=1}^N l_{ng}}. \end{aligned}$$

The Σ_g can be decomposed using an eigen decomposition into form,

$$\Sigma_g = \lambda_g D_g A_g D_g^T, \quad (3.3)$$

where λ_g is a constant of proportionality, D_g an orthogonal matrix of eigenvectors and A_g is a diagonal matrix where the elements are proportional to the eigenvalues as described by Fraley and Raftery (2002).

The estimates of Σ_g depend on the constraints placed on the eigenvalue decomposition; details of the calculations are given by Bensmail and Celeux (1996) and Celeux and Govaert (1995).

The parameters λ_g , A_g and D_g have interpretations in terms of volume, shape and orientation of the scatter of the component. The parameter λ_g controls the volume while the matrices A_g and D_g control the shape of the scatter and the orientation respectively. Constraining the parameters to be equal across groups gives great modelling flexibility. Some of the options for constraining the covariance parameters are given in Table 3.1 and are illustrated for the three group case by Figure 3.1.

Table 3.1: Parametrizations of the covariance matrix Σ_g

Model ID	Decomposition	Volume	Shape	Orientation
EII	$\Sigma_g = \lambda I$	Equal	Identity	Identity
VII	$\Sigma_g = \lambda_g I$	Variable	Identity	Identity
EEI	$\Sigma_g = \lambda A$	Equal	Equal	Identity
VEI	$\Sigma_g = \lambda_g A$	Variable	Equal	Identity
EVI	$\Sigma_g = \lambda A_g$	Equal	Variable	Identity
VVI	$\Sigma_g = \lambda_g A_g$	Variable	Variable	Identity
EEE	$\Sigma_g = \lambda D A D^T$	Equal	Equal	Equal
EEV	$\Sigma_g = \lambda D_g A D_g^T$	Equal	Equal	Variable
VEV	$\Sigma_g = \lambda_g D_g A D_g^T$	Variable	Equal	Variable
VVV	$\Sigma_g = \lambda_g D_g A_g D_g^T$	Variable	Variable	Variable

The letters in Model ID denote the volume, shape and orientation respectively. For example, EEV represents equal volume and shape with variable orientation. EII and VII represent spherically shaped groups.

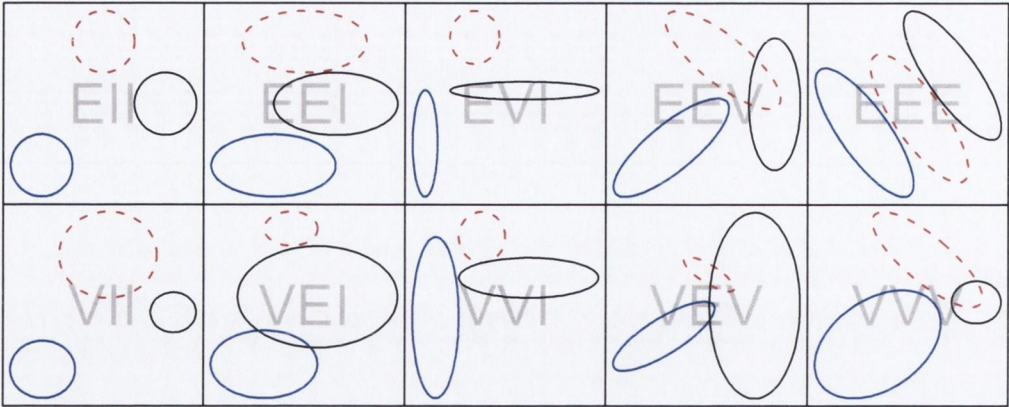


Fig. 3.1: Possible combinations of volume, shape and orientation for 3 covariance matrices

Model-based discriminant analysis is fitted to observations $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N$ by maximizing the likelihood (3.1) using the EM algorithm (Dempster et al., 1977). The resulting output from the EM algorithm includes estimates of the probability of group membership for each observation; these can be used to cluster the observations into their most probable groups.

The `mclust` (Fraley and Raftery, 2007, 2002, 1999, 1998; Banfield and Raftery, 1993) package for R (R Development Core Team, 2007) can be used to perform the model-based discriminant analysis. This allows for the possibility of the models mentioned in Table 3.1. It is worth noting that Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are special cases of model-based discriminant analysis and they correspond to the EEE and VVV models respectively.

3.4.1 Updating

Model-based discriminant analysis as developed in Bensmail and Celeux (1996) only uses the observations with known group membership in the model fitting procedure. Once the model is fitted, the observations with unknown group labels can be classified into their most probable groups.

An alternative approach is to model both the labelled data (\mathbf{w}, \mathbf{l}) and the unlabelled data \mathbf{y} and to maximize the resulting log-likelihood for the combined model. The mixture model assumes that observations come from one of G groups, that

observations within each group g are modelled by a density $f(\cdot|\theta_g)$ where θ_g are unknown parameters and that the probability of coming from group g is p_g .

Given unlabelled data \mathbf{y} with independent observations $\mathbf{y}_1, \dots, \mathbf{y}_M$, a mixture model with G groups has a likelihood function

$$\mathcal{L}_{\text{mix}}(\theta_1, \dots, \theta_G; p_1, \dots, p_G | \mathbf{y}) = \prod_{m=1}^M \sum_{g=1}^G p_g f(\mathbf{y}_m | \theta_g). \quad (3.4)$$

Maximising equation (3.4) using the EM algorithm (Dempster et al., 1977) is the basis of model-based clustering (Fraley and Raftery, 2002) and is easily implemented in the `mclust` library (Fraley and Raftery, 2007).

The likelihood function for the combined data is a product of the likelihood functions given in equations (3.1) and (3.4). This classification approach was developed in Dean et al. (2006) and was demonstrated to give improved classification performance over the classical model-based discriminant analysis in some food authenticity applications, using the NIR meats data.

With this modelling approach the likelihood function is of the form

$$\begin{aligned} \mathcal{L}_{\text{update}}(p, \theta | \mathbf{w}, \mathbf{1}, \mathbf{y}) &= \mathcal{L}_{\text{disc}}(p, \theta | \mathbf{w}, \mathbf{1}) \mathcal{L}_{\text{mix}}(p, \theta | \mathbf{y}) \\ &= \left[\prod_{n=1}^N \prod_{g=1}^G [p_g f(\mathbf{w}_n | \theta_g)]^{l_{ng}} \right] \left[\prod_{m=1}^M \sum_{g=1}^G p_g f(\mathbf{y}_m | \theta_g) \right]. \end{aligned} \quad (3.5)$$

The log of the likelihood (3.5) is maximized using the EM algorithm to find estimates for p (if estimated) and θ . Output from the EM algorithm includes estimates of the probability of group membership for the unlabelled observations \mathbf{y} , as given in equation (3.2). In a practical setting, test set data are the unlabelled observations while training data are labelled observations.

The EM algorithm for maximizing the log of the likelihood (3.5) proceeds iteratively substituting the unknown labels with their estimated expected values. At each iteration the estimated labels are updated and new parameter estimates are produced. By passing the estimated values of the unknown labels into the EM algorithm it is possible to “update” the classification results with some of the knowledge gained from fitting the model to all of the data. With small training sets updating has been shown to be beneficial in other studies Dean et al. (2006), thus the performance of updating techniques were also included for evaluation over other NIR datasets.

Updating techniques are especially useful when unlabelled observations can provide useful information about separation between groups. Figure 3.2 illustrates how the amount of information from an unlabelled observation can vary depending on the separation between the group means. The information provided by unlabelled observations decreases as the group means move closer together. Thus in Figure 3.2 points x , y and z are in the same positions, but the amount of information they provide about the groups varies according to the separation between groups.

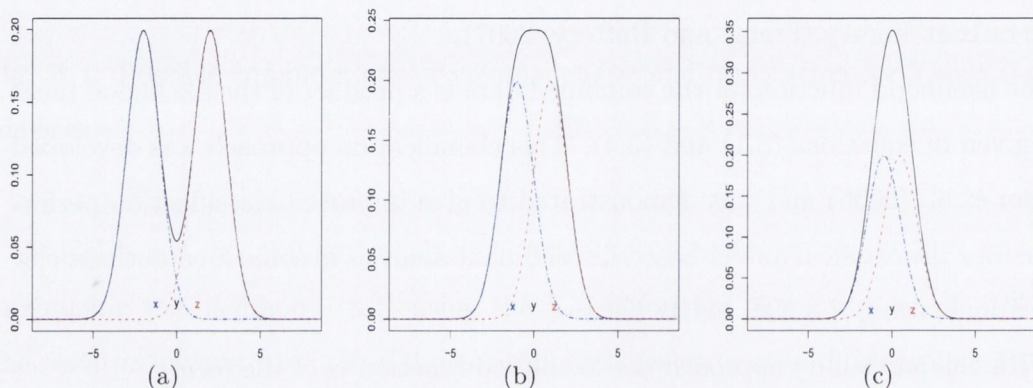


Fig. 3.2: Means Separated by: 4 Standard Deviations 3.2(a), 2 Standard Deviations 3.2(b) and 1 Standard Deviation 3.2(c)

3.4.2 Implementing the EM Algorithm

The EM algorithm (Dempster et al., 1977) is ideally suited to the problem of maximizing the log-likelihood function when some of the data have unknown group labels; this arises in the model-based clustering likelihood (3.4) and the model-based discriminant analysis with updating likelihood (3.5). In this section, the steps involved in the EM algorithm for model-based discriminant analysis with updating are illustrated; the model-based clustering steps are shown in Fraley and Raftery (2002).

Considering data to be classified as consisting of M multivariate observations consisting of two parts: known, \mathbf{y}_m and unknown \mathbf{z}_m . In this context the spectroscopic data, which are observed and thus known, are treated as the \mathbf{y}_m . The labels (pure or adulterated) are unknown and thus are treated as the \mathbf{z}_m . Additionally, N labelled observations are available, which consist of two parts: known \mathbf{w}_n and known labels \mathbf{l}_n .

The unobserved portion of the data, is a matrix of indicator functions, so that $\mathbf{z}_m = (z_{m1}, \dots, z_{mG})$, where $z_{mg} = 1$ if \mathbf{y}_m is from group g and $z_{mg} = 0$ otherwise.

Then the observed data likelihood can be written in the form

$$\mathcal{L}_O(p, \theta | w_N, \mathbf{1}_N, y_M) = \left[\prod_{n=1}^N \prod_{g=1}^G [p_g f(\mathbf{w}_n | \theta_g)]^{l_{ng}} \right] \left[\prod_{m=1}^M \sum_{g=1}^G p_g f_g(\mathbf{y}_m | \theta_g) \right]$$

and the complete data likelihood is

$$\mathcal{L}_C(p, \theta | w_N, \mathbf{1}_N, y_M, \mathbf{z}_M) = \left[\prod_{n=1}^N \prod_{g=1}^G [p_g f(\mathbf{w}_n | \theta_g)]^{l_{ng}} \right] \left[\prod_{m=1}^M \prod_{g=1}^G [p_g f(\mathbf{y}_m | \theta_g)]^{z_{mg}} \right]. \quad (3.6)$$

Initial estimates of \hat{p} (if estimated) and $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ are taken from classical model-based discriminant analysis, by maximizing (3.1).

The expected value of the unknown labels are calculated so that

$$\hat{z}_{mg}^{(k+1)} \leftarrow \frac{\hat{p}_g^{(k)} f(\mathbf{y}_m | \hat{\theta}_g^{(k)})}{\sum_{q=1}^G \hat{p}_q^{(k)} f(\mathbf{y}_m | \hat{\theta}_q^{(k)})}, \quad (3.7)$$

for $g = 1, \dots, G$ and $m = 1, \dots, M$ and the parameters p and $\theta = (\mu, \Sigma)$ can then be estimated by:

$$\begin{aligned} \hat{p}_g^{(k+1)} &\leftarrow \frac{\sum_{n=1}^N l_{ng} + \sum_{m=1}^M \hat{z}_{mg}^{(k+1)}}{N+M} && \text{if estimated} \\ \hat{\mu}_g^{(k+1)} &\leftarrow \frac{\sum_{n=1}^N l_{ng} \mathbf{w}_n + \sum_{m=1}^M \hat{z}_{mg}^{(k+1)} \mathbf{y}_m}{\sum_{n=1}^N l_{ng} + \sum_{m=1}^M \hat{z}_{mg}^{(k+1)}}. \end{aligned}$$

The estimates of Σ_g again depend on the constraints placed on the eigenvalue decomposition, details of the calculations are given in Bensmail and Celeux (1996); Celeux and Govaert (1995).

The iterative process continues until convergence is achieved. The use of an Aitken acceleration-based convergence criterion is discussed in Dean et al. (2006).

Updating can take two forms – soft and hard updating. In the case of soft updating (EM), updates of the missing labels are made using equation (3.7), so the unknown labels are replaced by probabilities rather than by 0 or 1 values. Whereas, hard updating (CEM) replaces the probabilities given in equation (3.7) with an indicator vector of the most probable group. The hard classification algorithm does not maximize equation (3.5) but actually tries to maximize equation (3.6) (although local maxima are a possibility). Local maxima are possible with both hard and soft updating, but more likely with hard updating. Using multiple random restarts of the algorithm is a common approach to compensate for this.

3.5 Dimension Reduction Techniques

Each NIR spectra examined contains at least 700 wavelengths with adjacent absorption values being highly correlated. Therefore before using model-based classification methods, a dimension reduction step is required – this avoids singular covariance matrices, improves computational efficiency and increases statistical modelling possibilities.

There are two main approaches to dimension reduction for NIR data – aim to approximate the entire spectrum using fewer variables, accomplished using wavelet analysis, or to select suitable wavelengths, either before using the model-based classification methods or as part of the classification process.

3.5.1 Wavelet Analysis

Wavelet analysis is a technique commonly used in image and signal processing in order to compress data. Here, it is used to decompose each NIR spectrum into a series of wavelet coefficients. Without any thresholding of these coefficients, the original spectra can be exactly reconstructed from the coefficients. However, many of the coefficients in the wavelet analysis are zero or close to zero. By thresholding the coefficients that are zero or close to zero, it is possible to dramatically reduce the dimensionality of the dataset. The resulting recomposed spectra are then approximations of each of the individual spectra. Ogden (1997) gives a good practical introduction to wavelet analysis.

Haar (1910) (a translation of which can be found in Heil and Walnut (2006)) wavelets (Figure 3.3(a)) were considered, mainly because of their simplicity of form but as Daubechies (1988) wavelet (Figure 3.3(b)) was used very effectively as a dimension reduction tool before Model Based Discriminant Analysis with applications to NIR data by Dean et al. (2006), it was natural to use it again in order to assess the relative robustness of the technique to different datasets.

Daubechies' wavelet is a consistently reliable type to use and is the default within `wavethresh` (Nason et al., 2006). Figure 3.4 illustrates how Daubechies' wavelet captures more of the structure of the NIR spectra than the Haar wavelet. To efficiently carry out wavelet analysis, the data dimension should be of the order 2^k , where

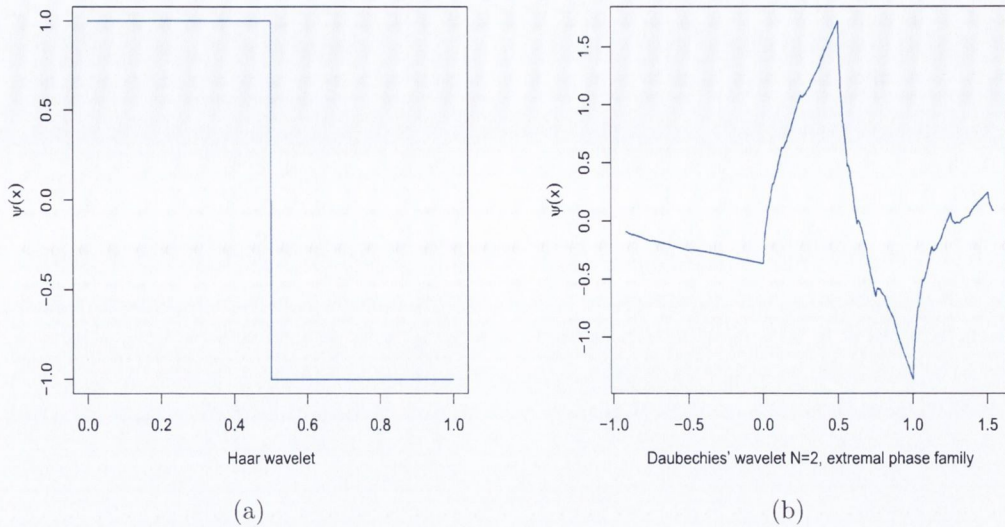


Fig. 3.3: Wavelet functions. Figure 3.3(a) is the Haar wavelet ; Figure 3.3(b) is the Daubechies' wavelet.

k is an integer. Unfortunately, this can result in quite a lot of information being set aside. Techniques of extending the data to bring them up to the nearest 2^k are available, but in this case these methods result in problems when carrying out the model-based discriminant analysis – the associated variance structures are often singular.

Over the range 400 nm – 2498 nm, this means that just $1050 - 2^{10} = 26$ wavelengths are dropped, which considering that spectral noise levels recorded by the scanning monochromator are seen to increase at extremes of the wavelength range studied is not an issue to be overly concerned about. However when only the range 1100 nm – 2498 nm is available, then only the central $2^9 = 512$ wavelengths are chosen – the range 1290 nm – 2312 nm.

While the goal is to reduce the dimensionality of the data, it is desirable that this reduction be achieved in a structured way, rather than by using an ad-hoc rule of the *central* 2^n points. Trying to determine the optimal window of the data to use would prove too computationally expensive and would require computation for each dataset.

As a default procedure, universal hard thresholding is used. When using universal thresholding, the threshold $\lambda = \hat{\sigma} \sqrt{2 \log 2^k}$, where $\hat{\sigma}$ is a robust estimate of the standard deviation of the coefficients and there are 2^k coefficients. Other thresh-

olding techniques may sometimes provide better approximations of the spectra but their use adds another decision into the process that may lead to over fitting rather than a general procedure.

Figure 3.4(a) illustrates the difference between the original NIR spectra of a pure honey sample alongside that of the reconstructed, post universal hard thresholding spectra, which is represented by 14 non-zero coefficients for Daubechies' wavelets and 12 non-zero coefficients for Haar wavelets. Figure 3.4(b) illustrates the difference between the original NIR spectra of a pure olive oil sample alongside that of the reconstructed, post universal hard thresholding spectra, which is represented by 17 non-zero coefficients for Daubechies' wavelets and 14 non-zero coefficients for Haar wavelets.

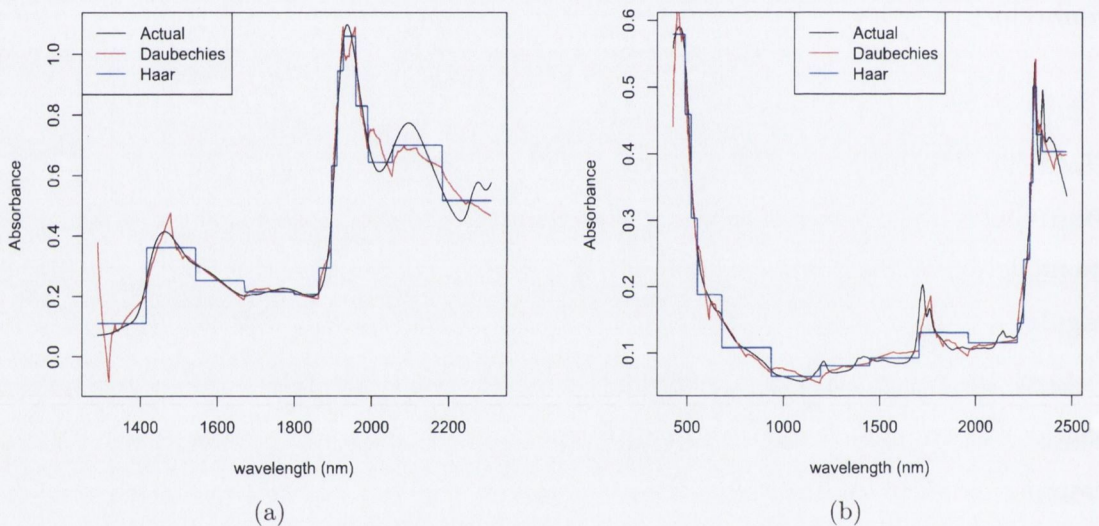


Fig. 3.4: Comparison of actual and reconstructed, universal hard thresholded honey spectra (Figure 3.4(a)) and olive oils spectra (Figure 3.4(b)) using both Daubechies' and Haar wavelets.

3.5.2 Wavelength Selection Methods

Wavelet thresholding is a dimension reduction technique that can summarise the entire spectra, or at least a significant proportion of it. An alternative approach is to use methods of selecting suitable individual wavelengths before undertaking model-based classification techniques. Prior to using such classification techniques,

3 types of wavelength selection were examined, as suggested by Indahl et al. (1999).

Assuming

$$\mu_g = \frac{1}{n_g} \sum_{j=1}^N z_{gj} x_j$$

where n_g is the number of observations in group g ($\sum_{j=1}^N z_{gj}$) and z_{gj} is an indicator variable for group membership for observation i belonging to group g and x_j is the j^{th} observation of X . Then the between groups covariance matrix B is

$$B = \sum_{g=1}^G (\mu_g - \mu)(\mu_g - \mu)'$$

and the within groups covariance matrix W is

$$W = \sum_{g=1}^G \sum_{j=1}^N z_{gj} (x_{gj} - \mu_g)(x_{gj} - \mu_g)'$$

Strategy 1: Between to Within Group Variances

Examine the univariate ratios of B to W to obtain what is called the *scatter curve* by Indahl et al. (1999). If B and W have been calculated as $p \times p$ matrices, then the diagonals of these matrices correspond to the univariate values. Find the local maxima of this curve and use these wavelengths in the further calculations.

Strategy 2: Between and Within Group Variances

Estimate the *variance curve* by taking the diagonal of the matrix $(B+W)/N$. Again find the local maxima of the resultant curve and use these wavelengths in the further calculations.

Strategy 3: Combining Strategies 1 and 2

Combine the wavelengths selected by strategies 1 and 2 in order to gain a more balanced representation of how the spectra vary – both in total and between groups.

Using local maxima in the curves is suitable when adjacent variables are highly correlated as is the case with NIR spectra. Figure 3.5 illustrates the wavelengths selected by these methods for the honey data. Strategy 1 selects 1102, 1216, 1702, 1792, 1922, 2454 and 2498 nm more than 75% of the time, while the second strategy selects 1100, 1104, 1580, 1666, 1834, 2170 and 2498 nm more than 75% of the time.

Using these strategies as a dimension reduction technique is most effective with 2 group type problems (*i.e* pure versus adulterated).

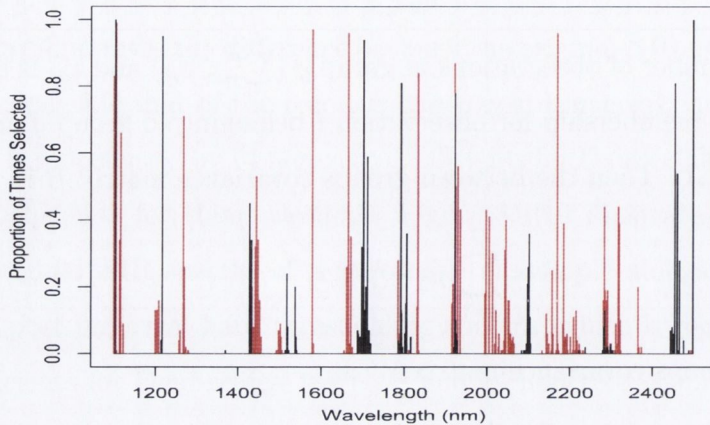


Fig. 3.5: Proportions of times wavelengths are selected using between and within group variances on honey spectra. The total height represents the proportion of times a wavelength was selected using strategy 3, the black represents wavelengths selected by strategy 1 and red represents those selected by strategy 2.

3.6 Model Selection Techniques

Cross validation and the Bayesian Information Criterion (BIC) were both used as model selection criteria.

The BIC penalizes models based on their complexity. Thus models that are considered to be too complex are rejected. Such a criterion is required to ensure that the same decision process is used for both models for each simulation, rather than a decision based on the subjective judgement of an analyst, in order to make a simulation study feasible.

5-fold cross validation was chosen as the cross validation selection method, with the cross validation being performed on the training data and then performance evaluated on the test data. It is also possible to choose models based on leave-one-out cross validation, but this is computationally expensive.

The certainty of the classification decisions are measured using Brier's score (1950).

Brier's score and the percentage error measure largely the same thing – with the percentage error equal to the Brier's score if all the probabilities are turned into hard classifications. Therefore the difference between the two is a measure of the uncertainty associated with each prediction. If the Brier's score is larger than the percentage error, this indicates that the errors made were associated with relatively certain probabilities or that most of the observations that were correctly classified had a relatively low probability of belonging to that group. On the other hand, if the Brier's score is smaller than the percentage error, this indicates that the errors made were associated with the relatively uncertain probabilities. Therefore, having a Brier's score smaller than the percentage error is a positive situation.

3.6.1 Bayesian Information Criterion (BIC)

The results given are models selected using the Bayesian Information Criterion (BIC), where the BIC of a function is

$$\text{BIC} = 2\log\text{likelihood} - d\log(N^*),$$

given that d is the number of parameters and N^* is the total number of observations used in fitting the model. Thus for discriminant analysis (without updating) and partial least squares regression (PLSR) $N^* = N$ and for updating discriminant analysis methods (EM and CEM) $N^* = N + M$.

3.6.2 F fold cross validation

The number of folds, F , is decided upon. Deciding on the number of folds is a bias variance trade off, as described by Hastie et al. (2001, p214-p217), but it can also be dependent on the size of the dataset and the relative sizes of each group within the dataset. After the data has been divided into training and test sets the training set is then randomly split into F subsets where the number of observations in each subset is not fixed. Each of these subsets is then treated in turn as a test set – withheld from calculations so that an estimate of out-of-sample performance can be obtained. The average performance across the subsets is then used to select which of the models under examination should be used on the test set data.

3.6.3 Brier's Score

Brier (1950) developed a method of producing a continuous performance measure where perfect prediction gives a Brier's score of zero. Given G groups and N samples and forecasted probabilities $\hat{z}_{n1}, \dots, \hat{z}_{nG}$ for sample n of belonging to group $1, \dots, G$ respectively then the Brier's score, B , is

$$B = \frac{100}{N \times G} \sum_{g=1}^G \sum_{n=1}^N (\hat{z}_{gn} - z_{true_{gn}})^2$$

where $z_{true_{gn}}$ is an indicator variable for the actual group membership. It is especially useful for determining the certainty of predictions. Some observations may be just barely put into the correct group, or indeed just miss out on correct classification. Observations that are barely classified correctly will add more to the Brier's score than those where a more certain classification is made.

A trait of PLSR is that some regression outputs may in fact be beyond the zero-one scale. For the purposes of calculating a pseudo-Brier score, such results were set to be equal to either zero or one so that these certain classifications do not add to the total.

3.7 Performance Comparison

In order to compare the performance of PLSR with and without a Savitzky-Golay filter (Savitzky and Golay, 1964) using both BIC and 5 fold cross validation, and model-based discriminant analysis, using the wavelength selection methods described in Section 3.5.2, 100 splits of the both the honey NIR data and the olive oil NIR data into training and test data was performed, so that the training to test ratios were: 50%:50%, 25%:75%, 10%:90%.

Savitzky-Golay filter

The Savitzky-Golay filter is a polynomial smoother that can also be used to differentiate curves. It was designed with the aim of reducing the signal-to-noise ratio of a spectrum. It assumes that curves formed by graphing the points are essentially smooth. While a moving average process only considers points "previous" on the measurement process, the Savitzky-Golay filter uses points on both sides of the

point under consideration, weighting these points in a symmetric fashion. The original paper (Savitzky and Golay, 1964) calculated the weightings for neighbouring points, but contained numerical errors, most of which were corrected by Steinier et al. (1972). Madden (1978) outlines how such numerical errors may be detected and provides a correction to one of the values given by Steinier et al. (1972). The Savitzky-Golay filter is commonly used in the spectroscopy field and was included for completeness.

3.7.1 PLSR

Using the Bayesian Information Criterion, up to a maximum of 40 components were considered for the model, when using 5 fold cross validation, up to a maximum of 20 components were considered.

As partial least squares regression was designed to be used with cross validation, it is unsurprising that the 5 fold cross validation is more effective when there are many observations in the training data (> 100 so that the expected number of observations in a fold is at least 20). However, once the number of observations available for inclusion in the training set is reduced, the advantages of using BIC as a model selection method, both in terms of its stability and its classification performance are evident in Tables 3.2 and 3.7. Due to the smaller number of observations in the olive oil dataset, the effectiveness of using BIC as a complexity criterion to determine the number of components rather than using cross validation is evident even at the 50% training and 50% test splits.

In Table 3.7 the extra robustness of using a method where all the training data are examined at once is evident when the training data sample size is small (10% in the training data set). 5 fold cross validation is unfeasible when there are a total of only 14 samples in the training data.

3.7.2 Model Based Methods

Comparing the model based classification techniques using the four different methods of dimension reduction discussed in Section 3.5. Tables 3.3, 3.4, 3.5 and 3.6 show the classification performance in terms of percentage error and Brier's scores, using both the BIC and five fold cross validation as model selection methods for the

honey data (Section 2.2.1) and the olive oil data (Section 2.2.3).

These illustrate that updating techniques can improve classification performance, especially when the sample size in the training set is small, if the original model assumptions are correct. However, if these assumptions are incorrect, updating exacerbates any problems, even causing a substantial disimprovement in classification performance as is evident in the 10%/90% split of the honey data in Table 3.3.

The effectiveness of the various dimension reduction techniques depend largely on the nature of the data. To carry out wavelet thresholding on the honey data, a large number of wavelengths had to be dropped. However, there are sufficiently many observations in each group so that, even when only 10% of the data was included in the training set, the sets were large enough to accurately determine appropriate wavelengths. For the olive oil data, the opposite was true – only a few wavelengths had to be dropped in advance of wavelet thresholding, but the smaller number of observations had an impact on the effectiveness of the methods of wavelength selection.

Cross validation using wavelet thresholding is effective on the olive oil data, even at the 10% training data level. However, the dimension reduction techniques that require information about the labels are more unstable – the potential for a very unrepresentative sample is much higher and the folds do not contain enough information to be useful.

Another issue arises in the implementation of cross validation with the wavelength selection methods. Each fold selects a different set of wavelengths, thus selecting the “best” model in terms of the covariance structure is unstable.

Table 3.2: Comparing the classification performance of PLSR with and without Savitzky Golay filter, using both BIC and cross validation at various training/test splits of the NIR honey data, using both the error rate and the Brier's score

Honey		BIC				5 fold CV			
		% Error		Brier		% Error		Brier	
50%/	SG	13.490	(2.119)	10.279	(1.419)	4.092	(1.310)	3.938	(0.489)
50%	None	10.540	(1.806)	8.167	(1.077)	4.690	(1.421)	4.450	(0.769)
25%/	SG	9.916	(2.058)	7.995	(1.180)	5.120	(1.603)	4.830	(0.943)
75%	None	9.212	(1.872)	7.366	(1.154)	6.593	(1.642)	6.034	(1.105)
10%/	SG	8.640	(2.379)	7.360	(1.724)	9.738	(4.697)	8.382	(3.078)
90%	None	11.110	(3.066)	9.081	(2.175)	11.980	(4.069)	10.011	(2.773)

Table 3.3: Comparing the classification performance of model-based discriminant analysis using the 3 strategies of wavelength selection given in Section 3.5.2 and Wavelet Thresholding on the NIR honey data, using both the error rate and the Brier's score and using BIC as the model selection method.

BIC	Honey	No Updating				EM				CEM			
		% Error		Brier		% Error		Brier		% Error		Brier	
	Wavelets	6.218	(1.340)	5.110	(1.104)	6.732	(1.629)	5.822	(1.502)	6.238	(1.159)	5.356	(1.024)
50%/	<i>B/W</i>	7.063	(1.889)	5.474	(1.317)	8.749	(4.066)	7.181	(3.217)	7.059	(1.841)	6.070	(1.606)
50%	<i>B + W</i>	6.531	(1.469)	5.206	(1.216)	7.439	(3.261)	6.471	(2.905)	6.335	(1.894)	5.644	(1.807)
	<i>B/W, B + W</i>	5.184	(1.254)	4.342	(1.175)	5.828	(1.389)	5.517	(1.352)	5.552	(1.379)	5.268	(1.339)
	Wavelets	7.513	(1.476)	6.257	(1.320)	11.487	(8.601)	10.353	(8.022)	7.543	(1.890)	6.798	(1.876)
25%/	<i>B/W</i>	7.830	(1.957)	6.214	(1.588)	20.660	(14.968)	18.891	(14.799)	9.919	(7.690)	8.962	(7.625)
75%	<i>B + W</i>	7.535	(1.501)	6.243	(1.323)	21.075	(14.199)	20.248	(14.175)	9.662	(5.828)	9.153	(5.776)
	<i>B/W, B + W</i>	6.189	(1.460)	5.461	(1.422)	8.955	(8.520)	8.845	(8.514)	6.504	(1.071)	6.331	(1.070)
	Wavelets	10.972	(4.348)	9.793	(4.544)	33.824	(20.853)	32.970	(20.925)	17.413	(17.397)	16.848	(17.322)
10%/	<i>B/W</i>	10.993	(4.104)	9.745	(4.194)	31.587	(19.413)	30.771	(19.483)	15.246	(13.289)	14.506	(13.416)
90%	<i>B + W</i>	13.107	(4.705)	12.162	(4.728)	17.366	(12.634)	16.762	(12.512)	10.135	(4.261)	9.623	(4.094)
	<i>B/W, B + W</i>	17.545	(6.716)	16.846	(6.588)	7.835	(2.826)	7.755	(2.709)	8.812	(4.294)	8.491	(3.908)

Table 3.4: Comparing the classification performance of model-based discriminant analysis using the 3 strategies of wavelength selection given in Section 3.5.2 and Wavelet Thresholding on the NIR honey data, using both the error rate and the Brier's score and using 5 fold cross validation as the model selection method.

CV	Honey	No Updating				EM				CEM			
		% Error		Brier		% Error		Brier		% Error		Brier	
	Wavelets	6.444	(1.396)	5.298	(1.010)	6.870	(1.429)	5.933	(1.278)	6.695	(1.355)	5.838	(1.276)
50%/	<i>B/W</i>	8.577	(1.892)	6.448	(1.017)	10.268	(1.900)	7.694	(1.216)	9.870	(2.102)	7.644	(1.469)
50%	<i>B + W</i>	8.088	(1.918)	6.195	(1.407)	9.372	(2.136)	7.406	(1.722)	9.054	(2.228)	7.362	(1.769)
	<i>B/W, B + W</i>	6.741	(1.547)	5.444	(1.166)	7.732	(1.593)	6.654	(1.430)	7.381	(1.710)	6.460	(1.513)
	Wavelets	7.595	(1.357)	6.331	(1.238)	7.785	(1.250)	6.997	(1.191)	7.626	(1.175)	6.843	(1.157)
25%/	<i>B/W</i>	9.846	(1.819)	7.568	(1.447)	12.179	(1.951)	9.544	(1.656)	11.749	(1.870)	9.607	(1.678)
75%	<i>B + W</i>	9.489	(2.358)	7.393	(1.725)	11.162	(4.620)	9.461	(4.740)	10.196	(2.079)	8.783	(1.751)
	<i>B/W, B + W</i>	7.944	(1.601)	6.643	(1.276)	8.310	(1.768)	7.547	(1.526)	8.271	(1.237)	7.497	(1.085)
	Wavelets	11.249	(4.093)	9.931	(3.767)	9.109	(2.749)	8.517	(2.520)	8.691	(2.131)	8.049	(2.014)
10%/	<i>B/W</i>	12.693	(2.862)	10.367	(2.392)	14.688	(5.633)	12.645	(6.118)	13.457	(2.938)	11.714	(2.615)
90%	<i>B + W</i>	12.900	(3.187)	10.486	(2.318)	13.807	(5.785)	12.282	(5.907)	12.326	(4.004)	11.047	(3.698)
	<i>B/W, B + W</i>	11.928	(3.601)	10.429	(3.119)	9.316	(3.782)	8.645	(3.629)	9.426	(3.763)	8.724	(3.600)

Table 3.5: Comparing the classification performance of model-based discriminant analysis using the 3 strategies of wavelength selection given in Section 3.5.2 and Wavelet Thresholding on the NIR olive oil data, using both the error rate and the Brier’s score.

BIC	Olive Oils	No Updating				EM				CEM			
		% Error		Brier		% Error		Brier		% Error		Brier	
	Wavelets	0	(0)	0.003	(0.026)	0	(0)	0	(0)	0	(0)	0	(0)
50%/	<i>B/W</i>	0.551	(0.869)	0.497	(0.770)	0.464	(0.744)	0.448	(0.688)	0.478	(0.744)	0.444	(0.688)
50%	<i>B + W</i>	0	(0)	2.2×10^{-9}	(2.2×10^{-8})	0	(0)	0	(0)	0	(0)	0	(0)
	<i>B/W, B + W</i>	0.145	(0.437)	0.149	(0.437)	0.145	(0.437)	0.145	(0.437)	0.145	(0.437)	0.145	(0.437)
	Wavelets	20.433	(12.151)	20.400	(12.182)	3.962	(11.479)	3.913	(11.337)	1.692	(7.770)	1.694	(7.772)
25%/	<i>B/W</i>	4.490	(6.842)	4.327	(6.845)	0.5	(0.859)	0.496	(0.851)	0.875	(1.504)	0.749	(1.236)
75%	<i>B + W</i>	1.125	(2.633)	1.118	(2.630)	0.106	(0.625)	0.106	(0.624)	0.375	(1.533)	0.333	(1.379)
	<i>B/W, B + W</i>	18.375	(2.090)	17.876	(2.167)	5.971	(5.052)	5.937	(5.018)	12.952	(4.115)	11.779	(3.748)
	Wavelets	27.440	(6.876)	25.929	(7.055)	5.064	(11.213)	4.903	(10.899)	17.632	(12.171)	15.766	(11.347)
10%/	<i>B/W</i>	21.376	(5.015)	20.108	(5.056)	4.296	(8.765)	4.228	(8.655)	10.160	(9.859)	9.085	(9.355)
90%	<i>B + W</i>	19.720	(4.545)	19.118	(4.669)	5.416	(9.306)	5.364	(9.233)	11.616	(8.797)	10.604	(8.575)
	<i>B/W, B + W</i>	20.216	(4.638)	19.794	(4.652)	14.768	(7.662)	14.714	(7.651)	17.808	(5.941)	17.009	(5.973)

Table 3.6: Comparing the classification performance of model-based discriminant analysis using the 3 strategies of wavelength selection given in Section 3.5.2 and Wavelet Thresholding on the NIR olive oil data, using both the error rate and the Brier's score.

CV	Olive Oils	No Updating				EM				CEM			
		% Error		Brier		% Error		Brier		% Error		Brier	
	Wavelets	0.043	(0.248)	0.044	(0.240)	0.014	(0.145)	0.014	(0.145)	0.116	(0.395)	0.115	(0.390)
50%/	<i>B/W</i>	0.725	(1.278)	0.672	(1.185)	0.522	(0.811)	0.506	(0.793)	0.580	(0.988)	0.529	(0.860)
50%	<i>B + W</i>	0	(0)	2.2×10^{-9}	(2.2×10^{-8})	0	(0)	0	(0)	0	(0)	0	(0)
	<i>B/W, B + W</i>	0.667	(3.086)	0.638	(2.924)	6.507	(11.071)	6.507	(11.071)	6.884	(11.751)	6.884	(11.751)
	Wavelets	0.625	(2.580)	0.533	(2.086)	0	(0)	0	(0)	0	(0)	0	(0)
25%/	<i>B/W</i>	4.414	(6.725)	4.042	(6.154)	1.683	(4.566)	1.658	(4.495)	2.183	(4.901)	2.012	(4.749)
75%	<i>B + W</i>	16.5	(5.636)	15.443	(5.348)	0.817	(4.117)	0.779	(3.901)	1.058	(4.144)	0.990	(3.960)
	<i>B/W, B + W</i>	19.183	(2.848)	18.364	(18.364)	7.087	(7.266)	7.055	(7.226)	13.990	(5.794)	12.884	(5.582)
	Wavelets	0.226	(2.258)	0.209	(2.095)	0	(0)	0	(0)	0.121	(1.210)	0.103	(1.035)
10%/	<i>B/W</i>	-	-	-	-	-	-	-	-	-	-	-	-
90%	<i>B + W</i>	-	-	-	-	-	-	-	-	-	-	-	-
	<i>B/W, B + W</i>	-	-	-	-	-	-	-	-	-	-	-	-

Table 3.7: Comparing the classification performance of PLSR with and without Savitzky Golay filter, using both BIC and cross validation at various training/test splits of the NIR olive oil data, using both the error rate and the Brier's score

Olive Oils	BIC				5 fold CV			
	% Error		Brier		% Error		Brier	
50%/ SG	0	(0)	0.517	(0.168)	1.652	(1.397)	3.671	(1.219)
50% None	0.449	(0.674)	0.582	(0.432)	0.188	(0.815)	0.769	(0.899)
25%/ SG	0.010	(0.096)	0.949	(0.331)	1.990	(2.047)	4.009	(1.682)
75% None	0.231	(0.413)	0.695	(0.256)	16.596	(3.358)	11.442	(1.604)
10%/ SG	0.792	(1.258)	1.858	(0.820)	-	-	-	-
90% None	0.704	(1.257)	1.898	(1.106)	-	-	-	-

3.8 Conclusions

When there are sufficient data available, using cross validation is a highly effective model selection technique. However, it relies on each of the folds in the training data being largely representative of the rest of the data. As the total number of samples available for the training data is decreased, the size of each of the folds also decreases, making it more likely that at least one group will not be represented in a given fold. As such, calculations become unstable and the process becomes infeasible.

To counteract this instability, the development and implementation of an alternative automatic model selection process for PLSR was required. Determining the effective number of parameters enabled BIC to be used as a method of selecting the number of components to include. BIC tended to select more components than cross validation without substantially effecting classification performance, indicating that the practice of selecting a low number of components to prevent over-fitting of the data may be too conservative.

The strength of using complexity criteria as a basis for model selection over cross validation becomes especially evident when using the olive oil data with 10% in the training set (14 samples). Most of the cross validation based methods completely

fail at this point. While this is an extreme scenario, it is worth noting that using BIC with PLSR remains effective.

Using wavelet thresholding as a dimension reduction technique for the spectra requires no information about the labelling, thus is more robust than the methods that require labelling information. The extra stability of the dimension reduction process requiring no labelling information is apparent in Table 3.4, where the wavelet thresholding method of dimension reduction provides consistently better classification results than methods requiring labelling information in order to select wavelengths and in Table 3.6 where the other methods are too unstable to provide results when the training data set is reduced to 10% of the overall data. If wavelet thresholding is to become a common analysis technique the range of variables scanned would have to increase from 1100 – 2498 nm to also include the colour spectrum (400 – 2498 nm) so that less information would need to be discarded. If the number of wavelengths available was routinely 1050, dropping the first and last 13 wavelengths removes a little of the border effect without losing too much information. Continuing using wavelet thresholding with only the region 1100 – 2498 nm potentially results in a loss of discriminatory regions of the spectrum in order to reduce the number of variables to 512 (2^9).

Dean et al. (2006) showed updating methods to be highly effective using the NIR meats data. However, using updating methods with model based classification methods rely on the initial model assumptions to be correct. If such conditions are not completely satisfied, any problems in the modelling process are exacerbated by incorporating the extra available unlabelled observations. The model assumptions are sufficiently satisfied by the olive oil data – where the updating methods (EM and CEM) consistently outperform the traditional methods. The support for updating on the honey data is not so clear as the updating methods only improve the classification performance at the 10%/90% split using the combination of B/W , $B+W$ for dimension reduction, but substantially disimprove classification performance with the other dimension reduction methods.

Using a hard classification rule at each iteration of the EM algorithm, as is the case with CEM, often results in faster convergence. Neither method consistently outperforms the other as shown by Tables (3.3, 3.4, 3.5 and 3.6). The hard updating

approach of CEM tends to converge to a solution faster, but the soft updating approach is more consistent when using the Brier's score as a performance metric in that it retains probabilities throughout. Using soft updating (EM) uncertain classification decisions can be deferred until the algorithm has converged to a solution rather than making the classification decision at an earlier point. This also can avoid the problem of an observation getting "*stuck*" in the wrong category at an early point of the algorithm.

Often the more important decision is whether to use any updating technique rather than which updating approach to use.

Toher et al. (2007) use the honey data to compare the performance of the model-based methods, using wavelet thresholding as dimension reduction technique, to that of PLSR, on a series of more extreme scenarios. The random splits of the data are constrained to have the same number of pure and adulterated samples in training set of each random split, with the proportion of pure to adulterated samples varied so that the training sets are extremely unrepresentative of the entire data. It demonstrates that in such extreme situations, the updating models that rely more heavily on the model assumptions are more likely to fail, with the soft updating (EM) version of updating more problematic than the hard updating (CEM). Using BIC as a model selection technique for PLSR proves to be extremely consistent in its performance under the extreme situations for the honey data studied in Toher et al. (2007).

Chapter 4

Group of Interest based Classification

4.1 General Concept

Group of Interest based classification is a variation of the general discriminant analysis techniques introduced in Chapter 3. It is motivated by food authenticity applications, where the information available about groups is unbalanced. Trying to verify that a product is what it claims to be on the label means that it can be compared to reference material matching the product claims, but what if the product is not what it is labelled to be? Rather than trying to identify exactly what a sample is, group of interest based classification focuses on the problem: is it what it claims to be or not? In such a situation the information available is unbalanced because while it is possible to obtain and examine fully samples that are what they claim to be, trying to account for all possible erroneous samples to the point at which they could all be correctly classified would be impractical.

Treating different groups in an asymmetric manner is examined in the context of finding suitable projections of very high dimensional datasets for visualisation purposes by Hennig (2004).

Here the group of interest itself (what the product claims to be) is treated as a homogeneous group and is modelled as a Gaussian distribution. Other observations are treated in a different fashion. Considering the other observations to follow the

distribution $f_o(x)$

$$f(x) = p_g \underbrace{\frac{1}{\sqrt{2\pi}|\Sigma_g|^{d/2}} \exp \left[-\frac{1}{2}(x - \mu_g)\Sigma_g^{-1}(x - \mu_g)^T \right]}_{\text{group of interest}} + \underbrace{p_o f_o(x)}_{\text{other}} \quad (4.1)$$

Dean and Raftery (2005) consider using a normal-uniform mixture model to represent differentially expressed and non-differentially expressed genes. Each gene is treated univariately, with no consideration given to the correlation between genes.

If the group of interest observations are distributed according to $f_g(x)$ and the other observations follow a $f_o(x)$ distribution then an observation x is classified as belonging to the group of interest if

$$f_g(x) > \frac{(1 - p_g)f_o(x)}{p_g} \quad (4.2)$$

and the expected values of the labels are

$$l = \frac{f_g(x)}{f_g(x) + \frac{(1-p_g)f_o(x)}{p_g}}. \quad (4.3)$$

This then leads immediately to the concept of thresholding. Setting the threshold, $\tau(x)$ to be $\frac{(1-p_g)f_o(x)}{p_g}$, an observation will be classified as belonging to the group of interest if $f_g(x)$ exceeds this threshold otherwise it will be classed as not belonging to the group of interest.

Two distribution types are considered to model the “other” observations: Poisson and a mixture of Gaussian distributions.

Considering the “other” observations as being Poisson noise is analogous to treating them as being randomly distributed over a defined subspace – “other” observations can in fact lie in the same part of the subspace as those belonging to the group of interest, or they can be dispersed over a defined potential space. This echoes the belief that those trying to commit commercial fraud would attempt to reproduce an approximate facsimile of the claimed product, thus any observation belonging to the “other” group should be relatively close to the original, with some deviation as a result of the product not being what it is claimed to be.

A mixture of Gaussian distributions is a very powerful modelling tool, especially when the number of Gaussian distributions is unconstrained. As many other distributions can be approximated using a Gaussian mixture, the Gaussian mixture approach provides a uniquely flexible framework to consider the “other” observations.

It also has a natural interpretation for chemists in that each of the components of the mixture could represent a different type of adulterant when trying to detect multiple adulterated samples when there are an unknown number of potential sources of adulteration.

$f_o(x)$ is Poisson noise

If $f_o(x)$ is Poisson noise on a set of volume V , then

$$\begin{aligned} f(x) &= p_g f_g(x|\mu_g, \Sigma_g) + p_o f(x|V) \\ &= p_g f_g(x|\mu_g, \Sigma_g) + (1 - p_g) f(x|V) \\ &= p_g f_g(x|\mu_g, \Sigma_g) + (1 - p_g) \frac{1}{V} \end{aligned}$$

Thus observations are classified as belonging to the group of interest if

$$\begin{aligned} f_g(x|\mu_g, \Sigma_g) &> \frac{(1 - p_g) f_o(x|V)}{p_g} \\ &> \frac{1 - p_g}{p_g V}. \end{aligned}$$

The volume V is calculated on the entire dataset – irrespective of the group membership of individual observations. If x is univariate, then $V = \max x - \min x$. If x is multivariate then the volume is estimated as the minimum of the volume of the ellipsoid hull spanning the data and the hypervolume of the data. As it does not depend on group membership, updating does not effect the value of V . However, the estimated values of p_g , μ_g and Σ_g can be influenced by updating methods. Considering that if

$$\begin{aligned} p_g &= \frac{\sum_{j=1}^N z_{jg}}{N}, & \hat{\mu}_g &= \frac{\sum_{j=1}^N z_{jg} x_j}{\sum_{j=1}^N z_{jg}}, \\ \hat{\Sigma}_g &= \frac{\sum_{j=1}^N z_{jg} (x_j - \mu_g)(x_j - \mu_g)'}{\sum_{j=1}^N z_{jg}}, \end{aligned}$$

and using z_{jg} as an indicator variable for the training data, following (4.3) the estimated probability of x_j belonging to the group of interest is:

$$z_{jg} = \frac{f(x_j|\hat{\mu}_g, \hat{\Sigma}_g)}{f(x_j|\hat{\mu}_g, \hat{\Sigma}_g) + \frac{1-p_g}{p_g V}}$$

which can then be used as weights in the calculation of μ_g , Σ_g , p_g which in turn are included in the calculation of the z_{jg} 's until the updating process reaches its stopping point.

However, using updating is found in Tables 4.1 and 4.3 to be ineffective when using $f_o(x) = 1/V$ as the threshold is raised too high and most observations are placed into the *other* group.

$f_o(x)$ is a mixture of Gaussian distributions

If $f_o(x)$ is a mixture of K Gaussian distributions then

$$\begin{aligned} f(x) &= p_g f_g(x|\mu_g, \Sigma_g) + p_o \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k) \\ &= p_g f_g(x|\mu_g, \Sigma_g) + (1 - p_g) \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k) \end{aligned}$$

and observations are classified as belonging to the group of interest if

$$\begin{aligned} p_g f_g(x|\mu_g, \Sigma_g) &> (1 - p_g) \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k) \\ f_g(x|\mu_g, \Sigma_g) &> \frac{(1 - p_g)}{p_g} \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k). \end{aligned}$$

The number of components, K , is determined using the BIC. This method is computationally very expensive, thus not practical for use with the variable selection procedure described in Section 4.2 on the entire NIR spectra. Therefore methods of dimension reduction described in Chapter 3 are used before variable selection is implemented.

No updating methods were examined for this method due mainly to the already high computational burden of using $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$, but also because of the extra complexity caused by updating increasing or decreasing K in each iteration.

As discussed in Chapter 3, due to the highly collinear nature of near infrared spectroscopic data, there is a need for some form of dimension reduction in order to have the invertible covariance matrices that are required for the calculation of $f_g(x)$ and, when $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$ also for calculating the Σ_k^{-1} 's. The dimension reduction techniques introduced in Chapter 3 give one approach to dimension

reduction – where the dimension reduction procedure is completed before the model is fitted to the data. Another approach is to consider the selection of individual wavelengths that optimise classification in a stepwise manner.

4.2 Variable Selection and Dimension Reduction

The dimension reduction techniques considered in Chapter 3 were considered as a method of reducing the computational burden for the NIR data sets. Also considered, where feasible, was using a method of variable selection described below.

The number of folds, F is set. The training set is then split into F separate folds. The Brier's score is calculated across each fold within the training set only so that variable selection is performed using only the training set data, but with consideration given to out of sample classification performance.

4.2.1 Variable Selection Procedure

Algorithm for variable selection:

Algorithm when $f_o(x) = 1/V$

- **Step 1** Each variable is proposed

$$f(x^{(p)}) = p_g f_g(x^{(p)} | \mu_g^{(p)}, \Sigma_g^{(p)}) + \frac{(1 - p_g)}{V(x^{(p)})} \text{ univariately.}$$

$$f_g(x^{(p)} | \mu_g^{(p)}, \Sigma_g^{(p)}) \sim N(\mu_g^{(p)}, \Sigma_g^{(p)})$$

where $V(x^{(p)}) = \max(x^{(p)}) - \min(x^{(p)})$.

The Brier's score is then calculated across all of the cross validation folds for each variable.

- **Step 2** The variable / wavelength that minimises the Brier's score across these folds is then selected. This forms the initial set, c , of selected variables.
- **Step 3** All remaining variables outside the set of currently selected are then proposed in turn:

$$f(x^{(c)}, x^{(p)}) = p_g f_g(x^{(c)}, x^{(p)} | \mu_g^{(c,p)}, \Sigma_g^{(c,p)}) + \frac{(1 - p_g)}{V(x^{(c)}, x^{(p)})}$$

univariately in the additional variable.

$$f_g(x^{(c)}, x^{(p)} |, \mu_g^{(c,p)}, \Sigma_g^{(c,p)}) \sim N(\mu_g^{(c,p)}, \Sigma_g^{(c,p)})$$

where $V(x^{(c)}, x^{(p)}) = \min(\text{volume of the ellipsoid hull of } (x^{(c)}, x^{(p)}), \text{hypervolume of the rectangle spanning of } (x^{(c)}, x^{(p)}))$.

The Brier's score is then calculated across all of the cross validation folds for each variable.

- **Step 4** The variable that minimises the Brier's score across the cross validation folds is added to c .
- **Step 5** The remaining variables are then returned to step 3, until the maximum allowable variables have been added to the model.

Algorithm when $f_o(x) = \sum_{k=1}^K q_k f(x | \mu_k, \Sigma_k)$

- **Step 1** Each variable is proposed

$$f(x^{(p)}) = p_g f_g(x^{(p)} |, \mu_g^{(p)}, \Sigma_g^{(p)}) + (1 - p_g) \sum_{k=1}^K q_k f(x^{(p)} | \mu_k^{(p)}, \Sigma_k^{(p)}) \text{ univariately.}$$

$$f_g(x^{(p)} |, \mu_g^{(p)}, \Sigma_g^{(p)}) \sim N(\mu_g^{(p)}, \Sigma_g^{(p)})$$

The number of components in the mixture describing the "other" group, K , is selected using the BIC.

The Brier's score is then calculated across all of the cross validation folds for each variable.

- **Step 2** The variable / wavelength that minimises the Brier's score across these folds is then selected. This forms the initial set, c , of selected variables.
- **Step 3** All remaining variables outside the set of currently selected are then proposed in turn:

$$f(x^{(c)}, x^{(p)}) = p_g f_g(x^{(c)}, x^{(p)} |, \mu_g^{(c,p)}, \Sigma_g^{(c,p)}) + (1 - p_g) \sum_{k=1}^K q_k f(x^{(c)}, x^{(p)} | \mu_k^{(c,p)}, \Sigma_k^{(c,p)})$$

univariately in the additional variable.

$$f_g(x^{(c)}, x^{(p)} |, \mu_g^{(c,p)}, \Sigma_g^{(c,p)}) \sim N(\mu_g^{(c,p)}, \Sigma_g^{(c,p)})$$

K , the number of components in the mixture, $\sum_{k=1}^K q_k f(x^{(c)}, x^{(p)} | \mu_k^{(c,p)}, \Sigma_k^{(c,p)})$, is selected using the BIC.

The Brier's score is then calculated across all of the cross validation folds for each variable.

- **Step 4** The variable that minimises the Brier's score across the cross validation folds is added to c .
- **Step 5** The remaining variables are then returned to step 3, until the maximum allowable variables have been added to the model.

The number of variables that minimises the Brier's score then determines the number of variables and hence which variables should be included in the final model.

The parameters for this model are then calculated across the entire training data and then applied to the test set data so that performance can be evaluated.

Illustration of Variable Selection Process

To illustrate the variable selection process in practice, the following example uses the wine data and considers the Barolo wines to be the group of interest. Splitting the data so that 50% of the data is in the training set with the remainder in the test set, 5 fold cross validation is used to determine the Brier's score within the training set. The model under consideration is

$$f(x) = p_g f(x | \mu_g, \Sigma_g) + (1 - p_g) \frac{1}{V}$$

The mean of the Brier's score across the 5 cross validation folds are plotted for each variable. In this case the variable with index **2** in Figure 4.1 – *Malic acid*, has the lowest average Brier's score and thus is the first to be added to the model. If this score remains unbeaten, *i.e.* is the lowest overall, then only one variable will be included in the model.

Plotting the difference in the Brier's scores for the remaining variables combined with the variable already selected (Malic Acid) and the Brier's score achieved in the single dimension case. In order to improve the model, the difference in the Brier's score would have to be less than 0, or below the green horizontal line in Figure

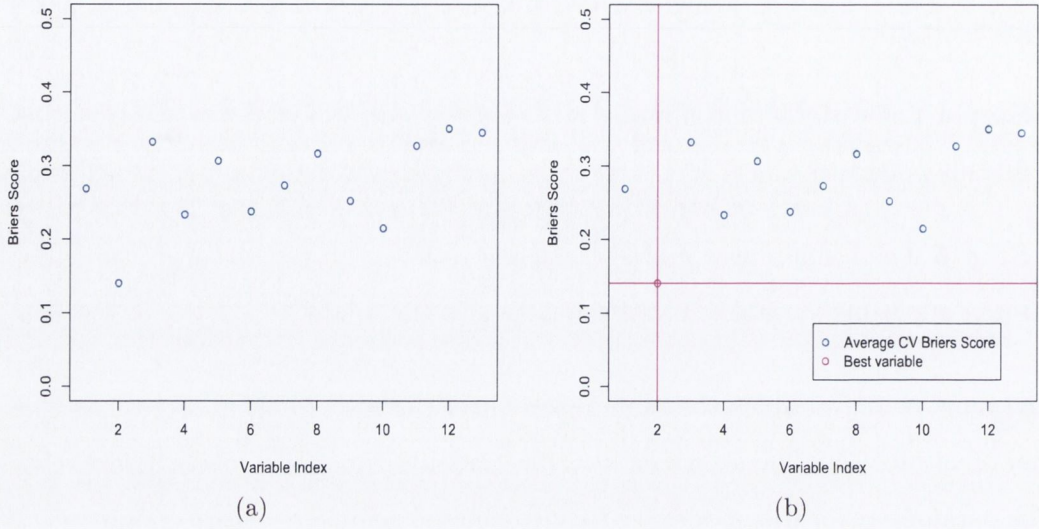


Fig. 4.1: Selecting the first variable

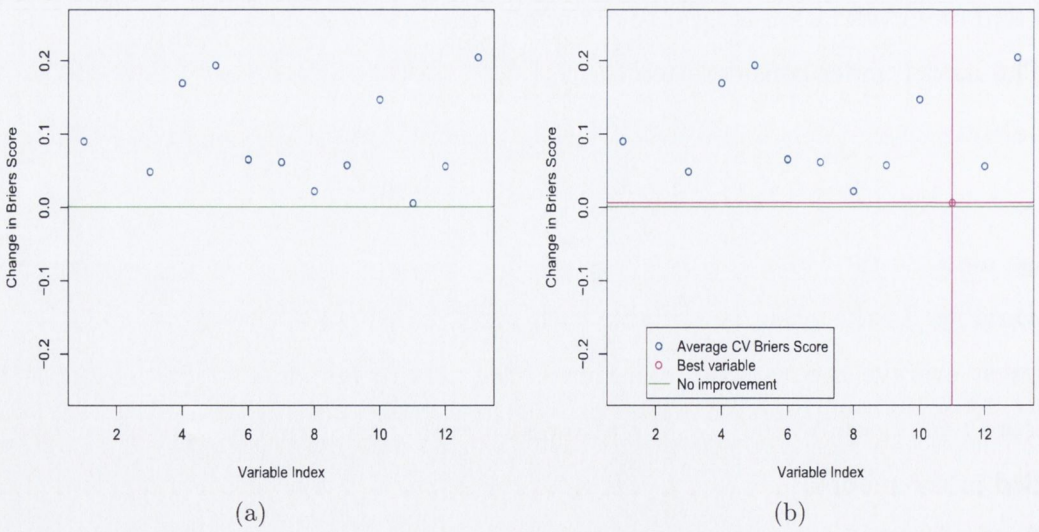


Fig. 4.2: Selecting the second variable

4.2(a). The second variable to be added to the model is that with index **11**, which is *Hue*. It is visible in Figure 4.2(b) that this does not improve on the Brier's score previously achieved. This variable is temporarily included in the model, so that if a lower Brier's score can be achieved by adding further variables, it will be included in the final model; otherwise the final model will only have one variable.

Continuing on to consider adding a third and fourth variable. Adding a third

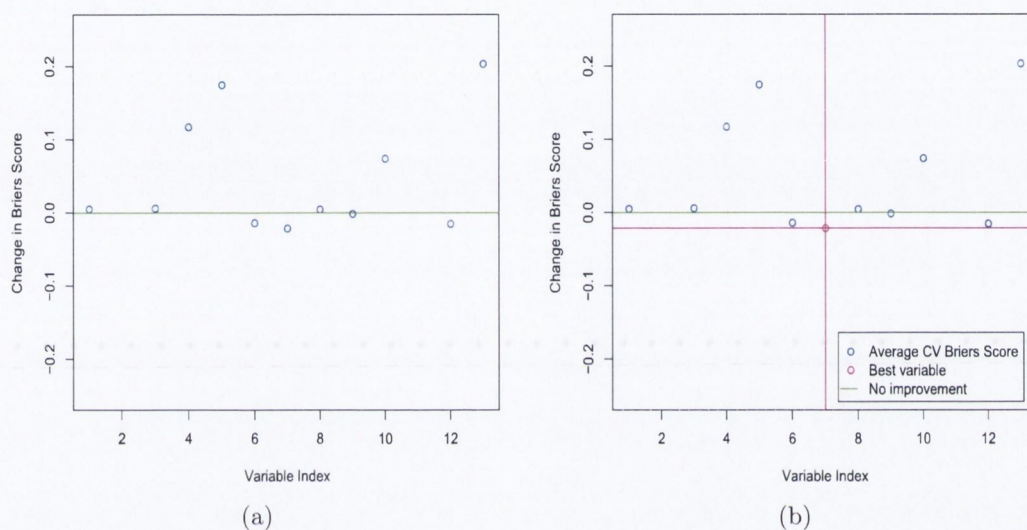


Fig. 4.3: Selecting the third variable

variable to the model reduces the Brier's score. The lowest Brier's score is achieved by adding the variable with index **7** in Figure 4.3, which are the *flavenoids*. Adding a fourth variable also reduces the Brier's score. Adding the *nonflavanoid phenols* (index **8** in Figure 4.4 reduces the Brier's score the most. This process is continued for the remaining variables. In this case, no further reduction in the Brier's score is achieved, so that the final model consists of *Malic acid*, *Hue*, *Flavenoids* and *Nonflavanoid Phenols*.

Once all of the variables for inclusion in the model have been selected, the entire training set is then used to calculate the values μ_g and Σ_g , V is calculated using all of the data, as it does not require group information.

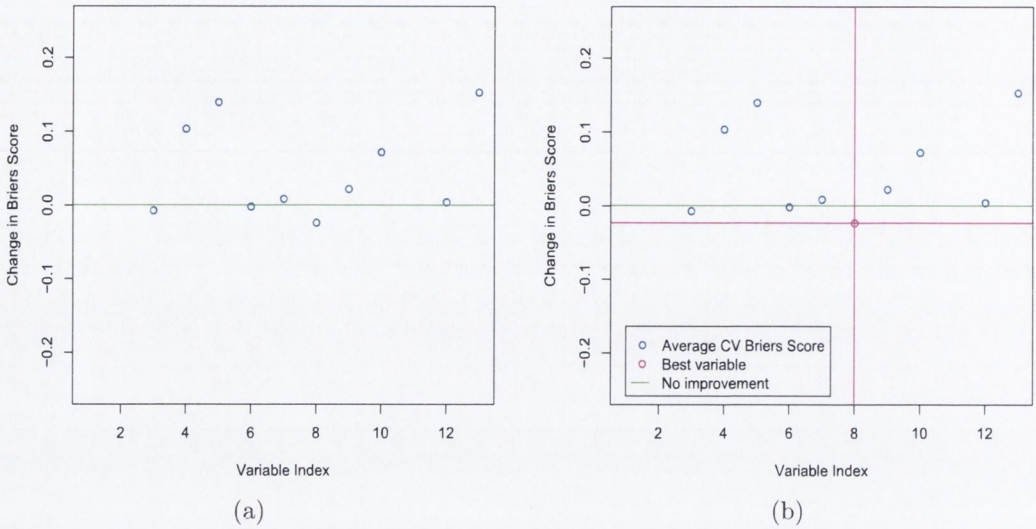


Fig. 4.4: Selecting the fourth variable

4.3 Varying the Threshold

Recall equation (4.2) the threshold $\tau(x)$ is:

$$\tau(x) = \frac{1 - p_g}{p_g} f_o(x)$$

and the decision becomes

$$\begin{aligned} f_g(x|\mu_g, \Sigma_g) &\leq \tau(x) \Rightarrow \text{classify as "other"} \\ &> \tau(x) \Rightarrow \text{classify as group of interest} \end{aligned}$$

If $f_o(x) = 1/V$, then

$$\tau(x) = \tau = \frac{1 - p_g}{p_g V}.$$

Figure 4.5 illustrates the effect of changing p_g and of changing the volume V on the value of this classification threshold. Observations in Figure 4.5 that fall below the horizontal lines would be classified as not belonging to the group of interest. Figure 4.5(a) fixes the volume to be $V = 10$ when $p_g = 0.9$, $\tau = 0.01$ (the red line) whereas when $p_g = 0.3$, $\tau = 0.23$ (the green line). This illustrates the dependence on the estimated proportions in each group. Figure 4.5(b) fixes the $p_g = 0.2$ and looks at how volume and the threshold, τ are related.

Continuing to assume that $f_g(\mathbf{x}_n|\theta_g) \sim N(\mu_g, \Sigma_g)$, it is apparent that the minimum possible value of $\tau(x)$ is 0 while the maximum possible value is at $f_g(\mu_g|\mu_g, \Sigma_g)$.

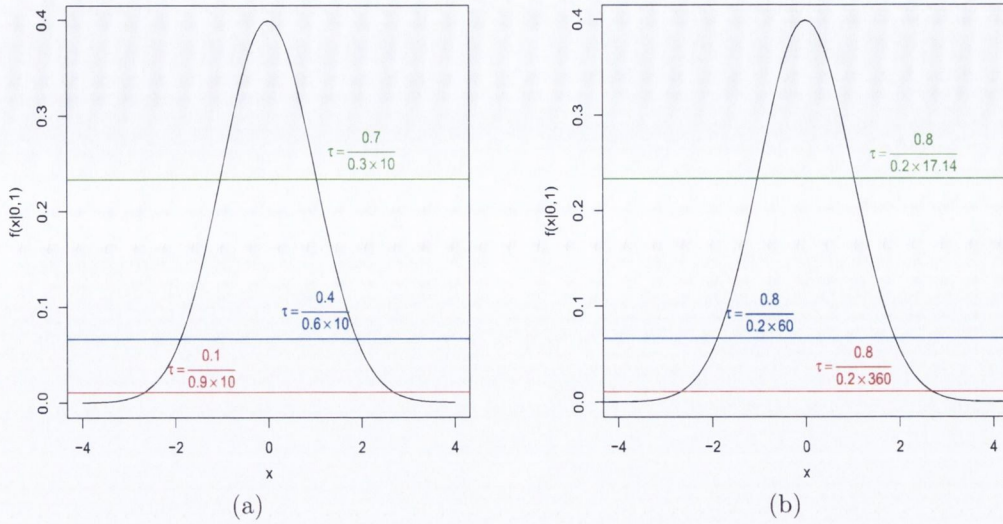


Fig. 4.5: One dimensional example of the behaviour of the threshold. Figure 4.5(a) varies p_g and Figure 4.5(b) varies the volume.

4.3.1 Evaluating the Threshold Directly

The threshold, $\tau(x)$ that determines if an observation is classified as belonging to the group of interest or not, is the primary mechanism for deciding on group membership. Determining an optimal value of τ that not dependent on the value of x , for classification purposes is of interest. Using a grid search with an initial range of τ from 0 to $f_g(\mu_g, \Sigma_g)$, the grid is made finer and finer until an optimal value of the threshold $\tau(x)$ for classification is found. Finding the optimal value of τ in this manner is computationally too expensive for practical ongoing use, especially as the number of variables in the model grow, therefore finding the relationship between τ and other, more accessible parameters is also of interest. As the volume V of the data is relatively simple to calculate and would correspond to $f_o(x)$ being Poisson noise, the volume of the data was considered as an estimator of $\tau(x)$. To do this, the relationship between the optimal $\tau(x)$ and V was calculated across several datasets, in order to determine if this relationship could be considered relatively independent of the actual data.

4.3.2 Behaviour of the Threshold

As the number of variables increases, the volume of the data becomes the dominant part of equation (4.2). As the predicted proportions remain relatively constant so does $\frac{1-p_g}{p_g}$. Therefore $f_o(x)$ becomes the dominant part of the threshold calculation. When $f_o(x) = 1/V$, the volume of the data decreases as the number of variables included in the model increases. This is illustrated using the NIR olive oil sample in Figure 4.6, using the first twenty random 50%/50% splits of the data.

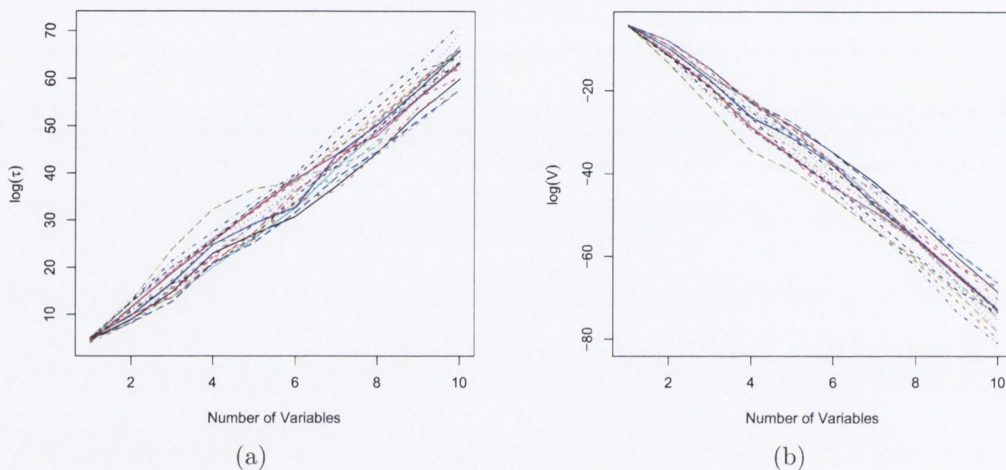


Fig. 4.6: Behaviour of $\log(\tau)$ and $\log(\text{volume})$ as the number of variables in the model increases.

4.4 Results

4.4.1 NIR datasets

In order to improve robustness, in this Chapter 3-fold cross validation rather than 5-fold cross validation is used for model selection purposes.

For the large NIR datasets, having a single fixed threshold for all observations, even when the optimal value for that threshold has been calculated, is poor as not enough flexibility exists in such models. To capture the variability within the observations not belonging to the group of interest, a more effective approach is to use $f_o(x) = \sum_{k=1}^K q_k f_k(x|\mu_k, \Sigma_k)$, where the number of components and the structure

of the covariance matrices as outlined in Table 3.1 are decided using BIC. However, due to the extra parameters to be estimate when using $f_o(x) = \sum_{k=1}^K q_k f_k(x|\mu_k, \Sigma_k)$, performance suffers as the number of observations in the training set is reduced.

The meats NIR data do not present an obviously one-sided problem, whereas both the olive oil and the honey NIR datasets do. Therefore, the results presented in this section are for the honey and olive oil datasets only.

Wavelets

No variable selection is used with the wavelets after hard thresholding has been used. This is because all remaining coefficients are required to summarise the spectral data.

NIR Honey Data

Using the NIR honey data, Table 4.1 illustrates the performance of using

$$f_g(x|\mu_g, \Sigma_g) > \frac{1 - p_g}{p_g V}$$

as the classification criterion while Table 4.2 illustrates the performance of using

$$f_g(x|\mu_g, \Sigma_g) > \frac{(1 - p_g) \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)}{p_g}$$

as the classification criterion. In Table 4.2 there are two parts. In the first part, once the initial dimension reduction technique is implemented, no further variable selection is used. This is then compared to using the same techniques to reduce the number of variables to be searched over using the variable selection technique described in Section 4.2. As using $f_o(x) = 1/V$ is much faster than $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$, it is feasible to search over the entire spectrum when $f_o(x) = 1/V$. It is worth noting that using the crude technique of selecting every 10th nm (equivalently every 5th wavelength) achieves similar classification results, but significantly reduces the computational time.

Figures 4.7, 4.8 and 4.9 examine the variables selected by the initial dimension reduction techniques referred to in Section 3.5.2 and how, when using these as an initial dimension reduction procedure, $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$ chooses more variables than $f_o(x) = 1/V$ across 100 random splits of the data into training and test sets. Both methods of describing f_o choose variables from similar parts of the

spectrum, indicating that using $f_o(x) = 1/V$ as an initial dimension reduction tool, not necessarily discarding the variables that are included after the optimum number of variables for $f_o(x) = 1/V$, may dramatically improve on the computational cost of performing $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$.

NIR Olive Oil Data

The same analysis is undertaken on the NIR olive oil data – the dimension reduction techniques described in Chapter 3 both as isolated methods and as methods of reducing the space over which the variable selection technique must search.

τ directly calculated

Using every 10th nm for both the NIR honey and the NIR olive oil datasets, τ was calculated directly. In Table 4.5 the performance of using this method is compared with that using $f_o(x) = 1/V$ – using the variables selected by calculating τ directly and using the variables selected by $f_o(x) = 1/V$.

With the exception of the 50%/50% split of the olive oil, using $f_o(x) = 1/V$ to select which variables to include, then calculating τ directly using these variables gives an improvement in classification performance. In Table 4.5, the number of variables to include in the model, out of those selected is determined by the same method as was used to select the variables. Table 4.6 compares the error rates when the number of variables to include in the model is chosen using the training data when τ is calculated directly against those when the $f_o(x) = 1/V$ determines the number of variables to include. In both cases, the set of variables to include (and the order in which to include them) were selected by $f_o(x) = 1/V$.

The classification performance achieved by both methods of determining the number of variables to include in the model for the honey data is very similar. However, using the combination of $f_o(x) = 1/V$ to determine which ordered subset of variables should be considered, then using the training set classification performance to determine the number of variables to include achieves a marked improvement in classification performance for the olive oil data.

Figure 4.12 gives an example of the strong linear relationship of $\log \tau$ to $\log 1/V$.

Table 4.1: Classification Performance of Group of Interest Based Methods on Honey Data: $f_o(x) = 1/V$. The performance of the various methods of reducing the search space are compared, as is the performance of using updating methods against not using updating methods.

Dimension		$f_o(x) = \frac{1}{V}$			
Reduction		No Updating		Updating	
Split	Method	% Error	Brier	% Error	Brier
50%/	None	20.343 (4.881)	14.577 (2.932)	22.301 (3.886)	16.816 (2.682)
50%	10 th nm	21.305 (5.608)	15.233 (3.424)	22.004 (4.779)	16.383 (2.967)
	B/W	24.640 (4.629)	17.363 (2.909)	28.088 (3.914)	20.573 (2.587)
	B+W	23.410 (4.717)	16.653 (3.018)	26.259 (3.938)	19.454 (2.657)
	B/W,B+W	22.661 (4.562)	16.045 (3.002)	24.778 (4.135)	18.451 (2.798)
25%/	None	22.316 (6.470)	16.063 (4.214)	30.070 (6.443)	22.564 (2.955)
75%	10 th nm	26.120 (6.235)	18.579 (4.052)	29.908 (6.042)	22.816 (2.969)
	B/W	30.961 (6.210)	21.073 (3.787)	31.975 (4.038)	24.989 (2.050)
	B+W	27.419 (6.263)	19.739 (4.086)	30.637 (3.534)	24.140 (2.592)
	B/W,B+W	26.715 (6.210)	18.958 (3.082)	31.045 (4.493)	24.265 (3.048)
10%/	None	27.816 (8.836)	20.003 (5.855)	35.695 (8.676)	28.656 (5.016)
90%	10 th nm	28.907 (8.540)	20.538 (5.797)	35.047 (8.134)	28.126 (4.411)
	B/W	33.700 (8.183)	23.142 (5.425)	39.433 (9.963)	30.758 (4.883)
	B+W	31.898 (8.269)	22.415 (5.452)	38.163 (9.393)	30.886 (5.160)
	B/W,B+W	30.058 (8.142)	21.112 (5.247)	38.144 (9.837)	30.255 (5.022)

Table 4.2: Classification Performance of Group of Interest Based Methods onHoney Data: $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$

Split	Method	$f_o(x) = \sum_{k=1}^K q_k f(x \mu_k, \Sigma_k)$	
		% Error	Brier
50%/50%	Wavelets	7.824 (2.202)	6.851 (1.996)
	B/W	7.025 (2.032)	5.924 (1.850)
	B+W	5.460 (1.783)	4.826 (1.625)
	1/V	6.531 (1.622)	5.176 (1.333)
25%/75%	Wavelets	9.578 (8.663)	8.567 (8.100)
	B/W	10.237 (12.017)	9.331 (12.175)
	B+W	9.528 (8.854)	8.751 (8.260)
	1/V	7.511 (2.066)	6.205 (1.698)
10%/90%	Wavelets	31.840 (20.144)	26.539 (16.877)
	B/W	37.742 (21.608)	33.034 (20.275)
	B+W	45.765 (19.925)	43.244 (20.165)
	1/V	34.498 (22.748)	31.001 (21.065)
Including Additional Variable Selection			
50%/50%	B/W	7.816 (1.964)	6.195 (1.423)
	B+W	6.423 (1.913)	5.067 (1.440)
	1/V	6.577 (1.604)	5.168 (1.296)
25%/75%	B/W	8.453 (2.305)	6.738 (1.860)
	B+W	7.503 (1.962)	6.015 (1.645)
	1/V	7.626 (2.412)	6.309 (1.972)
10%/90%	B/W	19.749 (15.794)	17.344 (15.812)
	B+W	19.216 (24.844)	16.830 (14.942)
	1/V	33.237 (22.195)	30.257 (21.030)

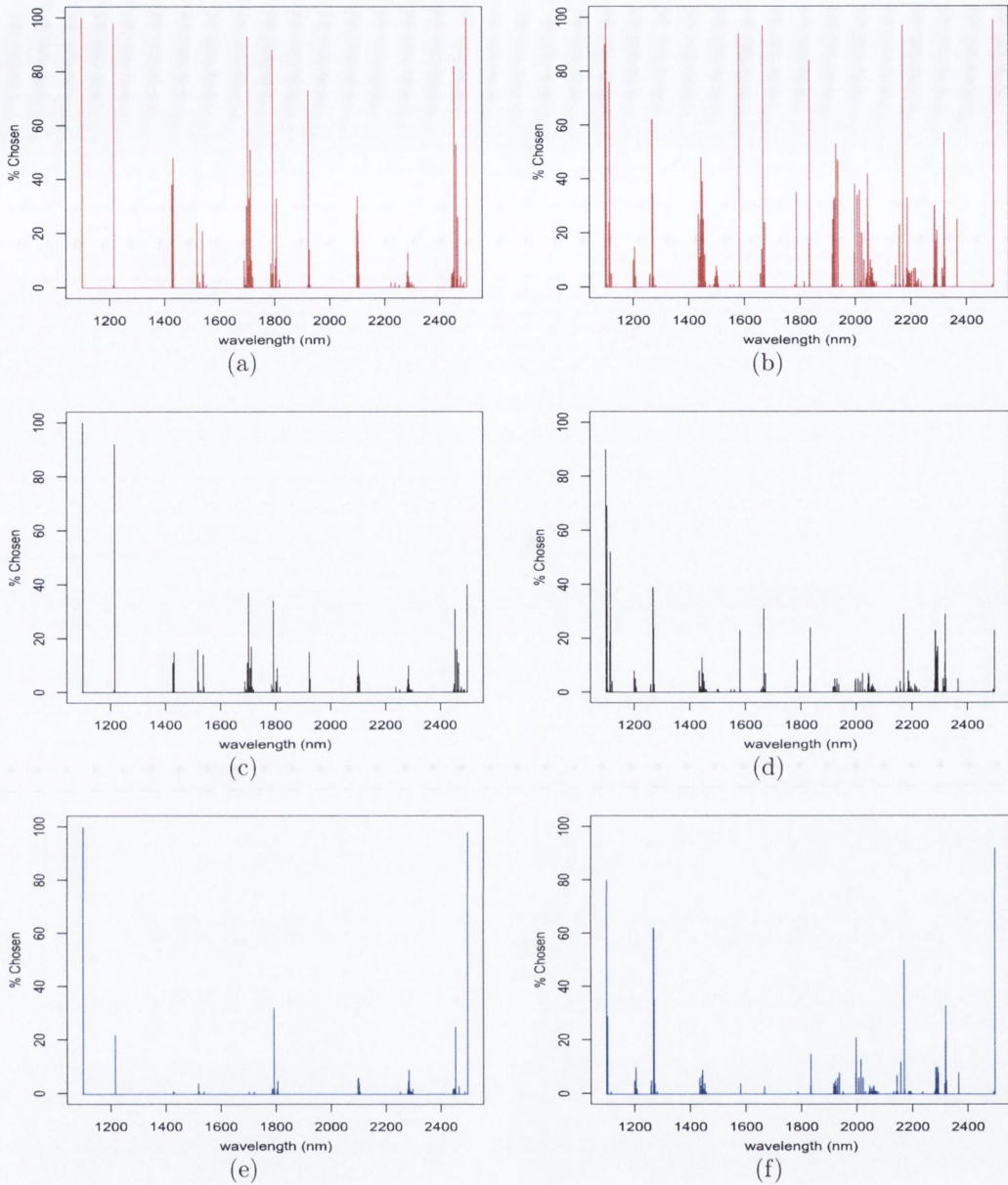


Fig. 4.7: Using a 50%/50% training test data split of the NIR honey data. Variables selected using B/W (Figure 4.7(a)) and $B + W$ (Figure 4.7(b)) as the prior dimension technique. The height of the red lines indicate the frequency that each of the wavelengths were selected using the initial dimension reduction technique. Then, using $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$ the height of the black lines in Figure 4.7(c) and in Figure 4.7(d) demonstrate the frequency that each wavelength is selected when B/W and $B + W$ respectively were used as a prior dimension reduction technique while the blue lines in Figures 4.7(e) and 4.7(f) do the same when using $f_o(x) = 1/V$ (using B/W and $B + W$ respectively).

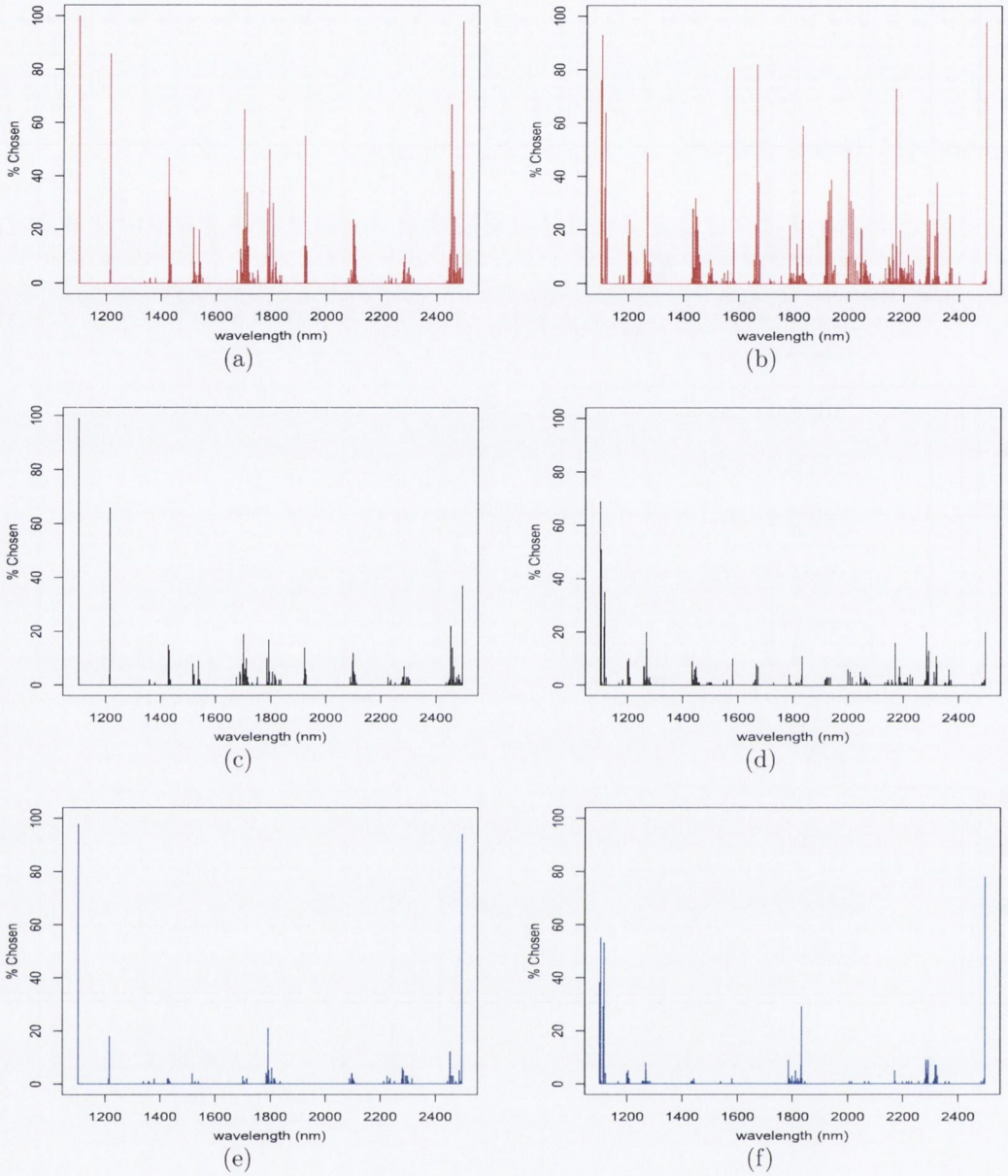


Fig. 4.8: Using a 25%/75% training test data split of the NIR honey data. Variables selected using B/W (Figure 4.8(a)) and $B + W$ (Figure 4.8(b)) as the prior dimension technique. The height of the red lines indicate the frequency that each of the wavelengths were selected using the initial dimension reduction technique. Then, using $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$ the height of the black lines in Figure 4.8(c) and in Figure 4.8(d) demonstrate the frequency that each wavelength is selected when B/W and $B + W$ respectively were used as a prior dimension reduction technique while the blue lines in Figures 4.8(e) and 4.8(f) do the same when using $f_o(x) = 1/V$ (using B/W and $B + W$ respectively).

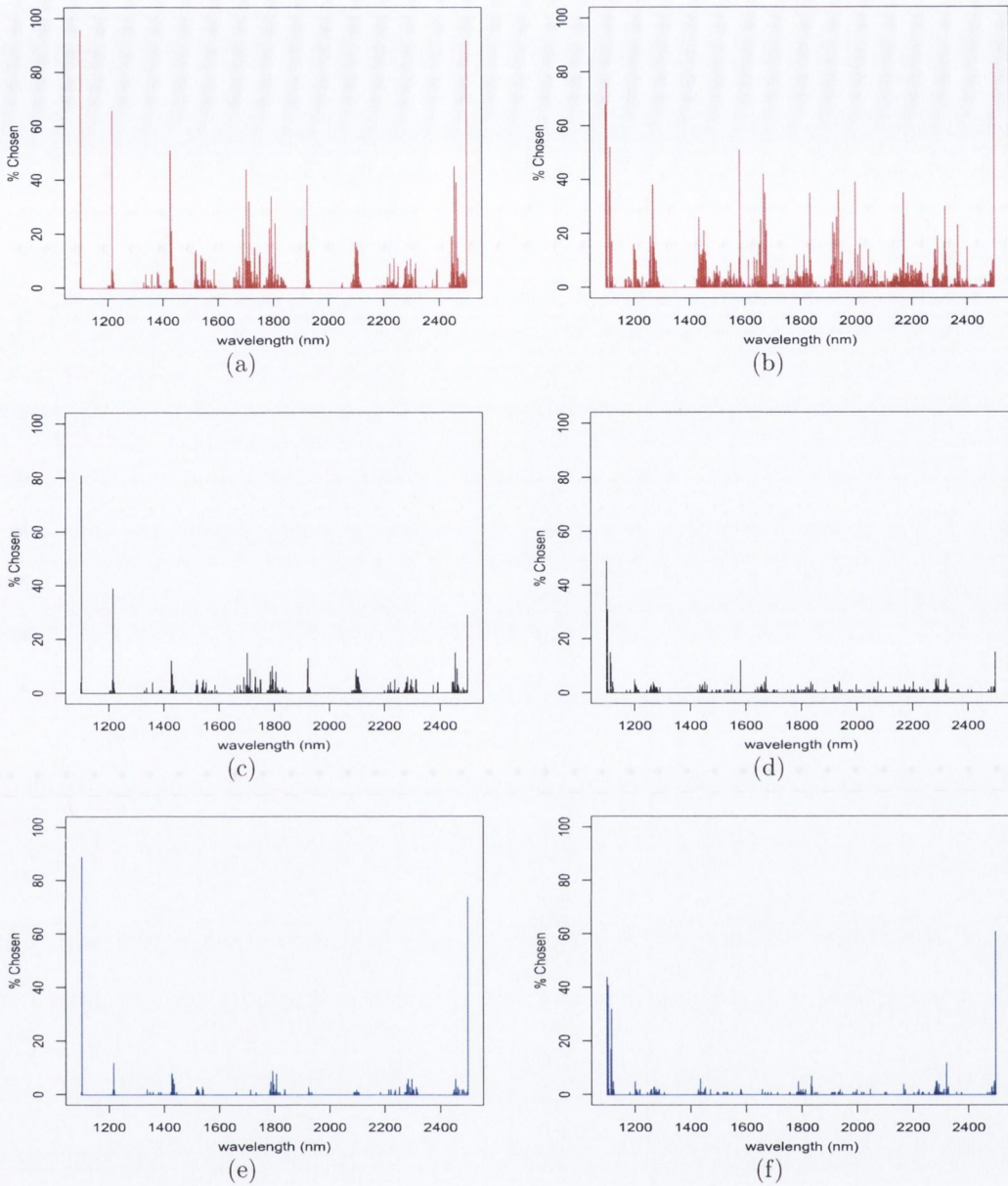


Fig. 4.9: Using a 10%/90% training test data split of the NIR honey data. Variables selected using B/W (Figure 4.9(a)) and $B + W$ (Figure 4.9(b)) as the prior dimension technique. The height of the red lines indicate the frequency that each of the wavelengths were selected using the initial dimension reduction technique. Then, using $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$ the height of the black lines in Figure 4.9(c) and in Figure 4.9(d) demonstrate the frequency that each wavelength is selected when B/W and $B + W$ respectively were used as a prior dimension reduction technique while the blue lines in Figures 4.8(e) and 4.8(f) do the same when using $f_o(x) = 1/V$ (using B/W and $B + W$ respectively).

Table 4.3: Classification Performance of Group of Interest Based Methods on NIR Olive Oil Data: $f_o(x) = 1/V$. The performance of the various methods of reducing the search space are compared, as is the performance of using updating methods against not using updating methods.

Dimension		$f_o(x) = \frac{1}{V}$			
Reduction		No Updating		Updating	
Split	Method	% Error	Brier	% Error	Brier
50%/	None	11.609 (3.779)	8.742 (2.538)	15.696 (5.378)	12.308 (3.601)
50%	10 th nm	12.551 (4.121)	9.675 (2.835)	16.957 (5.413)	12.874 (3.434)
	B/W	16.696 (4.029)	12.442 (2.948)	19.449 (5.220)	13.881 (3.374)
	B+W	13.536 (4.248)	10.335 (2.669)	16.420 (12.292)	12.292 (3.268)
	B/W,B+W	12.986 (4.311)	9.990 (2.837)	16.406 (5.339)	12.305 (3.609)
25%/	None	14.231 (5.201)	10.647 (3.446)	21.827 (5.908)	17.479 (5.308)
75%	10 th nm	15.913 (5.181)	11.663 (3.349)	21.327 (5.926)	16.929 (4.962)
	B/W	20.260 (5.467)	14.589 (3.374)	24.327 (8.352)	18.508 (6.646)
	B+W	15.413 (4.985)	11.500 (3.253)	21.760 (6.541)	16.590 (4.697)
	B/W,B+W	14.808 (4.732)	11.026 (3.270)	22.212 (5.510)	17.132 (4.407)

Table 4.4: Classification Performance of Group of Interest Based Methods on NIR

Olive Oil Data: $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$

Split	Method	$f_o(x) = \sum_{k=1}^K q_k f(x \mu_k, \Sigma_k)$	
		% Error	Brier
50%/50%	Wavelets	20.029 (12.056)	19.614 (12.151)
	B/W	20 (7.555)	19.623 (7.539)
	B+W	18.246 (4.084)	17.998 (3.976)
	1/V	3.391 (12.690)	3.361 (12.673)
25%/75%	Wavelets	64.173 (8.684)	63.659 (9.212)
	B/W	23.067 (6.485)	22.752 (6.492)
	B+W	21.962 (7.458)	21.466 (6.859)
	1/V	24.019 (23.627)	16.797 (15.901)
Including Additional Variable Selection			
50%/50%	B/W	4.536 (7.712)	3.740 (7.431)
	B+W	2.333 (7.159)	2.009 (6.994)
	1/V	1.884 (4.695)	1.632 (4.505)
25%/75%	B/W	23.567 (17.576)	22.239 (17.834)
	B+W	18.260 (20.432)	17.476 (20.550)
	1/V	15.212 (17.313)	12.952 (15.228)

Table 4.5: Comparison of Error Rates for using a direct calculation of a threshold, τ , versus using $\tau = (1 - p_g)/p_g V$, both when selecting which variables to include and as a method of calculating the threshold.

Data	Split	Variables Selected by τ		Variables Selected by $(1 - p_g)/p_g V$	
		τ direct	$\tau = (1 - p_g)/p_g V$	τ direct	$\tau = (1 - p_g)/p_g V$
Honey	50%/50%	16.854 (3.485)	27.690 (9.039)	14.803 (3.474)	21.305 (5.608)
	25%/75%	18.039 (3.809)	34.282 (10.151)	15.598 (3.382)	26.120 (6.235)
	10%/90%	20.826 (6.588)	38.209 (10.241)	17.981 (4.852)	28.907 (8.540)
Olive	50%/50%	5.507 (5.779)	24.435 (7.849)	9.058 (6.133)	12.551 (4.121)
Oils	25%/75%	13.202 (7.567)	24.846 (7.651)	12.183 (5.989)	15.913 (5.181)

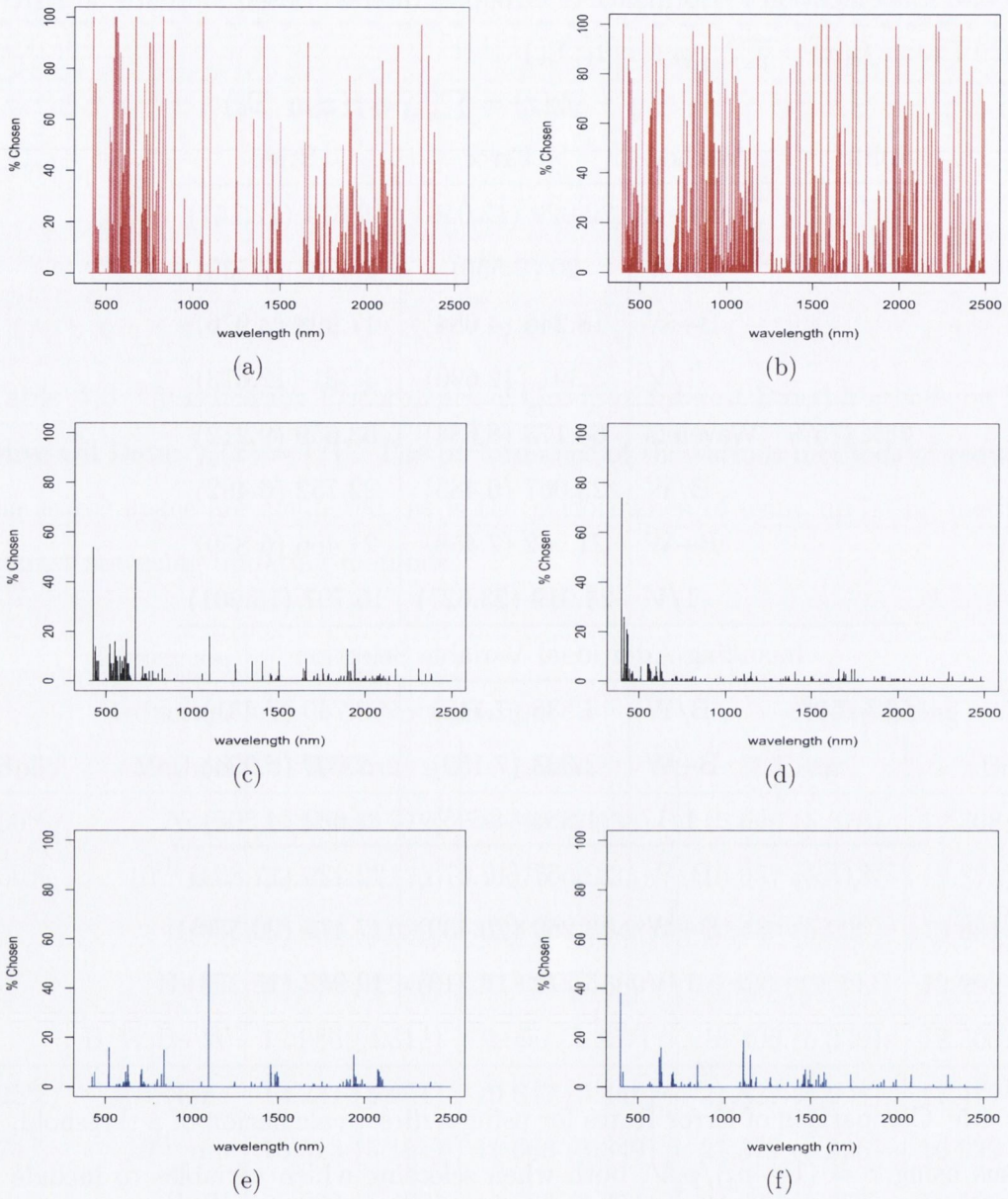


Fig. 4.10: Using a 50%/50% training test data split of the NIR olive oil data. Variables selected using B/W (Figure 4.10(a)) and $B + W$ (Figure 4.10(b)) as the prior dimension technique. The height of the red lines indicate the frequency that each of the wavelengths were selected using the initial dimension reduction technique. Then, using $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$ the height of the black lines in Figure 4.10(c) and in Figure 4.10(d) demonstrate the frequency that each wavelength is selected when B/W and $B + W$ respectively were used as a prior dimension reduction technique while the blue lines in Figures 4.10(e) and 4.10(f) do the same when using $f_o(x) = 1/V$ (using B/W and $B + W$ respectively).

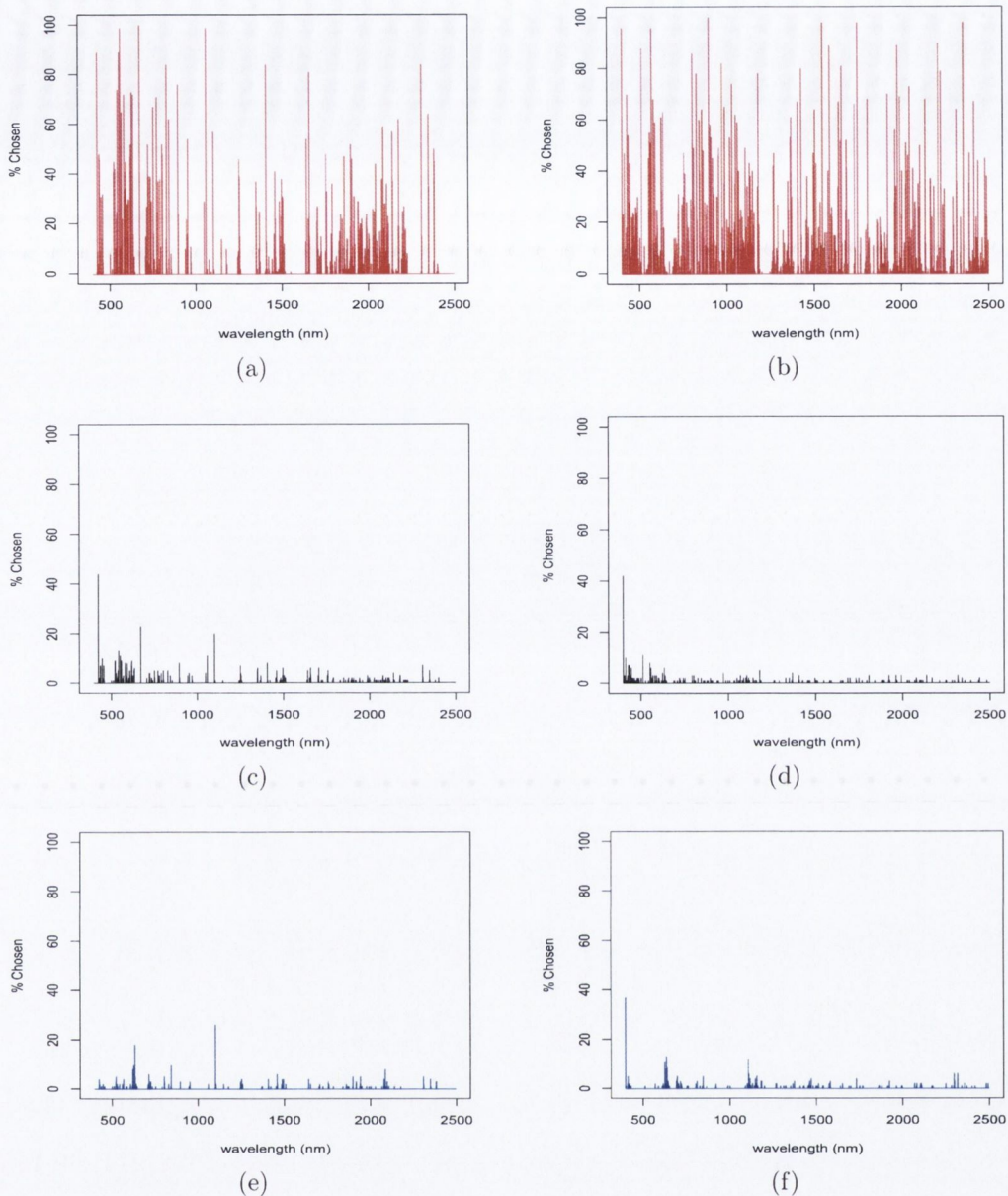


Fig. 4.11: Using a 25%/75% training test data split of the NIR olive oil data. Variables selected using B/W (Figure 4.11(a)) and $B + W$ (Figure 4.11(b)) as the prior dimension reduction technique. The height of the red lines indicate the frequency that each of the wavelengths were selected using the initial dimension reduction technique. Then, using $f_o(x) = \sum_{k=1}^K q_k f(x|\mu_k, \Sigma_k)$ the height of the black lines in Figure 4.11(c) and in Figure 4.11(d) demonstrate the frequency that each wavelength is selected when B/W and $B + W$ respectively were used as a prior dimension reduction technique while the blue lines in Figures 4.11(e) and 4.11(f) do the same when using $f_o(x) = 1/V$ (using B/W and $B + W$ respectively).

Table 4.6: Comparison of Error Rates for τ calculated directly, when the variables are selected by $(1 - p_g)/p_g V$, but the number of variables to include is determined by the training set performance of when calculating τ directly versus selected by $(1 - p_g)/p_g V$

Data	Split	Number of Variables To Include Selected by	
		τ directly calculated	$(1 - p_g)/p_g V$
Honey	50%/50%	14.573 (3.290)	14.803 (3.474)
	25%/75%	15.723 (3.154)	15.598 (3.382)
	10%/90%	18.133 (4.945)	17.981 (4.852)
Olive	50%/50%	3.449 (2.830)	9.058 (6.133)
Oils	25%/75%	9.125 (5.710)	12.183 (5.989)

By implementing a linear regression with an intercept at 0 and a slope b then:

$$\log \tau = b \log \left(\frac{1}{V} \right) = \log \left[\left(\frac{1}{V} \right)^b \right]$$

and therefore

$$\tau = \frac{1}{V^b}.$$

In practical terms this means that once an approximate value of b is determined using the first few dimensions, the grid search method of calculating τ directly can be more targeted. As the number of variables included grows, τ gets very large as is evident in Figure 4.6(a). Even by being able to estimate τ to an order of magnitude using a function of the volume before using the grid search technique would dramatically reduce the size of search space.

Modelling the alternative distribution $f_o(x)$ as Poisson noise is an imperfect solution, but one that has nice interpretation. b (as it is less than 1), represents an inflation factor – using the entire data (labelled and unlabelled) to calculate the volume V results in value for V that is too large, which in turn results in a value of $\frac{1}{V} = f_o(x)$ that is too small, especially as the number of additional variables included in the model is increased.

As the relationship is so strong, it makes sense to only calculate the “optimal τ ” for the first few dimensions, then use the relationship between that and $\frac{1}{V}$ to

Table 4.7: Analysis of the linear relationship of a directly calculated $\log(\tau)$ to $\log(1/V)$ when $(1 - p_g)/p_g V$ is used for variable selection and τ is directly calculated on these variables. This considers $\log(\tau) = b \log(1/V) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ and gives values for b and the corresponding values for the R^2 and adjusted R^2 . The results for the 25%/75% split of the olive oil data are heavily affected by one value of the directly calculated τ for one random split of the data. Omitting this random split of the data leads to the results in the row marked with the *.

Data	Split	b	R^2	Adjusted R^2
Honey	50%/50%	0.969 (0.008)	0.999 (0)	0.999 (0)
	25%/75%	0.969 (0.010)	0.999 (0)	0.999 (0)
	10%/90%	0.960 (0.029)	0.999 (0.001)	0.999 (0)
Olive	50%/50%	0.886 (0.019)	0.999 (0.001)	0.999 (0.001)
Oil	25%/75%	0.900 (0.103)	0.990 (0.095)	0.988 (0.106)
	25%/75%*	0.910 (0.023)	0.999 (0.001)	0.999 (0.001)

reduce the search space dramatically. This provides scope for dramatic reduction in computation time.

To quantify the relationship between $\log(1/V)$ and $\log(\tau)$ linear regression (forcing the intercept to be 0) was performed both when $(1 - p_g)/p_g V$ (Table 4.7) and a directly calculated τ (Table 4.8) were used as the variable selection criterion.

Using $1/V$ as a dimension reduction technique for Model-based Discriminant Analysis

Using $f_o(x) = 1/V$ as a preliminary method of reducing the dimensionality of the problem so that the dimension reduction techniques used for the model-based methods of Chapter 3 can be compared to using $f_o(x) = 1/V$ with variable selection. Table 4.9 gives the classification results based on the 1st 10 variables selected by $f_o(x) = 1/V$, irrespective of whether or not they were all to be included in the end model and also of only those variables included in the final model when $f_o(x) = 1/V$. Classification performance is similar (5 fold cross validation is used in Chapter 3 whereas 3 fold cross validation is used here), but the interpretation of the wavelengths selected is easier, as in general, fewer wavelengths are selected.

Table 4.8: Analysis of the linear relationship of a directly calculated $\log(\tau)$ to $\log(1/V)$ when a directly calculated τ is used for variable selection. This considers $\log(\tau) = b \log(1/V) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ and gives values for b and the corresponding values for the R^2 and adjusted R^2 .

Data	Split	b	R^2	Adjusted R^2
Honey	50%/50%	0.877 (0.119)	0.991 (0.014)	0.990 (0.015)
	25%/75%	0.816 (0.136)	0.990 (0.010)	0.989 (0.011)
	10%/90%	0.786 (0.140)	0.988 (0.014)	0.986 (0.015)
Olive	50%/50%	0.861 (0.026)	0.999 (0.001)	0.999 (0.001)
Oil	25%/75%	0.856 (0.063)	0.996 (0.009)	0.995 (0.010)

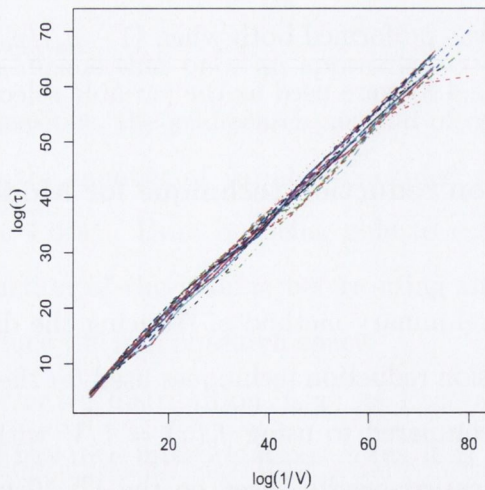


Fig. 4.12: The linear relationship between $\log(\tau)$ and $\log\left(\frac{1}{V}\right)$ is illustrated using the results of the first twenty random splits of the NIR olive oil data at 50%/50%.

Table 4.9: Comparing the classification performance of model-based discriminant analysis using the $f_o(x) = 1/V$ as the wavelength selection method on the NIR honey data, and the using both the error rate and the Brier's score, using 3-fold cross validation to determine the covariance structures.

No. of Variables		No Updating				EM				CEM			
Selected by		% Error		Brier		% Error		Brier		% Error		Brier	
Honey													
50%/50%	1 st 10 of 1/V	7.448	(1.486)	5.677	(1.075)	7.912	(1.319)	6.260	(1.164)	7.921	(1.383)	6.439	(1.209)
	1/V	9.305	(1.775)	7.046	(1.316)	9.812	(1.852)	7.629	(1.558)	9.653	(1.845)	7.756	(1.561)
25%/75%	1 st 10 of 1/V	8.101	(1.882)	6.421	(1.558)	8.816	(1.632)	7.585	(1.471)	8.879	(1.654)	7.642	(1.503)
	1/V	9.816	(1.621)	7.745	(1.512)	10.196	(1.831)	8.681	(1.662)	10.232	(1.897)	8.891	(1.697)
10%/90%	1 st 10 of 1/V	10.895	(3.035)	9.293	(2.822)	10.037	(2.471)	9.090	(2.266)	9.923	(2.396)	9.088	(2.223)
	1/V	11.498	(2.901)	9.382	(2.168)	11.840	(2.670)	10.233	(2.134)	11.774	(2.802)	10.406	(2.280)
Olive Oil													
50%/50%	1 st 10 of 1/V	0.058	(0.352)	0.051	(0.272)	0.043	(0.248)	0.028	(0.175)	0.029	(0.204)	0.029	(0.184)
	1/V	3.971	(3.971)	2.908	(2.957)	4.362	(4.496)	3.396	(3.462)	4.214	(4.418)	3.377	(3.565)
25%/75%	1 st 10 of 1/V	0.606	(1.780)	0.550	(1.620)	0.317	(1.238)	0.280	(1.609)	0.317	(1.282)	0.276	(1.114)
	1/V	6.625	(5.186)	5.283	(4.307)	7.577	(6.510)	6.174	(5.550)	7.462	(6.494)	6.225	(5.583)

4.5 Conclusions

Modelling a particular group using a single Gaussian distribution treats it as a homogeneous group. This follows the process of considering *pure* samples as being from some homogeneous population, whereas *adulterated* samples lie on the boundary of the pure population. This simplified viewpoint is complicated when adulteration occurs because of a concerted effort to perpetrate commercial fraud – as would be the situation in the honey adulteration scheme, especially with the adulteration by the fructose:glucose solutions that are developed to resemble constituent parts of a highly variable natural substance.

A single value for the threshold, while easy to interpret, does not provide good classification results. However, using $f_o(x) = 1/V$ as a method of searching through a high dimension space to find variables where the density of observations in the group of interest is more compact relative to other observations is useful. Trying to find an optimal value for this threshold dramatically improves classification performance as illustrated by Tables 4.5 and 4.6. It is interesting to note that using $f_o(x) = 1/V$ to first select the variables and then finding the optimal value for the threshold on these variables leads to better classification results than trying to find the set of variables using the optimal τ approach.

This is most likely due to increased influence of the cross validation process when finding the optimal τ over that when using $f_o(x) = 1/V$. The search strategy to find τ is interested in maximising classification performance on the cross validation sets, but the cross validation is also used at each stage to decide which variable to include in the model. As there is a strong relationship between this optimal value of the threshold and the volume of the data, the size of the search space in higher dimensions can be dramatically reduced by using the extra information obtained through this relationship. An alternative would be to use this relationship between the volume and the optimal τ to reduce the cross validation process to the selection of variables rather than the determination of the value of τ .

Modelling the adulterated olive oil samples with a mixture of Gaussian distributions without additional variable selection is shown by Table 4.4 to be inappropriate. This is due to the number of variables included by the initial variable selection techniques – the curves produced by B/W (*scatter curves*) and $B + W$ (*variance*

curves) are far quite jagged for the olive oil data, leading to more peaks that are then included as variables in the reduced set of variables. The effect of the additional variable selection on the honey data is not as obvious until the training data is reduced to 10% of the total data, as both the *scatter curves* and the *variance curves* are much smoother, leading to fewer local maxima and hence fewer variables included in the set from which the additional variable selection occurs.

The stepwise variable selection process used in this Chapter is by no means either the most efficient method or the method that produces the optimal set of variables for classification purposes. However, it is a compromise between efficiency and clarity – those using the method should be able to understand the logical process behind it. Any form of an *all subsets* approach is computationally infeasible, even when the number of variables are first reduced by alternative strategy. Raftery and Dean (2006) examine having inclusion and removal steps in a variable selection procedure in a clustering context, which has been extended by Murphy et al. (2008) to incorporate variable selection using a headlong search strategy for classification problems. While Tables 4.1 and 4.3 show that the dimension reduction techniques employed in this Chapter do not improve classification performance, they do dramatically reduce computation time.

Chapter 5

Updating Fisher's Linear Discriminant Analysis

5.1 Fisher's LDA

Fisher (1936) motivated his linear discriminant analysis technique using the iris dataset with the question: “*What linear function of the four measurements $X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$ will maximise the difference in the ratio of the difference between the specific means to the standard deviations within species?*”

To find the discriminant function between setosa and versicolor irises, Fisher first calculated a 4×4 matrix where the i^{th} entry is:

$$S_{ij} = \sum_{setosas} (x_i - \mu_{s,i})(x_j - \mu_{s,j}) + \sum_{versicolors} (x_i - \mu_{v,i})(x_j - \mu_{v,j})$$

where $\mu_{s,i}$ is the mean of the i^{th} variable for setosas and similarly $\mu_{v,i}$ is the mean of the i^{th} variable for versicolors.

The inverse of this matrix, when multiplied by the differences in the mean values for each species gives λ . These λ 's were then adjusted by Fisher so that $\lambda_1 = 1$.

In order to generalise this procedure to more than two groups Fisher's linear discriminant analysis assumes a common covariance matrix for all groups and then maximises the ratio of the variability between the groups relative to the common covariance within each group. If B measures the between group variance and W measures the within group variance, then Fisher's LDA finds the projection l that maximises the ratio in (5.1), subject to the constraint that $l' S_{pd} l = 1$, where $S_{pd} =$

$W/(N - G)$ is the pooled covariance matrix.

$$\frac{l'Bl}{l'Wl} \quad (5.1)$$

5.1.1 Theory

Assuming that there are data X (dimension $N \times p$) with G groups and group labels \mathbf{z} then:

$$\mu_g = \frac{1}{n_g} \sum_{j=1}^N z_{gj} x_j \quad (5.2)$$

where n_g is the number of observations in group g ($\sum_{j=1}^N z_{gj}$) and z_{gj} is an indicator variable for group membership for observation i belonging to group g and x_j is the j^{th} observation of X .

$$\mu = \frac{\sum_{j=1}^N x_j}{N}$$

Considering a linear combination $Y = l'X$, so that

$$\mu_{gY} \equiv \mathbb{E}[Y] = \mathbb{E}[l'X | \nu_g] = l' \mathbb{E}[X | \nu_g] = l' \mu_g \text{ for group } g$$

the transformed overall mean is then $\mu_Y \equiv l'\mu$ and

$$\text{Var}[Y] = \text{Var}[l'X] = l' \text{Cov}[X] l = l' \Sigma l$$

which is common for all groups, as each group is assumed to have an equal covariance matrix.

Thus

$$\begin{aligned} \frac{\text{Distance to overall mean of } Y}{\text{Variance of } Y} &= \frac{\sum_{g=1}^G (\mu_{gY} - \mu_Y)^2}{\text{Var}[Y]} \\ &= \frac{\sum_{g=1}^G (l' \mu_g - l' \mu)^2}{l' \Sigma l} \\ &= \frac{l' \left(\sum_{g=1}^G (\mu_g - \mu) (\mu_g - \mu)' \right) l}{l' \Sigma l} \end{aligned}$$

Now let

$$B = \sum_{g=1}^G (\mu_g - \mu) (\mu_g - \mu)'$$

and

$$W = \sum_{g=1}^G \sum_{j=1}^N z_{gj} (x_{gj} - \mu_g) (x_{gj} - \mu_g)' \quad (5.3)$$

Then to find the l 's that maximise (5.1):

- Find the eigenvectors e of $W^{-1}B$. There will be $s = \min(G - 1, p)$ nonzero eigenvalues.
- Scale the s non-zero eigenvectors so that $e'S_{pd}e = 1$ where S_{pd} is the pooled covariance matrix

$$S_{pd} = W/(N - G)$$

- These scaled eigenvectors are the linear discriminant functions

This algorithm assumes that the matrix W is invertible, which is often not the case, especially when p is large. The solution is to rotate W so that it takes the form of an identity matrix and then to use the same rotation on B . Details of how to find this rotation are given in Appendix A.1. Practically one uses the Singular Value Decomposition rather than finding the eigenvectors directly. Section 5.1.4 outlines how LDA is implemented in practice and then further develops the algorithm to outline how updating is incorporated into the algorithm.

5.1.2 Prediction

Given LDA coefficients and X_{new} is new data, then: For every group $g \in G$, let the Mahalanobis distance from Xa to the rotated group means $u_g a$ using the rotated within groups covariance matrix $a'\Sigma_w a$, be called M .

Then in order to convert this into probabilities, assuming that ν is a vector of the probabilities associated with each group find

$$\exp\left(-\frac{1}{2}M + \log(\nu)\right) \tag{5.4}$$

Then normalise the values in (5.4) so that for each observation they sum to 1 – turning them into probabilities of belonging to each group.

5.1.3 Updating

The idea of updating or semi-supervised methods is to use all of the data available in order to carry out classification. Therefore where complete information is available (including the labels) these are used, but those observations where the label is unobserved are also used. These extra observations are given partial group

membership in accordance with the probabilities of that observation belonging to each group. O'Neill (1978) examined using unclassified observations in estimating Fisher's Linear Discriminant Function in a two group problem (using the iris data as an example to discriminate between versicolor and setosa flowers).

There are several options of how to get the initial probabilities for the observations with unknown labels, including:

- Randomly assign into groups
- Use ν the total group probabilities
- First calculate the probabilities without the "extra" observations, then use these probabilities for initialising the updating procedure

The most consistent of the methods for the data sets examined in this thesis was using Fisher's LDA without the extra observations to assign initial probabilities and thus results given are for this method of initialization.

In order to combine Fisher's linear discriminant analysis with updating, simply change the z_{ig} 's of equations (5.2) and (5.3) so that for the extra observations where the labels are unknown, the z_{ig} 's are no longer indicator variables but rather are the probabilities of observation i belonging to group g .

5.1.4 Implementation of Semi-supervised Fisher's Linear Discriminant Analysis

LDA is implemented as follows:

Considering training data X ($N_1 \times p$), with G groups and group membership indicator variables z_{ig} . To find the rotation of the data:

- **Step 1** Find the group probabilities:

$$\nu = (\nu_1, \nu_2, \dots, \nu_G) \text{ where } \nu_g = \frac{\sum_{i=1}^{N_1} z_{ig}}{\sum_{g=1}^G \sum_{i=1}^{N_1} z_{ig}}.$$

- **Step 2** Find the group means: μ_g

$$\mu_g = (\mu_{1g}, \dots, \mu_{pg}) = \frac{\sum_{i=1}^{N_1} z_{ig} x_i}{\sum_{i=1}^{N_1} z_{ig}}.$$

- **Step 3** Find the overall variable means: $\bar{x} = (\bar{x}_1, \dots, \bar{x}_p)$
- **Step 4** Find the between groups variance:

$$\Sigma_b = \sum_{g=1}^G \nu_g (\mu_g - \bar{x})(\mu_g - \bar{x})'$$

- **Step 5** Find the within groups variance:

$$\Sigma_w = \sum_{g=1}^G \left[\frac{z_{ig}(x_i - \mu_g)(x_i - \mu_g)'}{\sum_{i=1}^{N_1} z_{ig}} \right]$$

- **Step 6** Find the singular value decomposition of $\Sigma_w = UDV'$ where U and V are orthogonal and D is a diagonal matrix with the singular values (d_{11}, d_{22}, \dots) on the diagonal.
- **Step 7** Let $r = \min(\sum_{g=1}^G \sum_{i=1}^{N_1} z_{ig} - G, p)$
- **Step 8** Denote $D_r^{-1/2}$ as the $r \times r$ diagonal matrix with entries $\left(\frac{1}{\sqrt{d_{ii}}}\right)$ for $i = 1, \dots, r$.
- **Step 9** Denote $V_{,r}$ as the first r columns of V .
- **Step 10** The rotation a is then $a = (D_r^{-1/2})(V_{,r})'$.
- **Step 11** Find $a\Sigma_b a' = (D_r^{-1/2})(V_{,r})'\Sigma_b(V_{,r})(D_r^{-1/2})$
- **Step 12** Find the first $\min(p, G - 1)$ eigenvectors of $a\Sigma_b a'$, denote these as e
- **Step 13** Rotate these eigenvectors again by $a = (D_r^{-1/2})(V_{,r})'$ to get the required rotation of the data:

$$A = (V_{,r})(D_r^{-1/2})e$$

Now include the new data (previously withheld as test data) – to have a dataset Y of dimension $(N \times p)$. The next 3 steps involve rotating the data, group means and the within group variance matrix so that they are in the new rotated space, in which the Mahalanobis distance will be calculated:

- **Step 14** Rotate the data by A : $YA = (Y)A$
- **Step 15** Rotate the group means by A : $\mu_g A$

- **Step 16** Rotate Σ_w by A : $A'\Sigma_w A$
- **Step 17** Find the Mahalanobis distance, M , on the rotated space (points YA , new means $\mu_g A$, new variance $A'\Sigma_w A$)
- **Step 18** Find $\exp(-\frac{1}{2}M + \log(\nu))$ and normalise to find the probabilities of group membership for each observation y_i .

Steps 1-13 indicate the procedure required to find the rotation of the data, Steps 14-18 the mechanism for finding the probabilities of group membership. Typical Fisher's Linear Discriminant Analysis stops at this point. However, in order to continue in the semi-supervised framework, revise the z_{ig} 's so that for the training data they remain indicator variables, but for the test data they are the predicted probabilities of observation i belonging to group g in Step 18 above. Now, including the new data (previously withheld as test data) – to have a dataset Y of dimension $(N \times p)$:

- **Step 0** $j = 0$
- **Step 1** Find the group probabilities:

$$\nu = (\nu_1, \nu_2, \dots, \nu_G) \text{ where } \nu_g = \frac{\sum_{i=1}^N z_{ig}}{\sum_{g=1}^G \sum_{i=1}^N z_{ig}}.$$

- **Step 2** Find the group means μ_g

$$\mu_g = (\mu_{1g}, \dots, \mu_{pg}) = \frac{\sum_{i=1}^N z_{ig} y_i}{\sum_{i=1}^N z_{ig}}.$$

- **Step 3** Find the overall variable means: $\bar{y} = (\bar{y}_1, \dots, \bar{y}_p)$
- **Step 4** Find the between groups variance:

$$\Sigma_b = \sum_{g=1}^G \nu_g (\mu_g - \bar{y})(\mu_g - \bar{y})'$$

- **Step 5** Find the within groups variance:

$$\Sigma_w = \sum_{g=1}^G \left[\frac{z_{ig} (y_i - \mu_g)(y_i - \mu_g)'}{\sum_{i=1}^N z_{ig}} \right]$$

- **Step 6** Find the singular value decomposition of $\Sigma_w = UDV'$ where U and V are orthogonal and D is a diagonal matrix with the singular values (d_{11}, d_{22}, \dots) on the diagonal.
- **Step 7** Let $r = \min(\sum_{g=1}^G \sum_{i=1}^N z_{ig} - G, p)$
- **Step 8** Denote $D_r^{-1/2}$ as the $r \times r$ diagonal matrix with entries $\left(\frac{1}{\sqrt{d_{ii}}}\right)$ for $i = 1, \dots, r$.
- **Step 9** Denote $V_{,r}$ as the first r columns of V .
- **Step 10** The rotation a is then $a = (D_r^{-1/2})(V_{,r})'$.
- **Step 11** Find $a\Sigma_b a' = (D_r^{-1/2})(V_{,r})'\Sigma_b(V_{,r})(D_r^{-1/2})$
- **Step 12** Find the first $\min(p, G - 1)$ eigenvectors of $a\Sigma_b a'$, denote these as e
- **Step 13** Rotate these eigenvectors again by $a = (D_r^{-1/2})(V_{,r})'$ to get the required rotation of the data:

$$A = (V_{,r})(D_r^{-1/2})e$$

- **Step 14** Rotate the data by A : $YA = (Y)A$
- **Step 15** Rotate the group means by A : $\mu_g A$
- **Step 16** Rotate Σ_w using by A : $A'\Sigma_w A$
- **Step 17** Find the Mahalanobis distance, M , on the rotated space (points YA , new means $\mu_g A$, new variance $A'\Sigma_w A$)
- **Step 18** Find $\exp\left(-\frac{1}{2}M + \log(\nu)\right)$ and normalise to find the probabilities of group membership for each observation y_i .
- **Step 19** $j = j + 1$
- **Step 20** Return to Step 1 until either there is no change in the values of z_{ig} 's, the change in the values of z_{ig} 's is sufficiently small (sufficiently small determined in advance) or until j equals some predetermined number.

5.2 Data

To illustrate why this updating is useful, the performance using 50% known (labelled) 50% unknown (unlabelled) splits, 25% known 75% unknown splits and 10% known 90% unknown splits will be shown on a number of data sets. For the example projections shown, circles are the training data (known labels), other symbols are the test data (unknown labels) and represent the group into which the observation was classified. The size of the symbols reflects the uncertainty of the prediction – the bigger the symbol in the figure, the greater the uncertainty that was associated with that prediction. Observations are coloured by actual group membership.

5.2.1 Fisher’s Iris Data

This data set comprises of 150 iris flowers with measurements of sepal length, sepal width, petal length and petal width, 50 from each of the following species: *setosa*, *versicolor*, and *virginica*. The goal is to be able to classify species using the sepal and petal measurements. In this data set, $p = 4$ and $G = 3$, so the number of discriminants is 2. This was the original data set used by Fisher in 1936 when he introduced his linear discriminant analysis technique.

Using an example 50% training 50% test random split of the data, Figures 5.1(a) and 5.1(b) illustrate the difference updating makes to the resultant projection of the data – making the projections more compact on the 2nd discriminant projection. Setosa are in black, versicolor are in red and virginica are in green.

As is evident in Table 5.1, using updating, even when 50% of the data is included in the training set, improves classification performance. As the percentage of observations included in the training set is decreased, the improvement becomes more pronounced.

Figure 5.2 illustrates how the differences in the projections created by LDA and updating come about. *Sepal Length* (Variable index 2) is largely unaffected by updating. For the other variables, the major source of difference is that updating balances the weighting more evenly between the variables. This effect is most evident when the original small training sample was not representative of the proportion of observations in each group.

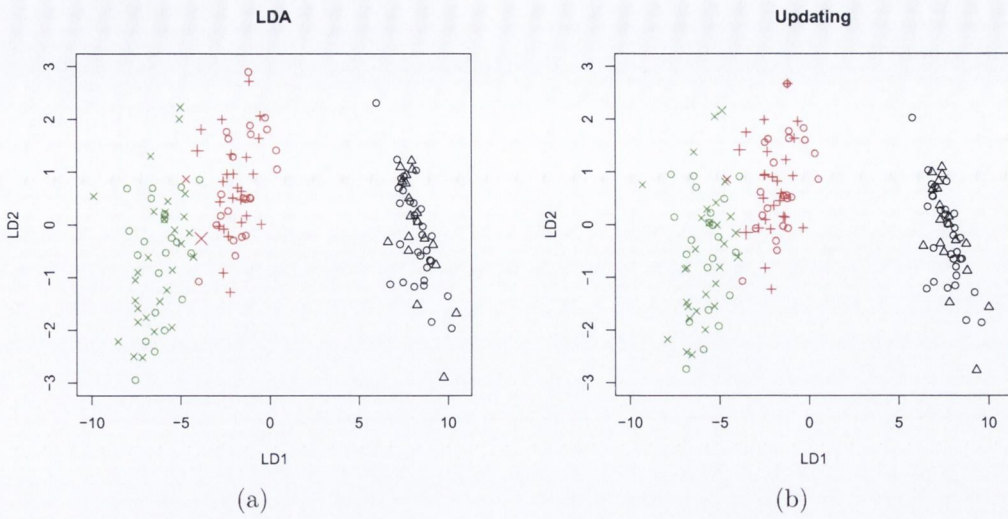


Fig. 5.1: Linear discriminant analysis projections of the famous iris dataset. Figure 5.1(a) is the projection for Fisher’s linear discriminant analysis; Figure 5.1(b) is the projection for the updating version, using a 50% training 50% test split of the data.

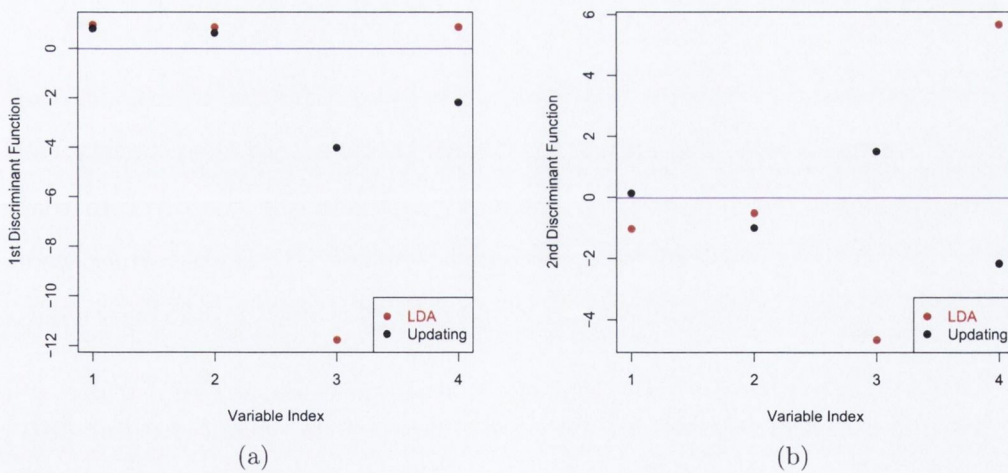


Fig. 5.2: The coefficients of the linear discriminant functions of the iris dataset. Figure 5.2(a) is the 1st discriminant function; Figure 5.2(b) is the 2nd discriminant function, using a 10% training 90% test split of the data.

Table 5.1: Comparing the classification performance of LDA and Updating at various training/test splits of the iris data.

Iris Data		LDA		Update	
50%/50%	% Error	2.333	(1.345)	2.133	(1.285)
	Brier	1.249	(0.610)	1.174	(0.576)
25%/75%	% Error	3.283	(1.670)	2.150	(0.952)
	Brier	1.691	(0.787)	1.145	(0.409)
10%/90%	% Error	5.681	(4.095)	2.681	(3.226)
	Brier	3.315	(2.620)	1.467	(2.039)

When a very small sample size is used by LDA the result can be a relatively poor estimation of the within group variation. As a result of the improved estimations of the group means and variances achieved by updating the new projections improve classification performance as shown in Table 5.1.

5.2.2 Wine Data

This dataset comprises of 13 variables and 178 observations. The variables are *Alcohol, Malic Acid, Ash, Alkalinity of Ash, Magnesium, Total Phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Colour Intensity, Hue, OD280/OD315 of diluted wines, and Proline*. The aim is to classify the wines to their variety (59 *Barolo*, 71 *Grignolino* and 48 *Barbera*). This is complicated by the problem that these wines did not all come from the same harvest (in fact the *Barolo* wines were from 1971-1974, the *Grignolino* wines were from 1970-1976 and the *Barbera* wines were from 1974-1979).

The extra information available in the variables *sugar-free extract, fixed acidity, tartaric acid, uronic acids, pH, potassium, calcium, phosphate, chloride, OD280/OD315 of flavanoids, glycerol, 2-3-butanediol, total nitrogen, methanol* do not add to the mean classification performance. However, the worst case performance in the 10%/90% split is significantly improved for the traditional LDA.

Again, using the updating technique improves classification performance for both the incomplete and full wine data sets. However, by including the extra variables,

Table 5.2: Comparing the classification performance of LDA and Updating at various training/test splits of the wine data.

Wine Data		LDA		Update	
50%/50%	% Error	2.112	(1.475)	1.191	(1.020)
	Brier	1.135	(0.722)	0.795	(0.652)
25%/75%	% Error	4.433	(2.242)	1.940	(1.387)
	Brier	2.650	(1.366)	1.273	(0.907)
10%/90%	% Error	24.710	(11.169)	3.354	(5.166)
	Brier	16.332	(7.471)	2.129	(3.162)

Table 5.3: Comparing the classification performance of LDA and Updating at various training/test splits of the full wine data.

Full Wine Data		LDA		Update	
50%/50%	% Error	2.562	(1.750)	1.157	(0.977)
	Brier	1.409	(1.040)	0.761	(0.642)
25%/75%	% Error	9.239	(3.957)	2.537	(2.210)
	Brier	5.878	(2.591)	1.672	(1.452)
10%/90%	% Error	14.880	(5.853)	4.565	(4.170)
	Brier	9.044	(3.711)	2.993	(2.685)

the improvement achieved by using updating over LDA is reduced. It is interesting to note that the reduced dataset actually provides better classification performance as the number of observations in the training set is reduced. This is mainly due to the reduction in the ability to accurately model the between and within group covariance matrices when the number of variables in the model is increased.

For both the wine datasets, the Brier's scores are much lower than the percentage error. This indicates that correct classifications are based on observations that have high probabilities associated with the correct group, whereas incorrect classifications are mainly due to observations that, while placed in the incorrect group, had relatively high probabilities of belonging to the correct group.

In projections for the incomplete wine dataset shown in Figure 5.3, where Barolo wines are in black, Grignolino are in red and Barbera are in green, again demonstrate while updating has little influence on the 1st discriminant projection, it causes the groups to be more compact in 2nd discriminant projection. Figure 5.4(b) confirms this – the values with updating for of each of the coefficients for the 2nd discriminant functions are closer to zero than the values of the coefficients without updating.

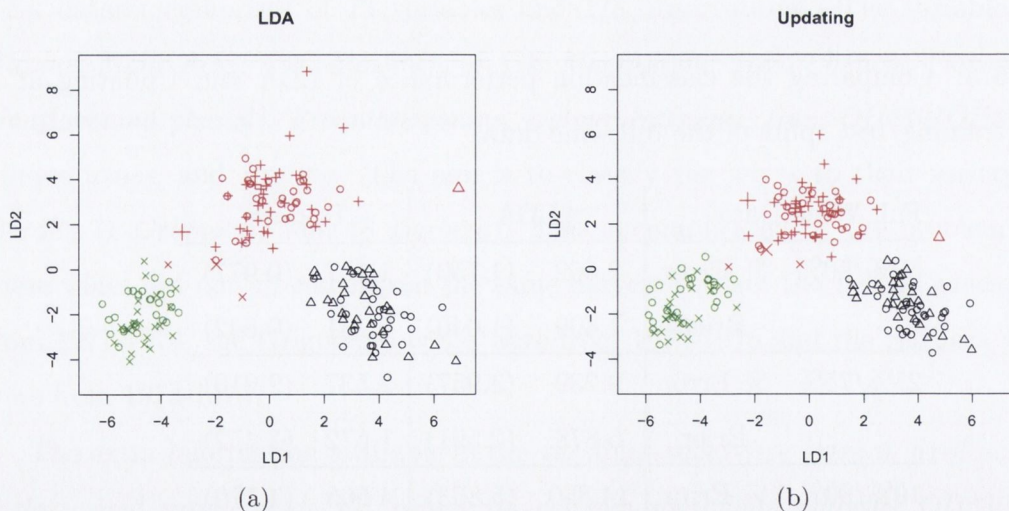


Fig. 5.3: Linear discriminant analysis projections of the wine dataset. Figure 5.3(a) is the projection for Fisher's linear discriminant analysis; Figure 5.3(b) is the projection for the updating version, using a 50% training 50% test split of the data.

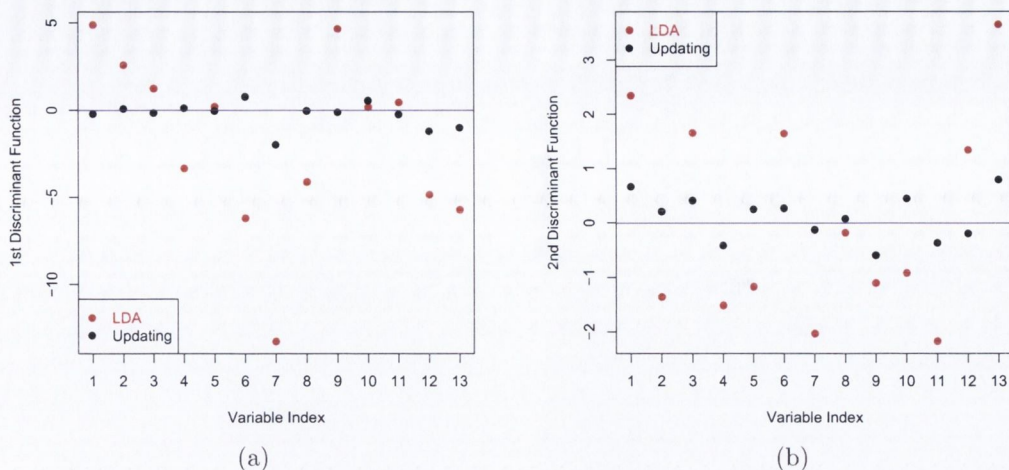


Fig. 5.4: The coefficients of the linear discriminant functions of the wine dataset. Figure 5.4(a) is the 1st discriminant function; Figure 5.4(b) is the 2nd discriminant function, using a 10% training 90% test split of the data.

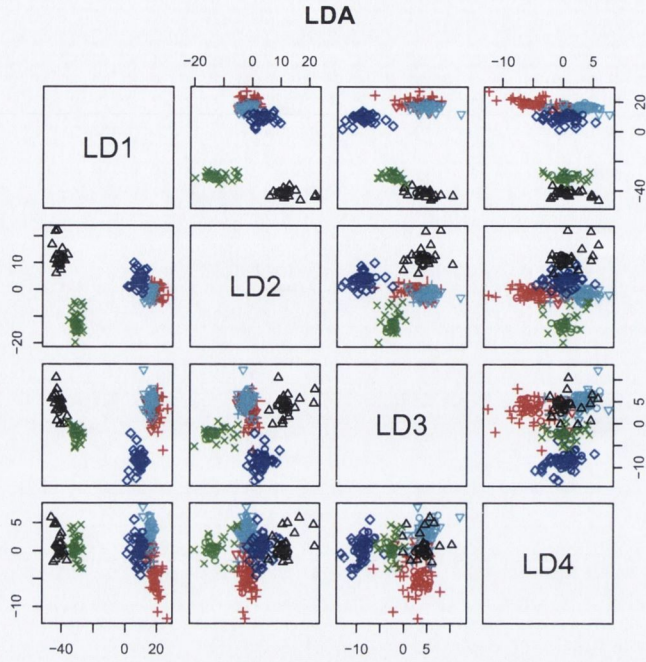
5.2.3 Meats Data

This dataset, containing 231 minced meat samples, was observed from 400-2498 nm (resulting in 1050 highly correlated variables). There were 32 Beef (*Black*), 55 Chicken (*Red*), 34 Lamb (*Green*), 55 Pork (*Blue*) and 55 Turkey (*Cyan*) samples, the spectra of which are illustrated in Figure 2.7.

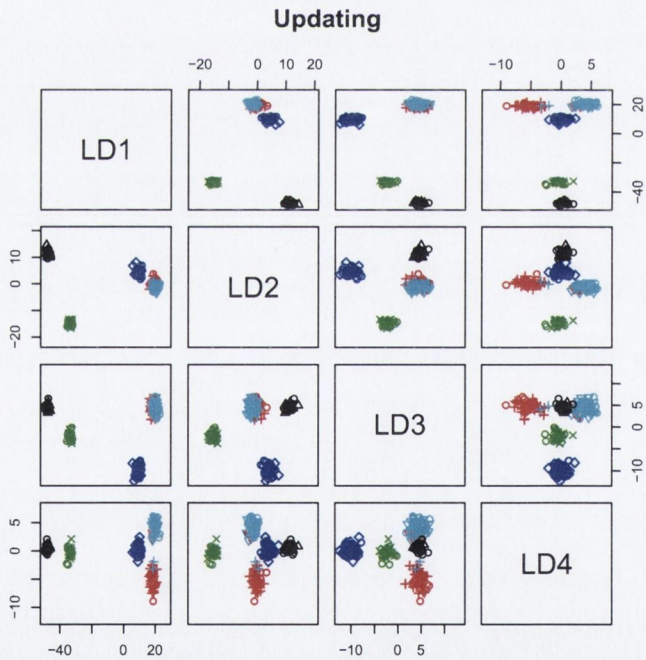
The performance of LDA with and without updating declines quite dramatically when the proportion of observations included in the training set is reduced to 10%. This 10% corresponds to a total of only 23 observations on which to build the original model. When this is split over the 5 groups, it is immediately apparent that, even when the training set is forced to contain at least one observation from each group, it is easily possible that the proportions of each meat type contained in the training set is unrepresentative of the entire data.

When comparing the projections in Figure 5.5 it is obvious that the main problem is distinguishing between *Chicken* (red) and *Turkey* (cyan) samples.

Examining the coefficients of the linear discriminant functions in Figure 5.6 it is apparent that while LDA focuses most of the weight on the visible part of the spectrum only, that updating places weight more equally between the visible part of



(a)



(b)

Fig. 5.5: Linear discriminant analysis projections of the NIR meat dataset. Figure 5.5(a) is the projection for Fisher's linear discriminant analysis; Figure 5.5(b) is the projection for the updating version, using a 50% training 50% test split of the data.

Table 5.4: Comparing the classification performance of LDA and Updating at various training/test splits of the meats data.

Meats Data		LDA		Update	
50%/50%	% Error	4.638	(2.005)	4.586	(1.956)
	Brier	1.758	(0.746)	1.834	(0.782)
25%/75%	% Error	7.609	(2.429)	7.506	(2.472)
	Brier	2.931	(0.951)	3.002	(0.989)
10%/90%	% Error	18.270	(6.016)	18.040	(6.061)
	Brier	7.028	(2.393)	7.216	(2.424)

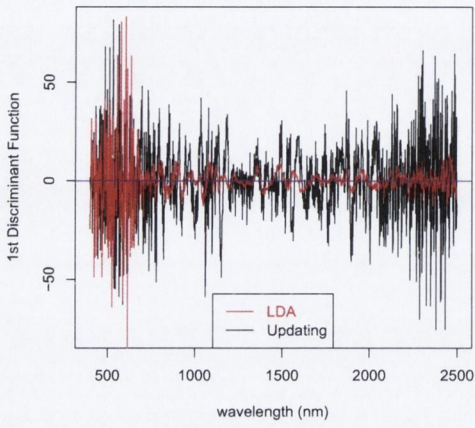
the spectrum and on the combinations region. For all four discriminant functions, the coefficients of updating indicate that much more of the spectrum is used for classification purposes.

5.2.4 Honey Data

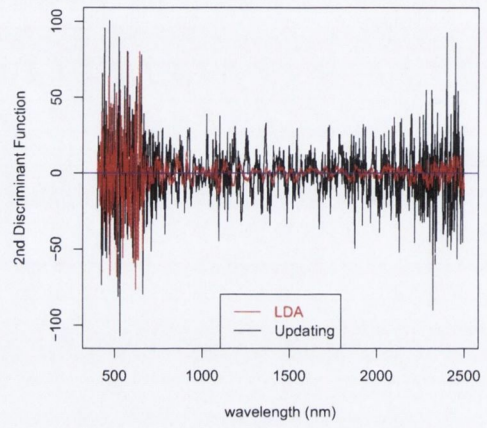
478 honey samples (157 pure, 321 adulterated) were measured over the NIR region 1100–2498 nm with the goal of separating pure from adulterated samples, with the type of adulteration not of interest. This is an example of a dataset that updating is not beneficial to classification performance.

Table 5.5: Comparing the classification performance of LDA and Updating at various training/test splits of the honey data.

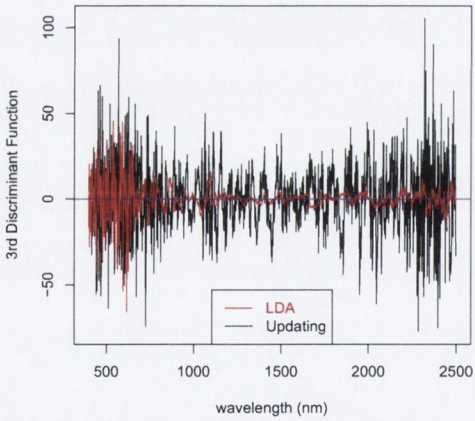
Honey Data		LDA		Update	
50%/50%	% Error	8.134	(1.568)	8.276	(1.571)
	Brier	7.756	(1.508)	8.276	(1.571)
25%/75%	% Error	6.089	(1.288)	6.251	(1.258)
	Brier	5.714	(1.272)	6.251	(1.258)
10%/90%	% Error	10.350	(2.801)	10.420	(2.774)
	Brier	9.781	(2.689)	10.420	(2.774)



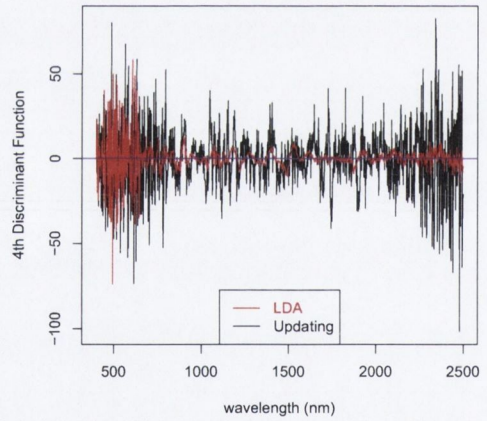
(a)



(b)



(c)



(d)

Fig. 5.6: The coefficients of the linear discriminant functions of the meats dataset. Figure 5.6(a) is the 1st discriminant function; Figure 5.6(b) is the 2nd discriminant function, Figure 5.6(c) is the 3rd discriminant function and Figure 5.6(d) is the 4th discriminant function using a 50% training 50% test split of the data.

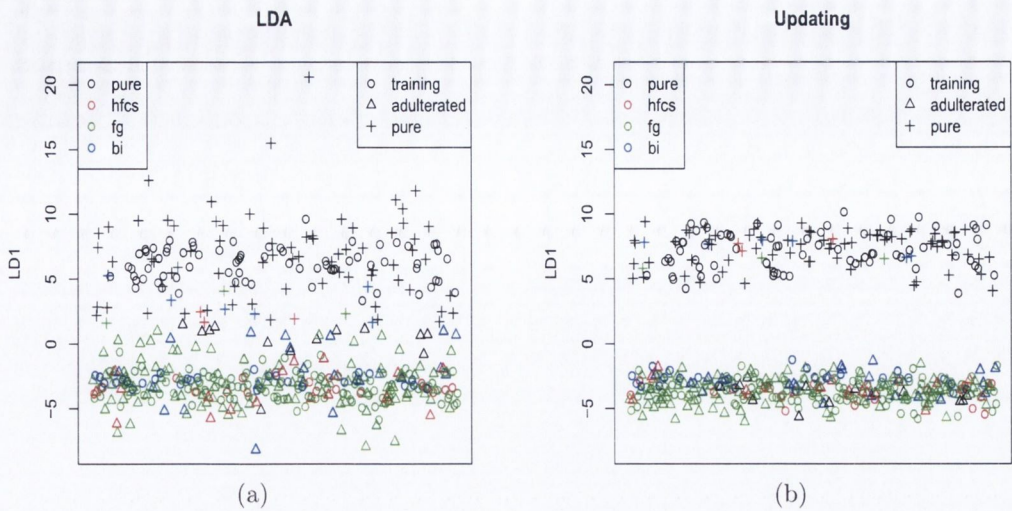


Fig. 5.7: Linear discriminant analysis projections of the NIR honey dataset. Figure 5.7(a) is the projection for Fisher’s linear discriminant analysis; Figure 5.7(b) is the projection for the updating version, using a 50% training 50% test split of the data. Observations have been placed in (the same) random order so that the spread of points can be more easily visualised.

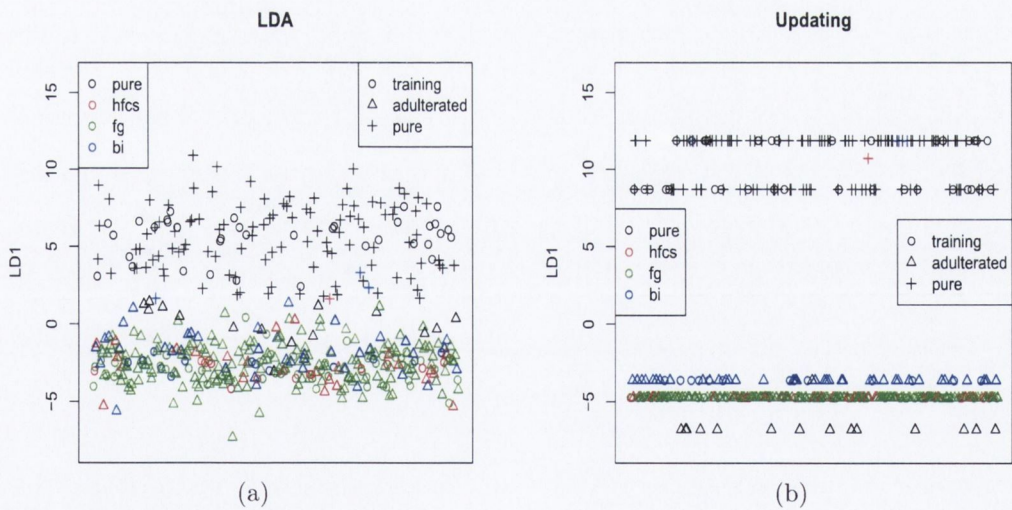


Fig. 5.8: Linear discriminant analysis projections of the NIR honey dataset. Figure 5.8(a) is the projection for Fisher’s linear discriminant analysis; Figure 5.8(b) is the projection for the updating version, using a 25% training 75% test split of the data. Again observations have been placed in (the same) random order so that the spread of points can be more easily visualised.

In this situation the type of adulterant are used for colours – blue is for beet invert syrups, red is for high fructose corn syrups, green is for fructose:glucose and black is for pure samples – but not for classification purposes. Again, the tightening effect of updating is noticeable in Figure 5.7, but in this case, the tightening is not beneficial for classification purposes – with some of the pure samples being much further away from most of the other pure samples than when LDA is applied without updating, leading to the big relative increase in the Brier’s score of incorrectly classified observations.

While not improving classification performance, comparing updating in Figure 5.8(b) with ordinary LDA in Figure 5.8(a) does provide insight into how the adulterated samples are broken into two rather than three distinct groups – with the samples adulterated with beet invert syrups separated from those adulterated by corn syrups and fructose:glucose solutions. This is of particular interest because the beet invert syrups and corn syrups were used in a different original study to the fructose:glucose solutions – so that these are not being grouped on the basis of the original pre-adulteration honey samples. This distinction is not evident in the 50%/50% split of the data shown in Figure 5.7, but becomes evident as the relative influence of updating increases. That samples adulterated with beet invert syrups are separated from those adulterated by corn syrups and fructose:glucose solutions rather than those adulterated with fructose:glucose solutions being separated from those adulterated by corn syrups and beet invert syrups is particularly surprising as the pure samples can be easily separated according to the original study in Figure 5.8(b).

The noticeable tightening of each of the groups in the projections illustrated by Figures 5.8(a) and 5.8(b) is also interesting as it demonstrates the main feature of using updating – each of the groups contract towards their respective means. Due to the scale of the separation between the groups in Figure 5.8(b) it appears that the projected data have identical values; the values are indeed extremely similar, but not identical. The effect is magnified when there is a two group classification problem as the projection is onto a single dimension.

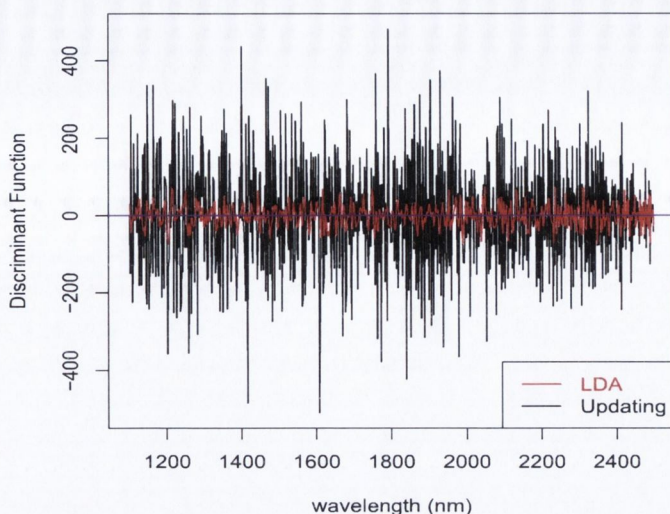


Fig. 5.9: The coefficients of the linear discriminant function of the honey dataset using a 50% training 50% test split of the data.

5.2.5 Olive Oil Data

Described in Section 2.2.3, this dataset comprises of a total of 138 observations – 46 pure olive oil and the remainder adulterated at either 1% or 5% with sunflower oil.

Classifying the NIR olive oil data into pure and adulterated samples in Table 5.6 provide interesting results. For the 50% and 25% training sets, updating does not alter the group probabilities – the Brier’s score and percentage error are identical with and without updating. There is enough evidence available when 25% of the data (34 observations) are included in the training set, however, when this is reduced to 10% of the data (14 observations), the classification performance dramatically declines.

In Figure 5.11 black circles are pure samples and red are adulterated samples. Circles are observations that were in the training data, triangles represent test set observations classified as pure, whereas crosses represent observations classified as adulterated. Only one observation is misclassified – put into the pure group, when it actually belongs in the adulterated group. While misclassified by both methods, as updating increases the separation between the groups, and reduces the separation within each group, the contribution to the Brier’s score for updating is greater.

Table 5.6: Comparing the classification performance of LDA and Updating at various training/test splits of the olive oil data.

Olive Oil Data		LDA		Update	
50%/50%	% Error	0.188	(0.490)	0.188	(0.490)
	Brier	0.188	(0.488)	0.188	(0.490)
25%/75%	% Error	0.019	(0.135)	0.019	(0.135)
	Brier	0.019	(0.135)	0.019	(0.135)
10%/90%	% Error	9.392	(6.139)	9.232	(5.960)
	Brier	8.849	(5.892)	9.231	(5.959)

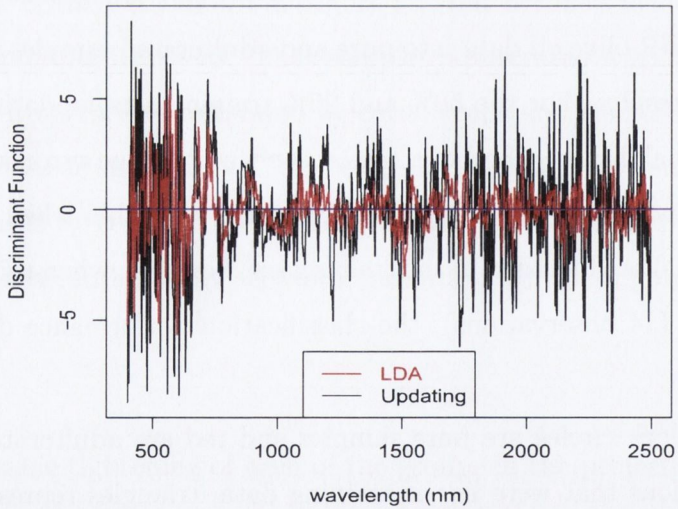


Fig. 5.10: The coefficients of the linear discriminant function of the olive oil dataset using a 50% training 50% test split of the data.

The olive oils were extended using 1% and 5% of sunflower oil. In Table 5.7 observations are classified according to the level of adulteration *i.e.* 0%, 1% and 5% sunflower oil in the olive oil.

Table 5.7: Comparing the classification performance of LDA and Updating at various training/test splits of the olive oil data, using the levels of adulteration.

Olive Oil Data		LDA		Update	
50%/50%	% Error	2.696	(2.039)	2.681	(2.043)
	Brier	1.704	(1.288)	1.787	(1.362)
25%/75%	% Error	5.000	(2.479)	4.913	(2.463)
	Brier	3.093	(1.568)	3.275	(1.642)
10%/90%	% Error	27.360	(7.455)	26.97	(7.430)
	Brier	17.052	(4.933)	17.973	(4.951)

5.3 Conclusions

Updating is a conceptually simple addition to Fisher's linear discriminant analysis. It can significantly improve classification performance, especially when the training set is small relative to the total data. It is especially useful in situations with multiple groups, where no single group can be identified as of particular importance relative to other groups.

However, updating methods should not be used blindly. When the assumptions of equal covariance matrices across groups fails and/or there is poor separation between groups (honey example), updating methods can hinder classification, even when the training set is small. For the examples in this Chapter the least extreme of the two discriminant functions provides the more reliable classifying tool when measured in terms of the Brier's score. Table 5.1 for the iris data and Tables 5.2 and 5.3 for the wine data are situations where the methods achieving the lowest percentage error also achieves the lowest Brier's score in all cases. In such situations the shrinking of the group boundaries do not result in many changes in classification and most observations are consistently pulled towards the correct group in the

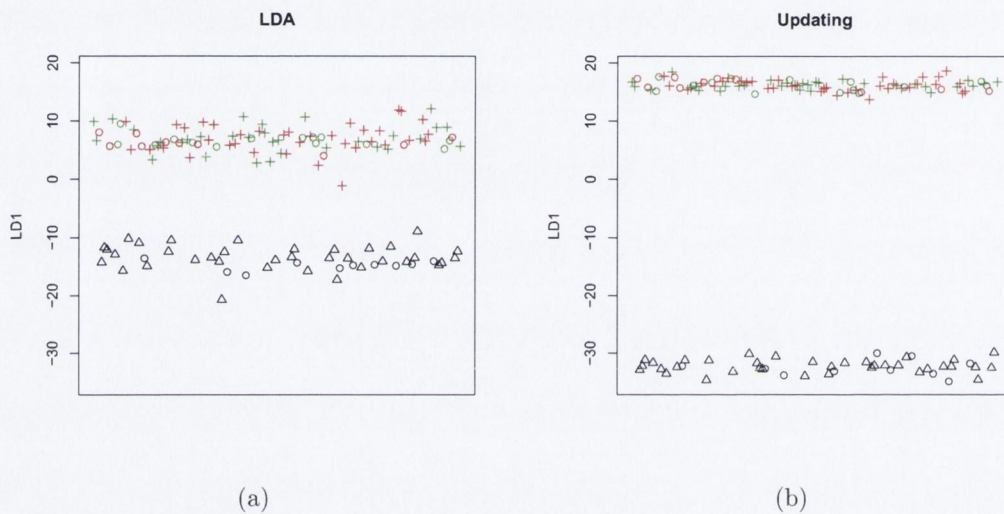


Fig. 5.11: Linear discriminant analysis projections of the NIR olive oil dataset. Figure 5.11(a) is the projection for Fisher’s linear discriminant analysis; Figure 5.11(b) is the projection for the updating version, using a 25% training 75% test split of the data. Both figures are coloured according to level of adulteration – black for 0% adulteration (pure), red for 1% adulteration and green for 5% adulteration with sunflower oil.

updating process. Figure 5.2 illustrates how the updating version yields discriminant functions less extreme than traditional LDA for the iris data while Figure 5.4 demonstrates the same for the wine data.

The improvement in classification performance achieved by using updating with a dataset as widely used and understood as the Iris data indicates that using semi-supervised linear discriminant analysis is a promising technique with possibilities for use in general applications.

The decision on whether or not to use a semi-supervised approach depends largely on the motivation behind the analysis. If the purpose of the analysis is for visualisation, then using updating is of more benefit - as the separation between groups is more evident and easier to visualise. This “tightening” of the groups increases the Brier’s score for any observation incorrectly classified. If uncertain about the model assumptions, plotting the projections resulting from LDA and comparing the relative size of each group can give a good indication about the appropriateness of using updating – if the groups in the resulting projected space are very different in size, then using updating is unlikely to improve classification performance.

For the NIR datasets, as the number of variables is much greater, the effect of each individual variable on the overall discriminant score is reduced. The four discriminant functions illustrated in Figure 5.6 associated with updating for the meats data are more extreme in the NIR part of the spectrum, but both methods place most of the weighting on wavelengths contained in the visible part of the spectrum. Performance-wise, it is extremely difficult to distinguish the two methods, Table 5.4 shows how updating reduces the percentage error, but traditional LDA has a lower Brier’s score. This is due to updating shrinking the size of each group, increasing the distances between the centre of each group. As the group boundaries become more clearly defined, any incorrectly classified observations are further away from their true group than before the updating process occurs.

The single discriminant function for the NIR honey data, both LDA and updating in Figure 5.9 show a very definite pattern. Both demonstrate that most of the available wavelengths are equally important, but the difference in the scale of the weighting assigned to each wavelength is noticeable, with updating being far more extreme throughout the NIR spectrum. Despite the scale of the difference in the two

discriminant function, the results shown in Table 5.5 show very similar classification performance, where LDA marginally outperforms updating. Even here, the updating process is not without its benefits as it provides much clearer group boundaries.

The refining effect of updating on the group boundaries is especially evident with the NIR datasets. Figure 5.5 illustrates some of the problems that may be associated with this refinement, namely that groups with ill-defined separation may be dragged closer together on some, if not all, of the projections, which may make visualisation of the separation difficult with many groups (and hence many discriminant projections). The two group problems posed by the NIR honey and olive oil datasets (pure versus adulterated) pose different visualisation problems – the single discriminant projection (Figures 5.7 and 5.11) can result in all the projected data points overlapping, making finding incorrectly classified observations more difficult to identify visually.

Chapter 6

Generalising Fisher's Linear Discriminant Analysis

6.1 Further Generalisation

Section 5.1 makes an important assumption by assuming that all the groups have the same covariance matrices. As noted by Zhu (2003) finding the l 's that maximise equation (5.1) is equivalent to maximising the likelihood ratio:

$$\begin{aligned}
 \arg \max_{\alpha} LR(\alpha) &= \arg \max_{\alpha} \log \left\{ \frac{\prod_{g=1}^G \prod_{j=1}^N \left[\frac{1}{\sqrt{2\pi\alpha'\Sigma_g\alpha}} \exp \left(-\frac{1}{2} \frac{\alpha'(x_j - \mu_g)(x_j - \mu_g)'\alpha}{\alpha'\Sigma_g\alpha} \right) \right]^{z_{jg}}}{\prod_{g=1}^G \prod_{j=1}^N \left[\frac{1}{\sqrt{2\pi\alpha'\Sigma_o\alpha}} \exp \left(-\frac{1}{2} \frac{\alpha'(x_j - \mu)(x_j - \mu)'\alpha}{\alpha'\Sigma_o\alpha} \right) \right]^{z_{jg}}} \right\} \\
 &= \arg \max_{\alpha} \left\{ \sum_{g=1}^G \sum_{j=1}^N z_{jg} \left[-\frac{1}{2} \log \alpha'\Sigma_g\alpha - \frac{1}{2} \frac{\alpha'(x_j - \mu_g)(x_j - \mu_g)'\alpha}{\alpha'\Sigma_g\alpha} \right. \right. \\
 &\quad \left. \left. - \frac{1}{2} \log 2\pi + \frac{1}{2} \log 2\pi + \frac{1}{2} \log \alpha'\Sigma_o\alpha + \frac{1}{2} \frac{\alpha'(x_j - \mu)(x_j - \mu)'\alpha}{\alpha'\Sigma_o\alpha} \right] \right\} \\
 &= \arg \max_{\alpha} \left\{ \sum_{g=1}^G \sum_{j=1}^N z_{jg} \left[-\frac{1}{2} \log \alpha'\Sigma_g\alpha - \frac{1}{2} \frac{\alpha'(x_j - \mu_g)(x_j - \mu_g)'\alpha}{\alpha'\Sigma_g\alpha} \right. \right. \\
 &\quad \left. \left. + \frac{1}{2} \log \alpha'\Sigma_o\alpha + \frac{1}{2} \frac{\alpha'(x_j - \mu)(x_j - \mu)'\alpha}{\alpha'\Sigma_o\alpha} \right] \right\} \tag{6.1}
 \end{aligned}$$

Considering the eigen decomposition of the covariance matrix (3.3), Zhu (2006) extends LDA beyond $\Sigma_g = \Sigma = \lambda DAD'$ to the common principal components model $\Sigma_g = \lambda_g DA_g D'$, of which $\Sigma_g = \lambda_g DAD'$ is special case.

6.1.1 $\Sigma_g = \Sigma = \lambda DAD' \forall g$

When the shape of the overall covariance matrix is unrestricted

$$\hat{\Sigma}_o = \frac{\sum_{j=1}^N (x_j - \mu)(x_j - \mu)'}{N} = \frac{\sum_{g=1}^G \sum_{j=1}^N z_{jg} (x_j - \mu)(x_j - \mu)'}{N}$$

so that

$$\sum_{g=1}^G \sum_{j=1}^N z_{jg} \frac{1}{2} \frac{\alpha'(x_j - \mu)(x_j - \mu)'\alpha}{\alpha'\Sigma_o\alpha} = \frac{N}{2} \frac{\alpha'\hat{\Sigma}_o\alpha}{\alpha'\hat{\Sigma}_o\alpha} = \frac{N}{2}.$$

This is not dependent on α , therefore can be ignored in the maximisation problem.

The maximisation problem can thus be reduced to:

$$\arg \max_{\alpha} LR(\alpha) = \arg \max_{\alpha} \left\{ \sum_{g=1}^G \sum_{j=1}^N \frac{z_{jg}}{2} \left[\log \left(\frac{\alpha'\Sigma_o\alpha}{\alpha'\Sigma_g\alpha} \right) - \frac{\alpha'(x_j - \mu_g)(x_j - \mu_g)'\alpha}{\alpha'\Sigma_g\alpha} \right] \right\} \quad (6.2)$$

For the case of LDA, when $\Sigma_g = \Sigma \forall g$, which is equivalent to $\Sigma_g = \lambda DAD'$, (6.2) can be further reduced using the identity:

$$\hat{\Sigma} = \sum_{g=1}^G \sum_{j=1}^N z_{jg} \frac{(x_j - \mu_g)(x_j - \mu_g)'}{N}$$

so that:

$$\sum_{g=1}^G \sum_{j=1}^N z_{jg} \frac{1}{2} \frac{\alpha'(x_j - \mu_g)(x_j - \mu_g)'\alpha}{\alpha'\Sigma_g\alpha} = \frac{N}{2} \frac{\alpha'\hat{\Sigma}_g\alpha}{\alpha'\hat{\Sigma}_g\alpha} = \frac{N}{2}.$$

Therefore, when $\Sigma_g = \Sigma \forall g$ (6.2) becomes:

$$\begin{aligned} \arg \max_{\alpha} LR(\alpha) &\equiv \arg \max_{\alpha} \left\{ \sum_{g=1}^G \sum_{j=1}^N \frac{z_{jg}}{2} \left[\log \left(\frac{\alpha'\hat{\Sigma}_o\alpha}{\alpha'\hat{\Sigma}\alpha} \right) \right] \right\} \\ &= \arg \max_{\alpha} \frac{N}{2} \log \left(\frac{\alpha'\hat{\Sigma}_o\alpha}{\alpha'\hat{\Sigma}\alpha} \right) \\ &= \arg \max_{\alpha} \log \left(\frac{\alpha'\hat{\Sigma}_o\alpha}{\alpha'\hat{\Sigma}\alpha} \right) \\ &= \arg \max_{\alpha} \left(\frac{\alpha'\hat{\Sigma}_o\alpha}{\alpha'\hat{\Sigma}\alpha} \right) \end{aligned}$$

which is equivalent to the maximisation problem posed in (5.1).

6.1.2 $\Sigma_g = \lambda I$

Assuming Σ_g to be a constant times the identity matrix for all groups means that the likelihood ratio of (6.1) is comparing $N(\mu_g, \lambda I)$ to $N(\mu, \gamma I)$. If $\hat{\Sigma}_o = \gamma I$, then

$\alpha' \hat{\Sigma}_o \alpha = \alpha' \gamma I \alpha = \gamma I \alpha' \alpha = \gamma$. Similarly, as $\hat{\Sigma}_g = \lambda I$, then $\alpha' \hat{\Sigma}_g \alpha = \lambda$. Thus the log term in (6.1) is not dependent on α . The maximisation problem posed by (6.1) then becomes:

$$\arg \max_{\alpha} \sum_{g=1}^G \sum_{j=1}^N \left\{ \frac{z_{jg}}{2} \left[\frac{\alpha'(x_j - \mu)(x_j - \mu)' \alpha}{\gamma} - \frac{\alpha'(x_j - \mu_g)(x_j - \mu_g)' \alpha}{\lambda} \right] \right\} \quad (6.3)$$

If $W = \sum_{g=1}^G \sum_{j=1}^N z_{jg} (x_j - \mu_g)(x_j - \mu_g)'$ and $T = \sum_{g=1}^G \sum_{j=1}^N z_{jg} (x_j - \mu)(x_j - \mu)'$, then (6.3) becomes:

$$\arg \max_{\alpha} \left\{ \frac{\alpha' T \alpha}{2\gamma} - \frac{\alpha' W \alpha}{2\lambda} \right\}.$$

Including the restraint that $\alpha' \alpha = 1$, to find α :

$$\begin{aligned} \frac{\partial}{\partial \alpha} \left[\frac{\alpha' T \alpha}{2\gamma} - \frac{\alpha' W \alpha}{2\lambda} + \phi(1 - \alpha' \alpha) \right] &= 0 \\ \frac{T}{\gamma} \alpha - \frac{W}{\lambda} \alpha - 2\phi \alpha &= 0 \\ \left(\left(\frac{T}{\gamma} - \frac{W}{\lambda} \right) - 2\phi I \right) \alpha &= 0 \end{aligned}$$

Since $\hat{\lambda} = \frac{\text{trace}(W)}{Np}$ (Bensmail and Celeux, 1996) and $\hat{\gamma} = \frac{\text{trace}(T)}{Np}$,

$$\begin{aligned} \left(\left(\frac{T}{\gamma} - \frac{W}{\lambda} \right) - 2\phi I \right) \alpha &= \left\{ Np \left[\frac{T}{\text{trace}(T)} - \frac{W}{\text{trace}(W)} \right] - 2\phi I \right\} \alpha \\ &= \left\{ \left[\frac{T}{\text{trace}(T)} - \frac{W}{\text{trace}(W)} \right] - 2\psi I \right\} \alpha \end{aligned} \quad (6.4)$$

where $\psi = \phi/Np$. Since $\alpha \neq 0$, this becomes $\left\{ \left[\frac{T}{\text{trace}(T)} - \frac{W}{\text{trace}(W)} \right] - 2\psi I \right\} \alpha = 0$. Find the eigenvectors of $\left[\frac{T}{\text{trace}(T)} - \frac{W}{\text{trace}(W)} \right]$, scale the eigenvalues appropriately in order to find the α 's.

6.1.3 $\Sigma_g = \lambda_g I$

Assuming Σ_g to be a (different) constant times the identity matrix for each of the groups means that the likelihood ratio of (6.1) is comparing $N(\mu_g, \lambda_g I)$ to $N(\mu, \gamma I)$. If $\hat{\Sigma}_o = \gamma I$, then $\alpha' \hat{\Sigma}_o \alpha = \alpha' \gamma I \alpha = \gamma I \alpha' \alpha = \gamma$. Similarly, as $\hat{\Sigma}_g = \lambda_g I$, then $\alpha' \hat{\Sigma}_g \alpha = \lambda_g$. Thus the log term in (6.1) is not dependent on α as it becomes

$$\frac{N}{2} \log \alpha' \alpha + \frac{N}{2} \log \gamma - \sum_{g=1}^G \frac{n_g}{2} \log \alpha' \alpha - \sum_{g=1}^G \frac{n_g}{2} \log \lambda_g$$

which is equivalent to

$$\frac{N}{2} \log \gamma - \sum_{g=1}^G \frac{n_g}{2} \log \lambda_g.$$

The maximisation problem posed by (6.1) then becomes:

$$\arg \max_{\alpha} \sum_{g=1}^G \sum_{j=1}^N \left\{ \frac{z_{jg}}{2} \left[\frac{\alpha'(x_j - \mu)(x_j - \mu)'\alpha}{\gamma} - \frac{\alpha'(x_j - \mu_g)(x_j - \mu_g)'\alpha}{\lambda_g} \right] \right\} \quad (6.5)$$

If $W_g = \sum_{j=1}^N z_{jg}(x_j - \mu_g)(x_j - \mu_g)'$ and $T = \sum_{g=1}^G \sum_{j=1}^N z_{jg}(x_j - \mu)(x_j - \mu)'$, then (6.5) becomes:

$$\arg \max_{\alpha} \left\{ \frac{\alpha'T\alpha}{2\gamma} - \sum_{g=1}^G \frac{\alpha'W_g\alpha}{2\lambda_g} \right\}.$$

Including the restraint that $\alpha'\alpha = 1$, to find α :

$$\begin{aligned} \frac{\partial}{\partial \alpha} \left[\frac{\alpha'T\alpha}{2\gamma} - \sum_{g=1}^G \frac{\alpha'W_g\alpha}{2\lambda_g} + \phi(1 - \alpha'\alpha) \right] &= 0 \\ \frac{T}{\gamma}\alpha - \sum_{g=1}^G \frac{W_g}{\lambda_g}\alpha - 2\phi\alpha &= 0 \\ \left(\left(\frac{T}{\gamma} - \sum_{g=1}^G \frac{W_g}{\lambda_g} \right) - 2\phi I \right) \alpha &= 0 \end{aligned}$$

Since $\alpha \neq 0$, $\hat{\lambda}_g = \frac{\text{trace}(W_g)}{n_g p}$ and $\hat{\gamma} = \frac{\text{trace}T}{Np}$, similarly to (6.4) this becomes a matter of finding the eigenvectors of $\left[\frac{NT}{\text{trace}(T)} - \sum_{g=1}^G \frac{n_g W_g}{\text{trace}(W_g)} \right]$ and scaling appropriately in order to find the α 's

6.1.4 $\Sigma_g = \lambda_g D_g A_g D_g'$

If Σ_g is unconstrained across all groups, then:

$\sum_{g=1}^G \sum_{j=1}^N z_{jg}(x_j - \mu)(x_j - \mu) = T$ and $\sum_{j=1}^N z_{jg}(x_j - \mu_g)(x_j - \mu_g) = W_g$. In this case $\hat{\Sigma}_o = \frac{T}{N}$ and $\hat{\Sigma}_g = \frac{W_g}{n_g}$ then (6.1) becomes:

$$\begin{aligned} \arg \max_{\alpha} LR(\alpha) &= \arg \max_{\alpha} \left\{ \sum_{g=1}^G \frac{n_g}{2} [\log(\alpha'\Sigma_o\alpha) - \log(\alpha'\Sigma_g\alpha)] \right. \\ &\quad \left. + \frac{N}{2} \frac{\alpha'T\alpha}{\alpha'T\alpha} - \sum_{g=1}^G \frac{n_g}{2} \frac{\alpha'W_g\alpha}{\alpha'W_g\alpha} \right\} \\ &= \arg \max_{\alpha} \left\{ \sum_{g=1}^G \frac{n_g}{2} [\log(\alpha'\Sigma_o\alpha) - \log(\alpha'\Sigma_g\alpha)] \right\} \\ &= \arg \max_{\alpha} \left\{ \frac{N}{2} \log(\alpha'\Sigma_o\alpha) - \sum_{g=1}^G \frac{n_g}{2} [\log(\alpha'\Sigma_g\alpha)] \right\} \\ &= \arg \max_{\alpha} \left\{ \frac{N}{2} \log(\alpha' \frac{T}{N} \alpha) - \sum_{g=1}^G \frac{n_g}{2} \left[\log(\alpha' \frac{W_g}{n_g} \alpha) \right] \right\} \end{aligned}$$

$$\begin{aligned}
&= \arg \max_{\alpha} \left\{ \frac{N}{2} \log(\alpha' T \alpha) - \frac{N}{2} \log N \right. \\
&\quad \left. - \sum_{g=1}^G \frac{n_g}{2} \log(\alpha' W_g \alpha) + \sum_{g=1}^G \frac{n_g}{2} \log n_g \right\} \\
&= \arg \max_{\alpha} \left\{ \frac{N}{2} \log(\alpha' T \alpha) - \sum_{g=1}^G \frac{n_g}{2} \log(\alpha' W_g \alpha) \right\} \\
&= \arg \max_{\alpha} \left\{ \sum_{g=1}^G \frac{n_g}{2} [\log(\alpha' T \alpha) - \log(\alpha' W_g \alpha)] \right\} \quad (6.6)
\end{aligned}$$

Maximising (6.6) subject to $\alpha' \alpha = 1$ (or $\phi(1 - \alpha' \alpha) = 0$) gives:

$$\sum_{g=1}^G n_g \left[\frac{T \alpha}{\alpha' T \alpha} - \frac{W_g \alpha}{\alpha' W_g \alpha} \right] - 2\phi \alpha = 0$$

Solving for α is problematic, so using numerical methods, such as Newton Raphson algorithm, provide a more feasible approach. Section 6.1.5 outlines how the Newton Raphson algorithm would be applied when no information about the structure of the data is used.

6.1.5 Numerical Approximations

Complications arise when it is not possible to eliminate the denominators involving both α and covariance matrices from the maximisation problem posed by (6.1). This is more evident when one attempts to maximise (6.1) under general conditions.

$$\begin{aligned}
\frac{\partial}{\partial \alpha} LR(\alpha) &= \frac{\partial}{\partial \alpha} \left\{ \sum_{g=1}^G \sum_{j=1}^N z_{jg} \left[-\frac{1}{2} \log \alpha' \Sigma_g \alpha - \frac{1}{2} \frac{\alpha' (x_j - \mu_g)(x_j - \mu_g)' \alpha}{\alpha' \Sigma_g \alpha} \right. \right. \\
&\quad \left. \left. + \frac{1}{2} \log \alpha' \Sigma_o \alpha + \frac{1}{2} \frac{\alpha' (x_j - \mu)(x_j - \mu)' \alpha}{\alpha' \Sigma_o \alpha} \right] \right\} \\
&= N \frac{\Sigma_o \alpha}{\alpha' \Sigma_o \alpha} + \frac{[(\alpha' \Sigma_o \alpha) T \alpha - (\alpha' T \alpha) \Sigma_o \alpha]}{(\alpha' \Sigma_o \alpha)^2} \\
&\quad - \sum_{g=1}^G \left\{ n_g \frac{\Sigma_g \alpha}{\alpha' \Sigma_g \alpha} + \frac{[(\alpha' \Sigma_g \alpha) W_g \alpha - (\alpha' W_g \alpha) \Sigma_g \alpha]}{(\alpha' \Sigma_g \alpha)^2} \right\}
\end{aligned}$$

Including the constraint that $\alpha' \alpha = 1$ so that $\frac{\partial}{\partial \alpha} \phi(1 - \alpha' \alpha) = -2\phi \alpha$:

$$\begin{aligned}
\frac{\partial}{\partial \alpha} LR(\alpha) &= -2\phi \alpha + \frac{(N \Sigma_o + T) \alpha}{\alpha' \Sigma_o \alpha} - \frac{(\alpha' T \alpha) \Sigma_o \alpha}{(\alpha' \Sigma_o \alpha)^2} \\
&\quad - \sum_{g=1}^G \left\{ \frac{(n_g \Sigma_g + W_g) \alpha}{\alpha' \Sigma_g \alpha} - \frac{(\alpha' W_g \alpha) \Sigma_g \alpha}{(\alpha' \Sigma_g \alpha)^2} \right\} \quad (6.7)
\end{aligned}$$

Full details of these calculations are contained in Appendix B.1.1. However, as (6.7) gives an expression for the gradient for α , finding an expression for the Hessian will facilitate the use of Newton Raphson to find a solution for α .

$$\begin{aligned}
\frac{\partial^2}{\partial\alpha\partial\alpha'}LR(\alpha) &= -2\phi + \frac{(N\Sigma_o + T)}{\alpha'\Sigma_o\alpha} - \frac{2(N\Sigma_o + T)\alpha\alpha'\Sigma_o}{(\alpha'\Sigma_o\alpha)^2} \\
&\quad - \frac{((\alpha'T\alpha)\Sigma_o + 2\Sigma_o\alpha\alpha'T)}{(\alpha'\Sigma_o\alpha)^2} + \frac{4(\alpha'T\alpha)\Sigma_o\alpha\alpha'\Sigma_o}{(\alpha'\Sigma_o\alpha)^3} \\
&\quad + \sum_{g=1}^G \left[\frac{2(n_g\Sigma_g + W_g)\alpha\alpha'\Sigma_g}{(\alpha'\Sigma_g\alpha)^2} - \frac{(n_g\Sigma_g + W_g)}{\alpha'\Sigma_g\alpha} \right] \\
&\quad + \sum_{g=1}^G \left[\frac{((\alpha'W_g\alpha)\Sigma_g + 2\Sigma_g\alpha\alpha'W_g)}{(\alpha'\Sigma_g\alpha)^2} + \frac{4(\alpha'W_g\alpha)\Sigma_g\alpha\alpha'\Sigma_g}{(\alpha'\Sigma_g\alpha)^3} \right]
\end{aligned} \tag{6.8}$$

Further details of this calculation are in Appendix B.1.2. (6.7) and (6.8) are then used in the algorithm:

While $|g(\alpha^i)| > \epsilon$ where $g(\alpha)$ is the gradient of α , ϵ is a predetermined tolerance value and $H(\alpha)$ is the Hessian at α :

1. $\alpha^i = \alpha^{i-1} - [H(\alpha^{i-1})]^{-1}g(\alpha^{i-1})$
2. $i = i + 1$

Then return α^i .

6.2 Conclusions

Linear discriminant analysis constrains the covariance matrices to be equal across groups, with no further constraints imposed on the structure of the covariance matrices. By considering it in a likelihood ratio context further possibilities for imposing structures on the covariance matrices are explored. Most of such structures can not be easily solved analytically, but a framework for a numerical approach is explored. As the equations in Section 6.1 all use z_{jg} as an indicator variable for group membership, using the same approach as Section 5.1.3 of replacing these indicator variables with z_{jg} becoming the probability of observation j belonging to group g ; the mathematics of Section 6.1 can also be extended to include a semi-supervised framework.

LDA has shown itself to be remarkably flexible, depending only mildly on the underlying model assumptions. As the extra computational burden involved in finding approximate solutions to the maximisation for the covariance structures examined by Celeux and Govaert (1995) is significant, using this generalisation is only appropriate when the assumptions of LDA fail dramatically.

As currently implemented, the extensions of LDA in this chapter are not suited to NIR data - the dimensionality of the problem is just too high, the matrix inversions required are too unstable for ongoing practical use. Future work will include the optimisation of code to enable the more general covariance structures to be considered for higher dimensional data.

Chapter 7

Conclusions and Further Work

7.1 Conclusions

The main emphasis of the work undertaken towards this thesis was the development of statistical techniques for Near Infrared data, with particular focus towards food authentication applications with some smaller datasets were used for illustration purposes. As the goal was to develop methods for “real-life” datasets, no results for simulated datasets were presented in this thesis. However, simulated data was used in the development of the methods when trying to understand the complexities of when the various methods studied worked and what was the main reason for when they failed.

Various approaches for the utilisation of the discriminatory information contained in NIR spectra have been examined in this thesis. Given the large number of variables in such datasets, even if it is not required for a given classification technique, dimension reduction aids in the visualisation of the groups and hence the identification of unusual observations.

7.1.1 Dimension Reduction Techniques

Dimension reduction in some form, be it as part of the classification process as with the projection methods in Chapter 5 or with variable selection as in Chapter 4 or before the classification process begins in earnest as in the model-based methods of Chapter 3, is required before the highly collinear variables produced through NIR spectroscopy can be utilised in a model-based classification framework.

Each of these dimension reduction approaches is found to have advantages and disadvantages in this thesis.

Wavelet Thresholding:

To avoid discarding much of the spectra, the visible spectrum must be combined with the NIR spectrum, so that there are close to, but slightly in excess of, 2^n wavelengths. Classification using wavelet thresholding as a dimension reduction tool does not enable the identification of which particular wavelengths provide the discriminating information. However, wavelet thresholding is suitable for situations with more than two groups and does not require labelling information, so that it can be applied to all data before the classification process commences.

Variable Selection prior to classification:

In Chapter 3 using B/W , $B + W$ and $(B/W, B + W)$ to select variables before the classification process was found to be particularly effective when used in association with BIC as a model selection tool. However, determining what is to be considered a local maximum is open to some interpretation – if the local region is too small and the curves are jagged, this can lead to poor dimension reduction, as in the olive oil example of Chapter 4 – where the local region was the same number of wavelengths as with the honey data, but as the curves were far more jagged, more wavelengths were selected than would typically be desired.

Searching the Space:

The process of searching the entire spectra illustrated in Chapter 4 is not practical when applied to most classification techniques with large numbers of variables. However, it provides a valuable insight into which wavelengths can be used to distinguish groups – additional information that chemists can use to cross-check the classification method.

Projection onto a Lower Dimensional Space:

PLS and LDA (with and without updating) both use a form of projection so that the entire spectra is projected onto a lower dimensional space as part of the classification

procedure. Both have the advantage over wavelet thresholding in that they can use all available wavelengths and dimension reduction occurs alongside classification, both suffer from the same problem as wavelet thresholding by not providing easily identifiable individual wavelengths where discrimination occurs.

7.1.2 Identification of Adulterated Samples

Both the NIR honey and olive oil datasets require the separation of pure and adulterated samples. When the training set was sufficiently large (25% of the data or 35 samples) there was enough evidence available to reliably detect the adulteration of olive oil with sunflower oil. In such cases the method of dimension reduction mattered little. The honey data proved more problematic – although more observations were available on which to build the models, the adulteration mechanism used meant that composition of honey samples, already extremely variable in nature, were echoed by the adulterants used. Partial Least Squares Regression proved to be the most consistent classification tool for the honey data, closely followed by using the combination of $(B/W, B + W)$ for dimension reduction with model-based discriminant analysis techniques.

The projections provided by Semi-Supervised Linear Discriminant Analysis give further insight into the behaviour of the honey samples, with the beet invert syrups separated from the other types of adulterants, despite the labelling of *pure* or *adulterated* being used for classification purposes. This indicates that the semi-supervised LDA is able to detect underlying groups without the additional labelling information being provided. This difference between the types of adulterated samples is not apparent when using traditional LDA.

7.1.3 Classification for Multiple Groups

Applying the semi-supervised framework to the all of the examples with multiple groups (iris, wine and NIR meats datasets) improves classification results, even when, as discussed in Section 5.3, the Brier's score is increased. The NIR meats dataset is a good example of a multiple group problem, where no single group can be identified as being more important than others. The projections of LDA (with and without updating) allow for the easy visualisation of the various groups. Although

updating only marginally improves classification performance, the group boundaries are more clearly defined and hence more easily identifiable.

7.2 Further Work

7.2.1 R package

As R is freely available and non-platform dependent, it is the ideal candidate to use as a base package reach beyond those with current access to chemometric software. Creating an R package to implement the semi-supervised methods developed throughout this thesis will thus enable the methods to be used by a broader audience. The development of such an R package will require the thorough documentation of the associated functions.

7.2.2 Implementation of generalised Fisher's LDA

The generalisation of Fisher's LDA to relax the assumption of equal covariance matrices across groups is a computationally difficult process, which may not be rewarded by a corresponding improvement in classification performance. Incorporating a semi-supervised framework into this generalisation process adds to the computational cost, requiring additional emphasis to be placed on improving computational efficiency for a practical implementation of the method. This improvement in computational efficiency is most likely to be obtained by implementing some of the more computationally burdensome part of the algorithms in a compiled language such as C.

7.2.3 Variable Selection

A further exploration of the use of the relationship between $1/V$ and optimal τ to select wavelengths on the basis of $f_o = 1/V^b$ when one group is of particular interest is needed. The dependency of the variables selected on the variable selection technique used is also of further interest. Including variable selection into the model-based discriminant analysis methods of Chapter 3, especially using the subset of variables chosen using $f_o = 1/V$ for two group classification problems would enable

further insight to be gained as to what chemical bonds enable discrimination between groups.

7.3 Final Comments

The goal of the work undertaken in this thesis was to develop statistical methods for NIR datasets that could be efficiently undertaken without specialist computing facilities. Not only have such methods to be computationally efficient, but also must be easily understood by chemists if they are to be adopted over existing methods. Thus methods developed in this thesis were designed to be as simple to explain to chemists as possible.

Chemists are currently content to use methods such as PLSR where they have little understanding of the underlying algorithm. The algorithms used in the methods introduced in this thesis are more generally more transparent than those in current use, with the possible exception of using wavelet thresholding to reduce the dimensionality of the space rather than using a variable selection approach.

The improvement in computation time achieved by using a complexity criterion rather than using cross validation for selecting the number of components to use with PLSR is significant. With the possibility of improving both speed and accuracy of prediction using a familiar technique, this combination offers the most likely candidate for early and widespread adoption within the spectroscopy community of the methods developed in this thesis.

Using a constant threshold, as investigated in Chapter 4, rather than a threshold that is a function of the data offered an especially simple concept that appealed to chemists, but did not provide a robust classifier for the discrimination problems investigated in this thesis.

Semi-supervised linear discriminant analysis is the most promising of the new techniques in that it combines a simple concept with a useful contribution towards improved classification and / or visualisation of groups.

In order for alternative methods to be adopted by the chemists who use NIR, they should also present some obvious benefit over currently used methods - measuring the uncertainty of individual classifications and improving on computation time

being areas that are of particular relevance.

As all of the methods developed were implemented in R, the techniques can be freely used on a variety of different computational platforms, and offer a significant cost saving over existing procedures, while allowing extra modelling flexibility to be incorporated by the user if desired. There is a learning curve involved with using R for chemists, however, as many are using MATLAB and similar command line programs, the additional flexibility should compensate for the effort involved. As part of a fully documented R package, the implementation of the methods introduced in this thesis provide a faster, more flexible and less expensive option than the methods currently used by chemists to analyse NIR data.

The methods developed in this thesis are designed to be used by chemists. Therefore, extensions of existing methods were the main focus of the thesis – the familiarity of the methods on which they are based serves to encourage the early adoption of the newer methods.

Appendix A

Details of Calculations

A.1 Rotation

In order to find a suitable rotation of W so that it takes the form of an identity matrix recall that:

If A is a $k \times k$ positive definite matrix with the spectral decomposition:

$$A = \sum_{i=1}^k \lambda e_i e_i' = P \Lambda P'$$

where the normalised eigenvectors are the columns of matrix P and where Λ is a diagonal matrix, with the eigenvalues λ as the diagonal entries then

$$A^{-1} = P \Lambda^{-1} P' = \sum_{i=1}^k \frac{1}{\lambda_i} e_i e_i' \quad (\text{A.1})$$

Now similarly if $\Lambda^{1/2}$ is the diagonal matrix with $\sqrt{\lambda_i}$ as it's i th diagonal element;

$$A^{1/2} = \sum_{i=1}^k \sqrt{\lambda_i} e_i e_i' = P \Lambda^{1/2} P'$$

and where $\Lambda^{-1/2}$ has $1/\sqrt{\lambda_i}$ as it's i th diagonal entry

$$A^{-1/2} = \sum_{i=1}^k \frac{1}{\sqrt{\lambda_i}} e_i e_i' = P \Lambda^{-1/2} P' \quad (\text{A.2})$$

Now for a covariance matrix Σ , using the transformation $a = \Sigma^{1/2} l$ and recalling that since Σ is symmetric, so is $\Sigma^{1/2}$, as is $\Sigma^{-1/2}$. Thus

$$a' \Sigma^{-1/2} B \Sigma^{-1/2} a = l' \Sigma^{1/2} \Sigma^{-1/2} B \Sigma^{-1/2} \Sigma^{1/2} l = l' B l$$

and the maximisation problem then becomes

$$\max \frac{l'Bl}{l'\Sigma l} = \max \frac{a'\Sigma^{-1/2}B\Sigma^{-1/2}a}{a'a}.$$

This is maximised when the ratio is λ_1 , which occurs when $a = e_1$. Thus $e_1 = a = \Sigma^{1/2}l_1$ so $l_1 = \Sigma^{-1/2}e_1$. Similarly for the remaining eigenvectors $l_k = \Sigma^{-1/2}e_k$. Considering e_i and λ_i as an eigenvector and eigenvalue pair of $\Sigma^{-1/2}B\Sigma^{-1/2}$, then

$$\Sigma^{-1/2}B\Sigma^{-1/2}e_i = \lambda_i e_i. \quad (\text{A.3})$$

Then multiplying (A.3) by $\Sigma^{-1/2}$ and using the results of (A.1) and (A.2)

$$\begin{aligned} \Sigma^{-1/2}\Sigma^{-1/2}B\Sigma^{-1/2}e_i &= \lambda_i \Sigma^{-1/2}e_i \\ \Sigma^{-1}B\Sigma^{-1/2}e_i &= \lambda_i (\Sigma^{-1/2}e_i) \\ P\Lambda^{-1}P'BP(\Lambda^{-1/2}P'e_i) &= \lambda_i (P\Lambda^{-1/2}P'e_i) \end{aligned}$$

So that the inverse of the covariance matrix W does not need to be found directly.

W can be written in the form $W = A'A$, where A has dimension $N \times p$. Finding the singular value decomposition of A is more efficient than finding that of W if $N \times G < p$

The rotation required is then ABA' and recalling that $s = \min(G - 1, p)$ so that $e_{1\dots s}$ are the largest s eigenvectors of ABA' , then the coefficients for Fisher's Linear Discriminant Analysis are

$$a = A'e_{1\dots s}$$

Appendix B

Details of Numerical Approximations

B.1 Numerical Approximation of $\hat{\alpha}$

B.1.1 Gradient for α

$$\begin{aligned}
 \frac{\partial}{\partial \alpha} LR(\alpha) &= \frac{\partial}{\partial \alpha} \left\{ \sum_{g=1}^G \sum_{j=1}^N z_{jg} \left[-\frac{1}{2} \log \alpha' \Sigma_g \alpha - \frac{1}{2} \frac{\alpha' (x_j - \mu_g)(x_j - \mu_g)' \alpha}{\alpha' \Sigma_g \alpha} \right. \right. \\
 &\quad \left. \left. + \frac{1}{2} \log \alpha' \Sigma_o \alpha + \frac{1}{2} \frac{\alpha' (x_j - \mu)(x_j - \mu)' \alpha}{\alpha' \Sigma_o \alpha} \right] \right\} \\
 &= \frac{\partial}{\partial \alpha} \left\{ \sum_{g=1}^G \left[\frac{n_g}{2} \log \alpha' \Sigma_o \alpha + \frac{1}{2} \frac{\alpha' T \alpha}{\alpha' \Sigma_o \alpha} - \frac{n_g}{2} \log \alpha' \Sigma_g \alpha - \frac{1}{2} \frac{\alpha' W_g \alpha}{\alpha' \Sigma_g \alpha} \right] \right\} \\
 &= \sum_{g=1}^G \left\{ \frac{\partial}{\partial \alpha} \left[\frac{n_g}{2} \log \alpha' \Sigma_o \alpha + \frac{1}{2} \frac{\alpha' T \alpha}{\alpha' \Sigma_o \alpha} - \frac{n_g}{2} \log \alpha' \Sigma_g \alpha - \frac{1}{2} \frac{\alpha' W_g \alpha}{\alpha' \Sigma_g \alpha} \right] \right\} \\
 &= \sum_{g=1}^G \left\{ \frac{\partial}{\partial \alpha} \left[\frac{n_g}{2} \log \alpha' \Sigma_o \alpha \right] + \frac{\partial}{\partial \alpha} \left[\frac{1}{2} \frac{\alpha' T \alpha}{\alpha' \Sigma_o \alpha} \right] \right. \\
 &\quad \left. - \frac{\partial}{\partial \alpha} \left[\frac{n_g}{2} \log \alpha' \Sigma_g \alpha \right] - \frac{\partial}{\partial \alpha} \left[\frac{1}{2} \frac{\alpha' W_g \alpha}{\alpha' \Sigma_g \alpha} \right] \right\} \\
 &= \sum_{g=1}^G \left\{ \frac{n_g}{2} \frac{\partial}{\partial \alpha} [\log \alpha' \Sigma_o \alpha] + \frac{1}{2} \frac{\partial}{\partial \alpha} \left[\frac{\alpha' T \alpha}{\alpha' \Sigma_o \alpha} \right] \right. \\
 &\quad \left. - \frac{n_g}{2} \frac{\partial}{\partial \alpha} [\log \alpha' \Sigma_g \alpha] - \frac{1}{2} \frac{\partial}{\partial \alpha} \left[\frac{\alpha' W_g \alpha}{\alpha' \Sigma_g \alpha} \right] \right\} \tag{B.1}
 \end{aligned}$$

Considering each of the components of (B.1):

$$\frac{\partial}{\partial \alpha} [\log \alpha' \Sigma_o \alpha] = \frac{2 \Sigma_o \alpha}{\alpha' \Sigma_o \alpha}$$

$$\frac{\partial}{\partial \alpha} [\log \alpha' \Sigma_g \alpha] = \frac{2 \Sigma_g \alpha}{\alpha' \Sigma_g \alpha}$$

$$\begin{aligned} \frac{\partial}{\partial \alpha} \left[\frac{\alpha' T \alpha}{\alpha' \Sigma_o \alpha} \right] &= \frac{1}{(\alpha' \Sigma_o \alpha)^2} \left[(\alpha' \Sigma_o \alpha) \frac{\partial}{\partial \alpha} (\alpha' T \alpha) - (\alpha' T \alpha) \frac{\partial}{\partial \alpha} (\alpha' \Sigma_o \alpha) \right] \\ &= \frac{[2(\alpha' \Sigma_o \alpha) T \alpha - 2(\alpha' T \alpha) \Sigma_o \alpha]}{(\alpha' \Sigma_o \alpha)^2} \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \alpha} \left[\frac{\alpha' W_g \alpha}{\alpha' \Sigma_g \alpha} \right] &= \frac{1}{(\alpha' \Sigma_g \alpha)^2} \left[(\alpha' \Sigma_g \alpha) \frac{\partial}{\partial \alpha} (\alpha' W_g \alpha) - (\alpha' W_g \alpha) \frac{\partial}{\partial \alpha} (\alpha' \Sigma_g \alpha) \right] \\ &= \frac{[2(\alpha' \Sigma_g \alpha) W_g \alpha - 2(\alpha' W_g \alpha) \Sigma_g \alpha]}{(\alpha' \Sigma_g \alpha)^2} \end{aligned}$$

So that (B.1) becomes:

$$\begin{aligned} \frac{\partial}{\partial \alpha} LR(\alpha) &= \sum_{g=1}^G \left\{ \frac{n_g}{2} \frac{2 \Sigma_o \alpha}{\alpha' \Sigma_o \alpha} + \frac{1}{2} \frac{[2(\alpha' \Sigma_o \alpha) T \alpha - 2(\alpha' T \alpha) \Sigma_o \alpha]}{(\alpha' \Sigma_o \alpha)^2} \right. \\ &\quad \left. - \frac{n_g}{2} \frac{2 \Sigma_g \alpha}{\alpha' \Sigma_g \alpha} - \frac{1}{2} \frac{[2(\alpha' \Sigma_g \alpha) W_g \alpha - 2(\alpha' W_g \alpha) \Sigma_g \alpha]}{(\alpha' \Sigma_g \alpha)^2} \right\} \\ &= \sum_{g=1}^G \left\{ n_g \frac{\Sigma_o \alpha}{\alpha' \Sigma_o \alpha} + \frac{[(\alpha' \Sigma_o \alpha) T \alpha - (\alpha' T \alpha) \Sigma_o \alpha]}{(\alpha' \Sigma_o \alpha)^2} \right. \\ &\quad \left. - n_g \frac{\Sigma_g \alpha}{\alpha' \Sigma_g \alpha} - \frac{[(\alpha' \Sigma_g \alpha) W_g \alpha - (\alpha' W_g \alpha) \Sigma_g \alpha]}{(\alpha' \Sigma_g \alpha)^2} \right\} \\ &= N \frac{\Sigma_o \alpha}{\alpha' \Sigma_o \alpha} + \frac{[(\alpha' \Sigma_o \alpha) T \alpha - (\alpha' T \alpha) \Sigma_o \alpha]}{(\alpha' \Sigma_o \alpha)^2} \\ &\quad - \sum_{g=1}^G \left\{ n_g \frac{\Sigma_g \alpha}{\alpha' \Sigma_g \alpha} + \frac{[(\alpha' \Sigma_g \alpha) W_g \alpha - (\alpha' W_g \alpha) \Sigma_g \alpha]}{(\alpha' \Sigma_g \alpha)^2} \right\} \end{aligned}$$

Including the constraint that $\alpha' \alpha = 1$ so that $\frac{\partial}{\partial \alpha} \phi(1 - \alpha' \alpha) = -2\phi\alpha$:

$$\begin{aligned} \frac{\partial}{\partial \alpha} LR(\alpha) &= -2\phi\alpha + N \frac{\Sigma_o \alpha}{\alpha' \Sigma_o \alpha} + \frac{[(\alpha' \Sigma_o \alpha) T \alpha - (\alpha' T \alpha) \Sigma_o \alpha]}{(\alpha' \Sigma_o \alpha)^2} \\ &= -2\phi\alpha + \frac{(N \Sigma_o + T) \alpha}{\alpha' \Sigma_o \alpha} - \frac{(\alpha' T \alpha) \Sigma_o \alpha}{(\alpha' \Sigma_o \alpha)^2} \\ &\quad - \sum_{g=1}^G \left\{ \frac{(n_g \Sigma_g + W_g) \alpha}{\alpha' \Sigma_g \alpha} - \frac{(\alpha' W_g \alpha) \Sigma_g \alpha}{(\alpha' \Sigma_g \alpha)^2} \right\} \end{aligned} \tag{B.2}$$

Solving for α here is non-trivial. However, (B.2) is the gradient for α . Calculating the Hessian will facilitate the use of Newton Raphson to find the solution for α .

B.1.2 Hessian for α

$$\begin{aligned} \frac{\partial^2}{\partial\alpha\partial\alpha'} LR(\alpha) = & -2\phi + \frac{\partial}{\partial\alpha'} \left\{ \frac{(N\Sigma_o + T)\alpha}{\alpha'\Sigma_o\alpha} - \frac{(\alpha'T\alpha)\Sigma_o\alpha}{(\alpha'\Sigma_o\alpha)^2} \right. \\ & \left. - \sum_{g=1}^G \left\{ \frac{(n_g\Sigma_g + W_g)\alpha}{\alpha'\Sigma_g\alpha} - \frac{(\alpha'W_g\alpha)\Sigma_g\alpha}{(\alpha'\Sigma_g\alpha)^2} \right\} \right\} \quad (B.3) \end{aligned}$$

Examining each of the sections of (B.3):

$$\frac{\partial}{\partial\alpha'} \left\{ \frac{(N\Sigma_o + T)\alpha}{\alpha'\Sigma_o\alpha} \right\} = \frac{1}{(\alpha'\Sigma_o\alpha)^2} [(\alpha'\Sigma_o\alpha)(N\Sigma_o + T) - 2(N\Sigma_o + T)\alpha\alpha'\Sigma_o]$$

$$\frac{\partial}{\partial\alpha'} \left\{ \frac{(n_g\Sigma_g + W_g)\alpha}{\alpha'\Sigma_g\alpha} \right\} = \frac{1}{(\alpha'\Sigma_g\alpha)^2} [(\alpha'\Sigma_g\alpha)(n_g\Sigma_g + W_g) - 2(n_g\Sigma_g + W_g)\alpha\alpha'\Sigma_g]$$

$$\frac{\partial}{\partial\alpha'} [(\alpha'T\alpha)\Sigma_o\alpha] = (\alpha'T\alpha)\Sigma_o + 2\Sigma_g\alpha\alpha'T$$

$$\frac{\partial}{\partial\alpha'} [(\alpha'\Sigma_o\alpha)^2] = 4(\alpha'\Sigma_o\alpha)\alpha'\Sigma_o$$

$$\frac{\partial}{\partial\alpha'} \left[\frac{(\alpha'T\alpha)\Sigma_o\alpha}{(\alpha'\Sigma_o\alpha)^2} \right] = \frac{(\alpha'\Sigma_o\alpha)^2 [(\alpha'T\alpha)\Sigma_o + 2\Sigma_o\alpha\alpha'T] - 4(\alpha'T\alpha)(\alpha'\Sigma_o\alpha)\Sigma_o\alpha\alpha'\Sigma_o}{(\alpha'\Sigma_o\alpha)^4}$$

$$\frac{\partial}{\partial\alpha'} \left[\frac{(\alpha'W_g\alpha)\Sigma_g\alpha}{(\alpha'\Sigma_g\alpha)^2} \right] = \frac{(\alpha'\Sigma_g\alpha)^2 [(\alpha'W_g\alpha)\Sigma_g + 2\Sigma_g\alpha\alpha'W_g] - 4(\alpha'W_g\alpha)(\alpha'\Sigma_g\alpha)\Sigma_g\alpha\alpha'\Sigma_g}{(\alpha'\Sigma_g\alpha)^4}$$

leads to the Hessian being:

$$\begin{aligned} \frac{\partial^2}{\partial\alpha\partial\alpha'} LR(\alpha) = & -2\phi + \frac{(N\Sigma_o + T)}{\alpha'\Sigma_o\alpha} - \frac{2(N\Sigma_o + T)\alpha\alpha'\Sigma_o}{(\alpha'\Sigma_o\alpha)^2} \\ & - \frac{((\alpha'T\alpha)\Sigma_o + 2\Sigma_o\alpha\alpha'T)}{(\alpha'\Sigma_o\alpha)^2} + \frac{4(\alpha'T\alpha)\Sigma_o\alpha\alpha'\Sigma_o}{(\alpha'\Sigma_o\alpha)^3} \\ & + \sum_{g=1}^G \left[\frac{2(n_g\Sigma_g + W_g)\alpha\alpha'\Sigma_g}{(\alpha'\Sigma_g\alpha)^2} - \frac{(n_g\Sigma_g + W_g)}{\alpha'\Sigma_g\alpha} \right] \\ & + \sum_{g=1}^G \left[\frac{((\alpha'W_g\alpha)\Sigma_g + 2\Sigma_g\alpha\alpha'W_g)}{(\alpha'\Sigma_g\alpha)^2} + \frac{4(\alpha'W_g\alpha)\Sigma_g\alpha\alpha'\Sigma_g}{(\alpha'\Sigma_g\alpha)^3} \right] \end{aligned}$$

Bibliography

- Banfield, J. D. and Raftery, A. E. (1993), 'Model-based Gaussian and non-Gaussian clustering', *Biometrics* **49**, 803–821.
- Bensmail, H. and Celeux, G. (1996), 'Regularized Gaussian discriminant analysis through eigenvalue decomposition', *Journal of the American Statistical Association* **91**, 1743–1748.
- Brier, G. W. (1950), 'Verification of forecasts expressed in terms of probability', *Monthly Weather Review* **78**, 1–3.
- Celeux, G. and Govaert, G. (1995), 'Gaussian parsimonious clustering models', *Pattern Recognition* **28**, 781–793.
- Daubechies, I. (1988), 'Orthonormal bases of compactly supported wavelets', *Communications on Pure and Applied Mathematics* **41**, 909–996.
- De Maesschalck, R., Candolfi, A., Massart, D. L. and Heuerding, S. (1999), 'Decision criteria for soft independent modelling of class analogy applied to near infrared data', *Chemometrics and Intelligent Laboratory Systems* **47**, 65–77.
- Dean, N., Murphy, T. B. and Downey, G. (2006), 'Using Unlabelled Data To Update Classification Rules With Applications In Food Authenticity Studies', *Journal of the Royal Statistical Society, Series C* **55**, 1–14.
- Dean, N. and Raftery, A. E. (2005), 'Normal uniform mixture differential gene expression detection for cDNA microarrays', *BMC Bioinformatics* **6**, 173.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood for incomplete data via the EM algorithm (with discussion)', *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

- Downey, G., Fouratier, V. and Kelly, J. D. (2003), 'Detection of honey adulteration by addition of fructose and glucose using near infrared transreflectance spectroscopy', *Journal of Near Infrared Spectroscopy* **11**, 447–456.
- Downey, G., McIntyre, P. and Davis, A. N. (2002), 'Detecting and quantifying sunflower oil adulteration in extra virgin olive oils from the Eastern Mediterranean by visible and near infrared spectroscopy', *Journal of Agricultural and Food Chemistry* **50**, 5520–5525.
- European Commission (2002), 'Council Directive 2001/110/EC of 20 June 2001, relating to honey'.
- Fisher, R. A. (1936), 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics* **7**, 179–188.
- Forina, M., Armanino, C. and Ubigli, M. (1986), 'Multivariate data analysis as a discriminating method of the origin of wines', *Vitus* **25**, 189–201.
- Forina, M. and Tiscornia, E. (1982), 'Classification of olive oils from their fatty acid composition', *Annali di Chimica* **72**, 127–141.
- Fraley, C. and Raftery, A. E. (1998), 'How many clusters? Which clustering method? - Answers via model-based cluster analysis', *Computer Journal* **41**, 578–588.
- Fraley, C. and Raftery, A. E. (1999), 'MCLUST: Software for model-based cluster analysis', *Journal of Classification* **16**, 297–306.
- Fraley, C. and Raftery, A. E. (2002), 'Model-based clustering, discriminant analysis, and density estimation', *Journal of the American Statistical Association* **97**, 611–631.
 URL: <http://www.stat.washington.edu/www/research/reports/2000/tr380.pdf>
- Fraley, C. and Raftery, A. E. (2007), *mclust: Model-Based Clustering / Normal Mixture Modeling*. R package version 3.1-1.
 URL: <http://www.stat.washington.edu/mclust>
- Frank, I. E. and Friedman, J. H. (1989), 'Classification: oldtimers and newcomers', *Journal of Chemometrics* **3**, 463–475.

- Frank, I. E. and Friedman, J. H. (1993), 'A statistical view of some chemometrics regression tools', *Technometrics* **35**, 109–135.
- Haar, A. (1910), 'Zur theorie der orthogonalen funktionensysteme', *Mathematische Annalen* **69**, 331–371.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer.
- Heil, C. and Walnut, D. F., eds (2006), *Fundamental papers in wavelet theory*, Princeton University Press.
- Helland, I. S. (1990), 'Partial least squares regression and statistical models', *Scandinavian Journal of Statistics* **17**, 97–114.
- Helland, I. S. (2001), 'Some theoretical aspects of partial least squares regression', *Chemometrics and Intelligent Laboratory Systems* **58**, 97–107.
- Hennig, C. (2004), 'Asymmetric linear dimension reduction for classification', *Journal of Computational and Graphical Statistics* **13**, 930–945.
- Hurley, C. (2004), *gclus: Clustering Graphics*. R package version 1.2.
- Indahl, U. F., Sahni, N. S., Kirkhus, B. and Naes, T. (1999), 'Multivariate strategies for classification based on NIR-spectra – with application to mayonnaise', *Chemometrics and Intelligent Laboratory Systems* **49**, 19–31.
- Madden, H. (1978), 'Comments on the Savitzky-Golay Convolution Method for Least-Squares Fit Smoothing and Differentiation of Digital Data', *Analytical Chemistry* **50**, 1383–1386.
- McDonald, J. (2008a), *China launches nationwide baby formula probe*. Associated Press, 13th September 2008.
- McDonald, S. (2008b), *Nearly 53000 Chinese children sick from milk*. Associated Press, 21st September 2008.
- McElhinney, J., Downey, G. and Fearn, T. (1999), 'Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats', *Journal of Near Infrared Spectroscopy* **7**, 145–154.

- Murphy, T. B., Dean, N. and Raftery, A. E. (2008), Variable selection and updating in model-based discriminant analysis for high-dimensional data, Technical Report 536, Department of Statistics, University of Washington, Seattle.
- Nason, G., Kovac, A. and Maechler, M. (2006), *wavethresh: Software to perform wavelet statistics and transforms*. R package version 2.2-9, ported by Guy Nason.
- Newton, P. N., Green, M. D., Fernandez, F. M., Day, N. P. J. and White, N. J. (2006), 'Counterfeit anti-infective drugs', *Lancet Infectious Diseases* **6**, 602–613.
- Ogden, R. T. (1997), *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhauser.
- O'Neill, T. J. (1978), 'Normal discrimination with unclassified observations', *Journal of the American Statistical Association* **73**, 821–826.
- R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- URL:** <http://www.R-project.org>
- Raftery, A. E. and Dean, N. (2006), 'Variable selection for model-based clustering', *Journal of the American Statistical Association* **101**, 168–178.
- Savitzky, A. and Golay, M. J. E. (1964), 'Smoothing and differentiation of data by simplified least squares procedures', *Analytical Chemistry* **36**, 1627–1639.
- Steinier, J., Termonia, Y. and Deltour, J. (1972), 'Comments on Smoothing and Differentiation of Data by Simplified Least Square Procedure', *Analytical Chemistry* **44**, 1906–1909.
- Stone, M. and Brooks, R. J. (1990), 'Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression', *Journal of the Royal Statistical Society, Series B* **52**, 237–269.
- Toher, D., Downey, G. and Murphy, T. B. (2007), 'A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data

- in food authentication studies', *Chemometrics and Intelligent Laboratory Systems* **89**, 102–115.
- Van Der Voet, H. (1999), 'Pseudo-degrees of freedom for partial least squares', *Journal of Chemometrics* **13**, 195–208.
- WHO (1999), 'Counterfeit drugs - guidelines for the development of measures to combat counterfeit drugs'.
- Wold, H. (1966a), Estimation of principal components and related models by iterative least squares, in 'Multivariate Analysis', pp. 391–420.
- Wold, H. (1966b), Nonlinear estimation by iterative least square procedures, in 'Research Papers in Statistics: Festschrift for J. Neyman', pp. 411–444.
- Wold, S. (1976), 'Pattern recognition by means of disjoint principal components models', *Pattern Recognition* **8**, 127–139.
- Zhu, M. (2003), 'Feature Extraction for Nonparametric Discriminant Analysis', *Journal of Computational and Graphical Statistics* **12**, 101–120.
- Zhu, M. (2006), 'Discriminant Analysis with Common Principal Components', *Biometrika* **93**, 1018–1024.