

# Using Unlabelled Data To Update Classification Rules With Applications In Food Authenticity Studies

Nema Dean<sup>1</sup>, Thomas Brendan Murphy<sup>2</sup> and Gerard Downey<sup>3</sup>.

Technical Report no. 444  
Department of Statistics  
University of Washington

February 22, 2004

<sup>1</sup>Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, USA. Email: [nemad@stat.washington.edu](mailto:nemad@stat.washington.edu). This research was partially supported by NIH grant 8 R01 EB002137-02.

<sup>2</sup>Department of Statistics, Trinity College, Dublin 2, Ireland. The author is very grateful to Adrian Raftery for his help and advice with this work.

<sup>3</sup>TEAGASC, The National Food Centre, Ashtown, Dublin 15, Ireland. This work was supported by the Irish Department of Agriculture under the Food Sub-programme of the Operational Programme for Industrial Development and the FIRM programme.

## Abstract

A classification method is developed to classify samples when both labelled and unlabelled samples are available. The classification rule is estimated using both the labelled and unlabelled data, in contrast to many classical methods which only use the labelled data for estimation.

This methodology models the data as arising from a Gaussian mixture model with parsimonious covariance structure, as is done in model-based clustering (Fraley and Raftery (2002)). A missing-data formulation of the mixture model is used and the models are fitted using the EM and CEM algorithms.

A comparison of the performance of model-based discriminant analysis and the proposed method of classification is given.

The methods are applied to the analysis of spectra of foodstuffs recorded over the visible and near-infrared wavelength range in food authenticity studies. The aim of this study is to classify the foodstuffs using their spectra. The proposed classification method is shown to yield very good misclassification rates. The correct classification rate was observed to be as much as 15% higher than the correct classification rate for model-based discriminant analysis.

*Keywords:* Classification, discriminant analysis, model-based clustering, food authenticity studies, near-infrared spectroscopy

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Model-based Clustering and Discriminant Analysis</b>	<b>2</b>
<b>3</b>	<b>Updating Classification Rules</b>	<b>3</b>
3.1	The EM Algorithm . . . . .	5
3.2	The CEM Algorithm . . . . .	6
3.3	Determining Convergence of EM/CEM Algorithms . . . . .	6
3.4	Classifying Observations . . . . .	7
3.5	Model Choice . . . . .	7
<b>4</b>	<b>Application In Food Authenticity Studies</b>	<b>7</b>
4.1	Near-infrared Spectra . . . . .	8
4.2	Data Reduction . . . . .	9
4.2.1	Wavelets . . . . .	9
4.2.2	Wavelet Thresholding . . . . .	10
4.3	Meat Samples . . . . .	10
4.4	Greek Olive Oils . . . . .	14
<b>5</b>	<b>Conclusions</b>	<b>14</b>

## List of Tables

1	The covariance restrictions available as part of the <code>mclust</code> library . . . . .	3
2	Average correct classification rates for the five meat groups . . . . .	11
3	Example of the classifications of the five group meats data . . . . .	12
4	Average correct classification rates for the four meat groups . . . . .	13
5	Example of the classifications of the four group meats data . . . . .	13
6	Average correct classification rates for the three olive oil groups . . . . .	14

## List of Figures

1	Examples of cluster shapes allowed by the covariance restrictions available in the <code>mclust</code> library . . . . .	4
2	Examples of infrared spectra for each meat type . . . . .	8
3	A plot of an infrared spectrum and the wavelet coefficients . . . . .	10
4	A plot of selected wavelet coefficients after thresholding . . . . .	11

# 1 Introduction

Discriminant analysis is used to classify unlabelled samples into groups (or classes) when labelled (or classified) samples are available (McLachlan (1992)). Two of the most popular methods are linear (LDA) and quadratic (QDA) discriminant analysis, although many other methods exist. These methods develop a classification rule using the labelled data and use this rule to classify the unlabelled data.

Cluster analysis uses unlabelled observations to discover groupings in data and to assign the observations to these groupings (Hartigan (1975)). Many clustering methods exist but some of these are developed from some heuristic reasoning. Some common clustering methods include hierarchical methods and  $k$ -means clustering.

Both discriminant analysis and cluster analysis methods can be developed from a statistical modelling viewpoint. This approach typically assumes that data within group  $g$  are generated from some density  $f_g(\cdot)$  and that the proportion of the population that belongs to group  $g$  is  $p_g$ ; that is, the data comes from a mixture model (Titterton, Smith, and Makov (1985), McLachlan and Peel (2000)). Therefore, the density of the data is given by  $f(x) = \sum_{g=1}^G p_g f_g(x)$ , where  $G$  is the total number of groups.

Mixture models are the basis of model-based clustering (Banfield and Raftery (1993), Fraley and Raftery (1998), Fraley and Raftery (2002)) which uses mixtures of normal distributions to develop a flexible suite of clustering methods; Bensmail and Celeux (1996) develop a suite of discriminant analysis methods in a similar way.

Model-based clustering and discriminant analysis uses constraints on the eigenvalue decomposition of the group covariance matrices to impose shape restrictions on the groups in the data. Further details of model-based clustering and discriminant analysis are given in Section 2.

In classical discriminant analysis, the unlabelled data is not used in the model fitting procedure. These data contain information that is potentially important, especially when there is very little labelled data. Following model-based clustering and discriminant analysis, we can model the labelled and unlabelled data as coming from a mixture model. In this framework, we assume that the unlabelled data has the labelling variable as missing data. The unlabelled data can then be used in the model fitting procedure.

We fit the mixture model using the EM and CEM algorithms; this is described in Section 3. This idea is very similar to ideas concerning updating classification rules that have appeared in Ganesalingam and McLachlan (1978) and O'Neill (1978), amongst others; these

related results are briefly reviewed in Section 3.

The proposed updated classification method provides an improved misclassification rate over that of discriminant analysis. The improvement in misclassification rate is especially good when there is very little labelled data available.

The proposed methods are applied to the analysis of spectra of foodstuffs recorded over the visible and near-infrared wavelength range (400–2498 nm) in food authenticity studies. The aim of the application is to classify foodstuffs using their spectra.

Two particular food authenticity studies are considered: the classification of raw homogenized meat samples into individual species (Chicken, Turkey, Pork, Beef, Lamb) and the classification of Greek olive oils according to their geographical origin (Crete, Peleponese and Other Regions). The proposed methods are shown to have excellent classification rates for both of these problems (Section 4).

## 2 Model-based Clustering and Discriminant Analysis

Normal mixtures are the basis of model-based clustering (Banfield and Raftery (1993), Fraley and Raftery (1998), Fraley and Raftery (2002)) and discriminant analysis (Bensmail and Celeux (1996)). This approach uses constraints on the eigenvalue decomposition of the covariance matrix to impose shape restrictions on the groups in the data.

The eigenvalue decomposition of each group covariance matrix  $\Sigma_g$  can be written as  $\Sigma_g = \lambda_g D_g A_g D_g^T$  where the  $D_g$  is an orthogonal matrix of eigenvectors of  $\Sigma_g$ , the  $A_g$  is a diagonal matrix with entries proportional to the eigenvalues of  $\Sigma_g$  and  $\lambda_g$  is a proportionality constant. Each component of the eigenvalue decomposition has a different morphologic interpretation in terms of the shape of data in a group. The matrix  $D_g$  governs the orientation of the group, the  $A_g$  matrix controls the shape and the  $\lambda_g$  controls the volume of the group.

The components of the eigenvalue decomposition can be unrestricted or they can be constrained to be identical for the different groups. In addition, the  $D_g$  and  $A_g$  matrices can be forced to be identity matrices to give even further modelling options.

A model-based clustering and discriminant analysis library called `mclust` has been developed for the statistical packages `S-Plus` and `R` (Fraley and Raftery (1999), Fraley and Raftery (2003)). The covariance constraints available in this library are described in Table 1 and Figure 1.

Model-based clustering and discriminant analysis allows for parsimonious modelling using a mixture of normals with constrained covariance matrices that use potentially less param-

eters than a mixture of normals with unconstrained covariance matrices (see Table 1).

Table 1: The covariance restrictions available as part of the `mclust` library. The first letter represents the volume constraint, the second letter represents the shape constraint and the third letter represents the orientation constraint ( $E$ =Equal,  $V$ =Variable,  $I$ =Identity). The number of groups is  $G$  and the data has dimension  $p$ .

MODELID	VOLUME	SHAPE	ORIENTATION	DECOMPOSITION	COVARIANCE PARAMETERS
EII	Equal	Spherical	—	$\Sigma_g = \lambda I$	1
VII	Variable	Spherical	—	$\Sigma_g = \lambda_g I$	$G$
EEI	Equal	Equal	Axis Aligned	$\Sigma_g = \lambda DD^T$	$p$
VEI	Variable	Equal	Axis Aligned	$\Sigma_g = \lambda_g DD^T$	$p + G - 1$
EVI	Equal	Variable	Axis Aligned	$\Sigma_g = \lambda D_g D_g^T$	$pG - G + 1$
VVI	Variable	Variable	Axis Aligned	$\Sigma_g = \lambda_g D_g D_g^T$	$pG$
EEE	Equal	Equal	Equal	$\Sigma_g = \lambda DAD^T$	$p(p+1)/2$
EEV	Equal	Equal	Variable	$\Sigma_g = \lambda D_g AD_g^T$	$Gp(p+1)/2$ $-(G-1)p$
VEV	Variable	Equal	Variable	$\Sigma_g = \lambda_g D_g AD_g^T$	$Gp(p+1)/2$ $-(G-1)(p-1)$
VVV	Variable	Variable	Variable	$\Sigma_g = \lambda_g D_g A_g D_g^T$	$Gp(p+1)/2$

### 3 Updating Classification Rules

The model-based clustering formulation is used to develop a classification method that updates the discriminant analysis classification rule using the unlabelled samples for which classifications are sought.

The proposal to update the classification rule using the unlabelled data stems from the idea that the unlabelled data may contain useful information with respect to the group parameters even if the group membership of these observations is unknown. Hence, the updated classification rule approach uses both the labelled and unlabelled data to estimate the parameters of the mixture model. The parameter estimation can be completed using the EM algorithm and/or its variants. The resulting fitted mixture model provides a classification rule for the unlabelled data.

The use of unlabelled data to update classification rules has been considered previously by McLachlan (1975), McLachlan (1977), Ganesalingam and McLachlan (1978) and O'Neill

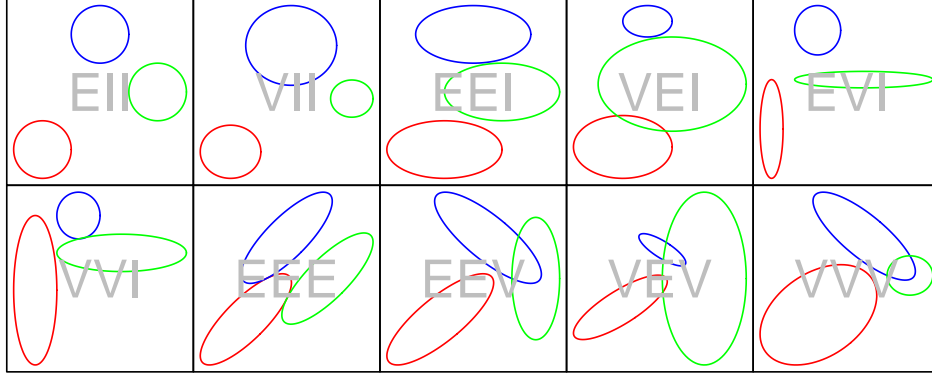


Figure 1: Examples of cluster shapes allowed by the covariance restrictions available in the `mclust` library in the bivariate case.

(1978) in the context of two component normal mixtures with equal or unequal covariance matrices. Ganesalingam and McLachlan (1978) and O’Neill (1978) compute the efficiency of using a sample with a fixed proportion of the data being unlabelled compared to a fully labelled data set. These results show that the benefit of updating the classification rule using unlabelled data depends heavily on the separation between the mixture component means. In addition, these results are verified using simulations by McLachlan and Ganesalingam (1982). A related problem is considered in Titterton (1976), where the problem of sequentially updating a classification rule as unlabelled observations arrive is considered.

General reviews of the topic of updating classification rules using unlabelled data are given in Titterton, Smith, and Makov (1985), McLachlan and Basford (1988), McLachlan (1992) and McLachlan and Peel (2000).

Let the labelled data be denoted by  $\mathbf{x}_N = (x_1, x_2, \dots, x_N)$  and their associated label variables  $\mathbf{l}_N = (l_1, l_2, \dots, l_N)$ , where  $l_{ng} = 1$  if observation  $n$  comes from group  $g$  and  $l_{ng} = 0$  otherwise. Let the unlabelled data be denoted by  $\mathbf{y}_M = (y_1, y_2, \dots, y_M)$  and their unknown labels (defined in a similar fashion to the known labels) be  $\mathbf{z}_M = (z_1, z_2, \dots, z_M)$ .

The observed-data likelihood is

$$L(\pi, \theta | \mathbf{x}_N, \mathbf{l}_N, \mathbf{y}_M) = \prod_{n=1}^N \prod_{g=1}^G [\pi_g f(x_n | \theta_g)]^{l_{ng}} \prod_{m=1}^M \sum_{g=1}^G \pi_g f(y_m | \theta_g),$$

where  $\theta_g$  are the parameters of the  $g$ th mixture component. Note that some elements of the  $\theta = (\theta_1, \theta_2, \dots, \theta_G)$  may be constrained.

The complete-data likelihood is

$$L_C(\pi, \theta | \mathbf{x}_N, \mathbf{l}_N, \mathbf{y}_M, \mathbf{z}_M) = \prod_{n=1}^N \prod_{g=1}^G [\pi_g f(x_n | \theta_g)]^{l_{ng}} \prod_{m=1}^M \prod_{g=1}^G [\pi_g f(y_m | \theta_g)]^{z_{mg}}.$$

The EM algorithm (Dempster, Laird, and Rubin (1977)) can be used to maximize the log-likelihood  $l(\pi, \theta) = \log L(\pi, \theta)$  to find maximum likelihood estimates for the unknown parameters (see Section 3.1). The fitted mixture model can be used to classify the observations with unknown labels (see Section 3.4). The CEM algorithm (Celeux and Govaert (1992)) can be used to maximize the complete-data likelihood and the estimates of  $\hat{\mathbf{z}}_M$  from the output can be used to classify the unlabelled samples (see Section 3.2).

### 3.1 The EM Algorithm

The EM algorithm for fitting the mixture model using labelled and unlabelled data involves the following steps:

0. Set  $k = 0$ . Find starting values by using the model-based discriminant analysis estimates of the parameters in the model. Call these estimates  $\hat{\pi}^{(0)}, \hat{\theta}^{(0)}$  (these can be found using `mc1ust`; Fraley and Raftery (1999), Fraley and Raftery (2003));
1. Calculate the expected value of the unknown labels by

$$\hat{z}_{mg}^{(k+1)} = \frac{\hat{\pi}_g^{(k)} f(y_m | \hat{\theta}_g^{(k)})}{\sum_{g'=1}^G \hat{\pi}_{g'}^{(k)} f(y_m | \hat{\theta}_{g'}^{(k)})} \text{ for } g = 1, 2, \dots, G \text{ and } m = 1, 2, \dots, M.$$

2. Using the data, the known labels and current estimates for the unknown labels (from step 1) estimate the parameters by

$$\hat{\pi}_g^{(k+1)} = \frac{\sum_{n=1}^N l_{ng} + \sum_{m=1}^M \hat{z}_{mg}^{(k+1)}}{N + M} \text{ for } g = 1, 2, \dots, G$$

$$\hat{\mu}_g^{(k+1)} = \frac{\sum_{n=1}^N l_{ng} \mathbf{x}_n + \sum_{m=1}^M \hat{z}_{mg}^{(k+1)} \mathbf{y}_m}{\sum_{n=1}^N l_{ng} + \sum_{m=1}^M \hat{z}_{mg}^{(k+1)}} \text{ for } g = 1, 2, \dots, G.$$

The estimation of  $\Sigma_g$  depends on the constraints put on the eigenvalue decomposition of the matrix. Details are given in Bensmail and Celeux (1996) and Celeux and Govaert (1995).

3. Check for convergence using the chosen stopping criterion (see Section 3.3). If converged; stop. If not, set  $k = k + 1$  and repeat steps 1 to 3.

### 3.2 The CEM Algorithm

The procedure for fitting mixture models using the CEM algorithm is almost identical to the EM algorithm, except that Step 1 is replaced by the following step:

1. Calculate the expected value of the unknown labels using the formula

$$w_{mg} = \frac{\hat{\pi}_g^{(k)} f(y_m | \hat{\theta}_g^{(k)})}{\sum_{g'=1}^G \hat{\pi}_{g'}^{(k)} f(y_m | \hat{\theta}_{g'}^{(k)})} \text{ for } g = 1, 2, \dots, G \text{ and } m = 1, 2, \dots, M,$$

and let

$$\hat{z}_{mg}^{(k+1)} = \begin{cases} 1, & \text{if } w_{mg} > w_{mg'} \text{ for all } g' \neq g \\ 0, & \text{otherwise} \end{cases}.$$

Note that the denominator of  $w_{mg}$  need not be calculated in practice.

That is, a discrete classification is made where each object is assigned to a unique group. For example, if Step 1 of the EM produces  $\hat{z}_m^{(k+1)} = (0.80, 0.05, 0.15)$  then the equivalent CEM step would transform this to  $\hat{z}_m^{(k+1)} = (1, 0, 0)$ . These values of  $\hat{z}_m^{(k+1)}$  are used in the subsequent calculations of the CEM algorithm.

### 3.3 Determining Convergence of EM/CEM Algorithms

We use the Aitken acceleration estimate of the final converged maximized log-likelihood to determine convergence of the EM algorithm (Böhning, Dietz, Schaub, Schlattmann, and Lindsay (1994)).

That is, at iteration  $k$ , an estimate of the final converged value of the log-likelihood was estimated using

$$l_\infty^{(k)} = l^{(k)} + \frac{1}{1 - a^{(k)}} (l^{(k+1)} - l^{(k)}), \text{ where } a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}},$$

where  $l^{(k)}$ ,  $l^{(k-1)}$  and  $l^{(k-2)}$  are the log-likelihood values on the last three iterations of the EM algorithm.

The EM algorithm was stopped when  $|l_\infty^{(t)} - l^{(t)}| < \epsilon = 10^{-5}$ , that is, when the current value of the log-likelihood is very close to the estimated final converged value; alternative stopping criteria using  $l_\infty^{(k)}$  are given in McLachlan and Peel (2000).

The CEM algorithm is stopped when the  $\hat{z}_{mg}^{(k)}$  values are equal on two consecutive iterations.

### 3.4 Classifying Observations

When the mixture model is fitted using the EM algorithm, the values of  $\hat{z}_m$  produced as part of the output are estimates of the posterior probability of group membership for each observation given the observed data value.

Observations can be classified into the group for which they have the maximum *a posteriori* group membership probability. That is, observation  $m$  is classified into group  $g$  if  $\hat{z}_{mg} > \hat{z}_{mg'}$  for all  $g' \neq g$ .

In the case where the mixture is fitted using the CEM algorithm, then the values of  $\hat{z}_{mg}$  provide a classification of the observations into groups.

### 3.5 Model Choice

Model-based clustering offers up to ten different covariance structures (Table 1). To decide which covariance structure to use, we select the model with the highest *BIC* value, where

$$BIC = 2 \times (\text{Maximized Log-Likelihood}) - 2 \log(n)(\text{Number of Parameters}).$$

For the data under consideration the EEE covariance structure was almost always chosen. While choosing the model with the highest BIC does not guarantee the highest correct classification rate, we found that, in practice, it gets very close to the best classification rate; this agrees with results contained in Biernacki and Govaert (1999).

When the model was fitted using the CEM algorithm, the model with the highest value for

$$2 \times (\text{Maximized Complete Data Log-Likelihood}) - 2 \log(n)(\text{Number of Parameters})$$

was selected.

## 4 Application In Food Authenticity Studies

Food authenticity studies are concerned with establishing if foodstuffs are actually what their labelling suggests them to be. For example, if a bottle of olive oil is labelled as being from Greece, then food authenticity studies are concerned with whether the oil is actually from Greece or from some other nation.

The proposed classification methods are applied to the analysis of spectra of foodstuffs recorded over the visible and near-infrared wavelength range (400–2498 nm) in food authenticity studies. The aim of the application is to classify foodstuffs using their spectra.

Two particular food authenticity problems are considered:

1. Using the spectra of raw homogenized meat samples to classify samples into individual species (Chicken, Turkey, Pork, Beef, Lamb); this work develops upon previous work by McElhinney, Downey, and Fearn (1999) who used other classification methods for this problem.
2. Using the spectra of Greek olive oils to classify the samples into their geographical origins; these data have previously been analyzed in Downey, McIntyre, and Davies (2003).

## 4.1 Near-infrared Spectra

The spectra for the meat samples were collected using the following spectroscopic process.

Combined visible and near infrared spectra were collected in reflectance mode using an NIRSystems 6500 instrument over the wavelength range 400–2498 nm at 2 nm intervals. Twenty five separate scans were collected during a single passage of the spectrophotometer and averaged, after which the sample mean spectrum of a reference ceramic tile (16 scans) were recorded and subtracted from the mean spectrum. Further details of how these data were collected are given in McElhinney, Downey, and Fearn (1999). Examples of spectra for the meat samples are shown in Figure 2.

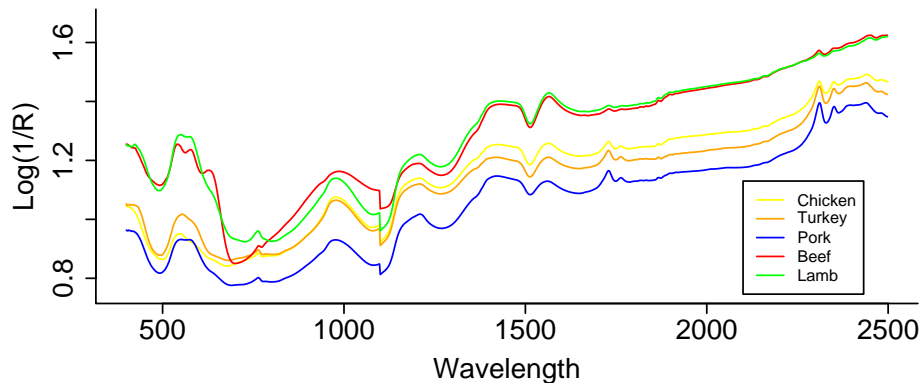


Figure 2: Examples of infrared spectra for each meat type. The discontinuity at 1100 nm is due to a change in detector at that wavelength.

The olive oil spectra were collected using a similar process which is described in Downey, McIntyre, and Davies (2003).

Detailed accounts of near-infrared spectroscopy and commonly used methods of analysis for these spectra are given in Næs, Isaksson, Fearn, and Davies (2002) and Osborne, Fearn, and Hindle (1993).

## 4.2 Data Reduction

The near-infrared spectrum for a particular food sample records 1050 reflectance values; this means that we have very high dimensional data. We can view the spectrum as recording the heights of a continuous function at discrete points. We propose reducing the dimension of the data by approximating each spectrum using a lower dimensional representation. We propose using the wavelet decomposition for each spectrum for this data reduction step.

### 4.2.1 Wavelets

Wavelets are functions that can express the frequency and time/space localization of a curve (for example, a spectrum). Wavelets are twinned with dual functions called scaling functions. Scaling function coefficients can be expressed as functions of wavelet coefficients and vice versa which means that the algorithms used for wavelet decompositions are quick and efficient. The wavelet coefficients can be viewed as the detail removed from the scaling function decomposition of smoothed data.

One requirement for the pyramidal algorithm (Mallat (1989)) for completing the discrete wavelet transform is that the length of the data vector must be a power of 2. In this analysis, the first and last 13 wavelengths in the data were removed and the remaining 1024 spectral values were used (424–2472 nm).

The discrete wavelet transformation of the spectrum does not reduce the dimension of the data per se, however most of the wavelet coefficients tend to be very small and only a few wavelet coefficients capture the overall shape of the spectrum (see Figure 3).

The wavelet used in our analysis was the Daubechies orthonormal compactly supported extremal phase family of wavelets (Daubechies (1988), Daubechies (1992)). The boundary handling was periodic and the algorithm used was Mallat’s pyramid algorithm. The wavelet decomposition was completed using the `wavethresh` library in R (Nason (1993), Nason and Silverman (1994)). A good introduction to wavelet decomposition is given in Ogden (1997).

We consider a method for selecting a low-dimensional subset of the wavelet coefficients based on a slight variant of wavelet thresholding (Section 4.2.2).

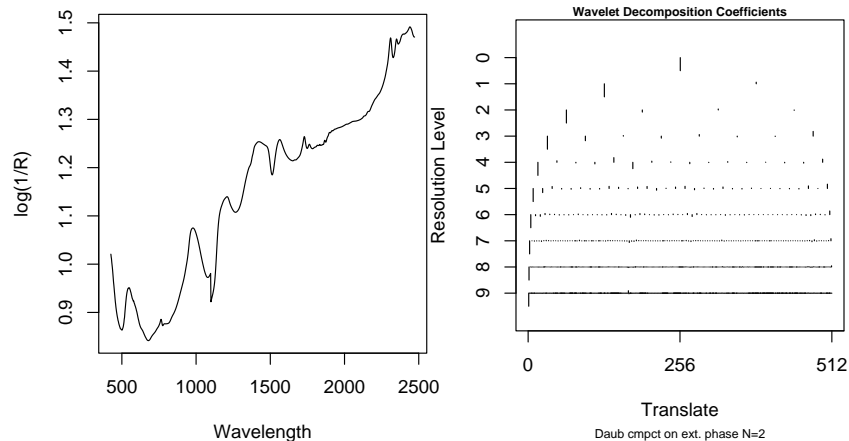


Figure 3: A plot of an infrared spectrum and the wavelet coefficients. The coefficients near the top of the plot capture the coarse features of the data whereas those near the bottom of the plot capture the fine structure.

#### 4.2.2 Wavelet Thresholding

Wavelet thresholding (Donoho and Johnstone (1994)) of a single spectrum would select those wavelet coefficients that exceed a threshold and set the remaining coefficients equal to zero. The universal hard thresholding policy sets the threshold level at  $\sqrt{2 \log(m)} \hat{\sigma}$  where  $\hat{\sigma}$  is a robust estimate of the scale of the coefficients. Other thresholding possibilities are described in Ogden (1997).

We cannot use thresholding directly for the current application because the thresholding policy may select different coefficients for the different spectra. We used a slight variant on thresholding where coefficients that were selected by universal thresholding for *any* spectrum were not set to zero. Therefore, the only coefficients set equal to zero were those not selected by universal thresholding for any spectrum.

An example of a spectrum reconstructed from the coefficients after thresholding is given in Figure 4.

### 4.3 Meat Samples

Two hundred and thirty one homogenized meat samples were used in the study (55 Chicken, 55 Turkey, 55 Pork, 32 Beef and 34 Lamb). Combined visible and near infrared spectra were collected in reflectance mode over the wavelength range 400–2498 nm at 2 nm intervals.

To assess the performance of the proposed updating classification procedures for classify-

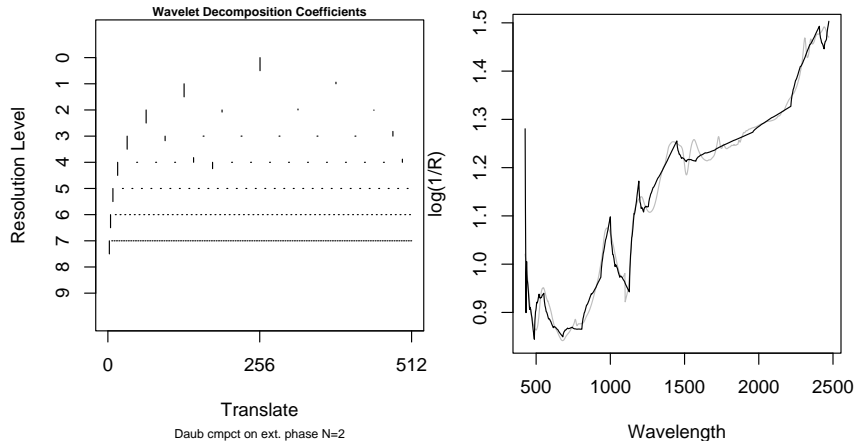


Figure 4: A plot of selected wavelet coefficients after thresholding. The wavelet reconstruction of the spectrum from these coefficients and the original spectrum are also included. The dark line is the wavelet reconstruction of the spectrum and the grey line is the original spectrum.

ing the meat samples, we divided the data into a training (labelled) sample and a validation (unlabelled) sample. We investigated the effect of there being differing proportions of the data in the training and validation samples.

Model-based discriminant analysis (DA), updating the classification rule using the EM algorithm (EM) and updating using the CEM algorithm (CEM) were used to find a classification rule and to classify the unlabelled samples. The results of this analysis are given in Table 2.

Table 2: Average correct classification rates for the five meat groups (after data reduction using thresholding) for 100 random splits into training and validation data. Model selection was completed using BIC. Standard deviations are given in parentheses.

DATA	OBSERVATIONS	DA	EM	CEM
Training	(28,28,28,16,17)	98.8% (0.8)	97.5% (1.0)	97.5% (1.1)
Validation	(27,27,27,16,17)	94.2% (1.9)	94.4% (2.0)	94.4% (2.0)
Training	(14,14,14,8,9)	99.7% (0.7)	97.5% (1.6)	97.5% (1.7)
Validation	(41,41,41,24,25)	90.7% (2.6)	93.2% (1.8)	92.4% (2.2)
Training	(7,7,7,7,7)	100% (0.0)	97.7% (2.2)	97.8% (2.2)
Validation	(48,48,48,25,27)	77.2% (7.1)	90.9% (4.7)	90.9% (4.7)

Clearly, all of the classification methods achieve good correct classification rates when

the number of observations in the training sample is equal to the number in the validation sample. The correct classification rate of 94.4% for the two updating procedures (EM and CEM) is slightly higher than for model-based discriminant analysis (DA). However, when the number of training sample observations is very small, then the updating procedures (EM and CEM) have much higher correct classification rates 90.9% compared with 77.2% for DA.

McElhinney, Downey, and Fearn (1999) report a best correct classification rate, when the training and validation samples sizes were equal, of 92.17% using factorial discriminant analysis after scatter correction using second derivatives and restricting only to wavelengths from 400-1100 nm. The best results achieved on the whole spectrum without scatter correction were a correct classification rate of 86.10%. Our method achieves a considerable improvement over the classification rates in this case.

We can look closer at the classification rates to see what types of misclassifications are happening. To do this, for each meat type we examined what percentage of classifications were attributed to the different meat types (Table 3). Clearly, most of the misclassifications are due to misclassifying chicken as turkey and to a lesser extent classifying turkey as chicken and beef as lamb.

Table 3: The classifications of the five group meats data in the case where there were the training data consisted of (28,28,28,16,17) observations and the validation data consisted of (27,27,27,16,17) observations.

		CLASSIFICATION					
	Chicken	Turkey	Pork	Beef	Lamb	TOTAL	
Chicken	81.7%	16.7%	1.6%			100%	
Turkey	3.0%	96.8%	0.2%			100%	
Pork		0.1%	99.9%			100%	
Beef				96.6%	3.4%	100%	
Lamb				0.4%	99.6%	100%	

One proposal adopted, in McElhinney, Downey, and Fearn (1999), is to combine the chicken and turkey into a poultry group and to investigate the performance of the classifications on this reduced problem.

We repeated our analysis with this reduced data and the correct classification rates are given in Table 4. Results showing the misclassifications that occur for this classification problem are given in Table 5.

We can see that the classification rates are greatly improved when we consider the four

group problem. The updating procedure (EM) achieved a correct classification rate of 99.0% when the training and validation sample sizes were equal and a rate of 95.3% when there was very little training data. These classification rates compare well with McElhinney, Downey, and Fearn (1999) whose best rate was 97.39% after scatter correction using second derivatives and only wavelengths from 400-1100 nm. Their best correct classification rate using the whole spectrum without scatter correction was 95.65%. Again, our results compare very favorably with these.

Table 4: Average correct classification rates for the four meat groups (after data reduction using thresholding) for 100 random splits into training and validation data. Model selection was completed using BIC. Standard deviations are given in parentheses.

DATA	OBSERVATIONS	DA	EM	CEM
Training	(55,28,16,17)	100.0% (0.2)	99.6% (0.5)	99.6% (0.5)
Validation	(55,27,16,17)	98.8% (1.1)	99.0% (0.9)	98.9% (1.1)
Training	(28,14,8,9)	100.0% (0.2)	99.8% (0.5)	99.8% (0.5)
Validation	(82,41,24,25)	97.2% (1.6)	98.3% (1.1)	98.3% (1.1)
Training	(8,8,8,8)	100.0% (0.0)	99.3% (1.6)	99.3% (1.6)
Validation	(102,47,24,26)	80.0% (9.2)	95.3% (6.1)	95.3% (6.1)

Table 5: The classifications of the four group meats data in the case where there were the training data consisted of (55,28,16,17) observations and the validation data consisted of (55,27,16,17) observations.

CLASSIFICATION					
	Poultry	Pork	Beef	Lamb	TOTAL
Poultry	99.1%	0.9%			100%
Pork		100.0%			100%
Beef			96.6%	3.4%	100%
Lamb		0.2%	0.3%	99.5%	100%

The above results show the utility of updating classification rules using unlabelled data within the model-based clustering framework. The EM updating procedure offers a small advantage over the CEM updating procedure in terms of classification rates. We also found that the EM procedure is more stable than the CEM procedure when the number of training sample observations is small.

## 4.4 Greek Olive Oils

Sixty five samples of virgin olive oil were collected from different locations in Greece. There were 18 samples from Crete, 28 samples from Peloponese and 19 from some other locations. A near infrared spectrum was recorded for each sample, giving reflectance values for wavelengths in the range 400–2098 nm in 2 nm intervals. Further details on the data collection are given in Downey, McIntyre, and Davies (2003).

In this case, we wish to investigate the use of near infrared spectra to classify the olive oils according to geographical origin.

We, again, compared the use of model-based discriminant analysis (DA) and the EM and CEM updating procedures. Results of this investigation are given in Table 6.

Table 6: Average correct classification rates for the three olive oil groups (after data reduction using thresholding) for 100 random splits into training and validation data. Model selection was completed using BIC. Standard deviations are given in parentheses.

DATA	OBSERVATIONS	DA	EM	CEM
Training	(9,14,9)	99.1% (1.7)	97.7% (2.0)	97.7% (2.0)
Validation	(9,14,10)	87.3% (5.5)	88.1% (6.3)	88.1% (6.3)
Training	(7,7,7)	99.7% (1.3)	97.6% (3.1)	97.4% (3.2)
Validation	(11,21,12)	73.2% (10.6)	80.2% (9.4)	80.2% (9.4)

We find that the updating procedure gives better classification results especially when the number of training observations is small. The best classification rate when the number of training observations was equal to the number of validation observations was 88.1%. Downey, McIntyre, and Davies (2003) report a classification rate of 93.9% when the training and validation samples are of equal size and factorial discriminant analysis is used, this correct classification rates is better than our classification rate. However, the other methods that Downey, McIntyre, and Davies (2003) used gave classification rates of under 81% when the whole near infrared spectrum was used.

## 5 Conclusions

In general, the two approaches to updating the classification rule provide an improvement over the ordinary discriminant analysis in terms of classification rates. If there is a lot of labelled data, then the gain over ordinary discriminant analysis is not very large. However, if

there is very little labelled data, then the classification rates are superior for the approaches that update the classification rule.

The updating approach that maximizes the likelihood and uses the fitted model to classify the unlabelled samples has superior classification rates over the method that maximizes the complete-data likelihood. However, if the mixture components are well separated, then there is little to choose between the two methods.

Wavelet decomposition followed by selection of coefficients provided a very quick method of data reduction. The variant of universal thresholding approach offers an automatic method of selecting the wavelet coefficients. The number of coefficients retained by the proposed thresholding procedure restricted the range of models that could be considered in the model-based clustering framework; this problem was most severe for model-based discriminant analysis. However, for larger sample sizes this would not be a problem.

Model selection using BIC selected models with very good classification rates. In most cases the BIC selected model chose the model with the highest classification rate. The value for the BIC of a particular model can be computed very quickly from the fitted mixture model.

It appears that for spectroscopic data that we used, the EEE model performs very well; this is the model that assumes equal, but general, covariances for all mixture components. Therefore, for these applications, the best updated classification rule has similar modelling assumptions to linear discriminant analysis.

## References

- Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49(3), 803–821.
- Bensmail, H. and G. Celeux (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association* 91, 1743–1748.
- Biernacki, C. and G. Govaert (1999). Choosing models in model-based clustering and discriminant analysis. *J. Statistical Computation and Simulation* 14, 49–71.
- Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. Lindsay (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics* 46, 373–388.

- Celeux, G. and G. Govaert (1992). A classification EM algorithm for clustering and two stochastic versions. *Comput. Statist. Data Anal.* 14(3), 315–332.
- Celeux, G. and G. Govaert (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.
- Daubechies, I. (1992). *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 39(1), 1–38. With discussion.
- Donoho, D. L. and I. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81(3), 425–455.
- Downey, G., P. McIntyre, and A. N. Davies (2003). Geographical classification of extra virgin olive oils from the eastern Mediterranean by chemometric analysis of visible and near infrared spectroscopic data. *Appl. Spectrosc.* 57(2), 158–163.
- Fraley, C. and A. E. Raftery (1998). How many clusters? which clustering method? - answers via model-based cluster analysis. *Computer Journal* 41, 578–588.
- Fraley, C. and A. E. Raftery (1999). Mclust: Software for model-based clustering. *Journal of Classification* 16, 297–306.
- Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* 97(458), 611–631.
- Fraley, C. and A. E. Raftery (2003). Enhanced software for model-based clustering, discriminant analysis, and density estimation: MCLUST. *Journal of Classification* 20, 263–286.
- Ganesalingam, S. and G. J. McLachlan (1978). The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika* 65(3), 658–662.
- Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, New York-London-Sydney. Wiley Series in Probability and Mathematical Statistics.
- Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11,

674–693.

- McElhinney, J., G. Downey, and T. Fearn (1999). Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. *J. Near Infrared Spectrosc.* 7, 145–154.
- McLachlan, G. J. (1975). Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *J. Amer. Statist. Assoc.* 70, 365–369.
- McLachlan, G. J. (1977). Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. *J. Amer. Statist. Assoc.* 72(358), 403–406.
- McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.
- McLachlan, G. J. and K. E. Basford (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker Inc.
- McLachlan, G. J. and S. Ganesalingam (1982). Updating a discriminant function on the basis of unclassified data. *Comm. Statist. B—Simulation Comput.* 11, 753–767.
- McLachlan, G. J. and D. Peel (2000). *Finite Mixture models*. New York: John Wiley & Sons.
- Næs, T., T. Isaksson, T. Fearn, and T. Davies (2002). *A user-friendly guide to Multivariate Calibration and Classification*. Chichester: NIR Publications.
- Nason, G. P. (1993). The `wavethresh` package: wavelet transform and thresholding software for s.
- Nason, G. P. and B. W. Silverman (1994). The discrete wavelet transform in s. *J. Comp. Graph. Statist.* 3, 163–191.
- Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Boston: Birkhäuser.
- O’Neill, T. J. (1978). Normal discrimination with unclassified observations. *J. Amer. Statist. Assoc.* 73(364), 821–826.
- Osborne, B. G., T. Fearn, and P. H. Hindle (1993). *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis*. Harlow, UK: Longman Scientific & Technical.

Titterington, D. M. (1976). Updating a diagnostic system using unconfirmed cases. *Appl. Statist.* 25(3), 238–247.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical analysis of finite mixture distributions*. Chichester: Wiley.