# Variable Selection and Updating In Model-Based Discriminant Analysis for High-Dimensional Data*

Thomas Brendan Murphy
School of Mathematical Sciences
University College Dublin, Ireland

Nema Dean
Department of Statistics
University of Glasgow, Scotland

Adrian E. Raftery
Department of Statistics
University of Washington, Seattle, USA.

June 5, 2008

**Abstract**

A model-based discriminant analysis method that includes variable selection is presented. The discriminant analysis model is fitted in a semi-supervised manner using both labeled and unlabeled data. The method is shown to give excellent classification performance on several high-dimensional multiclass datasets with more variables than observations. The variables selected by the proposed method provide information about which variables are meaningful for classification purposes. A headlong search strategy for variable selection is shown to be efficient in terms of computation and achieves excellent classification performance. In applications to several food classification datasets, our proposed method outperformed default implementations of Random Forests, AdaBoost and Bayesian Multinomial Regression by substantial margins.
**Keywords:** Headlong search, model-based discriminant analysis, normal mixture models, semi-supervised learning, updating classification rules, variable selection.

*Murphy was supported by Science Foundation of Ireland Basic Research Grant (04/BR/M0057) and Research Frontiers Programme Grant (2007/RFP/MATF281). Raftery was supported by NICHD grant R01 HD054511 and NSF grant ATM 0724721. All three authors were supported by NIH grant 8 R01 EB002137-02.

# 1 Introduction

Discriminant analysis is used to classify observations into predefined groups. Typically, a discriminant function is developed using observations with known group membership and this is then used to classify observations with unknown group membership.

Model-based discriminant analysis (Bensmail and Celeux 1996; Fraley and Raftery 2002) provides a framework for discriminant analysis based on parsimonious normal mixture models. This approach to discriminant analysis has been shown to be effective in practice and being based on a statistical model it allows for uncertainty to be treated appropriately.

In many applications, only a subset of the variables in a discriminant analysis contain any group membership information and including variables which have no group information increases the complexity of the analysis, potentially degrading the classification performance. Therefore, there is a need for including variable selection as part of any discriminant analysis procedure.

Variable selection can be completed as a preprocessing step prior to discriminant analysis (a filter approach) or as part of the analysis procedure (a wrapper approach). Completing variable selection prior to the discriminant analysis can lead to variables that have poor individual classification performance being omitted from the subsequent analysis. However, such variables could be important for classification purposes when jointly considered with others. Hence, performing variable selection as part of the discriminant analysis procedure is preferred.

Combining variable selection and linear or quadratic discriminant analysis has been considered previously in the literature; see McLachlan (1992, Chapter 12) for a review. Many of these methods are based on measuring the Mahalanobis distance between groups before and after the inclusion of a variable into the discriminant analysis model. In the machine learning literature, Kohavi and John (1997) developed a *wrapper* approach for combining variable selection in *supervised* learning, of which discriminant analysis is a special case.

Variable selection is of particular importance in situations where there are more variables than observations available; that is, large $p$, small $n$ ($n \ll p$) problems (West 2003). These situations arise with increasing frequency in statistical applications, including genetics, proteomics, image processing and food science. The two food science applications considered in Section 2 involve data sets with more variables than observations.

In this paper, a version of model-based discriminant analysis is developed by adapt-

ing the model-based clustering with variable selection method of Raftery and Dean (2006). This method of discriminant analysis builds a discriminant rule in a stepwise manner by considering the inclusion of extra variables into the model and also considering removing existing variables from the model based on their importance for classification. The procedure is iterated until convergence.

A brief review of model-based clustering and discriminant analysis is given in Section 3. The underlying model for model-based clustering with variable selection is reviewed in Section 3.1 and this model is extended to model-based discriminant analysis with variable selection in Section 3.2. In Section 3.3, the fitting of the discriminant analysis model is extended to incorporate semi-supervised updating using both the labeled and unlabeled observations (Dean et al. 2006) in order to improve the classification performance.

Search strategies for selecting the variables for inclusion and exclusion are discussed in Section 3.4. A headlong search strategy is proposed that combines good classification performance and computational efficiency. The proposed methodology is applied to the high dimensional datasets in Section 4 and the methodology and results are discussed in Section 5.

# 2 Data

## 2.1 Food Authenticity Studies

An authentic food is one that is what it claims to be. Important aspects of food description include its process history, geographic origin, species/variety and purity. Food producers, regulators, retailers and consumers need to be assured of the authenticity of food products.

Food authenticity studies are concerned with establishing whether foods are authentic or not. Many analytical chemistry techniques are used in food authenticity studies, including gas chromatography, mass spectroscopy, and vibrational spectroscopic techniques (Raman, ultraviolet, mid-infrared, near-infrared and visible). All of these techniques have been shown to be capable of discriminating between certain sets of similar biological materials. Downey (1996) provides a review of food authenticity studies with an emphasis on the use of near-infrared spectroscopy in these studies.

We consider two food authenticity data sets which consist of near-infrared spectroscopic measurements from food samples of different types. The aim of the food authenticity study is to classify the food samples into known groups.

Near-infrared spectroscopy (NIR) is used as a quick and efficient method of collect-

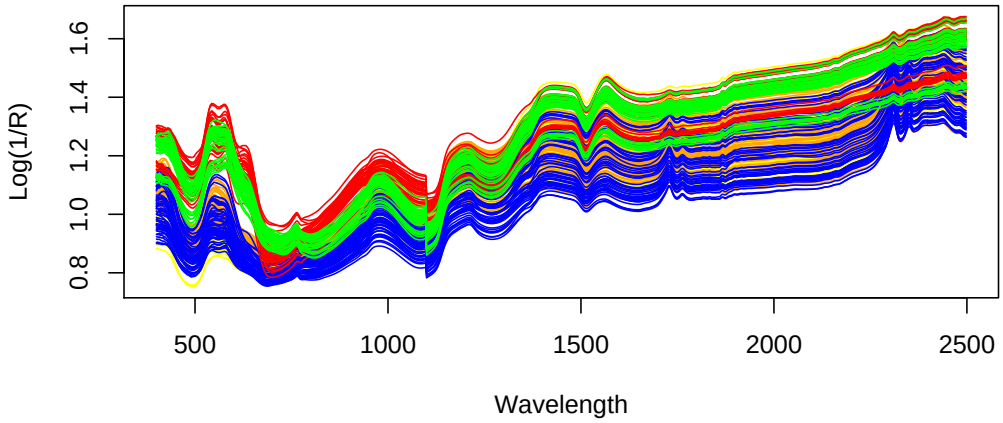


Figure 1: The near-infrared spectra recorded for all of the 231 meat samples in the study. The discontinuity at 1100 nm is due to a sensor change at that value. The samples are colored as Beef=red, Lamb=green, Pork=blue, Turkey=orange, Chicken=yellow.

ing data for use in food authenticity studies (Downey 1996). Two particular authenticity studies are considered in Sections 2.2 and 2.3:

- Classifying meats into species (Beef, Chicken, Lamb, Pork, Turkey)
- Classifying olive oils into geographic origin (Crete, Peloponese, Other).

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. The visible portion of the spectrum is the range 400–800 nm and the near-infrared region is the range 800–2498 nm.

For the meat samples, twenty five separate scans were collected during a single passage of the spectrophotometer and averaged, after which the mean spectrum of a reference ceramic tile (16 scans) was recorded and subtracted from the mean spectrum. A similar process was used for the olive oil data, but fewer scans were used.

## 2.2 Homogenized Meat Data

A total of 231 homogenized meat samples were collected for this study (55 Chicken, 55 Turkey, 55 Pork, 32 Beef and 34 Lamb). Details of the data collection process are given in McElhinney et al. (1999).

Each spectrum consists of reflectance readings at different wavelengths and consists of 1050 reflectance measurements. The spectra of all the samples are shown in Figure 1.


Figure 2: Regions of Greece where the olive oil samples were collected.

### 2.3 Greek Olive Oils Data

A total of 65 olive oil samples were collected from three different regions in Greece (18 Crete, 28 Peloponese, 19 Other). Each data value consists of 1050 reflectance values over the visible and near-infrared range. Here we want to use the spectra to determine the geographical origin (see Figure 2) of the oils. Details of these data and a previous analysis are given in Downey et al. (2003).

## 3 Model-based Clustering and Discriminant Analysis

Model-based clustering (Banfield and Raftery 1993; Fraley and Raftery 1998, 2002; McLachlan and Peel 2000) uses mixture models as a framework for cluster analysis. The underlying model in model-based clustering is a normal mixture model with $G$ components, that is,

$$f(\mathbf{x}) = \sum_{g=1}^{G} \tau_g f(\mathbf{x}|\mu_g, \Sigma_g),$$

where $f(\cdot|\mu_g, \Sigma_g)$ is a multivariate normal density with mean $\mu_g$ and covariance $\Sigma_g$.

A central idea in model-based clustering is the use of constraints on the group covariance matrices $\Sigma_g$; these constraints use the eigenvalue decomposition of the covariance matrices to impose shape restrictions on the groups. The decomposition is of the form, $\Sigma_g = \lambda_g D_g A_g D_g^T$, where $\lambda_g$ is the largest eigenvalue, $D_g$ is an orthonormal matrix of eigenvectors, and $A_g$ is a diagonal matrix of scaled eigenvalues. Interpretations for the parameters in the covariance decomposition are: $\lambda_g$= Volume; $A_g$= Shape; $D_g$= Orientation. These parameters can be constrained in various ways to be

Table 1: Constrained covariance structures in model-based clustering as implemented in the `mclust` package for R.

| **ModelID** | **Volume** | **Shape** | **Orientation** | **Covariance** ($\Sigma_g$) |
|---|---|---|---|---|
| EII | Equal | Equal Spherical | NA | $\lambda I$ |
| VII | Variable | Equal Spherical | NA | $\lambda_g I$ |
| EEI | Equal | Equal | Axis Aligned | $\lambda A$ |
| VEI | Variable | Equal | Axis Aligned | $\lambda_g A$ |
| EVI | Equal | Variable | Axis Aligned | $\lambda A_g$ |
| VVI | Variable | Variable | Axis Aligned | $\lambda_g A_g$ |
| EEE | Equal | Equal | Equal | $\lambda DAD^T$ |
| EEV | Equal | Equal | Variable | $\lambda D_g A D_g^T$ |
| VEV | Variable | Equal | Variable | $\lambda_g D_g A D_g^T$ |
| VVV | Variable | Variable | Variable | $\lambda_g D_g A_g D_g^T$ |

equal or variable across groups. Additionally, the shape and orientation matrices can be set equal to the identity matrix.

Bensmail and Celeux (1996) developed model-based discriminant analysis methods using the same covariance decomposition. An extension of model-based discriminant analysis that allows for updating of the classification rule using the unlabeled data was developed by Dean et al. (2006) and will be described in further detail in Section 3.3. Model-based clustering and discriminant analysis can be implemented in the statistics package R (R Development Core Team 2007) using the `mclust` package (Fraley and Raftery 1999, 2003, 2007).

## 3.1 Model-based Clustering with Variable selection

We argue that variable selection needs to be part of the discriminant analysis procedure, because completing variable selection prior to discriminant analysis may lose important grouping information. This argument is supported by the result of Chang (1983), who showed that the principal components corresponding to the larger eigenvalues do not necessarily contain information about group structure. This suggests that the commonly used filter approach of selecting the first few principal components to explain a minimum percentage of variation can be suboptimal. A similar argument can be made that selecting discriminating variables without reference to the grouping variable may miss important variables. In addition, some variables may contain strong group information when used in combination with other variables, but not on their own. Another critique of completing a variable (or feature) selection step before supervised

learning was given by Kohavi and John (1997, Section 2.4).

Raftery and Dean (2006) developed a version of model-based clustering that includes variable selection. With their method, variables are selected in a stepwise manner. Their method involves the stages:

- Find the variable with the greatest evidence of clustering given the already selected variables and add it to the set of chosen variables.
- Find the variable with the least evidence of clustering from the set of selected variables and remove it from the set of selected variables if it no longer has evidence of clustering.

This process is iterated until no further variables are added or removed. This approach, that combines variable selection and cluster analysis, avoids the problems of completing variable selection independently of the clustering.

## 3.2 Model-based Discriminant Analysis With Variable Selection

We adapt the ideas of Raftery and Dean (2006) to produce a discriminant analysis technique that includes variable selection. This discriminant analysis method uses a stepwise variable selection procedure to find a subset of variables that gives good classification results.

Each stage of the algorithm involves two steps:

- Determine if a variable (not already selected) would contribute to the discriminant analysis model. If the variable improves the model, it is added to the model; the procedure for searching for variables to add to the model is given in Section 3.4
- Determine if any selected variables should be removed from the discriminant analysis model. If a selected variable does not contribute to the model, then it is removed; the procedure for searching the variables to remove from the model is given in Section 3.4.

Let $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ be the observed data values and let $(\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n)$ be the group indicator variables for these observations where $l_{ig} = 1$ if observation $i$ belongs to group $g$ and $l_{ig} = 0$ otherwise.

Suppose that the observation $\mathbf{x}_i$ is partitioned into three parts: $\mathbf{x}_i^{(c)}$ are the variables already chosen; $\mathbf{x}_i^{(p)}$ is the variable being proposed; $\mathbf{x}_i^{(o)}$ are the remaining variables. The decision on whether to include or exclude a proposed variable is based ade on the comparison of two models:

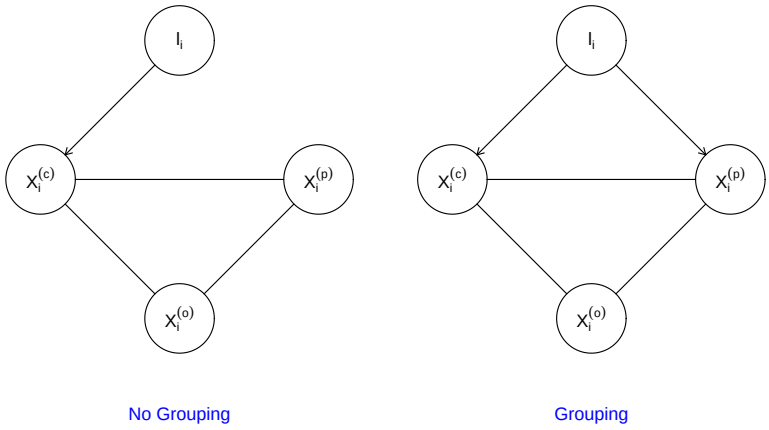


Figure 3: A graphical model representation of the Grouping and the No Grouping models.

- Grouping: $p(\mathbf{x}_i|\mathbf{l}_i) = p(\mathbf{x}_i^{(c)}, \mathbf{x}_i^{(p)}, \mathbf{x}_i^{(o)}|\mathbf{l}_i) = p(\mathbf{x}_i^{(o)}|\mathbf{x}_i^{(p)}, \mathbf{x}_i^{(c)})p(\mathbf{x}_i^{(p)}, \mathbf{x}_i^{(c)}|\mathbf{l}_i)$.
- No Grouping: $p(\mathbf{x}_i|\mathbf{l}_i) = p(\mathbf{x}_i^{(c)}, \mathbf{x}_i^{(p)}, \mathbf{x}_i^{(o)}|\mathbf{l}_i) = p(\mathbf{x}_i^{(o)}|\mathbf{x}_i^{(p)}, \mathbf{x}_i^{(c)})p(\mathbf{x}_i^{(p)}|\mathbf{x}_i^{(c)})p(\mathbf{x}_i^{(c)}|\mathbf{l}_i)$.

Figure 3 shows the difference between the "Grouping" and "No Grouping" models for $\mathbf{x}_i$. If the Grouping model holds, $\mathbf{x}_i^{(p)}$ provides information about which group the data value belongs to beyond that provided by $\mathbf{x}_i^{(c)}$, while if the No Grouping model holds, $\mathbf{x}_i^{(p)}$ provides no extra information.

The Grouping and No Grouping models are specified as follows:

- Grouping: We let $p(\mathbf{x}_i^{(p)}, \mathbf{x}_i^{(c)}|\mathbf{l}_i)$ be a normal density with parsimonious covariance structure as described in Table 1. That is,

$$\begin{aligned}(\mathbf{x}_i^{(p)}, \mathbf{x}_i^{(c)})|(l_{ig} = 1) &\sim N(\mu_g^{(p,c)}, \Sigma_g^{(p,c)}),\\ \mathbf{l}_i &\sim \text{Multinomial}(1, \tau).\end{aligned}$$

- No Grouping: We let $p(\mathbf{x}_i^{(c)}|\mathbf{l}_i)$ be a normal density with parsimonious covariance structure. In addition, $p(\mathbf{x}_i^{(p)}|\mathbf{x}_i^{(c)})$ is assumed to have a linear regression model structure. That is,

$$\begin{aligned}\mathbf{x}_i^{(c)}|(l_{ig} = 1) &\sim N(\mu_g^{(c)}, \Sigma_g^{(c)}),\\ \mathbf{l}_i &\sim \text{Multinomial}(1, \tau),\\ \mathbf{x}_i^{(p)}|\mathbf{x}_i^{(c)} &\sim N(\alpha + \beta^T \mathbf{x}_i^{(c)}, \sigma^2).\end{aligned}$$

The same model structure is assumed for $p(\mathbf{x}_i^{(o)}|\mathbf{x}_i^{(c)}, \mathbf{x}_i^{(p)})$ in the Grouping model as in the No Grouping model. Therefore, this part of the model does not influence the choice to include $\mathbf{x}_i^{(p)}$ in the model or not.

The decision as to whether the Grouping or No Grouping model is appropriate is made using the BIC approximation of the log Bayes factor. The logarithm of the Bayes factor is

$$\log(\text{Bayes Factor}) = \log \frac{p(\mathbf{x}_i|\mathcal{M}_G)}{p(\mathbf{x}_i|\mathcal{M}_{NG})}, \tag{1}$$

where $\mathcal{M}_G$ is the Grouping model, $\mathcal{M}_{NG}$ is the No Grouping model and

$$p(\mathbf{x}_i|\mathcal{M}_k) = \int p(\mathbf{x}_i|\theta_k, \mathcal{M}_k) p(\theta_k|\mathcal{M}_k) d\theta_k$$

is the integrated likelihood of model $\mathcal{M}_k$. We use the BIC approximation of the integrated likelihood in the form

$$\text{BIC } = \log \text{ maximized likelihood} - \frac{d}{2}\log(n),$$

where $d$ is the number of parameters in the model and $n$ is the sample size (Schwarz 1978). Following Raftery and Dean (2006), the log Bayes factor (1) can be reduced to

$$\begin{aligned} \log(\text{Bayes Factor}) &= \log \frac{p(\mathbf{x}_i^{(p)}|\mathbf{x}_i^{(c)}, \mathcal{M}_G) p(\mathbf{x}_i^{(c)}|\mathcal{M}_G)}{p(\mathbf{x}_i^{(p)}, \mathbf{x}_i^{(c)}|\mathcal{M}_{NG})} \\ &\approx \text{BIC(Grouping)-BIC(No Grouping)}, \end{aligned} \tag{2}$$

which only involves $(\mathbf{x}_i^{(c)}, \mathbf{x}_i^{(p)})$ and not $\mathbf{x}_i^{(o)}$. Variables with a positive difference in BIC(Grouping)-BIC(No Grouping) are candidates for being added to the model.

At each stage, we also check if an already chosen variable should be removed from the model. This decision is made on the basis of the BIC difference in a similar way to previously. In this case, $\mathbf{x}_i^{(p)}$ takes the role of the variable to be dropped, $\mathbf{x}_i^{(c)}$ takes the role of the remaining chosen variables and $\mathbf{x}_i^{(o)}$ are the other variables. The variables with a positive difference in BIC(No Grouping)-BIC(Grouping) are candidates for removal from the model.

## 3.3 Discriminant Analysis with Updating

In standard discriminant analysis, the unlabeled data is not used in the model fitting procedure. However, these data contain information that is potentially important, especially when very few labeled data values are available. We can model both the labeled and unlabeled data as coming from the same model, but where the unlabeled data is missing the labeling variable; this leads to a mixture model for the unlabeled data. Hence, the unlabeled data can then be used to help fit a model to the data. This idea has been investigated by many authors including Ganesalingam and McLachlan (1978) and O'Neill (1978) and more recently by Dean et al. (2006), Chapelle et al. (2006), Toher et al. (2007) and Liang et al. (2007).

Let $(\mathbf{x}_1, \mathbf{l}_1), (\mathbf{x}_2, \mathbf{l}_2), \ldots, (\mathbf{x}_N, \mathbf{l}_N)$ be the labeled data and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M$ be the unlabeled data. We let $\underline{\mathbf{z}} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M)$ be the unobserved (missing) labels for the unlabeled data. In this framework, the Grouping and No Grouping models for the observed data are of the form:

- Grouping: We let $p(\mathbf{x}_i^{(p)}, \mathbf{x}_i^{(c)} | \mathbf{l}_i)$ be a normal density with parsimonious covariance structure as described in Table 1, namely

$$\begin{aligned} (\mathbf{x}_i^{(p)}, \mathbf{x}_i^{(c)}) | (l_{ig} = 1) &\sim N(\mu_g^{(p,c)}, \Sigma_g^{(p,c)}), \\ \mathbf{l}_i &\sim \text{Multinomial}(1, \tau). \end{aligned}$$

  Also, $p(\mathbf{y}_j^{(p)}, \mathbf{y}_j^{(c)})$ is a mixture of normals with parsimonious covariance structures, namely

$$(\mathbf{y}_j^{(p)}, \mathbf{y}_j^{(c)}) \sim \sum_{g=1}^{G} \tau_g N(\mu_g^{(p,c)}, \Sigma_g^{(p,c)})$$

- No Grouping: We let $p(\mathbf{x}_i^{(c)} | \mathbf{l}_i)$ be a normal density with parsimonious covariance structure, namely

$$\begin{aligned} \mathbf{x}_i^{(c)} | (l_{ig} = 1) &\sim N(\mu_g^{(c)}, \Sigma_g^{(c)}), \\ \mathbf{l}_i &\sim \text{Multinomial}(1, \tau). \end{aligned}$$

  We also let $p(\mathbf{y}_j^{(c)})$ be a mixture of normal densities with parsimonious covariance structure, namely

$$\mathbf{y}_j^{(c)} \sim \sum_{g=1}^{G} \tau_g N(\mu_g^{(c)}, \Sigma_g^{(c)}).$$

  In addition, we assume a linear regression model for $p(\mathbf{x}_i^{(p)} | \mathbf{x}_i^{(c)})$ and $p(\mathbf{y}_j^{(p)} | \mathbf{y}_j^{(c)})$, namely

$$\mathbf{x}_i^{(p)} | \mathbf{x}_i^{(c)} \sim N(\alpha + \beta^T \mathbf{x}_i^{(c)}, \sigma^2) \text{ and } \mathbf{y}_j^{(p)} | \mathbf{y}_j^{(c)} \sim N(\alpha + \beta^T \mathbf{y}_j^{(c)}, \sigma^2).$$

In both models, we assume an identical model structure for $p(\mathbf{x}_i^{(o)} | \mathbf{x}_i^{(c)}, \mathbf{x}_i^{(p)})$ and $p(\mathbf{y}_j^{(o)} | \mathbf{y}_j^{(c)}, \mathbf{y}_j^{(p)})$, and this doesn't affect the choice to include a variable in the model or not.

This model can be fitted using the EM algorithm (Dempster et al. 1977) by introducing the missing labels $\underline{\mathbf{z}}$ into the model. The calculations involved in fitting the model including the labeled and unlabeled data follows those outlined in Dean et al. (2006). The maximum likelihood estimates for the regression part of the model correspond to least squares estimates of the regression parameters.

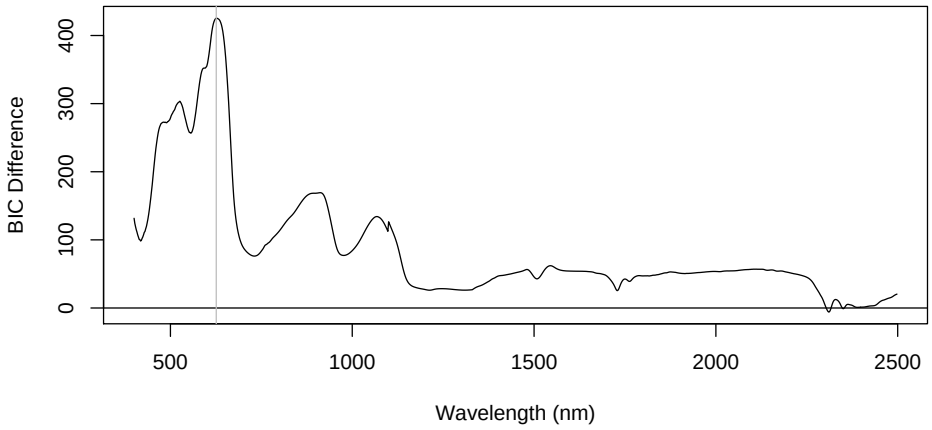


Figure 4: A plot of the BIC difference for each wavelength. The wavelength with the greatest difference is 626 nm.

The final estimates of the posterior probability of group memberships produced by the EM algorithm are used to classify the unlabeled observations. Thus each observation is classified into the group that maximizes $\hat{z}_{jg}$, where

$$\hat{z}_{jg} = \frac{\hat{\tau}_g p(\mathbf{y}_j^{(c)} | \hat{\mu}_g^{(c)}, \hat{\Sigma}_g^{(c)})}{\sum_{g'=1}^{G} \hat{\tau}_{g'} p(\mathbf{y}_j^{(c)} | \hat{\mu}_{g'}^{(c)}, \hat{\Sigma}_{g'}^{(c)})},$$

$\mathbf{y}_j^{(c)}$ is the set of chosen variables, and $\{(\hat{\tau}_g, \hat{\mu}_g^{(c)}, \hat{\Sigma}_g^{(c)}) : g = 1, 2, \ldots, G\}$ are the maximum likelihood estimates for the unknown model parameters for this set of chosen variables.

### 3.3.1 Example

An illustrative example of the BIC calculations when the proposed algorithm is applied to the meat spectroscopy data is shown in Figures 4–6; half the data of each type were randomly selected as training data in this example.

The variable selection algorithm begins by selecting 626 nm as the wavelength with the greatest difference between the Grouping and No Grouping models (Figure 4). Subsequently, the 814 nm wavelength is added to the model (Figure 5). It is worth noting that wavelengths close to 626 nm still have strong evidence of grouping even though the spectra are smoothly varying. At the third stage, the 774 nm wavelength is selected (Figure 6). The procedure continues until thirteen wavelengths are selected (details of the iterations are given in Table 2). Interestingly, many of the chosen wavelengths are in the visible range (400–800 nm) of the spectrum indicating that color is important when separating the meat samples.

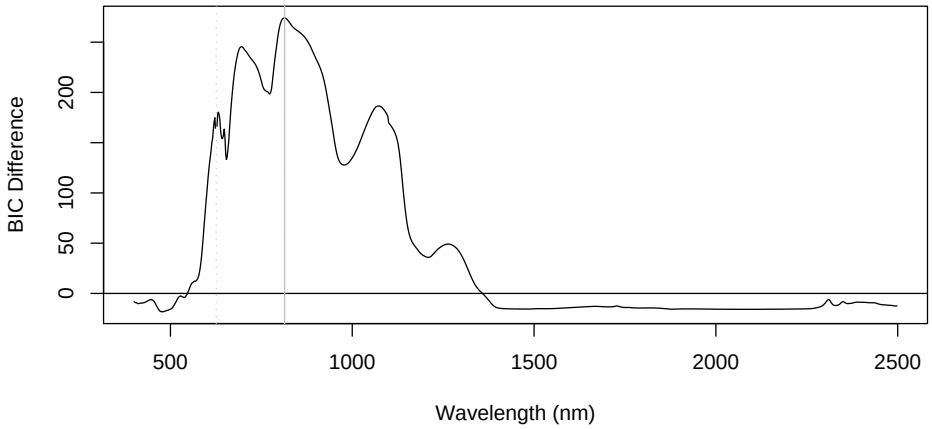


Figure 5: A plot of the BIC difference for each wavelength given that wavelength 626 nm is already accepted. The wavelength with the greatest BIC difference is 814 nm. Note that wavelengths close to 626 nm still have positive BIC difference values.

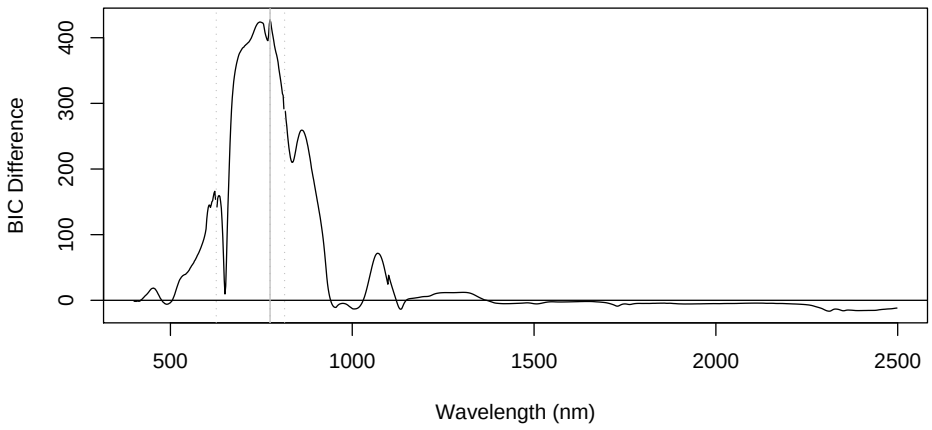


Figure 6: A plot of the BIC difference for each wavelength given that the wavelengths (626 nm and 814 nm) are already accepted. The wavelength with the greatest BIC difference is 774 nm.

Table 2: A full example of the variable selection procedure being used to classify the meat samples into five types. The updating procedure was used in this example.

| Iteration | Proposal | BIC Diff. | Decision | Proposal | BIC Diff. | Decision |
|---|---|---|---|---|---|---|
| 1 | Add 626 nm | 425.4 | Accepted | | | |
| 2 | Add 814 nm | 274.1 | Accepted | | | |
| 3 | Add 774 nm | 427.4 | Accepted | Remove 774 nm | -427.4 | Rejected |
| 4 | Add 664 nm | 142.6 | Accepted | Remove 626 nm | -120.1 | Rejected |
| 5 | Add 680 nm | 220.1 | Accepted | Remove 774 nm | -78.8 | Rejected |
| 6 | Add 864 nm | 165.2 | Accepted | Remove 774 nm | -91.7 | Rejected |
| 7 | Add 602 nm | 118.9 | Accepted | Remove 774 nm | -26.3 | Rejected |
| 8 | Add 794 nm | 118.3 | Accepted | Remove 774 nm | -86.2 | Rejected |
| 9 | Add 702 nm | 178.6 | Accepted | Remove 774 nm | -127.5 | Rejected |
| 10 | Add 1996 nm | 127.5 | Accepted | Remove 1996 nm | -127.5 | Rejected |
| 11 | Add 644 nm | 76.6 | Accepted | Remove 644 nm | -76.6 | Rejected |
| 12 | Add 2316 nm | 24.1 | Accepted | Remove 2316 nm | -24.1 | Rejected |
| 13 | Add 2310 nm | 103.2 | Accepted | Remove 702 nm | -26.1 | Rejected |
| 14 | Add 1936 nm | 10.8 | Accepted | Remove 702 nm | 4.4 | Accepted |
| 15 | Add 704 nm | -3.7 | Rejected | Remove 1936 nm | -41.3 | Rejected |

## 3.4 Headlong Model Search Strategy

The variable selection algorithm demonstrated in Section 3.3.1 is a greedy search strategy. At the variable addition stages of the algorithm, the variable with the greatest BIC difference is added and at variable removal stages the variable with the greatest BIC difference is removed. The process of finding the variable with the greatest BIC difference involves calculating the BIC difference for all variables under consideration; for the spectroscopic data there are typically about 1000 variables under consideration at the variable addition stages. Hence, this search strategy is computationally demanding.

A less computationally expensive alternative is to use a headlong search strategy (Badsberg 1992). The variable added or removed in the headlong search strategy need not be the best in terms of having the greatest BIC difference; it merely needs to be the first variable considered whose difference is greater than some pre-specified value (here *min.evidence*); we found that $min.evidence = 0$ gave good results for the applications in this paper. This means that instead of adding the variable with the greatest evidence for Grouping versus No Grouping, the first variable found to have a certain amount of evidence for Grouping versus No Grouping would be added. At the variable addition stages of the algorithm, the remaining variables are examined

in turn from an ordered list. The initial order of the list is based on the variables' original BIC differences at the univariate addition stage; this ordering was used in a similar context in Yeung et al. (2005). We experimented with the initial ordering and also tried using increasing wavelength and decreasing wavelength. The classification performance was not sensitive to the initial ordering but the selected variables did depend on the ordering. In the context of increasing and decreasing wavelength there was a bias towards selecting low and high wavelengths, respectively.

Here is a summary of the algorithm.

1. Select the first variable that is added to be the one that has the most evidence for Grouping versus No Grouping in terms of greatest BIC difference (the same as the first step of the greedy search algorithm). Create a list of the remaining variables in decreasing order of BIC differences.

2. Select the second variable that is added to be the first variable in the list of remaining variables with BIC difference for Grouping versus No Grouping, including the first variable selected, greater than *min.evidence*. Any variable checked and not used at this stage is placed at the end of the list of remaining variables.

3. Select the next variable that is added to be the first variable in the list of remaining variables with BIC difference for Grouping versus No Grouping, including the previous variables selected, greater than *min.evidence*. If no variable has BIC difference greater than *min.evidence* then no variable is added at this stage. Any variable checked and not used at this stage is placed, in turn, at the end of the list of remaining variables.

4. Check in turn each variable currently selected (in reverse order of inclusion) for evidence of No Grouping (versus Grouping), including the other selected variables, and remove the first variable with BIC difference greater than *min.evidence*. If no variable has BIC difference greater than *min.evidence* then no variable is removed at this stage. The removed variable is placed at the end of the list of other remaining variables.

5. Iterate steps 3 and 4 until two consecutive steps have been rejected, then stop.

# 4 Results

The proposed methodology was applied to the two food authenticity data sets described in Section 2.1. In each case, the data were split so that 50% of the data were used as labeled data and 50% as unlabeled. The methodology was applied to

Table 3: Classification performance on the Meats data for the variable selection algorithm with updating and for previous analyses of these data. Mean classification performance for the 50 random splits of the data are reported with standard deviations in parentheses.

| Method | Misclassification Rate |
|---|---|
| Variable Selection and Updating | 6.1% (3.5) |
| Dean et al. (2006) | 5.6% (2.0) |
| McElhinney et al. (1999) | 7.3–13.9% |
| Random Forests | 20.1% (3.8) |
| AdaBoost.M1 | 20.3% (4.8) |
| Bayesian Multinomial Regression | 34.2% (5.8) |

50 random splits of labeled and unlabeled data and the mean and standard deviation of the classification rate were computed. The results were compared to previously reported performance results for these data and several widely used alternative techniques: Random Forests (Breiman 2001), AdaBoost (Freund and Schapire 1997) and Bayesian Multinomial Regression (Madigan et al. 2005). We used the default settings in the R (R Development Core Team 2007) implementations of Random Forests (Liaw and Wiener 2002) and AdaBoost (Cortés et al. 2007) and for Bayesian Multinomial Regression we used cross validation to choose between the choice of prior variance values $\{10^p : p = -4, -3, -2, -1, 0, 1, 2, 3, 4\}$ as suggested in Genkin et al. (2005).

## 4.1 Meats Data

The results achieved on the homogenized meat data (Section 2.2) are reported in Table 3. These results show that the Variable Selection and Updating method gives comparable or better performance than previous analyses of these data; an improved classification rate has been achieved relative to those achieved by McElhinney et al. (1999) who used factorial discriminant analysis (FDA), k-nearest neighbors (kNN), discriminant partial least squares regression (PLS) and soft independent modeling of class analogy (SIMCA). Furthermore, a comparable classification performance has been achieved relative to Dean et al. (2006) who used model-based discriminant analysis with updating on a reduced form of the data derived from wavelet thresholding. The variable selection and updating procedure gave substantially better performance than other competing methods for classification.

An examination of the misclassification table (Table 4) for the Variable Selection and Updating method shows that many of the misclassifications were due to the difficulty in separating the chicken and turkey groups. Interestingly, no misclassifications

Table 4: Average classification results for the different meat types for the Variable Selection and Updating classification method.

| | *Predicted* | | | | |
|---|---|---|---|---|---|
| *Truth* | Beef | Lamb | Pork | Turkey | Chicken |
| Beef | 98.6 | 1.4 | 0.0 | 0.0 | 0.0 |
| Lamb | 1.4 | 98.6 | 0.0 | 0.0 | 0.0 |
| Pork | 0.0 | 0.0 | 99.2 | 0.5 | 0.3 |
| Turkey | 0.0 | 0.0 | 0.0 | 88.2 | 11.8 |
| Chicken | 0.0 | 0.0 | 0.0 | 11.1 | 88.9 |

were made between the red and white meats.

The chosen wavelengths show us which parts of the spectrum are of importance when classifying samples into different species. We recorded the chosen wavelengths for each of the 50 sets of results and these are shown in Figure 7. We can see that a large proportion (51%) of the chosen wavelengths are in the visible region (400 nm–800 nm) but some regions in the near-infrared spectrum are also chosen. Liu and Chen (2000, Table 1) assign many of the spectral features in the visible part of the spectrum to different forms of myoglobin such as deoxymyoglobin (430, 440, 445 nm), oxymyoglobin (545, 560, 575, 585 nm), metmyoglobin (485, 495, 500, 505 nm) and sulfmyoglobin (635 nm). Sulfmyoglobin is a product of the reaction of myoglobin with $H_2S$ generated by bacteria, and Arnalds et al. (2004) found the region of the spectrum close to 635 nm to be important when separating the red and white meat samples. The peak at 1100 nm is the wavelength where the sensor changes in the near-infrared spectrometer and the peak at 1068 nm can be attributed to third overtones of C-H stretch mode and C-H combination bonds from meat constituents other than oxymyoglobin (Liu et al. 2000). The near infrared region consisting of wavelengths near 1510 nm has been attributed to protein, and a cluster of chosen wavelengths is close to this region. In all cases, between 13 and 19 wavelengths were chosen for classification purposes.

Following McElhinney et al. (1999) and Dean et al. (2006), we combined the chicken and turkey groups into a poultry group to determine how well we can classify the homogenized meat samples into four types. The classification results are reported in Table 5 and the misclassifications from the variable selection method with updating are shown in Table 6. There is a significant improvement in classification performance from all of the methods. Again, the white and red meats are separated with zero error.

The wavelengths chosen for the four group classification problem (Figure 8) still

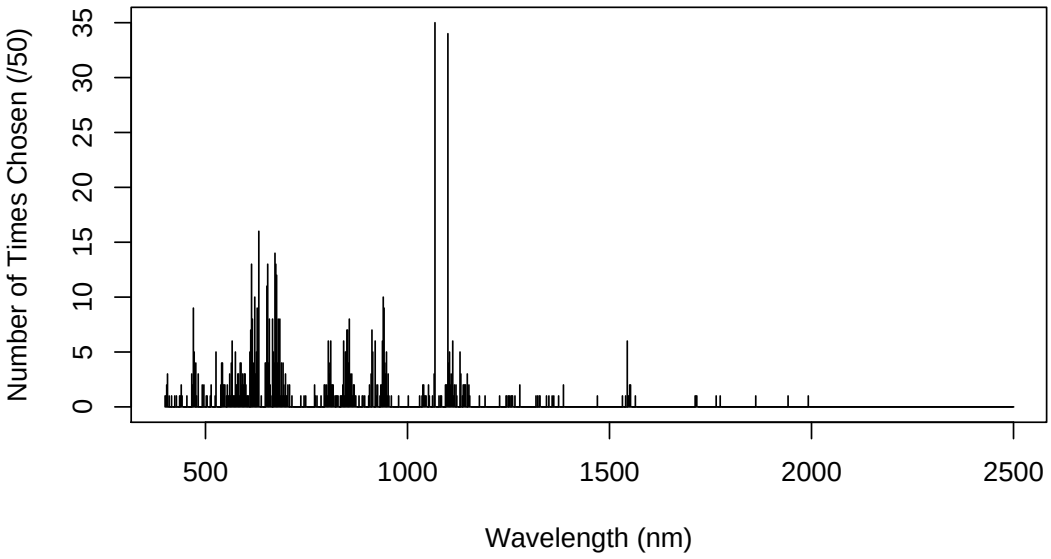


Figure 7: Wavelengths chosen in the five meat classification problem for the Variable Selection and Updating method. The height of the bars shows how many times the wavelength was chosen in 50 random splits of the data.

Table 5: Classification performance on the Meats data for the variable selection algorithm with updating and for previous analyses of these data after combining the chicken and turkey into a poultry group. Mean classification performance for the 50 random splits of the data are reported with standard deviations in parentheses.

| Method | Misclassification Rate |
|---|---|
| Variable Selection and Updating | 0.8% (1.3) |
| Dean et al. (2006) | 1.0% (0.9) |
| McElhinney et al. (1999) | 2.6–4.3% |
| Random Forests | 10.5% (3.3) |
| AdaBoost.M1 | 14.7% (3.7) |
| Bayesian Multinomial Regression | 17.2% (4.9) |

Table 6: Average classification results for the different meat types after combining the chicken and turkey into a poultry group. The results shown are for the Variable Selection and Updating method.

| | *Predicted* | | | |
|---|---|---|---|---|
| *Truth* | Beef | Lamb | Pork | Poultry |
| Beef | 98.2 | 1.8 | 0.0 | 0.0 |
| Lamb | 2.7 | 97.3 | 0.0 | 0.0 |
| Pork | 0.0 | 0.0 | 99.1 | 0.9 |
| Poultry | 0.0 | 0.0 | 0.0 | 100.0 |

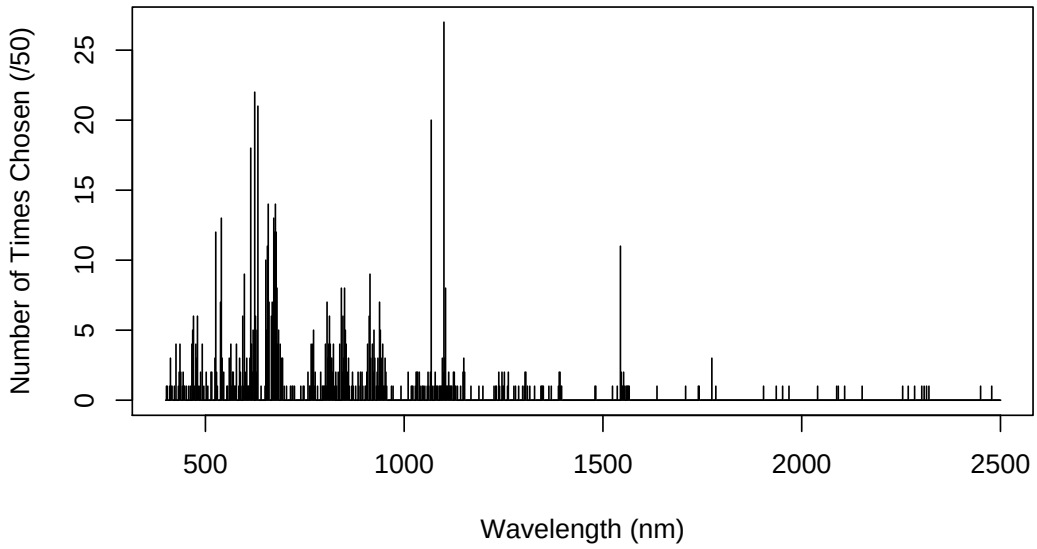


Figure 8: Wavelengths chosen in the four meat classification problem for the Variable Selection and Updating method.

have a substantial proportion chosen from the visible part of the spectrum (52%). In this application, between 13 and 21 wavelengths were chosen for classification purposes.

## 4.2 Greek Olive Oil Data

The methods were applied to the Greek olive oil data (Section 2.3) with 50% of the data being treated as training data and 50% as test data. Fifty random splits of training and test data were used. The misclassification rates achieved on these data are reported in Table 7. Variable selection and updating provides one of the best classification rates for these data. Downey et al. (2003) did report a better misclassification rate (6.1%) using factorial discriminant analysis (FDA) but the choice of a subset of wavelengths, data

Table 7: Classification performance on the Olive Oil data for the variable selection algorithm with updating and for previous analyses of these data. Mean classification performance for the 50 random splits of the data are reported with standard deviations in parentheses.

| Method | Misclassification Rate |
|---|---|
| Variable Selection and Updating | 6.9% (5.4) |
| Dean et al. (2006) | 11.9% (6.3) |
| Downey et al. (2003) | 6.1–19.0% |
| Random Forests | 19.3% (6.5) |
| AdaBoost.M1 | 34.1% (9.3) |
| Bayesian Multinomial Regression | 57.0% (1.2) |

Table 8: Average classification results for the olive oil groups. The results shown are for the Variable Selection and Updating method.

| | *Predicted* | | |
|---|---|---|---|
| *Truth* | Crete | Peleponese | Other |
| Crete | 90.0 | 8.7 | 1.3 |
| Peleponese | 1.0 | 92.9 | 6.1 |
| Other | 0.0 | 3.8 | 96.2 |

pre-processing method and classification method (from partial least squares, factorial discriminant analysis and k-nearest neighbors) was made with reference to the test data classification performance. In contrast, our model selection was done without reference to the test data classification performance.

A cross tabulation of the classifications with the true origin of the olive oils (Table 8) reveals the difficulty in classifying the oils.

In contrast to the meat classification problem, the chosen wavelengths for this problem (Figure 9) are concentrated in the near-infrared region (800–2498 nm) but some wavelengths in the visible region are also selected. The most commonly chosen wavelength is 2080 nm which has been attributed to an O-H stretching/O-H bend combination (Osborne et al. 1984). Wavelengths near 2310, 2346 and 2386 nm are due to C-H stretching vibrations and other vibrational modes. In particular, wavelengths in the 2310 nm region have previously been assigned to fat content. In all cases, between 6 and 29 wavelengths were selected with a mean of 15 wavelengths being chosen.

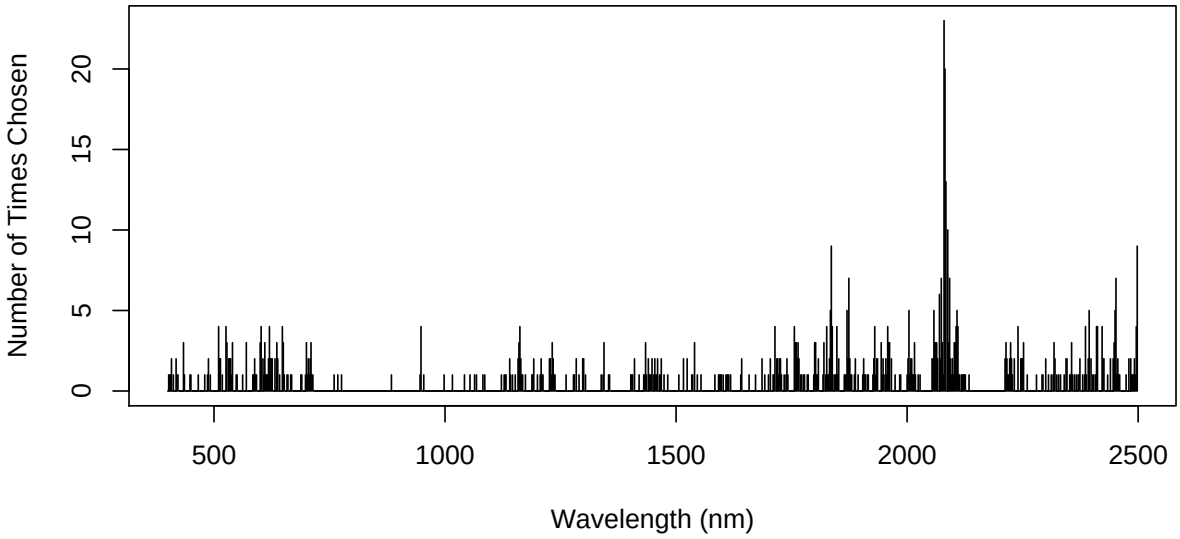


Figure 9: Wavelengths chosen in the olive oil classification problem using Variable Selection and Updating method. The height of the bars shows how many times the wavelength was chosen in 50 random splits of the data.

# 5 Discussion

The discriminant analysis method presented in this paper gave much better results than those given by other popular machine learning techniques such as Random Forests (Breiman 2001), AdaBoost (Freund and Schapire 1997) and Bayesian Multinomial Regression (Genkin et al. 2005; Madigan et al. 2005) for the high-dimensional food authenticity datasets analysed here. This improvement is further enhanced by the addition of the updating procedure for including the unlabeled data in the estimation method. It is clear from the results that the headlong search method for variable selection offers an efficient method for selecting wavelengths.

In addition to the improvement in classification results in the example data sets given, the number of variables needed for classification was substantially reduced. The variable selection results in the food authenticity application suggest the possibility of developing authenticity sensors that only use reflectance values over a carefully selected subset of the near-infrared and visible spectral range. In contexts such as gene expression data and document classification this could mean a substantial savings in terms of time for data collection and space for future data storage.

We have compared our method with three established leading classification methods from statistics and machine learning for which standard software implementations are available. One of these, AdaBoost, was identified by Leo Breiman as "the best off-the-

shelf classifier in the world" (Hastie et al. 2001).

A range of recent approaches to variable selection in a classification context include the DALASS approach of Trendafilov and Jolliffe (2007), variable selection for kernel Fisher discriminant analysis (Louw and Steep 2006), the stepwise stopping rule approach of Munita et al. (2006), a number of different search algorithms (proposed as alternatives to backward/forward/stepwise search) wrapped around different discriminant functions compared by Pacheco et al. (2006), and genetic search algorithms wrapped around Fisher discriminant analysis by Chiang and Pell (2004). Another example in the context of spectroscopic data is given by Indahl and Naes (2004).

In terms of other approaches, a good review of recent work on the problem of variable or feature selection in classification was given by Guyon and Elisseeff (2003) from a machine learning perspective. A good review of methods involving Support Vector Machines (SVMs) (along with a proposed criterion for exhaustive variable selection) is given by Mary-Huard et al. (2007). An extension allowing variable selection for the multiclass problem using SVMs is given by Wang and Xiatong (2007) and an approach for binary problems using SVMs within the framework of smoothing spline ANOVA models is given by Zhang (2006). An alternative approach for combining pairwise classifiers is given by Szepannek and Weihs (2006) based on Hastie and Tishirani (1998). For classifying streaming data a new approach has been proposed by Zhou et al. (2006). Greenshtein (2006) looks at theoretical aspects of the $n \ll p$ classification and variable selection problem in terms of empirical risk minimization subject to $l_1$ constraints. The Lasso was developed by Tibshirani (1996) and an alternative approach called the elastic net is given by Zou and Hastie (2005). Finally an alternative to single subset variable selection through Bayesian Model Averaging (Madigan and Raftery 1994) is given by Dash and Cooper (2004).

# References


Arnalds, T., McElhinney, J., Fearn, T., and Downey, G. (2004), "A Hierarchical Discriminant Analysis for Species Identification in Raw Meat by Visible and Near Infrared Spectroscopy," *Journal of Near Infrared Spectroscopy*, 12, 183–188.

Badsberg, J. H. (1992), "Model search in contingency tables by CoCo," in *Computational Statistics*, eds. Dodge, Y. and Whittaker, J., Heidelberg: Physica Verlag, vol. 1, pp. 251–256.

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.

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32.

Chang, W.-C. (1983), "On using principal components before separating a mixture of two multivariate normal distributions," *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 32, 267–275.

Chapelle, O., Schölkopf, B., and Zien, A. (eds.) (2006), *Semi-Supervised Learning*, Cambridge, MA: MIT Press.

Chiang, L. H. and Pell, R. J. (2004), "Genetic algorithms combined with discriminant analysis for key variable identification," *Journal of Process Control*, 14, 143–155.

Cortés, E. A., Martínez, M. G., and Rubio, N. G. (2007), *adabag: Applies Adaboost.M1 and Bagging*, R package version 1.1.

Dash, D. and Cooper, G. F. (2004), "Model Averaging for Prediction with Discrete Bayesian Networks," *Journal of Machine Learning Research*, 5, 1177–1203.

Dean, N., Murphy, T. B., and Downey, G. (2006), "Updating Classification Rules using Unlabelled Data with Applications in Food Authenticity Studies," *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 55, 1–14.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B: Methodological*, 39, 1–38, with discussion.

Downey, G. (1996), "Authentication of food and food ingredients by near infrared spectroscopy," *Journal of Near Infrared Spectroscopy*, 4, 47–61.

Downey, G., McIntyre, P., and Davies, A. N. (2003), "Geographical classification of extra virgin olive oils from the eastern Mediterranean by chemometric analysis of visible and near infrared spectroscopic data," *Applied Spectroscopy*, 57, 158–163.

Fraley, C. and Raftery, A. E. (1998), "How many clusters? Which clustering method? - Answers via Model-Based Cluster Analysis," *Computer Journal*, 41, 578–588.

— (1999), "MCLUST: Software for model-based clustering," *Journal of Classification*, 16, 297–306.

— (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–612.

— (2003), "Enhanced model-based clustering, density estimation and discriminant analysis software: MCLUST," *Journal of Classification*, 20, 263–296.

— (2007), *mclust: Model-Based Clustering / Normal Mixture Modeling*, R package version 3.1-1.

Freund, Y. and Schapire, R. E. (1997), "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, 55, 119–139.

Ganesalingam, S. and McLachlan, G. J. (1978), "The efficiency of a linear discriminant function based on unclassified initial samples," *Biometrika*, 65, 658–662.

Genkin, A., Lewis, D. D., and Madigan, D. (2005), "BMR: Bayesian Multinomial Regression Software," http://www.stat.rutgers.edu/~madigan/BMR/.

Greenshtein, E. (2006), "Best subset selection, persistence in high-dimensional statistical learning and optimization under $l_1$ constraint," *The Annals of Statistics*, 34, 2367–2386.

Guyon, I. and Elisseeff, A. (2003), "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, 3, 1157–1182.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001), *The Elements of Statistical Learning*, New York: Springer-Verlag.

Hastie, T. and Tishirani, R. (1998), "Classification by Pairwise Coupling," *Annals of Statistics*, 26, 451–471.

Indahl, U. and Naes, T. (2004), "A variable selection strategy for supervised classification with continuous spectroscopic data," *Journal of Chemometrics*, 18, 53–61.

Kohavi, R. and John, G. (1997), "Wrappers for feature selection," *Artificial Intelligence*, 91, 273–324.

Liang, F., Mukherjee, S., and West, M. (2007), "The Use of Unlabeled Data in Predictive Modeling," *Statistical Science*, 22, 189–205.

Liaw, A. and Wiener, M. (2002), "Classification and Regression by randomForest," *R News*, 2, 18–22.

Liu, Y. and Chen, Y. R. (2000), "Two-Dimensional Correlation Spectroscopy Study of Visible and Near-Infrared Spectral Variations of Chicken Meats in Cold Storage," *Applied Spectroscopy*, 54, 1458–1470.

Liu, Y., Chen, Y. R., and Ozaki, Y. (2000), "Two-Dimensional Visible/Near Infrared Correlation Spectroscopy Study of Thermal Treatment of Chicken Meat," *Journal of Agricultural and Food Chemistry*, 48, 901–908.

Louw, N. and Steep, S. J. (2006), "Variable selection in kernel Fisher discriminant analysis by means of recursive feature elimination," *Computational Statistics & Data Analysis*, 51, 2043–2055.

Madigan, D., Genkin, A., Lewis, D. D., and Fradkin, D. (2005), "Bayesian Multinomial Logistic Regression for Author Identification," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, eds. Knuth, K. H., Abbas, A. E., Morris, R. D., and Castle, J. P., vol. 803, pp. 509–516.

Madigan, D. and Raftery, A. E. (1994), "Model selection and accounting for model uncertainty in graphical models using Occam's window," *Journal of the American Statistical Association*, 89, 1535–1546.

Mary-Huard, T., Robin, S., and Daudin, J.-J. (2007), "A penalized criterion for variable selection in classification," *Journal of Multivariate Analysis*, 98, 695–705.

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.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley Interscience.

Munita, C. S., Barroso, L. P., and Oliveira, P. M. S. (2006), "Stopping rule for variable selection using stepwise discriminant analysis," *Journal of Radioanalytical and Nuclear Chemistry*, 269, 335–338.

O'Neill, T. J. (1978), "Normal discrimination with unclassified observations," *Journal of the American Statistical Association*, 73, 821–826.

Osborne, B. G., Fearn, T., Miller, A. R., and Douglas, S. (1984), "Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs," *Journal of the Science of Food and Agriculture*, 35, 99–105.

Pacheco, J., Casado, S., Núñez, L., and Gómez, O. (2006), "Analysis of new variable selection methods for discriminant analysis," *Computational Statistics & Data Analysis*, 51, 1463–1478.

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.

Raftery, A. E. and Dean, N. (2006), "Variable Selection for Model-Based Clustering," *Journal of the American Statistical Association*, 101, 168–178.

Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461–464.

Szepannek, G. and Weihs, C. (2006), "Variable selection for discrimination of more than two classes where data are sparse," in *From Data and Information Analysis to Knowledge Engineering*, eds. Spiliopoulou, M., Kruse, R., Borgelt, C., Nurnberger, A., and Gaul, W., Studies in Classification, Data Analysis and Knowledge Organization, pp. 700–707.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B*, 58, 267–288.

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.

Trendafilov, N. T. and Jolliffe, I. T. (2007), "DALASS: Variable selection in discriminant analysis via the LASSO," *Computational Statistics & Data Analysis*, 51, 3718–3736.

Wang, L. and Xiatong, S. (2007), "On $L_1$-Norm Multiclass Support Vector Machines: Methodology and Theory," *Journal of the American Statistical Association*, 102, 583–594.

West, M. (2003), "Bayesian factor regression models in the "large $p$, small $n$" paradigm," in *Bayesian Statistics 7*, Oxford University Press, pp. 723–732.

Yeung, K. Y., Bumgartner, R., and Raftery, A. E. (2005), "Bayesian Model Averaging: Development of an improved multi-class, gene selection and classification tool for microarray data," *Bioinformatics*, 21, 2394–2402.

Zhang, H. H. (2006), "Variable Selection for Support Vector Machines via Smoothing Spline ANOVA," *Statistica Sinica*, 16, 659–674.

Zhou, J., Foster, D. P., Stine, R. A., and Ungar, L. H. (2006), "Streamwise Feature Selection," *Journal of Machine Learning Research*, 7, 1861–1885.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 67, 301–320.