The Annals of Applied Statistics
2010, Vol. 4, No. 1, 396–421
DOI: 10.1214/09-AOAS279

VARIABLE SELECTION AND UPDATING IN MODEL-BASED DISCRIMINANT ANALYSIS FOR HIGH DIMENSIONAL DATA WITH FOOD AUTHENTICITY APPLICATIONS

BY THOMAS BRENDAN MURPHY[1,3], NEMA DEAN[3] AND ADRIAN E. RAFTERY[2,3]

University College Dublin, University of Glasgow and University of Washington, Seattle

Food authenticity studies are concerned with determining if food samples have been correctly labeled or not. Discriminant analysis methods are an integral part of the methodology for food authentication. Motivated by food authenticity applications, 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 food authenticity data sets 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 authenticity data sets, our proposed method outperformed default implementations of Random Forests, AdaBoost, transductive SVMs and Bayesian Multinomial Regression by substantial margins.

1. Introduction. Foods that are expensive are subject to potential fraud where rogue suppliers may attempt to provide a cheaper inauthentic alternative in place of the more expensive authentic food. Food authenticity studies are concerned with assessing the veracity of the labeling of food samples. Discriminant analysis methods are of prime importance in food authenticity studies where samples whose authenticity is being assessed are classified using a discriminant analysis method and the labeling and classification are compared. Samples determined to have potentially inaccurate labeling can be sent for further testing to determine if fraudulent labeling has been used.

Received April 2009; revised August 2009.

[1]Supported in part by the Science Foundation of Ireland Basic Research Grant 04/BR/M0057 and Research Frontiers Programme Grant 2007/RFP/MATF281.

[2]Supported in part by NICHD Grant R01 HD054511 and NSF Grant ATM 0724721.

[3]Supported in part by NIH Grant 8 R01 EB002137-02.

Key words and phrases. Food authenticity studies, headlong search, model-based discriminant analysis, normal mixture models, semi-supervised learning, updating classification rules, variable selection.

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. Additionally, if a subset of variables is found to be important for classification purposes, then it suggests the potential for collecting a smaller subset of variables using inexpensive methods rather than the full high dimensional data.

Variable selection can be completed as a preprocessing step prior to discriminant analysis (a filtering 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 weak 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 many more variables than observations.

In this paper a version of model-based discriminant analysis is developed by adapting 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. The stepwise selection 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, Murphy and Downey (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 data sets in Section 4 and the methodology and results are discussed in Section 5.

2. Data.

2.1. *Food authenticity and near infrared spectroscopy.* 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) and Reid, O'Donnell and Downey (2006) provide reviews of food authenticity studies with an emphasis on spectroscopic methods. Near infrared (NIR) spectroscopy provides a quick and efficient method of collecting data for use in food authenticity studies [Downey (1996)]. It is particularly useful because it requires very little sample preparation and is nondestructive to the samples being tested.

We consider two food authenticity data sets which consist of combined visible and 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. The two studies are outlined in detail 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).

In both studies, 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. Hence, the values collected for each food sample consist of 1050 reflectance values taken at 2 nm intervals (see, for example, Figure 1). 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

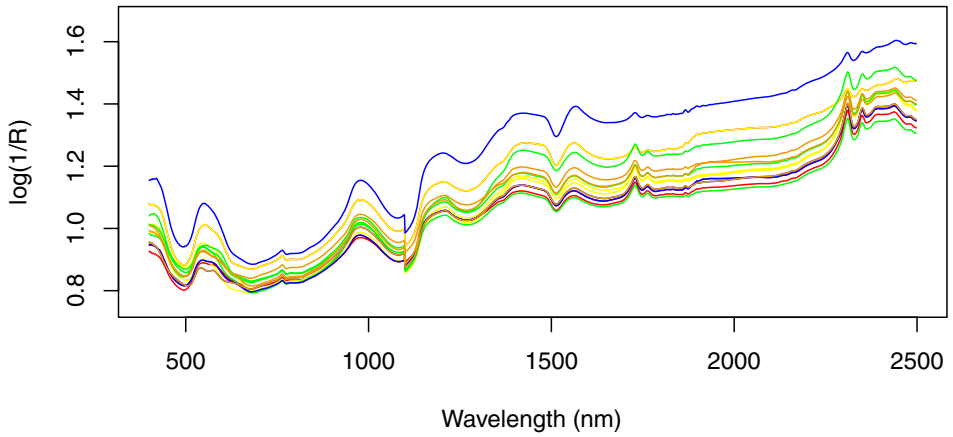

FIG. 1. *The near-infrared spectra recorded for three examples of each meat species in the study. The discontinuity at* 1100 *nm is due to a sensor change at that value. The samples are colored as Beef = red, Lamb = green, Pork = blue, Turkey = orange, Chicken = yellow.*

the olive oil data, but fewer scans were used. Full details of the spectral data collection process are given in McElhinney, Downey and Fearn (1999) and Downey, McIntyre and Davies (2003).

The reflectance values in the visible and near-infrared region are produced by vibrations in the chemical bonds in the substance being analyzed. The data are highly correlated due to the presence of a large number of overlapping broad peaks in this region of the electromagnetic spectrum and the presence of combinations and overtones. As a result, even though the data are very highly correlated, the reflectance values at adjacent wavelengths can have different sources and reflectance values at very different wavelengths can have the same source. So, the information encoded in each spectrum is recorded in a complex manner and spread over a range of locations. Osborne, Fearn and Hindle (1993) provide an extensive review of the chemical and technological aspects of near-infrared spectroscopy and its application. Further information on the combined spectra and their structure is given in Section 4 where the results of the analysis of the data are given.

Because of the complex nature of the combined spectroscopic data, there is interest in determining if a small subset of reflectance values contain as much information for authentication purposes as the whole spectrum does. If a small number of variables contain sufficient information for authentication purposes, then this indicates the possibility of developing portable sensors for food authenticity studies that are more rapid and have a lower cost than recording the combined visible and near-infrared spectrum. In fact, portable sensors have been developed on a commercial basis for the authentication of Scottish whiskys [Connolly (2006)] using ultraviolet spectroscopic technology. Hence, there are motivations for incorporating feature selection in the classification methods used on these data from the application and the modeling viewpoints.

The problem of feature selection is especially difficult because the number of possible subsets of wavelengths that could be selected in this problem is 2^{1050}. So, efficient search strategies need to be used so that a good set of features can be selected without searching over all possible subsets.

2.2. *Homogenized meat data.* McElhinney, Downey and Fearn (1999) constructed a collection of combined visible and near-infrared spectra from 231 homogenized meat samples in order to assess the effectiveness of visible and near-infrared spectroscopy as a tool for determining the correct species of the samples. The samples collected for this study consist of 55 Chicken, 55 Turkey, 55 Pork, 32 Beef and 34 Lamb samples. The samples were collected over an extended period of time and from a number of sources in order to ensure a representative sample of meats.

For each sample, a spectrum consisting of 1050 reflectance measurements was recorded (as outlined in Section 2.1). A plot of all of the spectra is shown in Figure 1. We can see that there is a discrimination between the red meats (beef and lamb) and the white meats (chicken, turkey and pork) over some of the visible region (400–800 nm), but discrimination within meat colors is less clear.

2.3. *Greek olive oils data.* Downey, McIntyre and Davies (2003) recorded near-infrared spectra from a total of 65 extra virgin olive oil samples that 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. The aim of their study was to assess the effectiveness of near-infrared spectroscopy in determining the geographical origin (see Figure 2) of the oils.

FIG. 2. *Regions of Greece where the olive oil samples were collected.*

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 μ_g and covariance Σ_g.

A central idea in model-based clustering is the use of constraints on the group covariance matrices Σ_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 λ_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 as follows: λ_g = volume; A_g = shape; D_g = orientation. These parameters can be constrained in various ways to be 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, Murphy and Downey (2006) and will be described in more 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 (filtering) is given by Kohavi and John (1997), Section 2.4.

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

- A variable is proposed for addition to the set of selected clustering variables. The Bayesian Information Criterion (BIC) is used to compare a model in which the variable contains extra information about the clustering beyond the information in the already selected variables versus a model where the variable doesn't contain additional information about the clustering beyond the information in the already selected variables. The variable with the greatest positive BIC difference is added to the model. If the proposed variable has a negative BIC difference, then no variable is added.
- BIC is used to consider whether a variable should be removed from the model; This step is the reverse of the variable addition step. If all of the selected variables contain clustering information, then none is removed from the set of selected clustering variables.

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. While the stepwise variable selection wrapper proposed in Raftery and Dean (2006) and other wrapper approaches can give excellent clustering results, there is a considerable computational burden with wrapper approaches when compared to filtering approaches; this is because the model needs to be fitted each time a variable is added or removed from the set of selected clustering variables.

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 a stepwise variable selection wrapper. 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. In order to do this, a model comparison using BIC is used to compare a discriminant analysis model where the variable contains group information beyond the information in the already selected variables versus a model where the variable does not contain group information beyond the information in the already selected variables. Variables where the BIC difference is positive are candidates for addition to the set of selected variables; 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. This step is the reverse of the variable addition step. Variables where the BIC model comparison suggests that the variable does not contain

group information are candidates for removing from the set of selected variables; the procedure for searching for variables to remove from the model is outlined 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 on the comparison of two 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: $p(\mathbf{x}_i|\mathbf{l}_i) = p(\mathbf{x}_i^{(c)}, \mathbf{x}_i^{(p)}, \mathbf{x}_i^{(o)}|\mathbf{l}_i) = p(\mathbf{x}_i^{(o)}|\mathbf{x}_i^{(p)}, \mathbf{x}_i^{(c)})p(\mathbf{x}_i^{(p)}|\mathbf{x}_i^{(c)}) \times p(\mathbf{x}_i^{(c)}|\mathbf{l}_i)$.

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,

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

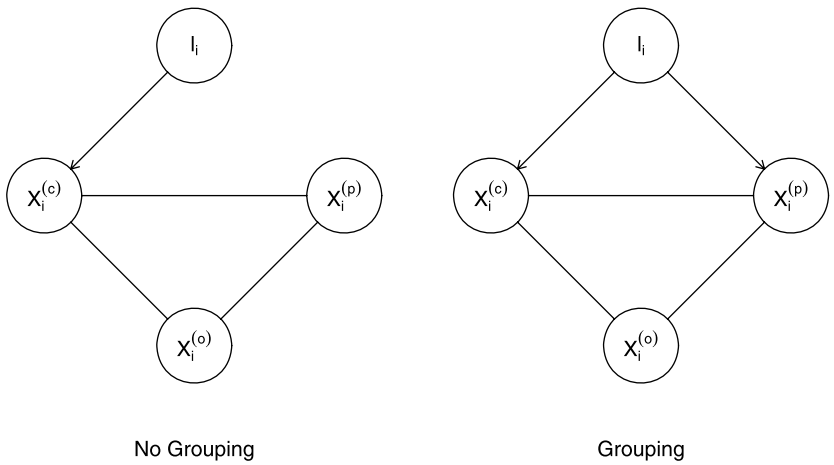

FIG. 3. *A graphical model representation of the Grouping and the No Grouping models.*

TABLE 1
Constrained covariance structures in model-based clustering as implemented in the `mclust` *package for* R

Model ID	Volume	Shape	Orientation	Covariance ($\boldsymbol{\Sigma}_g$)
EII	Equal	Equal spherical	NA	λI
VII	Variable	Equal spherical	NA	$\lambda_g I$
EEI	Equal	Equal	Axis aligned	λA
VEI	Variable	Equal	Axis aligned	$\lambda_g A$
EVI	Equal	Variable	Axis aligned	λA_g
VVI	Variable	Variable	Axis aligned	$\lambda_g A_g$
EEE	Equal	Equal	Equal	$\lambda D A D^T$
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	$\lambda_g D_g A_g D_g^T$

where $\tau = (\tau_1, \tau_2, \ldots, \tau_G)$.

- 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,

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

  where $\tau = (\tau_1, \tau_2, \ldots, \tau_G)$.

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

$$\log(\text{Bayes Factor}) = \log \frac{p(\mathbf{x}_i|\mathcal{M}_G)}{p(\mathbf{x}_i|\mathcal{M}_{NG})}, \tag{3.1}$$

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

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

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 (3.1) can be reduced to

$$\begin{aligned}\log(\text{Bayes Factor}) &= \log \frac{p(\mathbf{x}_i^{(p)}, \mathbf{x}_i^{(c)} | \mathcal{M}_G)}{p(\mathbf{x}_i^{(p)} | \mathbf{x}_i^{(c)}, \mathcal{M}_{NG}) p(\mathbf{x}_i^{(c)} | \mathcal{M}_{NG})} \\ &\approx \frac{1}{2}[\text{BIC(Grouping)} - \text{BIC(No Grouping)}],\end{aligned} \tag{3.2}$$

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 variable addition stage, the BIC of the grouping model is calculated using each of the ten covariance structures given in Table 1 and the model with the highest BIC is selected for the Grouping model for model comparison purposes.

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(Grouping) − BIC(No Grouping) are candidates for removal from the model; in this case, the BIC for the no grouping models are computed for all covariance structures from Table 1 and the model with the highest BIC is selected as the No Grouping model.

3.3. *Discriminant analysis with updating*. In standard discriminant analysis, the unlabeled data are 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, Murphy and Downey (2006), Chapelle, Schölkopf and Zien (2006), Toher, Downey and Murphy (2007) and Liang, Mukherjee and West (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,

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

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,

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

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, Laird and Rubin (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 follow those outlined in Dean, Murphy and Downey (2006). The maximum likelihood estimates for the regression part of the model correspond to least squares estimates of the regression parameters.

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 j is classified into the group g that maximizes $\hat{z}_{jg}$ over g, 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.

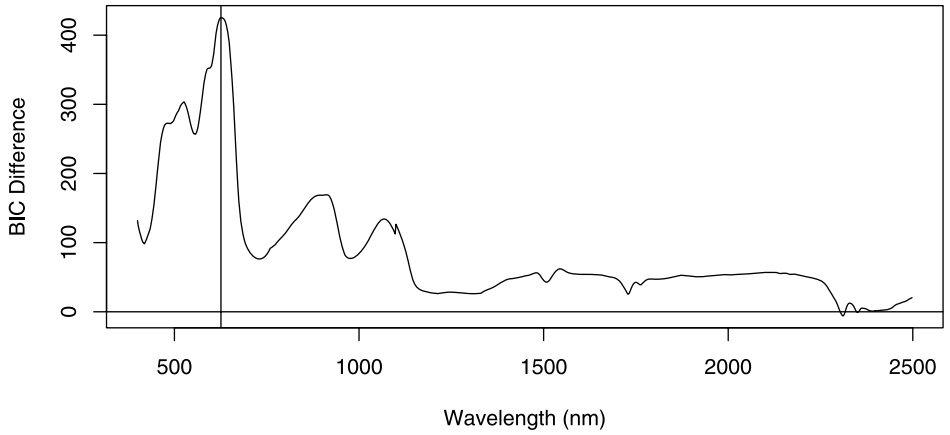

FIG. 4. *A plot of the BIC difference for each wavelength. The wavelength with the greatest BIC difference is* 626 *nm.*

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) and the E covariance structure was chosen. It is worth noting that wavelengths close to 626 nm still have strong evidence of grouping even though the spectra are smoothly varying. This phenomenon is due to the fact that the spec-

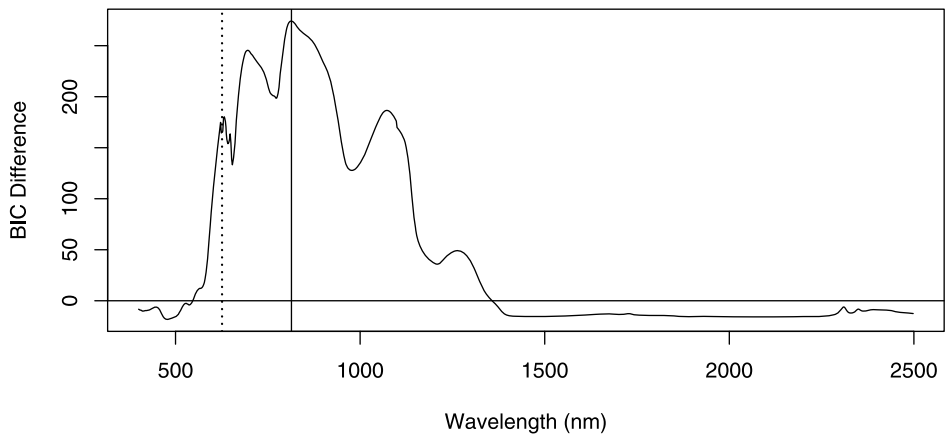

FIG. 5. *A plot of the BIC difference for each wavelength given that wavelength* 626 *nm is already accepted. The wavelength with the greatest BIC difference is* 814 *nm. Note that wavelengths close to* 626 *nm still have positive BIC difference values.*

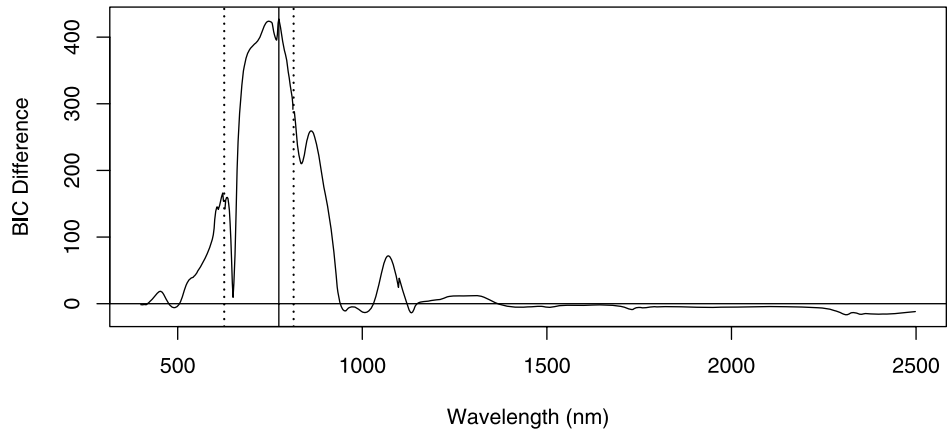

FIG. 6. *A plot of the BIC difference for each wavelength given that the first two wavelengths chosen* (626 *nm and* 814 *nm*) *are already accepted. The wavelength with the greatest BIC difference is* 774 *nm.*

trum consists of a number of overlapping peaks and the reflectances at adjacent locations can have different sources. As a result, extra grouping information can be available at wavelengths that are very close.

Subsequently, the 814 nm wavelength is added to the model (Figure 5) and the EEV covariance structure was chosen. At the third stage, the 774 nm wavelength is selected (Figure 6) and the VEV covariance structure was chosen. The procedure continues until thirteen wavelengths are selected (details of the iterations are given in Table 2) and the VEV covariance structure is chosen at all subsequent stages.

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. The closest two wavelengths that were selected were 2310 nm and 2316 nm and a number of wavelengths that were selected are approximately 20 nm apart. In summary, the selected wavelengths are spread out mainly in the visible region, but some wavelengths were selected in the near-infrared region.

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 just under 1050 variables under consideration at the variable addition stages. Hence, this search strategy is computationally demanding; this feature is shared by other wrapper variable selection methods too.

A less computationally expensive alternative is to use a headlong search strategy [Badsberg (1992)]. The variable added or removed in the headlong search strat-

TABLE 2
A full example of the variable selection procedure 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

egy 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 prespecified value (here *min.evidence*). We found that $min.evidence = 0$ gave good results for the applications in this paper. The headlong strategy has close connections to the "first-improvement" moves used in local search algorithms [e.g., Hoos and Stützle (2005), Chapter 2.1]. 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, Bumgarner and Raftery (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 toward 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 outlined 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 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)], Bayesian Multinomial Regression [Madigan et al. (2005)], and Transductive Support Vector Machines [Vapnik (1995), Joachims (1999), Collobert et al. (2006)].

We used the default settings in the `R` [R Development Core Team (2007)] implementations of Random Forests (`randomForest` version 4.5-30) [Liaw and Wiener (2002)] and AdaBoost (`adabag` version 1.1) [Cortés, Martínez and Rubio (2007)]. The use of various parameter settings was explored, but the results did not vary to a large extent with respect to the choice of parameter values. 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, Lewis and Madigan (2005). For the Transductive SVM analysis we used the UniverSVM software version 1.1 [Sinz and Roffilli (2007)] with a linear kernel and parameters $(c, s, z) = (100, -0.3, 0.1)$; other parameter values were considered, but the values reported yielded the best classifications.

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

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)
Variable selection (greedy) and updating	5.1% (1.9)
Variable selection only	9.3% (3.6)
Dean, Murphy and Downey (2006)	5.6% (2.0)
McElhinney, Downey and Fearn (1999)	7.3%–13.9%
Transductive SVMs	42.6% (5.7)
Random Forests	20.1% (3.8)
AdaBoost.M1	20.3% (4.8)
Bayesian Multinomial Regression	34.2% (5.8)

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, Downey and Fearn (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, Murphy and Downey (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 were made between the red and white meats.

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

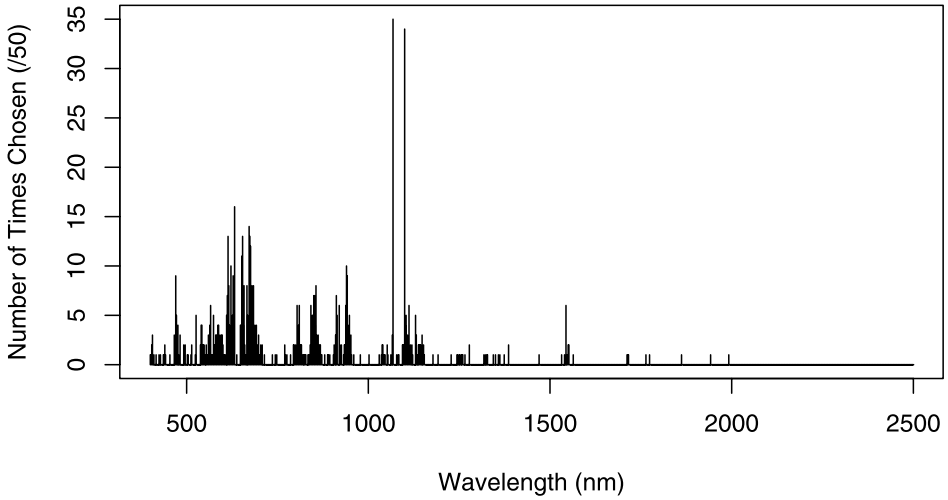

FIG. 7. *Wavelengths chosen in the five meat classification problem for the variable selection and updating method. The height of the bars shows how many times the wavelength was chosen in* 50 *random splits of the data.*

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 (400–800 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, Chen and Ozaki (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, Downey and Fearn (1999) and Dean, Murphy and Downey (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

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)
Variable selection (greedy) and updating	0.7% (0.7)
Variable selection only	1.8% (3.2)
Dean, Murphy and Downey (2006)	1.0% (0.9)
McElhinney, Downey and Fearn (1999)	2.6%–4.3%
Transductive SVMs	20.9% (8.0)
Random Forests	10.5% (3.3)
AdaBoost.M1	14.7% (3.7)
Bayesian Multinomial Regression	17.2% (4.9)

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 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. The VEV covariance structure was chosen in almost every run as the final model for both the four and five group meat classification problems.

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

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

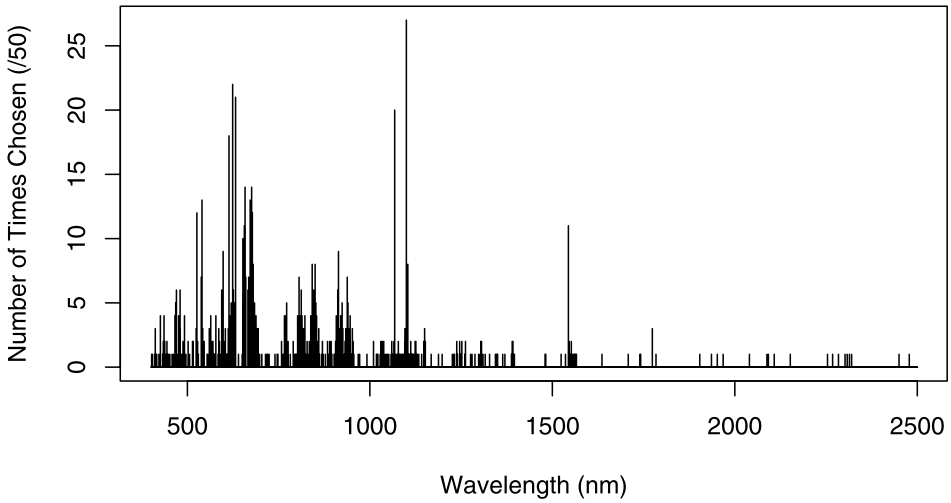

FIG. 8. *Wavelengths chosen in the four meat classification problem for the variable selection and updating method.*

and updating provides one of the best classification rates for these data. Downey, McIntyre and Davies (2003) did report a better misclassification rate (6.1%) using factorial discriminant analysis (FDA), but the choice of a subset of wavelengths, data preprocessing 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 any reference to the test data classification performance.

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. For the variable selection only results, the maximum number of selected wavelengths was restricted to be six to avoid degeneracies*

Method	**Misclassification rate**
Variable selection and updating	6.9% (5.4)
Variable selection (greedy) and updating	16.6% (11.3)
Variable selection only	17.9% (10.9)
Dean, Murphy and Downey (2006)	11.9% (6.3)
Downey, McIntyre and Davies (2003)	6.1%–19.0%
Transductive SVMs	12.4% (7.5)
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

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. The EEE covariance structure was chosen for every final model for the olive oil classification problem.

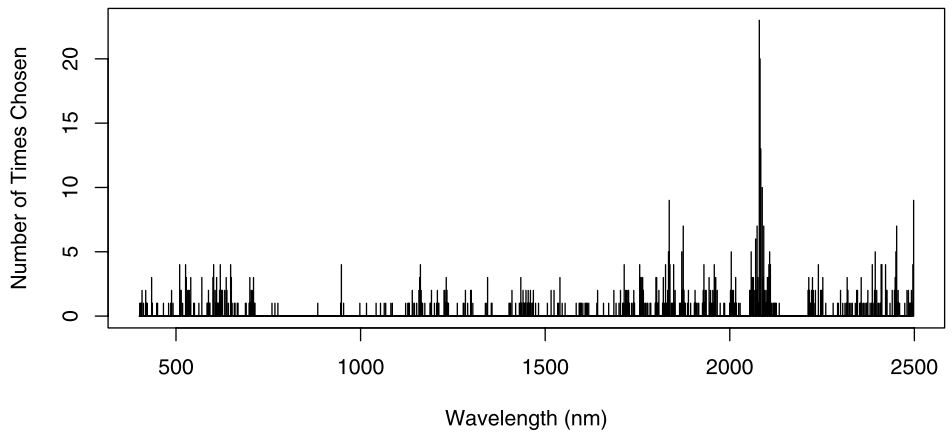

FIG. 9. *Wavelengths chosen in the olive oil classification problem using the variable selection and updating method. The height of the bars shows how many times the wavelength was chosen in* 50 *random splits of the data.*

TABLE 9
The change in classification performance for the variable selection and updating method as the number of adjacent wavelengths being aggregated increases

Aggregation level	Classification error
1	6.9%
2	9.7%
3	7.6%
5	7.9%
10	9.9%
15	8.5%
30	9.1%
50	13.2%
70	28.7%

4.3. *Sensitivity to spectral resolution*. In order to determine the sensitivity of the selected wavelengths to the resolution of the spectrometer used in this study, we investigated the effect of reducing the number of reflectance values by computing the mean reflectance value over sets of adjacent wavelengths and using these as inputs into the variable selection model. The results of this analysis are outlined for the olive oil authentication problem, and similar results were found for the meat species authenticity study.

We found that the classification error of the olive oil samples increases slightly as soon as any adjacent wavelengths are aggregated (Table 9). However, once the wavelengths are aggregated, the classification error remained steady for aggregating between 2 and 30 adjacent wavelengths. Thereafter, there was a serious deterioration in the classification performance when more than 30 adjacent wavelengths were aggregated. This suggests that a considerable amount of the group information is maintained at even low resolutions, but that there is more information in the raw data themselves.

The spectral regions selected when analyzing the data in aggregated form were found to be stable. In both applications, the selected regions were very similar for the aggregated data, but fewer variables tended to be selected because of the aggregation process. Figure 10 shows the chosen wavelengths when the raw spectra, two adjacent wavelengths and three adjacent wavelengths are aggregated and then analyzed for the olive oil classification problem. This shows that the selection procedure chooses very specific spectral regions on both the raw and aggregated scale.

5. Discussion. The discriminant analysis method presented in this paper gave much better results than those given by popular statistical and machine learning techniques such as Random Forests [Breiman (2001)], AdaBoost [Freund

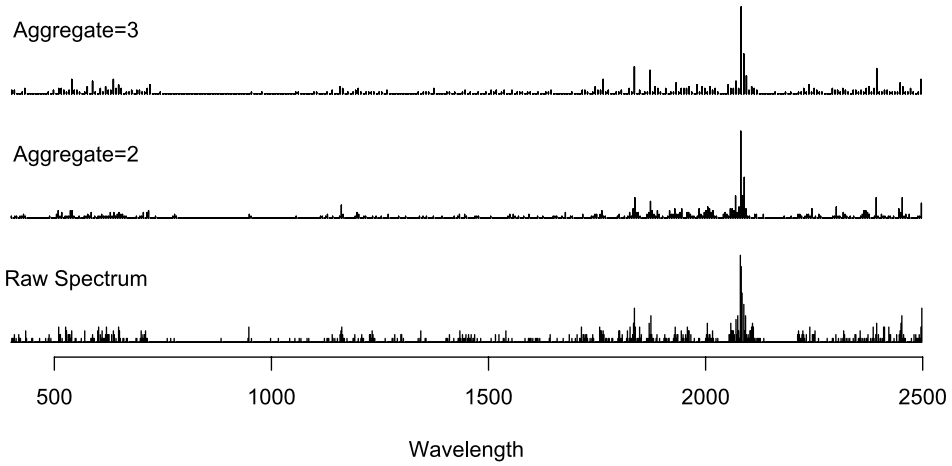

FIG. 10. *The chosen wavelengths when the raw olive oil spectra are analyzed and when adjacent wavelengths are aggregated.*

and Schapire (1997)] and Bayesian Multinomial Regression [Genkin, Lewis and Madigan (2005), Madigan et al. (2005)] and Transductive SVMs [Vapnik (1995), Joachims (1999)] for the high-dimensional food authenticity data sets analysed here. This improvement is further enhanced by the addition of the updating procedure for including the unlabeled data in the estimation method. The results show that the headlong search method for variable selection is 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 from 1050 to less than thirty. 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. The regions of the spectrum selected by the method can be interpreted in terms of the underlying chemical properties of the foods under analysis.

We have compared our method with four 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, Tibshirani and Friedman (2001)]. It is possible that the large improvement in performance of our method relative to the established methods we have compared it with is due to the fact that our data have many variables of which only a very small proportion (1%–3%) are useful. The variables that are not useful may introduce a great deal of noise and degrade performance, and so other methods that do not reduce the number of variables may suffer from this.

Although the methods were developed for the food authenticity application outlined herein, the method could be applied in contexts such as the analysis of gene

expression data and document classification. The results of the variable selection procedure could mean a substantial savings in terms of time for data collection and space for future data storage.

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)] and the stepwise stopping rule approach of Munita, Barroso and Oliveira (2006). A number of different search algorithms (proposed as alternatives to backward/forward/stepwise search) wrapped around different discriminant functions are compared by Pacheco et al. (2006), and genetic search algorithms wrapped around Fisher discriminant analysis are considered by Chiang and Pell (2004). Another example of variable selection methods in the context of classification using spectroscopic data is given by Indahl and Naes (2004).

In terms of other approaches to variable selection, 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, Robin and Daudin (2007). An extension allowing variable selection for the multiclass problem using SVMs is given by Wang and Xiatong (2007). An alternative approach for combining pairwise classifiers, based on Hastie and Tibshirani (1998), is given by Szepannek and Weihs (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. Finally, an alternative to single subset variable selection through Bayesian Model Averaging [Madigan and Raftery (1994)] is given by Dash and Cooper (2004).

Acknowledgments. We would like to thank the Editor, Associate Editor and Referees whose suggestions greatly improved this paper. We would also like to thank Gerard Downey for providing the food authenticity data and for help with interpreting the results of the analysis.

SUPPLEMENTARY MATERIAL

Supplement: Data (DOI: 10.1214/09-AOAS279SUPP; .zip). This zipfile [Murphy, Dean and Raftery (2009)] contains the data sets used in this paper. The original data source information and conditions for the use of the data are outlined in this file.

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* (Y. Dodge and J. Whittaker, eds.) **1** 251–256. Physica, Heidelberg.

BANFIELD, J. D. and RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803–821. MR1243494

BENSMAIL, H. and CELEUX, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *J. Amer. Statist. Assoc.* **91** 1743–1748. MR1439118

BREIMAN, L. (2001). Random Forests. *Mach. Learn.* **45** 5–32.

CHANG, W.-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *J. Roy. Statist. Soc. Ser. C.* **32** 267–275. MR0770316

CHAPELLE, O., SCHÖLKOPF, B. and ZIEN, A. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge. Available at http://www.kyb.tuebingen.mpg.de/ssl-book.

CHIANG, L. H. and PELL, R. J. (2004). Genetic algorithms combined with discriminant analysis for key variable identification. *J. Process Control* **14** 143–155.

COLLOBERT, R., SINZ, F., WESTON, J. and BOTTOU, L. (2006). Large scale transductive SVMs. *J. Mach. Learn. Res.* **7** 1687–1712. MR2274421

CONNOLLY, C. (2006). Spectroscopic and Analytical Developments Ltd fingerprints brand spirits with ultraviolet spectrophotometry. *Sensor Review* **26** 94–97.

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. *J. Mach. Learn. Res.* **5** 1177–1203. MR2248014

DEAN, N., MURPHY, T. B. and DOWNEY, G. (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *J. Roy. Statist. Soc. Ser. C* **55** 1–14. MR2224157

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

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.

FRALEY, C. and RAFTERY, A. E. (1999). MCLUST: Software for model-based clustering. *J. Classification* **16** 297–306.

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635

FRALEY, C. and RAFTERY, A. E. (2003). Enhanced model-based clustering, density estimation and discriminant analysis software: MCLUST. *J. Classification* **20** 263–296. MR2019797

FRALEY, C. and RAFTERY, A. E. (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. *J. Comp. System Sci.* **55** 119–139. MR1473055

GANESALINGAM, S. and MCLACHLAN, G. J. (1978). The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika* **65** 658–662. MR0521834

GENKIN, A., LEWIS, D. D. and MADIGAN, D. (2005). BMR: Bayesian multinomial regression software. Available at 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. *Ann. Statist.* **34** 2367–2386. MR2291503

GUYON, I. and ELISSEEFF, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3** 1157–1182.

HASTIE, T. and TIBSHIRANI, R. (1998). Classification by pairwise coupling. *Ann. Statist.* **26** 451–471. MR1626055

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2001). *The Elements of Statistical Learning*. Springer, New York. MR1851606

HOOS, H. H. and STÜTZLE, T. (2005). *Stochastic Local Search: Foundations and Applications*. Morgan Kaufmann, San Francisco.

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

JOACHIMS, T. (1999). Transductive inference for text classification using support vector machines. In *ICML'99: Proceedings of the Sixteenth International Conference on Machine Learning* 200–209. Morgan Kaufmann, San Francisco.

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. *Statist. Sci.* **22** 189–205. MR2408958

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. *Comput. Statist. Data Anal.* **51** 2043–2055. MR2307560

MADIGAN, D. and RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.

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* (K. H. Knuth, A. E. Abbas, R. D. Morris and J. P. Castle, eds.). *AIP Conf. Proc.* **803** 509–516. Institute of Physics, London.

MARY-HUARD, T., ROBIN, S. and DAUDIN, J.-J. (2007). A penalized criterion for variable selection in classification. *J. Multivariate Anal.* **98** 695–705. MR2322124

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*. Wiley, New York. MR1190469

MCLACHLAN, G. J. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York. MR1789474

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.

MURPHY, T. B., DEAN, N. and RAFTERY, A. E. (2009). Supplement to "Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications." DOI: 10.1214/09-AOAS279SUPP.

O'NEILL, T. J. (1978). Normal discrimination with unclassified observations. *J. Amer. Statist. Assoc.* **73** 821–826. MR0521330

OSBORNE, B. G., FEARN, T. and HINDLE, P. H. (1993). *Practical NIR Spectroscopy With Applications in Food and Beverage Analysis*. Longman Scientific & Technical, Harlow, UK.

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. *Comput. Statist. Data Anal.* **51** 1463–1478. MR2307519

R DEVELOPMENT CORE TEAM (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

RAFTERY, A. E. and DEAN, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc.* **101** 168–178. MR2268036

REID, L. M., O'DONNELL, C. P. and DOWNEY, G. (2006). Recent technological advances in the determination of food authenticity. *Trends in Food Science and Technology* **17** 344–353.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SINZ, F. and ROFFILLI, M. (2007). UniverSVM software. Version 1.1. Available at http://mloss.org/software/view/19/.

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* (M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nurnberger and W. Gaul, eds.) 700–707. Springer, Berlin.

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. *Comput. Statist. Data Anal.* **51** 3718–3736. MR2364486

VAPNIK, V. (1995). *The Nature of Statistical Learning Theory*, 2nd ed. Springer, New York. MR1367965

WANG, L. and XIATONG, S. (2007). On L_1-norm multiclass support vector machines: Methodology and theory. *J. Amer. Statist. Assoc.* **102** 583–594. MR2370855

WEST, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics* **7** 723–732. Oxford Univ. Press, Oxford. MR2003537

YEUNG, K. Y., BUMGARNER, 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.

T. B. MURPHY
SCHOOL OF MATHEMATICAL SCIENCES
UNIVERSITY COLLEGE DUBLIN
BELFIELD, DUBLIN 4
IRELAND
E-MAIL: brendan.murphy@ucd.ie

N. DEAN
DEPARTMENT OF STATISTICS
UNIVERSITY OF GLASGOW
GLASGOW, G12 8QQ
UNITED KINGDOM
E-MAIL: nema@stats.gla.ac.uk

A. E. RAFTERY
DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON, SEATTLE
BOX 354320
SEATTLE, WASHINGTON 98195-4320
USA
E-MAIL: raftery@stat.washington.edu