DIMENSIONALITY REDUCTION IN THE CREATION OF CLASSIFIERS AND THE EFFECTS OF CORRELATION, CLUSTER OVERLAP, AND MODELLING ASSUMPTIONS

A Thesis

Presented to

The Faculty of Graduate Studies

of

The University of Guelph

by

WILLIAM F. PETRCICH

In partial fulfilment of requirements

for the degree of

Master of Science

August, 2011

ABSTRACT

DIMENSIONALITY REDUCTION IN THE CREATION OF CLASSIFIERS AND THE EFFECTS OF CORRELATION, CLUSTER OVERLAP, AND MODELLING ASSUMPTIONS

William F. Petrcich
University of Guelph, 2011

Advisors:
Dr. P. McNicholas

Discriminant analysis and random forests are used to create models for classification. The number of variables to be tested for inclusion in a model can be large. The goal of this work was to create an efficient and effective selection program. The first method used was based on the work of others. The resulting models were underperforming, so another approach was adopted. Models were built by adding the variable that maximized new-model accuracy. The two programs were used to generate discriminant-analysis and random forest models for three data sets. An existing software package was also used. The second program outperformed the alternatives. For the small number of runs produced in this study, it outperformed the method that inspired this work. The data sets were studied to identify determinants of performance. No definite conclusions were reached, but the results suggest topics for future study.

Acknowledgments

Without the assistance of my advisor, Dr. Paul McNicholas, completing this project would have been much more difficult. Excellent examples helped with the process of learning LaTeX. Doing that efficiently, I accomplished and learned more through this work.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

A large amount of effort has been expended in the past conceiving of and studying the properties of various statistical algorithms for the classification of outcomes based on predictor variables. The ability to apply such algorithms is very useful in many fields. One such field is food authentication - the science of ensuring that a food product is as advertised.

Murphy et al. (2010) published a paper that served as a starting point for this work, reporting the results of comparisons of a number of classification techniques applied to a problem in food authentication. In particular, they described a method for step-wise construction of a classifier, and this served as an initial guide for the work herein. However, the results obtained using a wrapper program that selected and deleted variables based on changes in the Bayesian information criterion were not as hoped, and a search was conducted to both explain this and find approaches that would perform better. The results of that initial effort and the search are described in this paper.

Model-building wrappers using the Bayesian information criterion (BIC) and classification accuracy were created and used in conjunction with discriminant analysis and random forests to create models. In addition, an existing software pack-

age was used and the results compared. These methods were applied to three data sets. The model-building wrapper that selected variables based on a test accuracy criterion was the most successful method for all three data sets considered. Also, this wrapper was relatively easy to create and the method it uses is easily explained. The first data set studied in this work was one of the data sets used by Murphy et al. (2010). On this data set, the mean accuracy in testing of models built using the accuracy-based wrapper was higher than the mean accuracy of the BIC-based wrapper described and tested by Murphy et al. (2010).

The other two data sets have differing characteristics that allowed the testing of ideas about the performance of the different variable-selection methods and models used with them, and the effects of data structure on performance in general. It was hoped that useful lessons could be learned by linking data dissimilarities and differences in classification performance. No sweeping conclusions were reached, but at the very least, avenues for refinement of this work and future study have been identified.

Four sections follow this one. Theoretical material is introduced and explained in Chapter 2; the procedure used is then described in Chapter 3 with emphasis on links between theoretical material and specific issues in the implementation of the methods used. The results of the data analysis are presented next in Chapter 4; and, finally, the author's impression of the results are presented in Chapter 5.

Chapter 2

Background

Normal mixture models can be used to describe a population of interest when that population is assumed to consist of individuals that belong to distinct subgroups that can each be modelled using a multivariate-normal distribution . The normal mixture model has density

$$f(\mathbf{x}) = \sum_{g=1}^{G} \tau_g \phi(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{2.1}$$

where f() is the mixture pdf, and each ϕ, referred to as a component of the mixture, is one of G multivariate normal pdf's for a distribution with mean $\boldsymbol{\mu}_g$ and covariance $\boldsymbol{\Sigma}_g$. The τ_g are the proportions in which the components mix, with $\tau_g > 0$ for all g and $\sum_{g=1}^{G} = 1$.

Modelling using normal mixture models leaves choices for component means and covariance matrices. The use of models that allow constraints on covariance matrices allows for the creation of a model that has characteristics reflecting the potential nature of the data. This also allows for systematic model selection, which will be mentioned again in Chapter 3. Banfield and Raftery (1993) describe a system for producing such models. The covariance matrices are redefined in terms of their eigenvalue decompositions,

$$\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g^t, \tag{2.2}$$

where $\mathbf{\Sigma}_g$ is the group covariance matrix for the g-th group, λ_g is the largest eigenvalue for the decomposition, $\mathbf{D}_g$ is the matrix of eigenvectors, and $\mathbf{A}_g$ is the diagonal matrix having the corresponding scaled eigenvalues as entries. They also explain the function of each component in the decomposition: for the g-th component, $\mathbf{D}_k$ determines cluster orientation while λ_k and $\mathbf{A}_k$ determine cluster size and shape, respectively (Banfield and Raftery, 1993). The implementation of this scheme is discussed in Chapter 3.

A portion of the work presented in this paper makes use of an analysis method called model-based discriminant analysis. Fisher (1950) describes the method of linear discriminant analysis; the method is not model-based but makes use of data to obtain a specific linear function because "when two or more populations have been measured in several characters, $x_1, ..., x_s$, special interest attaches to certain linear functions of the measurements by which the populations are best discriminated".

The goal of linear discriminant analysis is to obtain a linear function for classifying individuals in a population. Within each sub-population, or species in the case of Fisher's example, this discriminant function will take on values and these will have a mean and variance. For this function, with a greater difference between within-group means and simultaneously smaller within- group variances, overlap in values that Fisher (1950) discusses are less likely to occur and so the function will be more likely to correctly classify observations. In this classic case, overlap is an important consideration.

Other methods for discriminant analysis exist. Also, there are other methods for classification. One of these, random forests, was used in this work. For back-

ground information about this method, the reader is referred to Breiman and Cutler (2011). Random forests creates an ensemble of decision trees for classification. A classification variable space and partitioning of it is selected based on best splits using a single variables. At each split, the algorithm is similar to linear discriminant analysis in that it is using a rule to separate observations into subgroups. It is different in that it is not searching for clustering, and it is limited to using splitting rules using a value for a single variable.

In a situation where the assumption of a mixture model like that defined by Equation 2.1 is valid, model-based discriminant analysis is appropriate. Classification is carried out based on the probability model, with an observation assigned membership in the group for which it has the highest probability of occurrence, based on the model. This method is used in the work presented here. In model-based discriminant analysis, there is no discriminant function for which separate clusters might take on overlapping ranges of values. It is the clusters defined by the probability model itself that may overlap and cause difficulties in classification.

A probability model such as Equation 2.1 for model-based discriminant analysis defines a distribution in terms of parameters and variable values - location in the space over which the pdf is defined. Murphy et al. (2010) assess the performance of a wrapper for dimensionality reduction in discriminant analysis, describing and implementing a practical method for selecting the variables used in the model based on the BIC,

$$\text{BIC} = 2l(\mathbf{x}, \hat{\boldsymbol{\Theta}}) - m \log n, \tag{2.3}$$

where $l(\mathbf{x}, \hat{\boldsymbol{\Theta}})$ is the natural log of the maximized likelihood for the observed data and parameter estimates, m is the number of parameters, and n is the number of

observations. The BIC is used in this context as an approximation to the logarithm of the ratio of two integrated likelihood values, one obtained using a model where the variable under consideration contains information about the grouping structure and one using a model were it does not (Murphy et al., 2010). If the Bayes factor is greater than zero, this ratio is greater than one, an outcome suggesting the variable is useful for modelling purposes. The logarithm of the Bayes factor is approximately equal to one half of the change in the BIC caused by the addition or deletion of a variable from the model (Murphy et al., 2010):

$$\log(\text{Bayes Factor}) \approx \frac{1}{2} \times [\text{BIC(Grouping)} - \text{BIC(No Grouping)}]. \tag{2.4}$$

A variable is selected or deleted based on this BIC difference; this is discussed further in Chapter 3.

Use of the BIC for the purposes of model selection is not novel. Its use in approximating the integrated likelihood is discussed by Biernacki and Govaert (1999); they also conclude that, for discriminant analysis, the BIC may be used as a theoretically-justified and efficient way to select among covariance models such as those given by 2.2. Steele and Raftery (2009) found that, for a simulation study, the BIC performed best among six methods tested for selection of the number of components in a mixture model. The BIC is used in these latter two ways in the software used for this work, `mclust` (Fraley and Raftery, 2009). In a theoretical treatment, Kerebin (2000) looks at the asymptotic properties of the BIC in estimating the number of mixture components and gives the results of an experiment from which he concludes that "the BIC criterion is seen to be very reliable". Implementation of the BIC in this work is mentioned again in Chapter 3.

Normal probability models can also be used in model-based agglomerative hierarchical clustering. However, the model for hierarchical clustering in the software package used for this work differs from that used for discriminant analysis. A normal distribution is assumed for each class. Unlike with model-based discriminant analysis, there is not a mixture of contributions of pdf's to determine density at a given data point. For this clustering model, this results in the classification likelihood given by Fraley and Raftery (2009),

$$\prod_{i=1}^{n} \phi_{l_i}(\mathbf{x}_i|\boldsymbol{\mu}_{l_i}, \boldsymbol{\Sigma}_{l_i}), \tag{2.5}$$

where the l_i each take on a unique value from 1 to G, the number of sub-groups in the population. They are indices defining cluster membership. The $\phi_{l_i}()$ are normal pdf's with parameters $\boldsymbol{\mu}_{l_i}$ and $\boldsymbol{\Sigma}_{l_i}$ (Fraley and Raftery, 2009). In the case of agglomerative hierarchical clustering, each observation is usually treated as an individual cluster initially, and at successive stages, the pair of clusters merged is the one for which Equation 2.5 is maximized. In this work, hierarchical clustering was used for initialization of the EM algorithm used to obtain the mixture models ultimately used for discrimination. This will be discussed in Chapter 3.

A number of functions in the software used for this work include instances of the EM algorithm, an iterative procedure for obtaining maximum likelihood estimates, in this case of model parameters. Robert and Casella (2010) describe a procedure that must be initialized and then consists of two repeated steps:

1. Initialization: pick starting values for parameters;

2. E-step: calculation of an expected value based on observed data and current parameter estimates;

3. M-step: re-calculation to obtain the new parameter estimates that maximize this expected value;

steps 2 and 3 are repeated until some "fixed point", a stopping condition, is reached. A theoretical treatment of the EM algorithm including discussions of its application with multivariate normal data and mixtures, including normal mixtures, is given by Dempster et al. (1977). As a practical example, for the software used in this work, Fraley and Raftery (2009) describe the mechanics of the EM algorithm:

> "an iteration of EM consists of an 'E'-step, which computes a matrix $\mathbf{z}$ such that $\mathbf{z}_{ik}$ is an estimate of the conditional probability that observation i belongs to group k given the current parameter estimates, and an 'M'-step, which computes parameter estimates given $\mathbf{z}$."

Practical issues need to be considered when using EM. Fraley and Raftery (2007) state that "the results of EM are highly dependent on the initial values, and model-based hierarchical clustering can be a good source of initial values". What EM is actually finding is most likely to be a local rather than global maximum and which of these it finds depends on how it is initialized. Biernacki and Govaert (1999) explain their use of the EM algorithm, stating that it was "started with the true underlying centers" and that solutions found may be suboptimal because EM is "started only once with the true centers". For the model-based analysis package used in this work, Fraley and Raftery (2009) treat initializing values as control parameters for EM and also discuss the use of a prior distribution in this context. A theoretical treatment of the priors available in the software can be found in Fraley and Raftery (2007).

For the BIC-based wrapper created as part of this work, a scheme described by Dean et al. (2006) for model updating was implemented. This procedure uses both data having class labels and unlabelled data in model training, and involves use of the EM algorithm. There is an initialization step plus three steps repeated for subsequent iterations (Dean et al., 2006):

1. Initialization by training a mixture model using data with labels and then using the resulting parameter estimates;
2. Calculation of the expected values of the unknown labels;
3. Recalculation of parameter estimates using both labelled and unlabelled data;
4. Convergence check using a stopping criterion.

The stopping procedure they give was used. First, after three complete iterations, the second-iteration estimate of the log-likelihood on convergence is estimated by

$$l_{\infty}^{(k)} = l^{(k)} + \frac{1}{1 - a^{(k)}}(l^{(k+1)} - l^{(k)}), \tag{2.6}$$

where

$$a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}}, \tag{2.7}$$

and k is the iteration number (Dean et al., 2006). $|l_{\infty}^{(k)} - l^{(k)}|$ is calculated and recalculated for each subsequent iteration with the procedure halted once this value becomes sufficiently small (Dean et al., 2006). They used the updating method to carry out a discriminant analysis of the same meats data set studied in this work, and they concluded that rule updating using unlabelled data resulted in improved classification performance over standard discriminant analysis (Dean et al., 2006).

O'Neill (1978) discusses the use of unlabelled observations in linear discriminant analysis. He shows that, under certain conditions, the information content of unlabelled data is about one-fifth to two-thirds that of labelled data for estimation of Fisher's linear discriminant rule (O'Neill, 1978). He argues first (O'Neill, 1978) that the Fisher information matrices of classified observations ($\mathbf{I}_C$) equal the sum of the Fisher information matrices for an unclassified instance of the same observation ($\mathbf{I}_{UC}$) plus an instance where the classification is treated as conditional given the observation ($\mathbf{I}_{LR}$),

$$\mathbf{I}_C = \mathbf{I}_{UC} + \mathbf{I}_{LR}. \tag{2.8}$$

The asymptotic relative efficiency of two unbiased estimators for the same quantity is the limiting value of the ratio of their variances as sample sizes become infinitely large. It allows comparison of two estimators based on their sampling distributions. He defines a measure of asymptotic relative efficiency for the discrimination procedure, in terms of asymptotic error rates in classification (AER), as the ratio of the AER for complete data divided by the AER for the data with unlabelled observations and then uses a more easily-found equivalent quantity based on efficiencies of estimators for discriminant function parameters to calculate values for this ratio (O'Neill, 1978). This allows quantification of the effects of using unlabelled data in the training of a model for linear discriminant analysis.

In summary, his argument centers on two ideas:

1. Because the information in a complete observation can be decomposed into a component arising from explanatory data alone and a component arising from the observation obtained by using this data to create a label, information in

unlabelled data is a defined fraction of the information in complete data;

2. Using a defined value for asymptotic relative efficiency, relative classification performance in linear discriminant analysis can be calculated.

An equivalent explanation for the case of model-based discriminant analysis has not been found in the literature, but it appears possible to extend this argument. Other approaches to the problem have also been proposed. In the context of developing a rule for discrimination between two populations, McLachlan (1975) discusses the use of risk in evaluating the relative performance of models constructed using complete data and data with unlabelled observations.

Chapter 3

Methodology

The original goal of this work was to create a variable selection wrapper for discriminant analysis based on the approach of Murphy et al. (2010) introduced in Chapter 2. Although some preliminary exploration of the data was done with MS Excel ®, the statistical software package R (R Development Core Team, 2010) was used for all of the calculations and results presented later in this paper. All work was done using a Hewlett-Packard G61 laptop with an AMD Athlon(tm) II Dual-Core M320 processor.

For the first model-generating algorithm, the BIC-based model evaluation function `mclustBIC` and the model-based discriminant analysis functions `mclustDA`, `mclustDAtrain`, and `mclustDAtest` of the add-on package `mclust` (Fraley and Raftery, 2002, 2006) were used, incorporated into a wrapper created using original script written in the built-in programming language. The `mclustDAtrain` and `mclustDAtest` functions split the model training and testing capabilities that are combined in `mclustDA`. The algorithm is described below.

initialization:

create list with all variables

get BIC for model containing each one

select the variable with highest BIC value for inclusion

remove this variable from the list

while test accuracy $<$ cut-off value **and** number of variables not in preset range:

check ΔBIC for adding each variable in list to model

add first variable in list for which ΔBIC $>$ cut-off

remove this variable from list

check ΔBIC for removal of each variable in model

remove first, in reverse order of inclusion, for which ΔBIC $>$ cut-off

use `mclustDA` to create and test model

end while

output: preliminary model

begin updating:

initialize: classify unlabelled data using model and `mclustDAtest`

while not converged

E-step: use all data to create new model with `mclustDAtrain`

M-step: use with `mclustDAtest` to obtain new labels for unlabelled data

end while

In an initialization step, the best variable in terms of BIC becomes the first variable in the model; it is removed from the list of variables to be checked. Then, each subsequent step consists of two sub-steps. First, the algorithm obtains changes in BIC values resulting from the addition of each of the remaining variables to the existing model and adds the first one in the list for which the BIC change is greater than a cut-off value set by the experimenter. The variable selected for addition is

then removed from the list of potential variables. Unlike in the initialization step, a search is then made of all variables in the current model to determine if one can be eliminated; changes in BIC due to the elimination of each variable in the model are obtained and the first, if any, taken in reverse order of inclusion, for which the BIC difference is greater than a cut-off value, is eliminated. As for the additive step, the cut-off value for this elimination step was set by the experimenter. This reverse cut-off value was set to encourage model growth; a relatively high value pushed the algorithmic process more quickly toward a locally-optimal model rather than a model potentially closer in performance to the BIC-optimal model.

The lower part of the pseudocode shows the algorithm used in updating the model using the unlabelled data. The procedure implemented is that of Dean et al. (2006) described in Chapter 2. The EM algorithm is initialized with a model built using only the training data to classify the unlabelled data and then the resulting labels are appended to the unlabelled explanatory data. This full data set is then used to create a new model, an M-step. This model is used to estimate new labels, an E-step, that replace the existing labels for the unlabelled subset of observations. This new estimate is then used as in the M-step described above and this process is repeated until convergence is obtained, using the criterion described in Chapter 2. Unfortunately, this algorithm does still suffer from the problem experienced by Biernacki and Govaert (1999) mentioned in Chapter 2.

Once it became apparent that this wrapper was underperforming in terms of accuracy in classifying test data, potential reasons were identified and a second wrapper was created. Pseudo-code for this second wrapper is given on the next page.

initialization:

create list with all variables

use `mclustDAtrain` to create model containing each variable

use `mclustDAtest` to get accuracy for each model

select the variable with highest accuracy for inclusion

remove this variable from the list

while test accuracy $<$ cut-off value **and** number of variables not in preset range:

create model with each new variable included (`mclustDAtrain`)

get accuracy for each of these models (`mclustDAtest`)

add variable for which accuracy is greatest

remove this variable from list

end while

In an initialization step, the best variable in terms of classification accuracy is selected from the initial list of 1050 available explanatory variables and becomes the first variable in the model. Each subsequent step involves the checking of classification accuracies for models created by adding each potential new variable to the existing model and selecting the new model for which classification accuracy is maximized. There is no elimination step. While model performance in classification of test data and the number of variables included in the model are both below thresholds set by the experimenter, the model is simply grown. This wrapper is also constrained to construct locally-optimal solutions. It may, occasionally and coincidentally, obtain the optimal solution, and matched it for one case studied in this work.

Having created this second wrapper and wanting to have other results for comparison, this second wrapper was also used to build models for the `randomForest` function in R; this is an implementation of the methodology described by Breiman and Cutler (2011) and mentioned in Chapter 2. The pseudo-code is shown below.

initialization:

create list with all variables

use `randomForest` to create model containing each one

use `predict.randomForest` to test accuracy for each model

select the variable with highest accuracy for inclusion

remove this variable from the list

while test accuracy $<$ cut-off value **and** number of variables not in preset range:

use `randomForest` to create model with each variable in list added

use `predict.randomForest` to test accuracy of each

add variable for which accuracy is greatest

remove this variable from list

end while

The best variable based on classification accuracy becomes the first variable in the model in an initialization step and is deleted from the list of remaining variables. At each subsequent step, the variables that remain for potential addition are each added to the existing model, and the model with the highest classification accuracy is adopted as the new model. Again, there is no elimination step. Although an accuracy cut-off was set, in practice it was usually not reached, so models were

grown until either improvements in accuracy began to diminish or they included a number of variables determined by experiment. The BIC wrapper generates sets of explanatory variables independently of the classification algorithm being used. So, to obtain BIC-generated `randomForest` models, the classifier was applied to the locally-optimal models already obtained using the wrapper for discriminant analysis.

For the reader's reference, Technical Report 504 (Fraley and Raftery, 2009) is a user's manual for the package `mclust`. Using the `mclust` package involved implementation of many of the ideas discussed in Chapter 2. For example, the first and second wrapper were built around the discriminant-analysis model training function `mclustDAtrain`, and use of this function involved consideration of a number of these issues. Some of these will be discussed now because they are fundamental or necessitated obvious departures from the default settings of the function.

From Technical Report 504: "The idea is to produce a density estimate for the training data which is a mixture model, in which each class is modeled by a single Gaussian term" (Fraley and Raftery, 2009); discriminant analysis in the `mclust` package makes use of the model given in Equation 2.1.

In `mclust`, the covariance decomposition scheme discussed in Chapter 2 is implemented. The models are listed, along with characteristics, in Table 3.1; this information is reproduced from Table 1 of the software user's manual (Fraley and Raftery, 2009). The use of these constraining models makes the search for the best model systematic because each of these is considered in turn and the best of them is selected. In practice, this also allows one to limit a search to certain models, if appropriate.

Table 3.1: Covariance structures available with `mclust`.

Name	Model	Distribution	Volume	Shape	Orientation
E	Univariate	Equal	-	-	-
V	Univariate	Variable	-	-	-
EII	$\lambda\mathbf{I}$	Spherical	Equal	Equal	NA
VII	$\lambda_k\mathbf{I}$	Spherical	Variable	Equal	NA
EEI	$\lambda\mathbf{A}$	Diagonal	Equal	Equal	Coord. axes
VEI	$\lambda_k\mathbf{A}$	Diagonal	Variable	Equal	Coord. Axes
EVI	$\lambda\mathbf{A}_k$	Diagonal	Equal	Variable	Coord. Axes
VVI	$\lambda_k\mathbf{A}_k$	Diagonal	Variable	Variable	Coord. Axes
EEE	$\lambda\mathbf{D}\mathbf{A}\mathbf{D}^t$	Ellipsoidal	Equal	Equal	Equal
EEV	${}_k\mathbf{A}\mathbf{D}_k^t$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^t$	Ellipsoidal	Variable	Equal	Variable
VVV	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^t$	Ellipsoidal	Variable	Variable	Variable

The BIC is used in the first wrapper described earlier in this section. The BIC is also a prominent tool in the `mclust` package and is used by a number of its functions. As mentioned in Chapter 2, it is used in covariance-model selection and in selection of the number of mixture components; this is discussed in Technical Report 504 (Fraley and Raftery, 2009).

Also from Technical Report 504: Function `hc` "implements fast methods based on the multivariate normal classification likelihood" (Fraley and Raftery, 2009). Again, this is the method described in Chapter 2. However, it was found that the results produced by `hc` did not closely reflect the data. As discussed previously, the results produced by the EM algorithm depend on how it is initialized. To force the initialization to produce a better resulting model, a program was written to produce a matrix with attached attributes mimicking `hc` output. This imitation defined the initial clustering present in each set of training data. It's use for EM initalization improved classification accuracies.

As discussed in Chapter 2, use of a prior distribution is an available means for controlling the results produced using EM. In addition to changing the default initialization where possible, use of a constraining prior distribution based on the data was adopted. The default settings for component means for a prior are the averages obtained from the data (Fraley and Raftery, 2009); these were used without modification. However, the default scale settings were changed to better reflect the data. The defaults are conjugate priors; an iteration of EM theoretically should produce similarly-distributed posterior estimates with which to start the next iteration.

As mentioned above, the R function `randomForest` from the package with the same name was also used. Default settings were used, as they were for the work by Murphy et al. (2010). Descriptions of `randomForest` and its features are available in the R documentation and online (Breiman and Cutler, 2011). While training data was used to create `randomForest` models, the function `predict.randomForest` in the same package was used to obtain classification accuracies based on model performance in predicting labels for the corresponding test data sets. Also, a fourth method was tested: the function `clustvarsel` from the R package with the same name. This function produces simultaneous estimates of the optimal number of clusters and their covariance structures, and also returns the subset of explanatory variables used in this model. To obtain classification results, the models obtained using `clustvarsel` with training data were used to create discriminant-analysis models in `mclust` and each of these was applied to its matched test data. For details regarding `clustvarsel`, the reader is referred to Raftery and Dean (2006).

For the purpose of making comparisons, it was assumed that all accuracies obtained using the methods described above are normally distributed. Thus, two-

sample t-tests were performed using R to compare mean accuracies for the methods. Because of small sample sizes, this was done using original script implementing a small-sample test described by Stock and Watson (2007). Results of these tests are also presented in the next section.

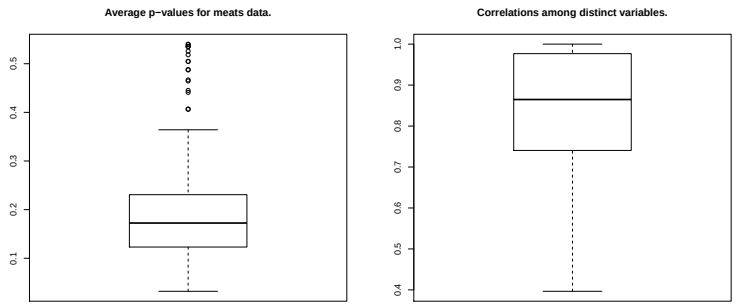
Chapter 4

Data Analysis

4.1 Data sets

Three data sets were studied. The first, originally introduced by McElhinney et al. (1999), was the five-meat dataset studied by Murphy et al. (2010). Reflectance is a measure of the fraction of incident electromagnetic radiation that is reflected, in this case from the samples of meat studied to create the data, expressed as a percentage. The data set contains 231 observations; for each of these, there are 1050 reflectance measurements taken at wavelengths from 400nm to 2498nm in increments of 2nm. The data set also contains a categorical variable, type; for each meat sample, type is either chicken, turkey, pork, beef, or lamb. There are 55 observations each for the first three meat types listed, 32 for beef, and 34 for lamb. The p-values obtained for group-wise Shapiro-Wilk testing of each variable revealed 29 of the 1050 potential explanatory variables had average p-values of less than 0.05. This was taken to suggest that non-normality within subgroups may not be a serious problem for most of the variables. A boxplot of average p-values for the 1050 potential explanatory variables is shown in Figure 4.1. Graphical checks of bivariate scatterplots also did not reveal systematic problems with the assumption of multivariate normality. In addition to this check, for random samples of 10-variable subsets, the within-group distributions of squared distances between group member and mean were checked

Figure 4.1: Meats data set.



using quantile-quantile plots to see if they were approximately chi-squared. Again, no obvious problems appeared.

The correlation matrix for the 1050 explanatory variables was examined. The average correlation between different variables is 0.845. A boxplot for the correlations between different variables is shown in Figure 4.1. Also, visual examination of a sample of side-by-side boxplots revealed overlapping of ranges among the variables. An example is shown in Figure 4.2.

The second data set consists of two observations collected on each of 60 samples of authenticated extra virgin olive oils from four different European countries. Each observation consists of a categorical variable identifying country of origin and 570 measures of reflectance at different infrared wavelengths. The data set is described by Tapp et al. (2003). A larger part of the data, 60 of the 570 potential explanatory variables had average p-values below 0.05 when subjected to group-wise Shapiro-Wilk testing. A boxplot of the average values is shown in Figure 4.3. Samples of bivariate scatterplots and chi-square quantile-quantile plots of within-group squared distances from group mean were difficult to interpret. Subjectively, there appeared to be more

Figure 4.2: Within group boxplots for measurements at $1690nm$.

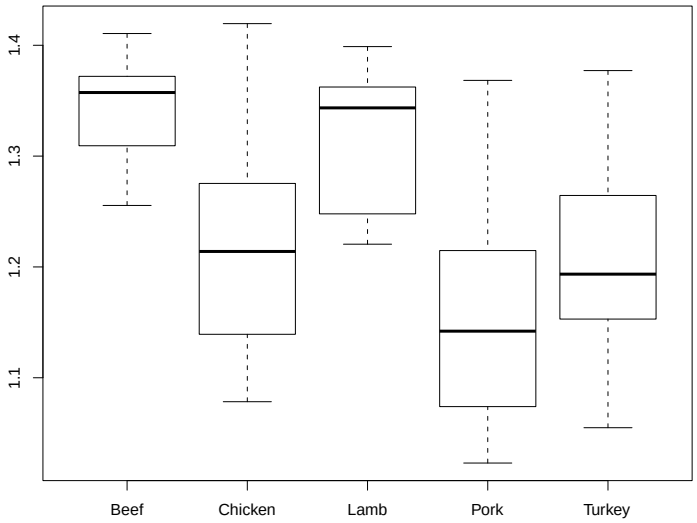


Figure 4.3: Olive oil data set.

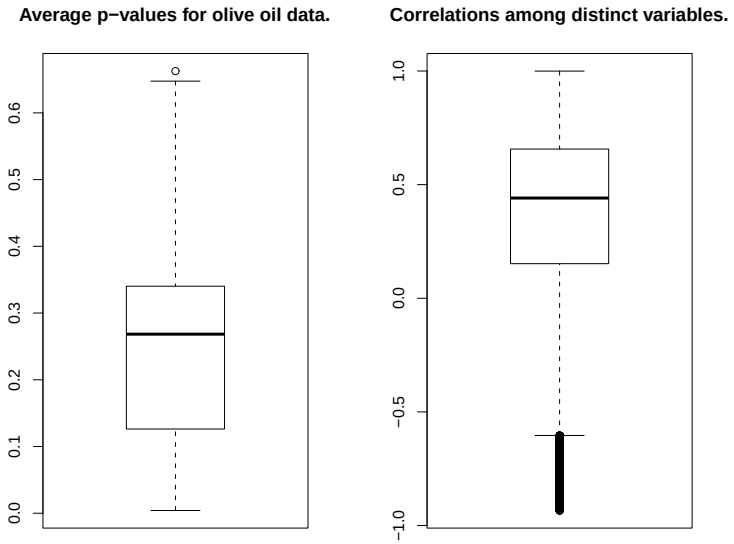
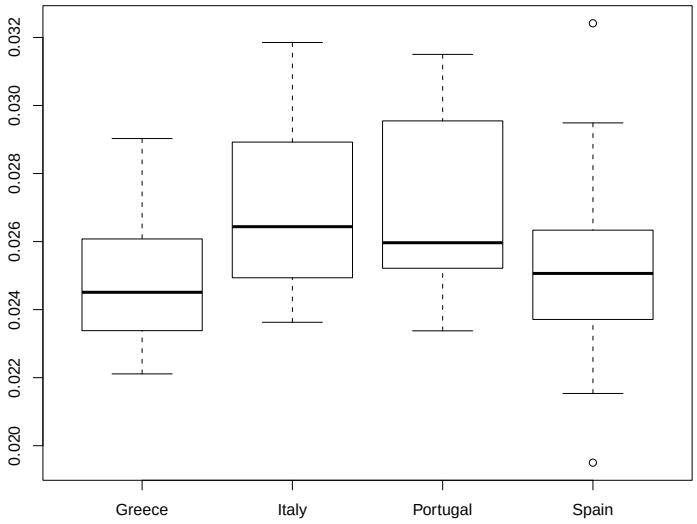


Figure 4.4: Within-group boxplots for measurements of variable 405.

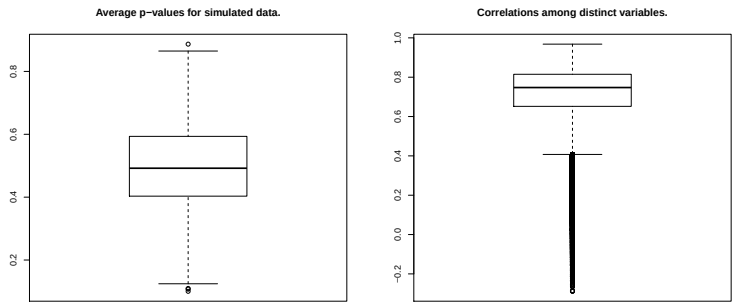


potential problems with an assumption of multivariate normality among the samples of variables tested.

The average absolute value of the correlation between different explanatory variables was found to be 0.46. A boxplot showing correlation values for the olive oil data is shown in Figure 4.3. Examination of 10 samples of boxplots for within-group variable values revealed overlapping within-group ranges. An example is shown in Figure 4.4.

The third data set was simulated with care taken to create a five-group sample with group proportions equal to those in the meats data and the potential for both minimal overlap of some groups and nesting of others. Data values were generated randomly from normal distributions. The minimum group-wise average

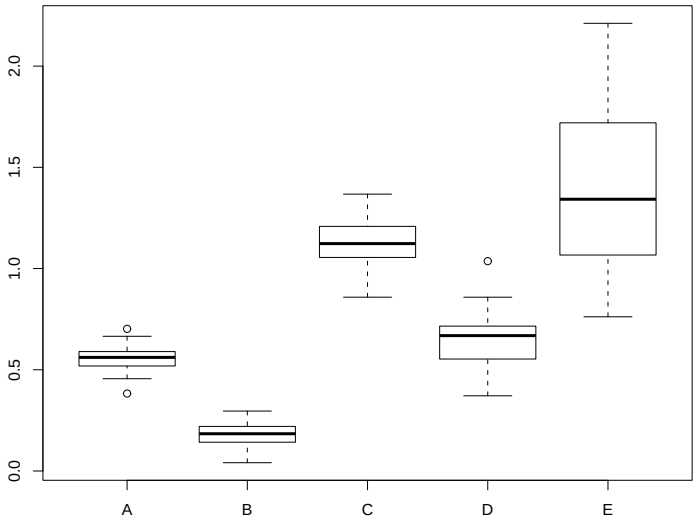
Figure 4.5: Simulated data set.



p-value for the 1050 potential explanatory variables, from Shapiro-Wilk testing for non-normality, was found to be 0.1. A boxplot for the average p-values is shown in Figure 4.5.

Checks of bivariate scatterplots and distribution of squared distances did not reveal obvious issues with an assumption of multivariate normality. The average absolute value of the correlation between variables is 0.715. A boxplot of correlations between different variables is given in Figure 4.5. Finally, overlap was also checked for this data set. Visual examination of a sample of group-wise boxplots suggested less overlap among sub-groups compared to the other two data sets. An example of group-wise boxplots is shown in Figure 4.6.

Figure 4.6: Within-group boxplots for variable 787.



4.2 Results

Results of applications of the selection methods for each data set are summarized in the tables that follow. For each data set, analysis type, selection method, number of replicates, mean accuracy, and standard deviation are shown.

Table 4.1: Results for meats data.

Analysis method	Wrapper type	n	Mean accuracy	Stan. dev.
Discriminant analysis	BIC	9	90.92	2.92
Discriminant analysis	Class. acc.	7	96.15	2.00
`randomForest`	BIC	9	74.01	2.36
`randomForest`	Class. acc.	5	82.06	3.04
`clustvarsel`	None	5	89.22	4.06

Where an assumption of equality of variances was not ruled out by testing of the results data, hypothesis tests were carried out using a significance level of 0.05.

Table 4.2: Results for olive oil data.

Analysis method	Wrapper type	n	Mean accuracy	Stan. dev.
Discriminant analysis	BIC	8	76.26	3.43
Discriminant analysis	Class. acc.	8	99.17	0.89
`randomForest`	BIC	8	67.71	9.26
`randomForest`	Class. acc.	8	94.79	0.59
`clustvarsel`	None	5	87.66	4.80

Table 4.3: Results for simulated data.

Analysis method	Wrapper type	n	Mean accuracy	Stan. dev.
Discriminant analysis	BIC	8	97.07	2.50
Discriminant analysis	Class. acc.	8	100.00	0.00
`randomForest`	BIC	8	65.87	13.72
`randomForest`	Class. acc.	8	73.47	14.68
`clustvarsel`	None	5	84.38	3.01

For the meats data set, with an attained significance of $p = 1.33 \times 10^{-5}$, the results indicate a difference for values of mean accuracy between the BIC- and accuracy-based wrappers with the sample estimate being higher for the accuracy-based wrapper. The results do not suggest a difference between mean accuracies for the BIC-wrapper and `clustvarsel`-generated discriminant-analysis models applied to the meats data ($p = 0.118$). Finally, for the meats data, the results indicate a statistically-significant difference in mean accuracies attained using `randomForest`, with ($p = 1.13 \times 10^{-6}$), between the BIC- and accuracy-based wrappers with the sample value again being higher for the accuracy-based wrapper.

While potential inequality of variances ruled out a t-test, there is a difference between sample estimates of mean accuracy of 22.91 percentage points for the BIC- and accuracy-based wrappers applied to the olive oil dataset for discriminant analysis with the accuracy-based wrapper attaining a higher mean. The results ($p = 6.73 \times 10^{-7}$) indicate rejection of equality of mean accuracies between the BIC-

wrapper and `clustvarsel`-generated discriminant-analysis models, with the latter achieving a higher value of mean accuracy. Again, a t-test could not be used, but the sample estimate of mean accuracy achieved by randomForest for the olive oil data was 27.08 percentage points higher for the accuracy-based wrapper than for the BIC wrapper.

For the simulated data, the performances of the BIC- and accuracy-based wrappers with discriminant analysis were much closer, with both achieving high success rates on the test data. Again, with an attained significance level of 4.46×10^{-8}, the results support rejection of equality of mean accuracies between the BIC-wrapper and `clustvarsel`-generated discriminant-analysis models with the BIC wrapper performing better. Over eight replicates, the best performer was the accuracy-based wrapper, which was perfect for all replicates. For `randomForest`, the results indicate that the mean accuracies attained using the BIC- and accuracy-based wrappers differed, with the accuracy-based wrapper achieving a greater mean accuracy. However, as will be discussed in the next section, their performance on this data set was erratic, so the assumption of normality underlying a small-sample t-test may not be valid. To allow, perhaps, a better comparison, the median accuracy for the BIC wrapper was 56.52 % and was 68.65 % for the accuracy-based wrapper; a Mann-Whitney test yielded a p-value of 0.004, indicating a difference in locations.

Tests were also conducted among data sets. The data suggest rejection of equality of mean accuracies for the BIC wrapper applied to the meats and olive-oil data ($p = 3.32 \times 10^{-11}$). The results do not support rejection of equality of mean accuracies for `clustvarsel` applied to these two data sets ($p = 0.277$). However, with a p-value of 0.001, the results allow rejection of equality of mean accuracies for

`clustvarsel` applied to the meats and simulated data with higher accuracy achieved on the meats data. The opposite occurred for the BIC wrapper applied to these two data sets; with $p = 1.43 \times 10^{-6}$, the results indicate a significant difference between mean accuracies with the BIC wrapper performing better on the simulated data.

For the meats data, on average, models selected for discriminant analysis using the BIC wrapper had correlation matrices with different means of absolute values of off-diagonal entries when compared with models built using the accuracy wrapper. While inequality of variance and small sample size precluded testing, the observed difference was 0.095 with the accuracy-selected models having the greater mean. Between the BIC wrapper and the accuracy wrapper, a difference in corresponding results was also observed with `randomForest`. Use of multiple tests gave conflicting results with the conservative approach indicating that the difference was not significant ($p = 0.13$) while a Mann-Whitney test suggested it was ($p = 0.003$). The mean value was higher for the accuracy-based wrapper again.

The BIC wrapper with both discriminant analysis and `randomForest` achieved lower mean accuracies when applied to the olive oil data. For the models selected, among the explanatory variables, the averages of absolute values of the non-diagonal entries of the correlation matrices ranged from 0.41 to 0.96 with a median of 0.96. This is appears to be in sharp contrast to the average of 0.46 for the entire dataset and the values for the accuracy-selected models, which averaged 0.398. Again, no hypotheses were tested, but qualitatively, this is a large difference. When the accuracy wrapper was used with `randomForest`, the average across the eight replicates of results was 0.281. This reverses what was observed in the meats data.

For the simulated data set, the averages of absolute values of non-diagonal entries of the correlation matrix for each model produced using the BIC were between 0.66 and 0.78 with a median of 0.66. For models produced for discriminant analysis using the accuracy wrapper, the median was 0.878. BIC-wrapper values do not differ as dramatically from the corresponding average of 0.715 for the entire data set as did the values for the olive-oil models. Again, for this data, values are higher in accuracy-selected models than BIC-selected models, as they were for the meats data. The corresponding values for models constructed using the accuracy-based wrapper and `randomForest` ranged from 0.69 to 0.86 with a median of 0.73. A Mann-Whitney test gave a p-value of 0.008, supporting rejection of equality of median values between the `randomForest` models created using the BIC- and accuracy-based wrappers. Here, the medians follow the pattern observed with the meats data as well.

The classification results and tests will be discussed further in Chapter 5. In summary, the accuracy-based wrapper performed best in all instances. The BIC-based wrapper's performance improved as prelminary checks suggested the data being analysed presented fewer problems for use of the normal mixture model of Equation 2.1. `clustvarsel` produced its best results on the meats data and did better on the olive oil data than the BIC-based wrapper that used higher cut-off values for variable addition and deletion. Finally, the use of models built using the BIC produced uniformly inferior average test accuracies in `randomForest` models. With the accuracy-based wrapper, `randomForest` models achieved highest test accuracies for the olive oil data, where correlations tended to be lower.

Chapter 5

Conclusions

As discussed in Chapter 4, for the meats data, the results of preliminary tests did not appear to rule out the fitting of many of the potential combinations of variables with a normal mixture. The BIC-based wrapper, making use of a normal mixture in the modelling process, did not perform as well as the wrapper of Murphy et al. (2010). It was also outperformed by the accuracy-based wrapper, which also achieved a higher mean accuracy in testing than the BIC-based wrapper of Murphy et al. (2010). It seems surprising that BIC underperforms in a situation where the modelling assumptions seem to fit the data quite well. The BIC wrapper selects for an increase in BIC and so is selecting for the addition of a variable that causes an increase in likelihood large enough to more than offset a penalty equal to the natural log of the number of observations in the data. It appears to be possible that this increase could come about through the selection of a variable that creates data points that more closely fit the model, for example. There doesn‘t appear to be any reason to expect that this will increase test accuracy. With overlap of variable values among clusters being a potential problem, this type of selection, rather than one for classification accuracy, might be creating problems. The results presented here indicate that model building that adds variables that maximize new-model accuracy appears to be the best approach among those compared.

With this data, `randomForest` models performed relatively poorly. A statistically-significant difference in performance between the BIC- and accuracy-based wrappers was found, with the accuracy-based wrapper again producing models that performed better in predicting test set labels.

There was not a significant difference between the performance of the BIC wrapper and `clustvarsel`. The cut-offs for variable addition for the BIC wrapper were higher than those set for `clustvarsel`. Considering the assessment of relatively good fit of the normal-mixture assumption to this data, with the BIC wrapper finding models closer to the BIC-optimal best than `clustvarsel` and the two performing similarly, these results are worthy of further investigation.

Based on the Shapiro-Wilk results and the graphical checks, the olive oil data appear to show more evidence for potential problems with variable values not supporting fit of a normal mixture. The BIC-based wrapper makes use of normal mixtures and it clearly underperformed here, achieving a mean accuracy of 76.26 %. Correlations within the data are typically much lower than for the meats data, but BIC-selected variables were very highly correlated relative to the average. Again, the accuracy-based wrapper performed extremely well, achieving an average accuracy of 99.17 % in discrimination of test data, better than all of the alternatives considered here. For this data set, correlations in the models built with this wrapper were actually lower than the average, and very different from corresponding values for BIC-selected models. For some reason, a reversal in nature of in-model correlation has occurred between the meats and olive oil data. This is also worthy of further investigation.

While the BIC wrapper did not perform well, there was not a significant change in average accuracy for `clustvarsel` compared to its application to the meats data. It is believed this occurred because, since cut-off values for selection based on BIC were higher for the BIC wrapper, the whole model-building process was biased to select a model that would be more likely to not adequately describe data exhibiting more potential for selection of non-normal modelling variables.

With the accuracy-based wrapper for model building, `randomForest` achieved a mean accuracy of 94.79 % for applications to test data, the only instance where `randomForest` performed relatively well. Interestingly, checking of correlation matrices revealed averages of absolute values of off-diagonal entries that were very low, with a mean of 0.281. This is a reversal of what occurred with the meats data were in-model correlations were higher with the accuracy wrapper. Perhaps this is why `randomForest` performed best here. However, when applied to the locally BIC-optimal models, the function did not do well, achieving a mean accuracy of 67.71 %. The average absolute values of correlations among modelling variables were typically much higher than the average for the whole dataset and suggest the BIC wrapper may have favoured inclusion of highly correlated variables. The second part of the effect discussed earlier is the potential for variables with overlapping values to be selected. Could such a combination be at play here? Pre-analysis samples of group-wise boxplots for variables suggested overlap was present, and inspection of boxplots for the first four results replicates obtained using the BIC wrapper revealed that the wrapper consistently selected variables with overlapping ranges. For the accuracy-based wrapper, inspection of boxplots indicated overlap was also present for the modelling variables. Subjectively, it appeared that overlap was less of a problem for variables in

the models built for accuracy; a method for quantifying this and comparing wrappers would have been extremely useful here.

For the simulated data set, pre-analysis checks did not suggest there would be any problems using a multivariate-normal mixture model. In these conditions, the BIC wrapper was able to build successful models using normal mixtures: it achieved a mean accuracy of 97.07 % on test data. However, it was still outperformed by the accuracy-based wrapper which averaged 100 % on applications to test data. There was a statistically significant difference between the mean accuracies of the BIC wrapper and `clustvarsel`, with `clustvarsel` achieving the lower mean accuracy at 84.38 %. The normal mixture model appeared to be a very good fit, and with the higher cut-offs for inclusion based on BIC favouring selection of a model causing a large likelihood increase, the BIC-wrapper has produced models that better describe the data. Conversely, the lower default cut-off settings of `clustvarsel` hindered its approach to the BIC-optimal model, a model which, in this case, might be close to, if not the best model. Manipulation of settings for `clustvarsel` might greatly improve performance.

The function `randomForest` achieved mean accuracies of 73.47 % and 65.87 %, for the accuracy- and BIC-based wrappers, respectively. Predictive performance was erratic, with higher standard deviations for accuracy. So far, results here suggest that high correlation can cause difficulties for `randomForest`. In this situation, the data building process created a result with the potential for relatively compact clusters to be nested in more dispersed clusters. However, with decreased correlation, as was observed in the replicate models constructed by the BIC wrapper, it may be possible for observations to be more diffusely scattered within their clusters

than they would be otherwise and this could create a situation where cluster nesting could cause classification difficulties. The performance of `randomForest` with each of the two wrappers was significantly different, and predictive accuracy did improve with the accuracy-based wrapper. The average absolute correlations among included variables were higher than for the BIC models. The observed difference between medians was 0.07 and a Mann-Whitney U-test supported rejection of equality of location ($p = 0.008$). In this case, with overlap, an increase in correlation has accompanied an apparent increase in predictive ability of `randomForest` models. This result suggests that generalizations linking performance to correlation alone may be dangerous.

It is not surprising that the performance of different classification algorithms and model-building methods depends on the nature of the data being analyzed. What is interesting in these results are relative performances of variable selectors with the two classification methods and the potential reasons for the differences. Even when the normal mixture model was highly applicable, and it appeared to be so for two of the three data sets, the BIC wrapper and `clustvarsel` where still outperformed here by a simple search for variables that maximize new-model classification accuracy. The accuracy-based wrapper, maximizing new-model accuracy and only adding variables until model completion, performed best on all three sets of data, with both classifiers. Based on these results, there is no reason to generalize this, but its occurence here suggests simply searching for a local approximation to the desired effect, highest predictive accuracy, may be the best option for variable selection.

The results suggest interactions between model fit, correlation, overlap, and performance in classification for the model selectors and classifiers considered here. Unfortunately, it does not seem possible or safe to make any generalizations.

The potential for normal models appeared to help the BIC wrapper, but never enough to beat the accuracy wrapper. At times, correlation may have been a cause of difficulty for `randomForest`. However, it may help in cases where correlation is relatively low and there is potential for overlap such that relatively small clusters are completely contained in larger ones. Overlap was always present in the data dealt with here. Its nature and varying levels of correlation appeared to be important in results. Finally, details of model building are important. The BIC wrapper and `clustvarsel` both use the BIC, but performed quite differently. It is possible to handicap a potentially good model-building technique by using settings that hinder its finding the best available models.

Examining the data and results, difficulties became apparent. Assessing normality and levels of correlation were difficult, as was assessing overlap of groups. Based on these results, visualization and assessment of data appear to be critical in planning and developping a successful classifier, so further work in testing existing visualization and assessment techniques and developping new ones has the potential to be beneficial. The data sets used here all exhibited some degree of potential to allow fit of a normal mixture. Presumably, not all data will be structured to allow this. Work examining the extension of these techniques to very different data sets could be fruitful. Perhaps, most interesting of all would be work to put the new technique, the accuracy-based wrapper, to the test. An attempt to find and theoretically describe data that would hurt the performance of the accuracy-based wrapper relative to the BIC wrapper and `clustvarsel` could be very informative.

Bibliography

Banfield, J. D. and A. E. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics 49*, 803–821.

Biernacki, C. and G. Govaert (1999). Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation 64(1)*, 49–71.

Breiman, L. and A. Cutler (2011). Random forests. Available at url: http://www.stat.berkeley.edu/ breiman/RandomForests.

Dean, N., T. B. Murphy, and G. Downey (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 55(1)*, 1–14.

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

Fisher, R. A. (1950). *Contributions to Mathematical Statistics*. New York, USA: John Wiley and Sons.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association 97*, 611–631.

Fraley, C. and A. E. Raftery (2006). MCLUST version 3 for R: Normal mixture modeling and model-based clustering". Technical Report 504, University of Washington, Department of Statistics. (revised 2009).

Fraley, C. and A. E. Raftery (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification 24*, 155–181.

Fraley, C. and A. E. Raftery (2009, December). `MCLUST` version 3 for R: Normal mixture modelling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics.

Kerebin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A 62(1)*, 49–66.

McElhinney, J., G. Downey, and T. Fearn (1999). Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. *Journal of Near Infrared Spectroscopy 7*, 145–154.

McLachlan, G. (1975). Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association 70(350)*, 365–369.

Murphy, T. B., N. Dean, and A. E. Raftery (2010). Variable selection and updating in discriminant analysis for high dimensional data with food authenticity applications. *The Annals of Applied Statistics 4(1)*, 396–421.

O'Neill, T. J. (1978). Normal discrimination with unclassified observations. *Journal of the American Statistical Association 73(364)*, 821–826.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Raftery, A. E. and N. Dean (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association 101*(473), 168–178.

Robert, C. P. and G. Casella (2010). *Introducing Monte Carlo Methods with* R. New York, USA: Springer.

Steele, R. J. and A. E. Raftery (2009, September). Performance of bayesian model selection criteria for gaussian mixture models. Technical Report 559, University of Washington, Department of Statistics.

Stock, J. H. and M. W. Watson (2007). *Introduction to Econometrics*. Boston, USA: Pearson/Addison Wesley.

Tapp, H. S., M. Defernez, and K. Kemsley (2003). FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils. *Journal of Agricultural and Food Chemistry 51(21)*, 6110–5.