Semi-Supervised Linear Discriminant Analysis

Deirdre Toher*, Gerard Downey† & Thomas Brendan Murphy‡

*Department of Mathematics and Statistics, University of the West of England, Bristol, BS16 1QY, United Kingdom

†Prepared Foods Department, Teagasc Ashtown Food Research Centre, Dublin 15, Ireland

‡School of Mathematical Sciences, University College Dublin, Dublin 4, Ireland

Abstract

Fisher's linear discriminant analysis is one of the most commonly used and studied classification methods in chemometrics. The method finds a projection of multivariate data into a lower dimensional space so that the groups in the data are well separated. The resulting projected values are subsequently used to classify unlabeled observations into the groups.

A semi-supervised version of Fisher's linear discriminant analysis is developed, so that the unlabeled observations are also used in the model fitting procedure. This approach is advantageous when few labeled and many unlabeled observations are available.

The semi-supervised linear discriminant analysis method is demonstrated on a number of data sets where it is shown to yield better separation of the groups and improved classification over Fisher's linear discriminant analysis.

1 Introduction

Fisher [1] motivated his linear discriminant analysis technique for analyzing the iris data set [2] by posing the question: *"What linear function of the four measurements* $X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$ *will maximize the difference in the ratio of the difference between the specific means to the standard deviations within species?"* Fisher proceeded to develop a method for discriminating between pairs of species of irises. Rao [3] extended Fisher's linear discriminant analysis procedure to more than two groups'.

Alternatively, the problem can be motivated by "*What projection of the data will maximize the ratio between the likelihood function assuming that the data come from G distinct groups and the likelihood function assuming that the data come from a single group?*". If it is assumed that data within groups are multivariate normal, then answering this question arises at the same projection of the data, provided an assumption of a common covariance matrix for each group is imposed.

Fisher's Linear Discriminant Analysis (LDA) is one of the most commonly used classification techniques in chemometrics [4, 5, 6]. In addition, the Fisher discriminant scores provide a visualization that shows the separation between the groups in the data. Fisher's LDA has also been studied and extended in a number of ways [7, 8]; in particular, a number of extensions combine LDA with variable selection methods to improve its performance.

The projections and classification rule used in Fisher's LDA are found using only data where the group membership is known. In practical situations, it is common that only a small subset of the data is fully labeled. For example, in food authentication studies [9] it is difficult and expensive to obtain reliably labeled observations whereas it is relatively inexpensive to collect unlabeled samples. So, the development of classification methods that can use unlabeled samples in the model fitting procedure are important in such situations.

Recently, in statistics and machine learning, there has been an increased interest in developing semi-supervised classification methods that use both labeled and unlabeled data in the model fitting procedure. Semi-supervised methods differ from many traditional methods in that both the labeled training data and the unlabeled test data are used in the model fitting rather than fitting a model using only the labeled training data and testing the model predictions on the unlabeled test data. These methods are especially useful when the number of labeled values is very small relative to the number of unlabeled samples. A selection of recently developed semi-supervised classification methods are outlined in [10] and [11].

In this paper, we develop a semi-supervised version of Fisher's LDA. The method iteratively updates the Fisher's linear discriminant function, where the model parameters are estimated by taking into account the estimated probability of each unlabeled observation

coming from each group. The discriminant function is updated until the classification probabilities converge, thus yielding a semi-supervised version of Fisher's linear discriminant analysis. The work outlined herein illustrates how semi-supervised methods can be utilized to extend standard classification methods like Fisher's LDA to yield methods that give improved performance when few training samples are available.

The paper is outlined as follows. In Section 2 the background theory of Fisher's LDA is reviewed and in Section 3 a semi-supervised version of Fisher's LDA is developed. In Section 4 a number of data sets where the semi-supervised method is demonstrated are introduced and the results of these analyses are presented in Section 5. We conclude, in Section 6, by discussing the semi-supervised Fisher's LDA method and possible extensions to this method.

2 Fisher's Linear Discriminant Analysis

Assume that we have data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ where each observation consists of P values recorded in a column vector. Suppose that the data come from G groups and we have group label vectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$ where $z_{jg} = 1$ if observation j comes from group g and $z_{jg} = 0$ otherwise.

We assume that the probability of an observation coming from group g is π_g and that observations within group g are modeled by a $N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$.

Hence, the likelihood function is of the form,

$$L(\pi_1, \pi_2, \ldots, \pi_G, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}) = \prod_{j=1}^{N} \left[\frac{\pi_g}{(2\pi)^{P/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_g) \right\} \right]^{z_{jg}},$$

where $\boldsymbol{\mu}_g = (\mu_{g1}, \mu_{g2}, \ldots, \mu_{gP})'$.

The maximum likelihood estimates of the model parameters are

$$\hat{\boldsymbol{\mu}}_g = \frac{1}{n_g} \sum_{j=1}^{N} z_{jg} \mathbf{x}_j$$

where $n_g = \sum_{j=1}^{N} z_{jg}$ is the number of observations in group g,

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{g=1}^{G} \sum_{j=1}^{N} z_{jg} (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_g)'}{N} \tag{1}$$

and $\hat{\pi}_g = n_g/n$. The unbiased estimates of the model parameters are of the same as the maximum likelihood except in the case of $\hat{\boldsymbol{\Sigma}}$ where the N in the denominator of (1) is replaced by $N - G$. Robust estimators of the mean and covariance [12] can be used in situations where it is suspected that the data contain outliers. In some applications, $\pi_1, \pi_2, \ldots, \pi_G$ are fixed *a priori* rather than being estimated from the data.

If the G groups are assumed to have equal mean and covariance, that is $\boldsymbol{\mu}_g = \boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$, then the data would be $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and the maximum likelihood estimate of the model parameters are $\hat{\boldsymbol{\mu}}_0 = \overline{\mathbf{x}}$ and

$$\hat{\boldsymbol{\Sigma}}_0 = \frac{\sum_{j=1}^{N}(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_0)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_0)'}{N}.$$

The unbiased estimate of $\boldsymbol{\Sigma}_0$ replaces the denominator of $\hat{\boldsymbol{\Sigma}}_0$ by $N - 1$.

It is well known that linear combinations of multivariate normal random variables are normally distributed. Hence, if $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{Y}'\boldsymbol{\alpha} \sim N(\boldsymbol{\mu}'\boldsymbol{\alpha}, \boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ is a column vector representing the coefficients of the linear combination. Hence, if we consider linear combinations of the data values, that is $\mathbf{x}_1'\boldsymbol{\alpha}, \mathbf{x}_2'\boldsymbol{\alpha}, \ldots, \mathbf{x}_N'\boldsymbol{\alpha}$, the ratio of likelihood when the group means are unconstrained versus when the group means are constrained to be equal is

$$LR(\boldsymbol{\alpha}) = \frac{\prod_{j=1}^{N}\left[\frac{\hat{\pi}_g}{(2\pi)^{1/2}(\boldsymbol{\alpha}'\hat{\boldsymbol{\Sigma}}\boldsymbol{\alpha})^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{\alpha}'\mathbf{x}_j - \boldsymbol{\alpha}'\hat{\boldsymbol{\mu}}_g)(\boldsymbol{\alpha}'\hat{\boldsymbol{\Sigma}}\boldsymbol{\alpha})^{-1}(\boldsymbol{\alpha}'\mathbf{x}_j - \boldsymbol{\alpha}'\hat{\boldsymbol{\mu}}_g)\right\}\right]^{z_{jg}}}{\prod_{j=1}^{N}\left[\frac{1}{(2\pi)^{1/2}(\boldsymbol{\alpha}'\hat{\boldsymbol{\Sigma}}_0\boldsymbol{\alpha})^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{\alpha}'\mathbf{x}_j - \boldsymbol{\alpha}'\hat{\boldsymbol{\mu}}_0)(\boldsymbol{\alpha}'\hat{\boldsymbol{\Sigma}}_0\boldsymbol{\alpha})^{-1}(\boldsymbol{\alpha}'\mathbf{x}_j - \boldsymbol{\alpha}'\hat{\boldsymbol{\mu}}_0)\right\}\right]}.$$

[13] shows that the ratio $LR(\boldsymbol{\alpha})$ is equivalent to the objective function in Fisher's linear discriminant analysis which is of the form

$$\frac{\boldsymbol{\alpha}'\mathbf{B}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\mathbf{W}\boldsymbol{\alpha}} \tag{2}$$

where

$$\mathbf{B} = \frac{\sum_{g=1}^{G} n_g(\hat{\boldsymbol{\mu}}_g - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_g - \hat{\boldsymbol{\mu}})'}{G-1}$$

and

$$\mathbf{W} = \frac{\sum_{g=1}^{G}\sum_{j=1}^{N} z_{jg}(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_g)'}{(N-G)}.$$

The value of $\boldsymbol{\alpha}$ that maximizes (2) subject to $\boldsymbol{\alpha}'\mathbf{W}\boldsymbol{\alpha} = 1$ is the eigenvector corresponding to the largest eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$; this $\boldsymbol{\alpha}$ gives the coefficients of the first Fisher linear discriminant function. The other eigenvectors corresponding to non-zero eigenvalues yield the other Fisher linear discriminant functions. There are at most $\min\{G-1, P\}$ non-zero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$, leading to at most this number of Fisher linear discriminant functions.

In many chemometric applications, the matrix $\mathbf{W}$ is not invertible because it is not of full rank due to the fact that $N \ll P$. This is particularly true in applications involving spectroscopic data where the number of spectral measurements can greatly exceed the number of observations. A solution to this problem is sphering the data [14] so that the value of $\mathbf{W}$ computed for the sphered data is the identity matrix. The sphering is implemented by computing an eigendecomposition of $\mathbf{W}$ and selecting the eigenvectors corresponding to the non-zero eigenvalues. The data are projected onto a low dimensional

subspace using linear combinations of the variables, where the weights are given by the eigenvectors divided by the square root of their corresponding eigenvalues. The within group covariance of these projected values is thus an identity matrix.

Furthermore, in the $N \ll P$ case, the number of non-zero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ is thus at most $\min\{G-1, P, N-G\}$. In practice, the number of eigenvalues with appreciable value can be even lower and the eigenvectors corresponding to very small eigenvalues can be discarded [15].

The values obtained when the data are projected using the Fisher linear discriminant functions provide a low-dimensional representation of the data, where the groups are well separated; these values are called the Fisher discriminant scores. If the coefficients of the linear discriminant functions are $\mathbf{A} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_M)$, where each column of $\mathbf{A}$ contains the coefficients of one of the discriminant functions, then the discriminant scores are given by $\mathbf{A}'\mathbf{x}_1, \mathbf{A}'\mathbf{x}_2, \ldots, \mathbf{A}'\mathbf{x}_N$.

2.1 Prediction and Classification

Given a new unlabeled observations $\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \ldots, \mathbf{x}_{N+M}$, we want to estimate their probabilities of belonging to each of the groups $1, 2, \ldots, G$. Note that, if the labeled data is sphered, then we use the same sphering transformation on $\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \ldots, \mathbf{x}_{N+M}$.

We can estimate the posterior probability of observation $\mathbf{x}_j$ coming from group g as

$$\hat{z}_{jg} = \frac{\hat{\pi}_g \exp\left\{-\frac{1}{2}(\mathbf{A}'\mathbf{x}_j - \mathbf{A}'\hat{\boldsymbol{\mu}}_g)(\mathbf{A}'\hat{\boldsymbol{\Sigma}}\mathbf{A})^{-1}(\mathbf{A}'\mathbf{x}_j - \mathbf{A}'\hat{\boldsymbol{\mu}}_g)'\right\}}{\sum_{g'=1}^{G} \hat{\pi}_{g'} \exp\left\{-\frac{1}{2}(\mathbf{A}'\mathbf{x}_j - \mathbf{A}'\hat{\boldsymbol{\mu}}_{g'})(\mathbf{A}'\hat{\boldsymbol{\Sigma}}\mathbf{A})^{-1}(\mathbf{A}'\mathbf{x}_j - \mathbf{A}'\hat{\boldsymbol{\mu}}_{g'})'\right\}}, \tag{3}$$

for $j = N+1, N+2, \ldots, N+M$. The observations can be classified into groups by finding the group that has the *maximum a posteriori* (MAP) group membership probability.

3 Semi-Supervised Approach

The idea behind semi-supervised learning methods is to use all of the data available in the model fitting process. Therefore both the fully labeled and the unlabeled data are used in the model fitting process. The use of unlabeled data when fitting a linear discriminant analysis model in the two group case was investigated in [16] and [17]; their work concentrated on the asymptotic properties of the method relative to when only labeled observations are used. More recent work on semi-supervised classification includes [18] who develop a semi-supervised version of model-based discriminant analysis, [19] who implement semi-supervised methods based on mixture models, [20] who develops a semi-supervised model-based classification method based on a parsimonious family of Gaussian mixture models and [21] who develop a semi-supervised model-based discriminant analysis

method that also includes variable selection. In addition, overviews of semi-supervised learning are available in [10], [22] and [11].

The implementation of a semi-supervised version of Fisher's LDA involves an iterative algorithm that proceeds as follows:

0. Apply Fisher's LDA model to the labeled data. Choose initial values for $\hat{z}_{jg}$ for the unlabeled observations using one of the following options:

   (a) Randomly assign the unlabeled observations into groups. This implies that for each j one value of $\hat{z}_{jg}$ is set to be equal to one and the remaining values are zero, where $j = N+1, N+2, \ldots, N+M$. Here, no information in the labeled data is used to provide an initial classification for the unlabeled samples.

   (b) Use $\hat{\pi}_g$ the probabilities from the initial model fit. This implies that $\hat{z}_{jg} = \hat{\pi}_g$ for all $j = N+1, N+2, \ldots, N+M$. Here, only group sizes for the labeled data are used to provide an initial classification for the unlabeled samples.

   (c) Use the estimated *a posteriori* probability of coming from each group from the Fisher's LDA model (3) for $j = N+1, N+2, \ldots, N+M$. Here, the classifier fitted to the labeled data used to provide an initial classification for the unlabeled data.

   The most consistent of the methods was to use (0c) but in some cases the other options yielded superior results. In particular, when very few labeled obervations are available (0c) can sometimes provide a poor initial estimate of the group parameters which is then reinforced in the iterative algorithm, whereas (0a) and (0b) were less likely to get trapped in this way. However, when a moderate number of labeled observations are available (0c) is the most preferred starting option; this option was used in all of the results presented herein.

1. Update all of the model parameter estimates $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \ldots, \hat{\boldsymbol{\mu}}_G$, $\pi_1, \pi_2, \ldots, \pi_G$, $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\Sigma}}_0$ using the equations given in Section 2, but with the calculations involving all $N+M$ observations and with $\hat{z}_{jg}$ being used in place of z_{jg} for $g = 1, 2, \ldots, G$ and $j = N+1, N+2, \ldots, N+M$. Also update the values $\mathbf{B}$ and $\mathbf{W}$ in a similar manner.

2. Update the linear discriminant coefficients $\mathbf{A}$ using the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$.

3. Re-estimate the *a posteriori* probabilities, $\hat{z}_{jg}$), of belonging to each group using (3) for $j = N+1, N+2, \ldots, N+M$.

4. Check if the $\hat{z}_{jg}$ values have converged. If so, stop. Otherwise, return to Step 1. In practice, we considered that the values had converged in the values changed by less than 10^{-5} in successive iterations.

3.1 Brier's Score

Classification performance can be assessed using percentage misclassification error and also using Brier's score [23]. Brier's score gives an indication of how accurate the classifications are in terms of probability of group membership rather than just on the hard classification results.

Brier [23] developed a method of producing a continuous performance measure where perfect prediction gives a Brier's score of zero. In the context of the the classification study outlined herein, given G groups, M unlabeled samples and *a posteriori* probabilities $\hat{z}_{j1}, \ldots, \hat{z}_{jG}$ for samples $j = N+1, N+2, \ldots, N+M$. Brier's score is defined as,

$$\text{Brier's Score} = \frac{100}{2M} \sum_{g=1}^{G} \sum_{j=N+1}^{N+M} (\hat{z}_{jg} - z_{\text{true}_{jg}})^2,$$

where $z_{\text{true}_{jg}}$ is an indicator of the true group membership.

Brier's score is useful for assessing the certainty of predictions. Some observations may have maximum probability for the correct group but the maximum probability is not much larger than the probability of membership of the other groups. Such observations contribute more to Brier's score than definitively correctly classified observations. Likewise, some observations may have high, but not maximum, probability for the correct group and these observations contribute less to Brier's score than definitively incorrectly classified observations. Hence, if Brier's score is much lower than the misclassification rate, this indicates that incorrectly classified observations have high uncertainty associated with them, whereas correctly classified observations have high probability of belonging to the correct group.

It is worth noting that if $\hat{z}_{jg}$ is a hard classification rather than the probability of observation j belonging to group g, then the Brier's score becomes equivalent to the percentage misclassification error.

4 Data

We demonstrate the semi-supervised Fisher's LDA (SSLDA) on a number of examples. First, in Section 4.1, we demonstrate how the method works on four simulated data sets. In the first two simulated data sets the LDA model assumptions are correct and in the second two data sets the common covariance assumption is not valid. We then demonstrate the method on the well known wine data set [24] (Section 4.2). Finally, we implement the method of a homogenized meat authentication problem [25] (Section 4.3); this problem is of particular interest because the total sample size ($N + M$=231) is much less than the data dimension ($P = 1050$).

4.1 Simulated Data

Four data sets were generated, each containing $N + M = 300$ observations with $P = 10$ variables, 100 observations from each of three groups. In all of the examples, two variables contain group information and the remaining eight variables are each *i.i.d.* $N(0, 1)$ and thus contain no group information. The first two data sets were generated from multivariate Gaussian distributions using a common covariance matrix $\mathbf{\Sigma}$ where

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 1 & 0 & \cdots & 0 \\ 1 & 2 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

For data set 1:

- $\mathbf{x}_j|\text{Group } 1 \sim N(\boldsymbol{\mu}_1, \mathbf{\Sigma})$ where $\boldsymbol{\mu}_1 = (-3, -2, 0, \ldots, 0)$
- $\mathbf{x}_j|\text{Group } 2 \sim N(\boldsymbol{\mu}_2, \mathbf{\Sigma})$ where $\boldsymbol{\mu}_2 = (8, 2, 0, \ldots, 0)$
- $\mathbf{x}_j|\text{Group } 3 \sim N(\boldsymbol{\mu}_3, \mathbf{\Sigma})$ where $\boldsymbol{\mu}_3 = (3, 7, 0, \ldots, 0)$

Hence, the first two variables contain all of the group information and the last eight have no group information.

For data set 2:

- $\mathbf{x}_j|\text{Group } 1 \sim N(\boldsymbol{\mu}_1, \mathbf{\Sigma})$ where $\boldsymbol{\mu}_1 = (0, -2, 0, \ldots, 0)$
- $\mathbf{x}_j|\text{Group } 2 \sim N((\boldsymbol{\mu}_2, \mathbf{\Sigma})$ where $\boldsymbol{\mu}_2 = (6, 2, 0, \ldots, 0)$
- $\mathbf{x}_j|\text{Group } 3 \sim N((\boldsymbol{\mu}_3, \mathbf{\Sigma})$ where $\boldsymbol{\mu}_3 = (3, 5, 0, \ldots, 0)$

For the next two simulated data sets, we use three groups with the same mean vectors as before, but now with different covariance matrices for each group:

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 4 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \mathbf{\Sigma}_2 = \begin{pmatrix} 3 & -1 & 0 & \cdots & 0 \\ -1 & 2 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \mathbf{\Sigma}_3 = \begin{pmatrix} 4 & 1 & 0 & \cdots & 0 \\ 1 & 2 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

For data set 3:

- $\mathbf{x}_j|\text{Group } 1 \sim N(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ where $\boldsymbol{\mu}_1 = (-3, -2, 0, \ldots, 0)$
- $\mathbf{x}_j|\text{Group } 2 \sim N(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$ where $\boldsymbol{\mu}_2 = (8, 2, 0, \ldots, 0)$
- $\mathbf{x}_j|\text{Group } 3 \sim N(\boldsymbol{\mu}_3, \mathbf{\Sigma}_3)$ where $\boldsymbol{\mu}_3 = (3, 7, 0, \ldots, 0)$

For data set 4:

- $\mathbf{x}_j|\text{Group } 1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ where $\boldsymbol{\mu}_1 = (0, -2, 0, \ldots, 0)$
- $\mathbf{x}_j|\text{Group } 2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ where $\boldsymbol{\mu}_2 = (6, 2, 0, \ldots, 0)$
- $\mathbf{x}_j|\text{Group } 3 \sim N(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$ where $\boldsymbol{\mu}_3 = (3, 5, 0, \ldots, 0)$

Scatter plots of realizations from the four simulation scenarios outlined are shown in Figure 1.

Figure 1 Goes Here

4.2 Wine Data

Forina et al. [24] collected and analyzed chemical and physical properties of wine samples from the Piedmont region of Italy. The aim is to study the ability to classify the wines to their correct variety (*Barolo*, *Grignolino* and *Barbera*). The most commonly available version of this data set contains $P = 13$ measurements per sample recording the *alcohol, malic acid, ash content, alkalinity, magnesium content, total phenols, flavonoids, non-flavonoid phenols, proanthocyanins, intensity, hue, OD280/OD315 of phenols and proline* levels for each of $N + M = 178$ wine samples. This data set is available in the gclus [26] package for R [27].

For the purposes of classifying wines solely into *Barolo*, *Grignolino* and *Barbera* the study was not well designed – the 59 Barolo wine samples come from the years 1971, 1973–4; the 71 Grignolino wines are from the years 1970–6 while the 48 Barbera wines come from the years 1974, 1976, 1978–9. Thus one cannot be certain that discrimination has not been made on the basis of year of production rather than solely by type, as intended.

4.3 Meats Data

The spectra from a total of $N + M = 231$ homogenized meat samples were measured over the range 400-2498 nm at intervals of 2 nm, leading to $P = 1050$ measurements per sample. These spectra encompass both the visible and near infrared part of the electromagnetic spectrum. The samples were of raw homogenized (minced) meat and there are a total of 32 beef, 55 chicken, 34 lamb, 55 pork and 55 turkey samples. The resultant spectra are illustrated in Figure 2, with beef samples in black, chicken in red, lamb in green, pork in blue and turkey as the cyan colored lines.

The meats were purchased over a period of 10-12 weeks in the form of breast meat (chicken and turkey), pork loin chops, round steak (beef) and lamb side loin chops. The samples were refrigerated overnight, then prepared in order to produce the greatest quantity of lean meat in each sample by removing skin, bone, fatty and connective tissue. Excess surface moisture was removed by patting the meat samples dry before the samples

were individually minced. The samples were then refrigerated again before being scanned later on the same day. The preparation process is explained more fully in [25].

Figure 2 Goes Here

5 Results

In all of the examples, the performance of Fisher's LDA and the semi-supervised Fisher's LDA are assessed using 100 random splits of the data into: 50% known (labeled) 50% unknown (unlabeled) data; 25% known 75% unknown data and 10% known 90% unknown data.

For the example projections shown (Figures 3 to 8), circles are the training data (known labels), other symbols are the test data (unknown labels) and represent the group into which the observation was classified. The size of the symbols reflects the uncertainty of the prediction – the bigger the symbol in the figure, the greater the uncertainty that was associated with that prediction. Observations are colored by actual group membership.

The results presented in this section are the mean and standard deviations (in brackets) based on 100 random splits of the data into training and test sets.

5.1 Simulated Data

Table 1 Goes Here

As illustrated in Figure 1(a), the first simulated data set not only has a common covariance matrix for each group, but the groups are also well separated. Therefore LDA would be expected to perform extremely well. When the assumptions that the groups are entirely separable on the basis of their group means, Table 1 illustrates that the semi-supervised version performs extremely well, with perfect classification results achieved, even when the training data only amounts to 10% (30 observations) of the entire data set.

Figure 3 illustrates the effect of updating on the resulting projections of the data. In the random split of the data illustrated, SSLDA results in no misclassifications and LDA results in only one misclassification. The point that is misclassified by LDA has high uncertainty associated with the predicted group.

Figure 3 Goes Here

Table 2 Goes Here

Figure 1(b) illustrates that while the assumption of a common covariance matrix holds for the second simulated data set, the groups are poorly separated at their boundaries. Table 2 shows how SSLDA improves on LDA both in terms of classification and in terms of Brier's score – with the improvement become more marked as the number of observations in the training set is reduced.

The relative size of the symbols in Figure 4, as with the other projection plots, indicate the relative uncertainty associated with the group membership assigned to a particular observation. In both Figure 4(a) and Figure 4(b) the large symbols indicate that while the observation was placed in a group, the probability of belonging to the group is relatively low, when compared to other observations. Such observations, if correctly classified, would contribute relatively more to the Brier's score than the other correctly classified observations. If those observations are incorrectly classified, they would contribute relatively less to the Brier's score relative to other incorrectly classified observations. These figures show that there is higher uncertainty in the classifications with LDA compared to SSLDA, especially at the boundaries of the groups.

Figure 4 Goes Here

For the remaining simulated data sets the group covariances are not equal, that is $\Sigma_g \neq \Sigma$. The impact in terms of classification rates is immediately apparent in Tables 3 and 4 where the group means are the same as in Tables 1 and 2, but the covariance matrices are different.

Both Table 3 and Figure 5 illustrate that although the assumption of equal covariance matrices is false, if there is sufficient separation between the groups on the basis of the group means, the detrimental effect on classification performance (either in terms of percentage error or in terms of Brier's score) is minor. However, if the group means are not sufficiently far apart, as is the case in the fourth simulated data set, then updating can harm the classification performance, as can be seen in the 25%/75% row of Table 4, for example. Figure 6 illustrates that while both approaches struggle to separate the groups, as updating has a tightening effect on the groups because each group's covariance is estimated using all of the data. Hence, the points have lower uncertainties associated with them.

Table 3 Goes Here

Figure 5 Goes Here

Table 4 Goes There

Figure 6 Goes Here

5.2 Wine Data

In this example, using the semi-supervised technique improves classification performance, especially as the number of observations used in the training set is reduced. In addition, the Brier's scores are much lower than the percentage error. This indicates that correct classifications are based on observations that have high probabilities associated with the correct group, whereas incorrect classifications are mainly due to observations that, while placed in the incorrect group, had relatively high probabilities of belonging to the correct

group.

Table 5 Goes Here

Updating makes little difference to the projections of the 50%/50% split of the data illustrated in Figure 7(a) and Figure 7(b). However, the difference that updating can make on the compactness of the groups is obvious when comparing Figure 7(c) and Figure 7(d). This is because the unlabeled values are also used in estimating the group means and covariance in the semi-supervised LDA approach whereas in Fisher's LDA only the labeled values are used in parameter estimation.

Figure 7 Goes Here

5.3 Meats Data

Table 6 Goes Here

The classification performance of both approaches is relatively similar across all training/test splits of the meats data (Table 6). However, the advantages of using a semi-supervised approach are evident in Figure 8, where the groups are more compact and better separated in Figure 8(b) than in Figure 8(a). It is also clearer how each of the discriminant functions contribute towards separating the groups. The first linear discriminant functions separates the white meats (chicken, pork, turkey) from the red meats (beef and lamb). The second discriminant function separates beef and lamb, while the third separates pork samples from the others. The fourth discriminant function then separates the two poultry meats – chicken and turkey. While this is also true for the Fisher's LDA where the groups are not as compact, so the effect of each discriminant function is not as clear.

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

Figure 8 Goes Here

When comparing the projections in Figure 8 it is obvious that the main difficulty in classifying the meat samples is distinguishing between *Chicken* (red) and *Turkey* (cyan) samples.

6 Conclusions

Combining semi-supervised learning with linear discriminant analysis can result in improved classification performance, more compact and more clearly separated groups in the low-dimensional discriminant score space. The more compact and well separated groups can result in better visualization of the differences between each group than the default Fisher's LDA approach. The difference in the visualizations between the two approaches is especially noticeable within the meats data – where the classification performance, both in terms of the percentage error and the Brier's score were comparable for both methods, but the resulting projections of the data were substantially different (Figure 8).

However, as found in [22], there are situations where classification performance degrades in the semi-supervised method. As identified using the simulated data, this generally occurs when the model assumptions are violated, for example when the equal covariance assumption is seriously violated. However, the violation of this assumptions can be detected in the visualization of the data in the discriminant score space. The discriminant scores are found in a manner such that the scores are normally distributed with spherical covariance (that is, proportional to the identity covariance matrix) within each group in the linear discriminant score space. This facilitates a quick visual check of the equal covariance assumption because the groups should appear as equally scattered sets of observations in the projection plot (contrast Figure 3 and Figure 6 for example).

The SSLDA approach yields superior results to the LDA approach when the groups are well separated. This is because the unlabeled values are highly informative and contain almost as much group information as a labeled value. In addition, the group boundaries are easier to determine in this case. This issue is discussed further in [22] who provide an illustration as to why unlabeled values are particularly useful when the groups are well separated.

A likelihood based generalization of Fisher's LDA for finding discriminant coordinates in cases that are more general than the case where each group is normally distributed with common covariance is given in [13] and [28]. Their approach is computationally demanding because the discriminant function can only be found by iterative methods. However, [29] develops an extension of Fisher's LDA by considering particular cases within that framework where the need for iterative methods is avoided. The approach for developing semi-supervised methods described herein could be adapted to develop semi-supervised versions of these extensions to Fisher's LDA and other classification methods also.

A Software

An R package implementing the methodology outlined in this paper will be submitted to CRAN (`http://www.r-project.org`) shortly.

Acknowledgements

The editor and reviewers made important suggestions which greatly improved this paper. The work reported in this paper is funded by Teagasc under the Walsh Fellowship Scheme. The work is also partially supported by Science Foundation Ireland Basic Research Grant (2007/RFP/MATF281). A substantial part of this work was completed when the first author was a Ph.D. student in the School of Computer Science and Statistics at Trinity College Dublin, Ireland.

References

[1] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[2] E. Anderson. The irises of the Gapse Peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

[3] C. Radhakrishna Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, Series B*, 10(2):159–203, 1948.

[4] M. S. Larrechi, M. R. Franques, M. Ferre, and F. X. Rius. Structure modelling and discrimination of Catalan white wines. *Journal of Chemometrics*, 3:261–274, 1989.

[5] U. F. Indahl, N. S. Sahni, B. Kirkhus, and T. Naes. Multivariate strategies for classification based on NIR-spectra – with application to mayonnaise. *Chemometrics and Intelligent Laboratory Systems*, 49:19–31, 1999.

[6] Zoltán Kovács, István Dalmadi, Larina Lukács, László Sipos, Katalin Szántai-Köhegyi, Zoltán Kókai, and András Fekete. Geographical origin identication of pure Sri Lanka tea infusions with electronic nose, electronic tongue and sensory prole analysis. *Journal of Classification*, 24:121–130, 2010.

[7] M. Barker and W. Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17:166–173, 2003.

[8] C. Reynès, S. de Souza, R. Sabatier, G. Figuères, and B. Vidal. Selection of discriminant wavelength intervals in NIR spectrometry with genetic algorithms. *Journal of Chemometrics*, 20:136–145, 2006.

[9] G. Downey. Authentication of food and food ingredients by near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, 4:47–61, 1996.

[10] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[11] X. Zhu and A. B. Goldberg. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2009.

[12] D. M. Rocke and D. L. Woodruff. Computation of robust estimates of multivariate location and shape. *Statistica Neerlandica*, 47(1):27–42, 2008.

[13] M. Zhu. *Feature Extraction and Dimension Reduction with Applications to Classification and the Analysis of Co-occurrence Data*. PhD thesis, Stanford University, 2001.

[14] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, 2003.

[15] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[16] T. J. O'Neill. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 73:821–826, 1978.

[17] S. Ganesalingam and G. J. McLachlan. The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika*, 65(3):658–662, 1978.

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

[19] S. J. Frame and S. R. Jammalamadaka. Generalized mixture models, semi-supervised learning, and unknown class inference. *Advances in Data Analysis and Classification*, 1(1):23–38, 2007.

[20] P. D. McNicholas. Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference*, 140:1175–1181, 2010.

[21] T. B. Murphy, N. Dean, and A. E. Raftery. Variable selection and updating in model-based discriminant analysis for high-dimensional data with food authenticity applications. *Annals of Applied Statistics*, 4:To appear, 2010.

[22] D. Toher, G. Downey, and T. B. Murphy. 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, 2007.

[23] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

[24] M. Forina, C. Armanino, and M. Ubigli. Multivariate data analysis as a discriminating method of the origin of wines. *Vitus*, 25:189–201, 1986.

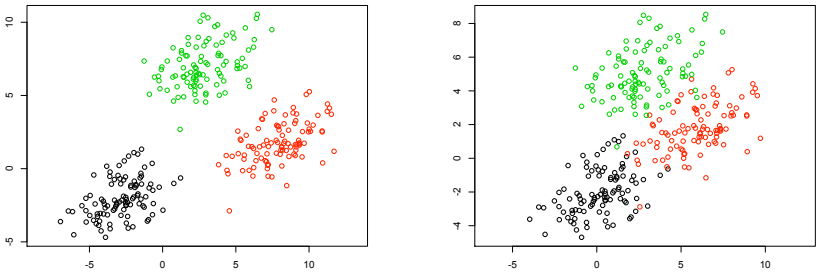
[25] J. McElhinney, G. Downey, and T. Fearn. Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. *Journal of Near Infrared Spectroscopy*, 7:145–154, 1999.

[26] C. Hurley. *gclus: Clustering Graphics*, 2004. R package version 1.2.

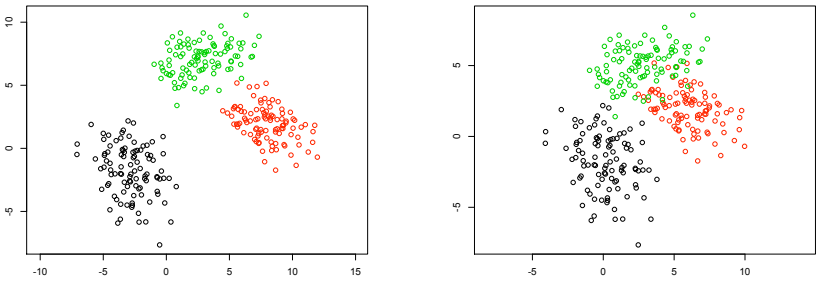
[27] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[28] M. Zhu and T. Hastie. Feature extraction for nonparametric discriminant analysis. *Journal of Computational and Graphical Statistics*, 12:101–120, 2003.

[29] M. Zhu. Discriminant analysis with common principal components. *Biometrika*, 93:1018–1024, 2006.



(a) (b)



(c) (d)

Figure 1: Simulated Data Sets. Figure 1(a) is the 1st data set, which assumes common covariance; Figure 1(b) is the 2nd data set, which assumes common covariance; Figure 1(c) is the 3rd data set which has the same group mean vectors as the 1st data set, but uses unequal group covariances; Figure 1(d) is the 4th data set which uses the group mean vectors of the 2nd data set, and the same unequal group covariances as the 3rd data set. Only the first two variables are plotted because the remaining ones show no separation between the groups.

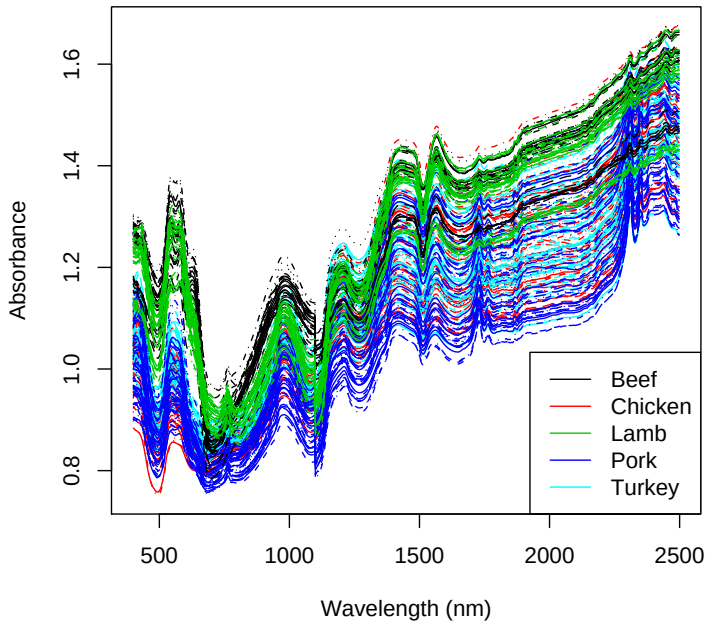
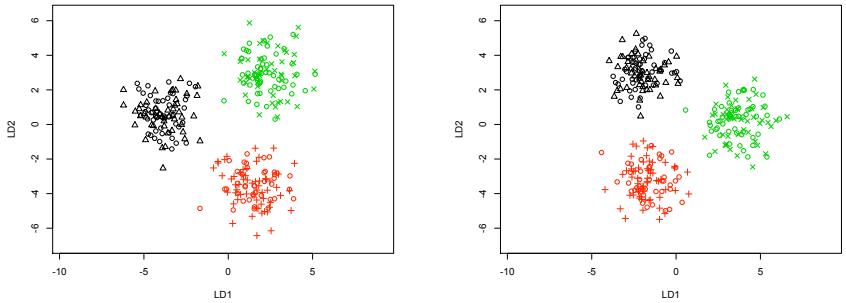
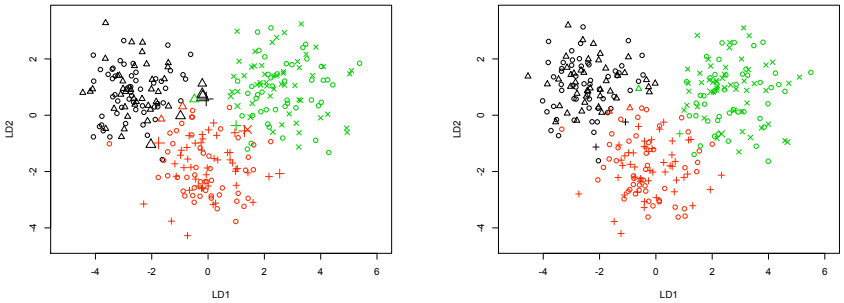


Figure 2: NIR spectra of raw homogenized meat samples



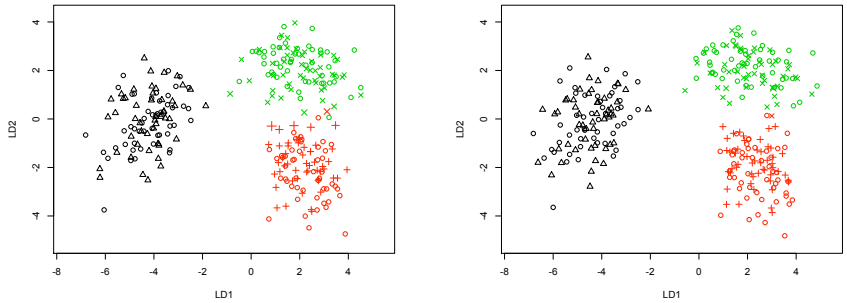
(a) (b)

Figure 3: Linear discriminant analysis projections of the first simulated data set. Figure 3(a) is the projection for Fisher's linear discriminant analysis; Figure 3(b) is the projection for the semi-supervised version, using a 50% training 50% test split of the data.



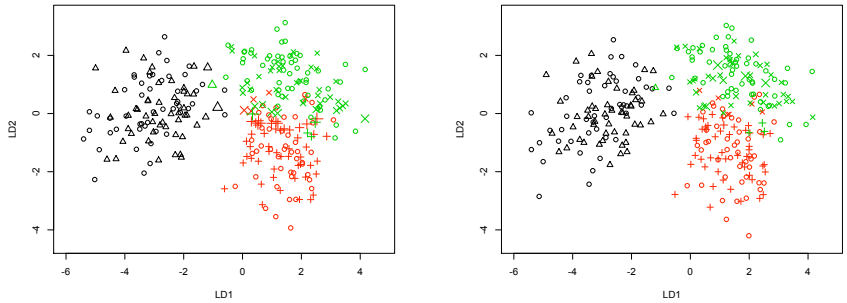
(a) (b)

Figure 4: Linear discriminant analysis projections of the second simulated data set. Figure 4(a) is the projection for Fisher's linear discriminant analysis; Figure 4(b) is the projection for the semi-supervised version, using a 50% training 50% test split of the data.



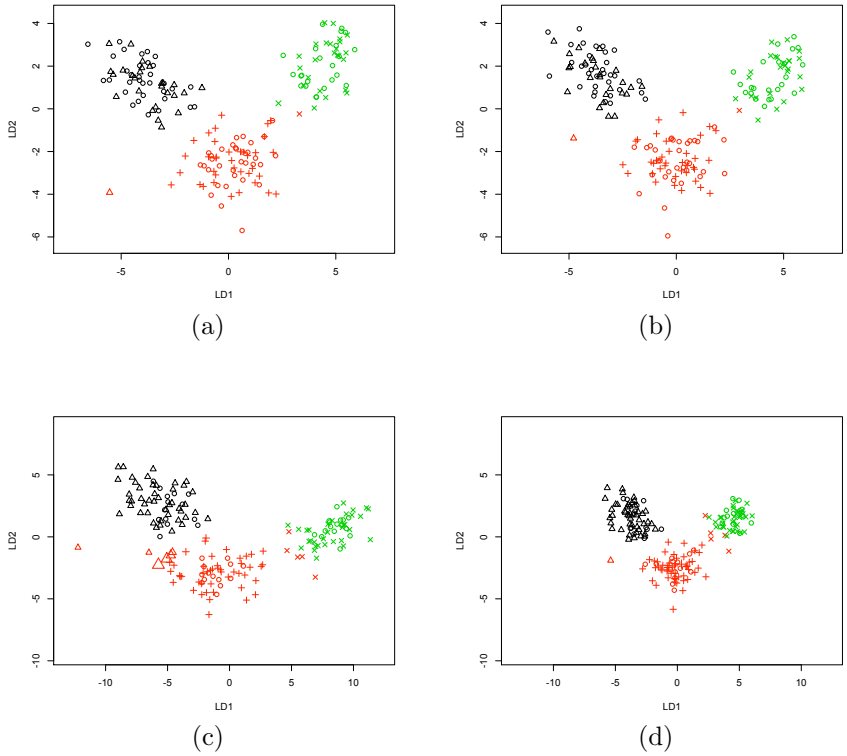
(a) (b)

Figure 5: Linear discriminant analysis projections of the third simulated data set. Figure 5(a) is the projection for Fisher's linear discriminant analysis; Figure 5(b) is the projection for the semi-supervised version, using a 50% training 50% test split of the data.



(a) (b)

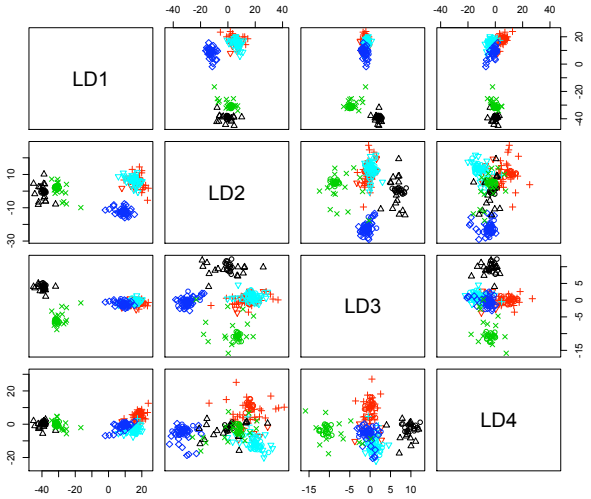
Figure 6: Linear discriminant analysis projections of the fourth simulated data set. Figure 6(a) is the projection for Fisher's linear discriminant analysis; Figure 6(b) is the projection for the semi-supervised version, using a 50% training 50% test split of the data.



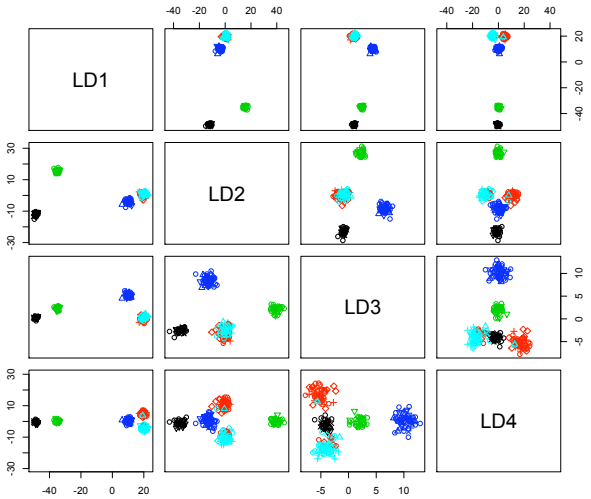
(a) (b)

(c) (d)

Figure 7: Linear discriminant analysis projections of the wine data set. Figure 7(a) is the projection for Fisher's linear discriminant analysis; Figure 7(b) is the projection for the semi-supervised version, using a 50% training 50% test split of the data. Figure 7(c) is the projection for Fisher's linear discriminant analysis; Figure 7(d) is the projection for the semi-supervised version, using a 25% training 75% test split of the data.



(a)



(b)

Figure 8: Linear discriminant analysis projections of the NIR meat data set. Figure 8(a) is the projection for Fisher's linear discriminant analysis; Figure 8(b) is the projection for the semi-supervised version, using a 50% training 50% test split of the data.

Table 1: Comparing the classification performance of LDA and SSLDA at various training/test splits of the first simulated data set.

Simulated Data Set 1		LDA		SSLDA	
50%/50%	% Error	0.173	(0.293)	0.000	(0.000)
	Brier	0.115	(0.157)	0.056	(0.048)
25%/75%	% Error	0.209	(0.263)	0.000	(0.000)
	Brier	0.162	(0.180)	0.043	(0.030)
10%/90%	% Error	1.385	(1.072)	0.000	(0.000)
	Brier	1.158	(0.933)	0.046	(0.019)

Table 2: Comparing the classification performance of LDA and SSLDA at various training/test splits of the second simulated data set.

Simulated Data Set 2		LDA		SSLDA	
50%/50%	% Error	5.527	(1.645)	5.180	(1.495)
	Brier	4.191	(1.071)	2.571	(0.646)
25%/75%	% Error	6.756	(1.503)	5.364	(0.981)
	Brier	5.181	(1.103)	2.656	(0.435)
10%/90%	% Error	12.737	(3.429)	5.481	(0.618)
	Brier	10.458	(2.910)	2.669	(0.258)

Table 3: Comparing the classification performance of LDA and SSLDA at various training/test splits of the third simulated data set.

Simulated Data Set 3		LDA		SSLDA	
50%/50%	% Error	0.573	(0.544)	0.613	(0.489)
	Brier	0.466	(0.271)	0.200	(0.125)
25%/75%	% Error	1.284	(0.791)	0.618	(0.267)
	Brier	0.914	(0.534)	0.209	(0.074)
10%/90%	% Error	3.778	(2.186)	0.663	(0.169)
	Brier	3.130	(1.918)	0.225	(0.049)

Table 4: Comparing the classification performance of LDA and SSLDA at various training/test splits of the fourth simulated data set.

Simulated Data Set 4		LDA		SSLDA	
50%/50%	% Error	7.807	(1.701)	7.620	(1.517)
	Brier	5.541	(1.003)	3.755	(0.625)
25%/75%	% Error	9.373	(1.922)	9.587	(1.670)
	Brier	6.743	(1.233)	4.483	(0.625)
10%/90%	% Error	16.359	(4.404)	11.800	(1.863)
	Brier	12.969	(3.803)	5.446	(0.923)

Table 5: Comparing the classification performance of LDA and SSLDA at various training/test splits of the wine data.

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

Table 6: Comparing the classification performance of LDA and SSLDA at various training/test splits of the meats data.

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