A dynamic linear model for heteroscedastic LDA under class imbalance

Gyamfi, S., Brusey, J., Hunt, A. & Gaura, E.

Original citation & hyperlink:
Gyamfi, S, Brusey, J, Hunt, A & Gaura, E 2019, 'A dynamic linear model for heteroscedastic LDA under class imbalance' Neurocomputing, vol. 343, pp. 65-75.
https://dx.doi.org/10.1016/j.neucom.2018.07.090

DOI 10.1016/j.neucom.2018.07.090
ISSN 0925-2312

Publisher: Elsevier

A dynamic linear model for heteroscedastic LDA under class imbalance

Kojo Sarfo Gyamfi[a,*], James Brusey[a], Andrew Hunt[a], Elena Gaura[a]

[a] *Faculty of Engineering and Computing, Coventry University, Coventry, CV1 5FB, United Kingdom*

Abstract

Linear Discriminant Analysis (LDA) yields the optimal Bayes classifier for binary classification for normally distributed classes with equal covariance. To improve the performance of LDA, heteroscedastic LDA (HLDA) that removes the equal covariance assumption has been developed. In this paper, we show using first and second-order optimality conditions that the existing approaches either have no principled computational procedure for optimal parameter selection, or underperform in terms of the accuracy of classification and the area under the receiver operating characteristics curve (AUC) under class imbalance. Using the same optimality conditions, we then derive a dynamic Bayes optimal linear classifier for heteroscedastic LDA that is optimised via an efficient iterative procedure, which is robust against class imbalance. Experimental work is conducted on two artificial and eight real-world datasets. Our results show that the proposed algorithm compares favourably with the existing heteroscedastic LDA procedures as well as the linear support vector machine (SVM) in terms of the error rate, but is superior to all the algorithms in terms of the AUC under class imbalance. The fast training time of the proposed algorithm also encourages its use for large-data applications that show high incidence of class imbalance, such as in human activity recognition.

Keywords: LDA, heteroscedasticity, class imbalance, AUC

*Corresponding author:

Email addresses: gyamfik@uni.coventry.ac.uk (Kojo Sarfo Gyamfi), j.brusey@coventry.ac.uk (James Brusey), ab8187@coventry.ac.uk (Andrew Hunt), csx216@coventry.ac.uk (Elena Gaura)

1. Introduction

Statistical classification is a fundamental task in many machine learning applications. It can take many forms, such as classifying an incoming email as either spam or non-spam, or categorising a person as standing or sitting (in the field of human activity recognition). More generally, classification involves assigning an object $\mathbf{x}$ to belong to one of K distinct classes. Arguably, one of the most well-known algorithms for performing such tasks is the linear support vector machine (SVM).

Being a kernel classifier, however, the SVM does not scale well with large amounts of training data. A low-complexity, but often sub-optimal alternative to the SVM for binary classification is Linear Discriminant Analysis (LDA). At its core, LDA makes assumptions on the data, namely that the data in each class has a multivariate normal distribution, and that the covariance matrices of these distributions are equal among the classes. Another assumption LDA makes is that the distributions of the data in the individual classes are non-overlapping (Izenman, 2009). This permits the construction of a linear boundary to discriminate between the classes. Being a linear model, all that is required is an inner product between the object $\mathbf{x}$ and a vector of weights $\mathbf{w}$ which is compared to a threshold w_0. When these assumptions are met, LDA yields the optimal Bayes linear classifier (Izenman, 2009; Hamsici & Martinez, 2008). While these assumptions are not often encountered in practice, the robustness provided by it being a linear classifier has encouraged the use of LDA for many applications (Mika et al., 1999). Moreover, the fact that many physical data tend to have distributions that are close to normal (Lyon, 2014) make the performance of LDA satisfactory for a lot of applications (Guo et al., 2007; Yu & Yang, 2001; Sharma & Paliwal, 2008).

However, experimental results, in work by, for example, Mika et al. (1999); Hastie & Tibshirani (1996); Marks & Dunn (1974); Zhao et al. (2009), have shown that if one accounts for the violation of the assumptions in the original procedure, the performance of LDA can be improved. Kernel Fisher Discriminant (KFD) (Mika et al., 1999), for instance, has been developed for the case where the patterns are overlapping, where a linear boundary will not be inappropriate. The KFD applies the kernel trick (Barber, 2012) to LDA in order to learn non-linear decision boundaries. Other work has focused on the violation of the normality assumption by modelling a non-normal distribution as a mixture of Gaussians (Hastie & Tibshirani, 1996). Other non-parametric approaches to LDA include using a local neighbourhood structures to construct a different similarity matrix rather than the covariance matrix in order to overcome the normality assumption

(Cai et al., 2007).

Accounting for the differences in covariance matrices in LDA has led to several heteroscedastic extensions of LDA, the most natural extension being Quadratic Discriminant Analysis (QDA) which makes use of a quadratic boundary to discriminate the two classes in binary classification. However, for reasons of robustness, shorter training and testing times, as well as the fact that linear classifiers can be kernelised, a linear approximation to the quadratic boundary in QDA, such as is indicated in Fig. 1, is often sought for. As there is a multitude of such linear approximations that can be made, the works by Marks & Dunn (1974), Anderson & Bahadur (1962) Peterson & Mattson (1966) and Fukunaga (2013) describe several heteroscedastic LDA procedures aimed at obtaining the linear approximation that minimises the Bayes error. Several other heteroscedastic extensions of LDA

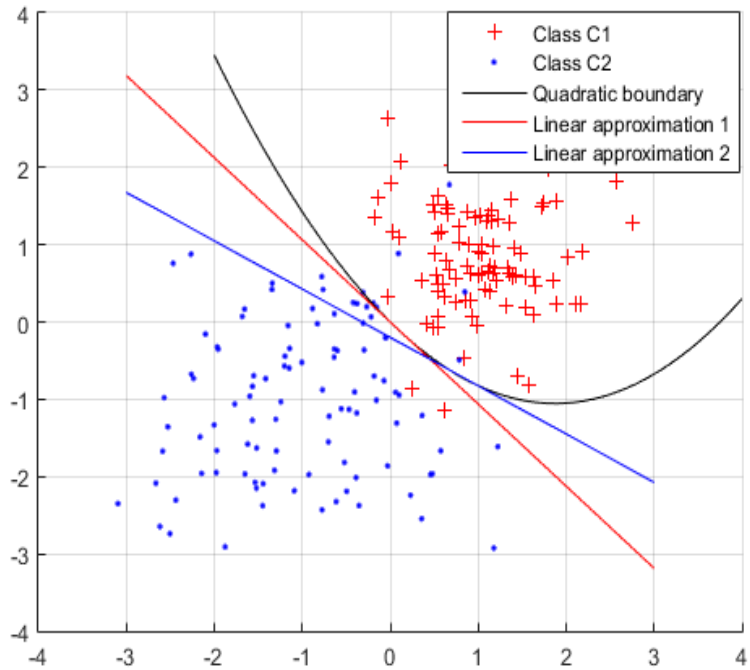

Figure 1: Quadratic boundary in heteroscedastic LDA and linear approximations

have been proposed (Duin & Loog, 2004; Decell et al., 1981; Malina, 1981; Loog & Duin, 2002; McLachlan, 2004; Decell Jr & Marani, 1976; Decell & Mayekar, 1977), but have mostly only been concerned with linear dimensionality reduction,

which involves finding a linear transformation that reduces the dimensionality of the dataset, while maximising the class discriminatory information. Our focus in this paper, however, is not on dimensionality reduction, but on the design of an optimum linear approximation to the quadratic boundary shown in Fig. 1 under heteroscedasticity and class imbalance.

The class imbalance problem arises when the number of objects in one class far exceeds the cardinality of the other classes. Such datasets are often found in anomaly detection applications like falls detection (Ojetola et al., 2015; Noury et al., 2007) in remote health monitoring, customer churn prediction in telecommunication systems (Burez & Van den Poel, 2009; Xie et al., 2009), or machine health monitoring (McBain & Timusk, 2011; Ashkezari et al., 2013), where a "fault" state is not as probable as the "normal" state of the system.

Most classifiers tend to perform poorly under class imbalance in terms of detecting the minority class, and LDA is no exception. A common approach to dealing with unbalanced data involves rebalancing the dataset by procedures such as random oversampling, random undersampling and SMOTE (Akbani et al., 2004; Chawla et al., 2002). However, it is known that rebalancing the data does not guarantee a better performance in LDA (Xue & Titterington, 2008). This is due to the fact that LDA, unlike SVM or logistic regression, is a generative classifier which attempts to learn the model that generates the data. Specifically, LDA relies on knowledge of the true class prior probabilities, which are best estimated from the empirical distribution of the classes in the dataset, so that rebalancing the dataset may result in poor estimates of the prior probabilities.

Another common approach to handling class imbalance is to bias the discriminating threshold so that more minority samples are detected (Akbani et al., 2004). Certainly, if the minority class is considered as the positive class, then shifting the discriminating threshold in such a way as to improve the correct classification of more positive samples, i.e., the true positive rate (TPR) will necessarily increase the majority negative samples that are wrongly classified as positive, i.e., the false positive rate (FPR). In this case, a measure of the goodness of the classifier is the level of FPR–TPR trade-off it is able to provide, such as is measured by the area under the so-called receiver operating characteristics (ROC) curve, AUC. Thus, a large AUC generally indicates a good classifier performance under class imbalance.

In this paper, we show that the optimality conditions for minimising the Bayes error under heteroscedasticity in LDA permit a dynamic linear model that shows a robust performance under class imbalance. The proposed model provides a much improved AUC over the existing LDA and heteroscedastic LDA procedures under

class imbalance. The model is dynamic in the sense that the vector of weights $\mathbf{w}$ can be optimally adjusted for any given discriminating threshold, as the threshold is varied to allow the detection of more minority samples. Therefore, the proposed model is more generalised than the existing heteroscedastic LDA procedures, as it can easily be employed in different applications with different goals in terms of either minimising the probability of false alarm (false positive rate) or maximising the probability of detection (true positive rate).

We evaluate the proposed classifier experimentally on 2 artificial datasets and 8 real world datasets from the University of California, Irvine (UCI) machine learning repository. We compare our algorithm to the original LDA procedure, QDA, the existing heteroscedastic LDA models by Marks & Dunn (1974); Fukunaga (2013), and the linear SVM (Hsu & Lin, 2002). The results of these experiments are presented in Section 4.

2. Background and Related Work

Consider a training dataset $\mathcal{X}$ made up of n feature vectors each of dimensionality d, i.e. $\mathcal{X} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\}$. We consider binary classification where there are only two distinct classes $\mathcal{C}_1$ and $\mathcal{C}_2$ in $\mathcal{X}$ (See (Hsu & Lin, 2002) for multi-class classification). We assume that the data in each class has a normal distribution with a mean of $\bar{\mathbf{x}}_1$ and covariance of $\mathbf{\Sigma}_1$ for $\mathcal{C}_1$, and mean of $\bar{\mathbf{x}}_2$ and covariance of $\mathbf{\Sigma}_2$ for $\mathcal{C}_2$.

The Bayes classifier assigns a given object to a class based on the *a posteriori* probability of the object. This is known as the maximum *a posteriori* (MAP) decision rule. For a given feature vector $\mathbf{x}$, the MAP decision rule may be expressed as:

$$\frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)} \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \frac{\pi_2}{\pi_1}, \tag{1}$$

where π_1, π_2 are the prior probabilities of classes $\mathcal{C}_1$ and $\mathcal{C}_2$. More often than not, these prior probabilities are unknown, but may be estimable from the relative frequencies of $\mathcal{C}_1$ and $\mathcal{C}_2$ in $\mathcal{X}$, i.e.,

$$\pi_1 = \frac{n_1}{n}, \quad \pi_2 = \frac{n_2}{n} \tag{2}$$

where n_1 and n_2 are the cardinalities of $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively.

A major limitation of the MAP decision rule is the difficulty in estimating the conditional distributions $p(\mathbf{x}|\mathcal{C}_1)$ and $p(\mathbf{x}|\mathcal{C}_1)$. For this reason, LDA proceeds from (1) with two basic assumptions (Izenman, 2009):

1. The conditional probabilities $p(\mathbf{x}|\mathcal{C}_1)$ and $p(\mathbf{x}|\mathcal{C}_2)$ have multivariate normal distributions.
2. The two classes have equal covariance matrices, an assumption known as homoscedasticity.

The normality assumption allows the conditional probabilities $p(\mathbf{x}|\mathcal{C}_1)$ and $p(\mathbf{x}|\mathcal{C}_2)$ to be expressed as:

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{\sqrt{(2\pi)^d \det(\mathbf{\Sigma}_k)}} \exp\left[-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_k)^\top \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\bar{\mathbf{x}}_k)\right], \quad \text{for} \quad k \in \{1,2\} \tag{3}$$

Given the above definitions of the conditional probabilities, one may evaluate the natural logarithm of (1), yielding the log-likelihood ratio given as:

$$\ln \lambda(\mathbf{x}) = \frac{1}{2}\ln\frac{\det \mathbf{\Sigma}_2}{\det \mathbf{\Sigma}_1} + \frac{1}{2}\left[(\mathbf{x}-\bar{\mathbf{x}}_2)^\top \mathbf{\Sigma}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2) - (\mathbf{x}-\bar{\mathbf{x}}_1)^\top \mathbf{\Sigma}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1)\right] \tag{4}$$

which is then compared against $\ln \tau$ so that $\mathcal{C}_1$ is chosen if $\ln \lambda(\mathbf{x}) \geq \ln \tau$, and $\mathcal{C}_2$ otherwise. Here, $\tau = \pi_2/\pi_1$. Therefore, the MAP decision rule for classifying a vector $\mathbf{x}$ becomes:

$$(\mathbf{x}-\bar{\mathbf{x}}_2)^\top \mathbf{\Sigma}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2) - (\mathbf{x}-\bar{\mathbf{x}}_1)^\top \mathbf{\Sigma}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \ln\frac{\tau^2 \det \mathbf{\Sigma}_1}{\det \mathbf{\Sigma}_2} \tag{5}$$

In general, this result is a quadratic discriminant. However, for the already mentioned reason of robustness and speed, a linear classifier is preferred and can be obtained in two ways from the above quadratic discriminant.

First, by calling on the assumption of homoscedasticity, i.e. $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_x$, the original quadratic discriminant given by (5) for classifying a given vector $\mathbf{x}$ decomposes into the following linear decision rule:

$$\mathbf{x}^\top \mathbf{\Sigma}_x^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \ln \tau + \frac{1}{2}(\bar{\mathbf{x}}_1^\top \mathbf{\Sigma}_x^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^\top \mathbf{\Sigma}_x^{-1}\bar{\mathbf{x}}_2) \tag{6}$$

which can be expressed as:

$$\mathbf{w}^\top \mathbf{x} \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} w_0 \tag{7}$$

where

$$\mathbf{w} = \mathbf{\Sigma}_x^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \quad w_0 = \ln \tau + \frac{1}{2}(\bar{\mathbf{x}}_1^\top \mathbf{\Sigma}_x^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^\top \mathbf{\Sigma}_x^{-1}\bar{\mathbf{x}}_2) \tag{8}$$

and $\mathbf{\Sigma}_x$ is the pooled within-class covariance matrix given by:

$$\mathbf{\Sigma}_x = \pi_1 \bar{\mathbf{\Sigma}}_1 + \pi_2 \bar{\mathbf{\Sigma}}_2, \tag{9}$$

with $\bar{\mathbf{\Sigma}}_1$ and $\bar{\mathbf{\Sigma}}_1$ being the sample covariance estimates of classes $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively.

Since many real-world data tend to be heteroscedastic, the second approach to linearising the quadratic discriminant of (5) is to obtain a linear approximation to the quadratic boundary as shown in Figure 1 by minimising the Bayes error or probability of misclassification p_e as given by 10 while accounting for heteroscedasticity.

$$p_e = \pi_1 p(y < w_0 | \mathcal{C}_1) + \pi_2 p(y \geq w_0 | \mathcal{C}_2) \tag{10}$$

Here, $y = \mathbf{w}^\top \mathbf{x}$. In the heteroscedastic case, the Bayes optimal weight vector $\mathbf{w}$ and threshold w_0 required for the linear decision rule in (7) are no longer as given in (8), as they do not minimise the Bayes error. Unfortunately, there is no closed-form analytical solution to the minimisation of (10) (Anderson & Bahadur, 1962) under heteroscedasticity, even though the optimal solution of $\mathbf{w}$ is known to be of the form:

$$\mathbf{w} = \left[s_1 \mathbf{\Sigma}_1 + s_2 \mathbf{\Sigma}_2 \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{11}$$

where s_1 and s_2 are unknown parameters (Fukunaga, 2013; Marks & Dunn, 1974).

Marks & Dunn (1974) and Anderson & Bahadur (1962) describe an iterative procedure that involves solving for the optimal $\mathbf{w}$ as given by (11) and w_0 as given by:

$$w_0 = \mu_1 - s_1 \sigma_1^2 = \mu_2 + s_2 \sigma_2^2, \tag{12}$$

where

$$\mu_1 = \mathbf{w}^\top \bar{\mathbf{x}}_1 \quad \mu_2 = \mathbf{w}^\top \bar{\mathbf{x}}_2 \quad \sigma_1^2 = \mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w} \quad \sigma_2^2 = \mathbf{w}^\top \mathbf{\Sigma}_2 \mathbf{w}. \tag{13}$$

Here, one obtains the optimal values of s_1 and s_2 via systematic trial and error. This heteroscedastic LDA procedure is referred to as random heteroscedastic linear discriminant (R-HLD) in this paper, for the reason that two parameters s_1 and s_2 are chosen at random. Thus, the approaches taken in Marks & Dunn (1974) and Anderson & Bahadur (1962) present no principled computational procedure for optimum parameter selection.

Peterson & Mattson (1966) and Fukunaga (2013) make the observation that if the weight vector $\mathbf{w}$ and the threshold w_0 are both multiplied by the same positive

scalar, the decision boundary remains unchanged. Therefore, by multiplying (11) and (12) through by the scalar $s_1 + s_2$, **w** and w_0 can be put in the form of:

$$\mathbf{w} = \ s\mathbf{\Sigma}_2 + (1-s)\mathbf{\Sigma}_1 \ ^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{14}$$

where s can be thought of as s_2 and $1 - s$ as s_1. Then, to avoid the unguided trial and error procedure by Marks & Dunn (1974); Anderson & Bahadur (1962), Peterson & Mattson (1966) and Fukunaga (2013) propose a theoretical approach described below:

1. Change s from 0 to 1 with small step increments Δs.
2. Evaluate **w** as given by:

$$\mathbf{w} = \ s\mathbf{\Sigma}_2 + (1-s)\mathbf{\Sigma}_1 \ ^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{15}$$

3. Evaluate w_0 as given by:

$$w_0 = \frac{s\mu_2\sigma_1^2 + (1-s)\mu_1\sigma_2^2}{s\sigma_1^2 + (1-s)\sigma_2^2} \tag{16}$$

4. Compute the probability of misclassification p_e.
5. Choose **w** and w_0 that minimise p_e.

This procedure is referred to as constrained heteroscedastic linear discriminant (C-HLD) in this paper, for the reason that the optimal s is constrained in the interval $[0, 1]$.

However, two main problems with the above C-HLD procedure are highlighted:

1. There is no obvious choice of the step rate Δs. Too small a value of Δs will demand too many matrix inversions in Step 2, as there will be too many s values, thus increasing the computational complexity especially for very-high dimensional datasets. On the other hand, if Δs is too large, the optimal s may not be refined enough, and the vector **w** obtained may not be optimal. Specifically, the change in **w** that results from a small change in s is given as:

$$\mathrm{d}\mathbf{w} = \big(s\mathbf{\Sigma}_2 + (1-s)\mathbf{\Sigma}_1\big)^{-1}(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)\big(s\mathbf{\Sigma}_2 + (1-s)\mathbf{\Sigma_1}\big)^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathrm{d}s \tag{17}$$

   Such a change in **w** may significantly affect the classification performance of the linear model especially if the two classes are not well-separated.

2. The solution obtained this way is only locally optimal as s is constrained in the interval $[0, 1]$. When there is a class imbalance (Xue & Titterington, 2008), the optimal s may be found outside the interval $[0, 1]$ so that the vector $\mathbf{w}$ found by this approach leads to poor classification accuracy.

Still, the existing heteroscedastic LDA procedures, as well as the original LDA procedure, do not directly address the class imbalance problem.

3. Proposed procedure

We seek to find a vector of weights $\mathbf{w} \in \mathbb{R}^d$ and a threshold $w_0 \in \mathbb{R}$ so that for a given feature vector $\mathbf{x}$ from a test set, the predicted class $\mathcal{C}^*(\mathbf{x})$ is such that:

$$\mathcal{C}^*(\mathbf{x}) = \begin{cases} \mathcal{C}_1 & \text{if} \quad y = \mathbf{w}^\top \mathbf{x} \geq w_0 \\ \mathcal{C}_2 & \text{if} \quad y = \mathbf{w}^\top \mathbf{x} < w_0 \end{cases} \tag{18}$$

Since $\mathbf{x}$ belongs to either class $\mathcal{C}_1$ or $\mathcal{C}_2$, y is normally distributed with a mean of μ_1 and a variance of σ_1^2 in class $\mathcal{C}_1$, and normally distributed with a mean of μ_2 and a variance of σ_2^2 in class $\mathcal{C}_2$ given as:

$$\mu_1 = \mathbf{w}^\top \bar{\mathbf{x}}_1 \quad \mu_2 = \mathbf{w}^\top \bar{\mathbf{x}}_2 \quad \sigma_1^2 = \mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w} \quad \sigma_2^2 = \mathbf{w}^\top \mathbf{\Sigma}_2 \mathbf{w} \tag{19}$$

The Bayes optimal linear classifier may be obtained by minimising the probability of misclassification as given by (10):

The individual misclassification probabilities can be expressed as:

$$p(y < w_0 | \mathcal{C}_1) = \int_{-\infty}^{w_0} \frac{1}{\sqrt{2\pi}\sigma_1} \exp \quad - \frac{(\zeta - \mu_1)^2}{2\sigma_1^2} \; d\zeta = 1 - Q\left(\frac{w_0 - \mu_1}{\sigma_1}\right) \tag{20}$$

and

$$p(y \geq w_0 | \mathcal{C}_2) = \int_{w_0}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_2} \exp \quad - \frac{(\zeta - \mu_2)^2}{2\sigma_2^2} \; d\zeta = Q\left(\frac{w_0 - \mu_2}{\sigma_2}\right) \tag{21}$$

where $Q(\cdot)$ is the Q-function. Therefore, the probability of misclassification, which is to be minimised, may be rewritten as:

$$p_e = \pi_1 \; 1 - Q(z_1) \; + \pi_2 \; Q(z_2) \tag{22}$$

where

$$z_1 = \frac{w_0 - \mu_1}{\sigma_1} \quad \text{and} \quad z_2 = \frac{w_0 - \mu_2}{\sigma_2} \tag{23}$$

3.1. Optimality conditions

First, we obtain the necessary first-order equations for the minimisation of p_e, i.e.,

$$\nabla p_e(\tilde{\mathbf{w}}) = \left[\frac{\partial p_e}{\partial \mathbf{w}^\top}, \frac{\partial p_e}{\partial w_0} \right]^\top = \mathbf{0} \tag{24}$$

where $\tilde{\mathbf{w}} = [\mathbf{w}^\top, w_0]^\top$. From (22), it can be shown that:

$$\frac{\partial p_e}{\partial \mathbf{w}} = \frac{1}{\sqrt{2\pi}} \left[-\pi_1 e^{-z_1^2/2} \left[\frac{\sigma_1 \bar{\mathbf{x}}_1 + z_1 \mathbf{\Sigma}_1 \mathbf{w}}{\sigma_1^2} \right] + \pi_2 e^{-z_2^2/2} \left[\frac{\sigma_2 \bar{\mathbf{x}}_2 + z_2 \mathbf{\Sigma}_2 \mathbf{w}}{\sigma_2^2} \right] \right] \tag{25}$$

It can similarly be shown from (22) that,

$$\frac{\partial p_e}{\partial w_0} = \frac{\pi_1}{\sqrt{2\pi}} \left[\frac{1}{\sigma_1} e^{-z_1^2/2} \right] - \frac{\pi_2}{\sqrt{2\pi}} \left[\frac{1}{\sigma_2} e^{-z_2^2/2} \right] \tag{26}$$

Now, equating the gradient $\nabla p_e(\mathbf{w}, w_0)$ to zero, the following set of equations are obtained:

$$\left[\frac{\pi_2 z_2}{\sigma_2^2} e^{-z_2^2/2} \mathbf{\Sigma}_2 - \frac{\pi_1 z_1}{\sigma_1^2} e^{-z_1^2/2} \mathbf{\Sigma}_1 \right] \mathbf{w} = \left[\frac{\pi_1}{\sigma_1} e^{-z_1^2/2} \bar{\mathbf{x}}_1 - \frac{\pi_2}{\sigma_2} e^{-z_2^2/2} \bar{\mathbf{x}}_2 \right] \tag{27}$$

$$\frac{\pi_1}{\sigma_1} e^{-z_1^2/2} = \frac{\pi_2}{\sigma_2} e^{-z_2^2/2} \tag{28}$$

Substituting (28) into (27) yields:

$$\left[\frac{z_2}{\sigma_2} \mathbf{\Sigma}_2 - \frac{z_1}{\sigma_1} \mathbf{\Sigma}_1 \right] \mathbf{w} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{29}$$

Then the vector $\mathbf{w}$ can be given by:

$$\mathbf{w} = \left[\frac{z_2}{\sigma_2} \mathbf{\Sigma}_2 - \frac{z_1}{\sigma_1} \mathbf{\Sigma}_1 \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{30}$$

It will be noted however that $\mathbf{w}$ as given by (30) is still in terms of w_0 through z_1 and z_2. Therefore, an explicit representation of w_0 in terms of $\mathbf{w}$ is needed from (28) to substitute in z_1 and z_2 in (30).

Now, solving for w_0 from (28) results in the following quadratic:

$$\frac{z_2^2}{2} - \frac{z_1^2}{2} - \ln \left[\frac{\tau \sigma_1}{\sigma_2} \right] = 0 \tag{31}$$

which can be simplified to:

$$\frac{w_0 - \mu_2}{\sigma_2}^{2} - \frac{w_0 - \mu_1}{\sigma_1}^{2} - 2\ln\frac{\tau\sigma_1}{\sigma_2} = 0, \tag{32}$$

where $\tau = \pi_2/\pi_1$. If τ is defined and not equal to zero, and $\sigma_1^2 \neq \sigma_2^2$, (32) can be shown to have the following solutions:

$$w_0 = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 \pm \sigma_1\sigma_2\sqrt{(\mu_1-\mu_2)^2 + 2(\sigma_1^2-\sigma_2^2)\ln\frac{\tau\sigma_1}{\sigma_2}}}{\sigma_1^2-\sigma_2^2} \tag{33}$$

i.e.,

$$w_0^+ = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 + \sigma_1\sigma_2\beta}{\sigma_1^2-\sigma_2^2} \tag{34}$$

and

$$w_0^- = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 - \sigma_1\sigma_2\beta}{\sigma_1^2-\sigma_2^2}, \tag{35}$$

where

$$\beta = \sqrt{(\mu_1-\mu_2)^2 + 2(\sigma_1^2-\sigma_2^2)\ln\frac{\tau\sigma_1}{\sigma_2}} \tag{36}$$

Nevertheless, since there are two solutions, a choice has to be made as to which of them is substituted into (30). To do this, we consider a second-order necessary condition for the minimisation of p_e, namely,

$$\frac{\partial^2 p_e}{\partial w_0^2} \geq 0 \tag{37}$$

From (26), it can be shown that:

$$\frac{\partial^2 p_e}{\partial w_0^2} = \frac{\pi_1}{\sqrt{2\pi}} - \frac{z_1}{\sigma_1^2}e^{-z_1^2/2} + \frac{\pi_2}{\sqrt{2\pi}} \frac{z_2}{\sigma_2^2}e^{-z_2^2/2} \tag{38}$$

If we plug (28) (from which we get the stationary points in (33)) into (38), we end up with the following inequality:

$$\frac{z_2}{\sigma_2} \geq \frac{z_1}{\sigma_1} \tag{39}$$

Thus, if $\tilde{\mathbf{w}}$ is to be a local minima of p_e, it is necessary that (39) is satisfied.

Now, when one considers the two solutions of w_0 in (33), only w_0^+ satisfies the inequality of (39), i.e., only this choice of w_0 corresponds to a local minimum. The proof of this is shown below.

Theorem 1. *Let w_0^+ and w_0^- be the two distinct solutions of* (33)*, then w_0^+ and w_0^- cannot both satisfy* (39) *given that $\sigma_1 = \sigma_2$.*

Proof. Let β be a positive scalar given by:

$$\beta = \sqrt{(\mu_1 - \mu_2)^2 + 2(\sigma_1^2 - \sigma_2^2) \ln \quad \frac{\tau\sigma_1}{\sigma_2}} \tag{40}$$

Note that when $\beta = 0$, (33) has a repeated root so that $w_0^+ = w_0^-$, which are not distinct. Moreover, when $\beta < 0$, (33) has no real solutions.

Also, let

$$w_0^+ = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 + \sigma_1\sigma_2\beta}{\sigma_1^2 - \sigma_2^2} \tag{41}$$

Then

$$z_2 = \frac{(\mu_2 - \mu_1)\sigma_2 + \beta\sigma_1}{\sigma_1^2 - \sigma_2^2}, \quad \text{and} \quad z_1 = \frac{(\mu_2 - \mu_1)\sigma_1 + \beta\sigma_2}{\sigma_1^2 - \sigma_2^2} \tag{42}$$

so that

$$\frac{z_2}{\sigma_2} = \frac{(\mu_2 - \mu_1)\sigma_2 + \beta\sigma_1}{\sigma_2(\sigma_1^2 - \sigma_2^2)}, \quad \frac{z_1}{\sigma_1} = \frac{(\mu_2 - \mu_1)\sigma_1 + \beta\sigma_2}{\sigma_1(\sigma_1^2 - \sigma_2^2)} \tag{43}$$

Suppose that w_0^+ satisfies (39), then

$$\frac{(\mu_2 - \mu_1)\sigma_2 + \beta\sigma_1}{\sigma_2(\sigma_1^2 - \sigma_2^2)} \geq \frac{(\mu_2 - \mu_1)\sigma_1 + \beta\sigma_2}{\sigma_1(\sigma_1^2 - \sigma_2^2)} \tag{44}$$

i.e.,

$$\frac{\beta\sigma_1^2}{\sigma_1^2 - \sigma_2^2} \geq \frac{\beta\sigma_2^2}{\sigma_1^2 - \sigma_2^2} \implies \beta\frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2 - \sigma_2^2} \geq 0, \tag{45}$$

Therefore $\beta \geq 0$.

Consider now w_0^- given as:

$$w_0^- = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 - \sigma_1\sigma_2\beta}{\sigma_1^2 - \sigma_2^2} \tag{46}$$

Then

$$z_2 = \frac{(\mu_2 - \mu_1)\sigma_2 - \beta\sigma_1}{\sigma_1^2 - \sigma_2^2}, \quad \text{and} \quad z_1 = \frac{(\mu_2 - \mu_1)\sigma_1 - \beta\sigma_2}{\sigma_1^2 - \sigma_2^2} \tag{47}$$

such that

$$\frac{z_2}{\sigma_2} = \frac{(\mu_2 - \mu_1)\sigma_2 - \beta\sigma_1}{\sigma_2(\sigma_1^2 - \sigma_2^2)}, \quad \frac{z_1}{\sigma_1} = \frac{(\mu_2 - \mu_1)\sigma_1 - \beta\sigma_2}{\sigma_1(\sigma_1^2 - \sigma_2^2)} \tag{48}$$

In order for (39) to be satisfied,

$$\frac{(\mu_2 - \mu_1)\sigma_2 - \beta\sigma_1}{\sigma_2(\sigma_1^2 - \sigma_2^2)} \geq \frac{(\mu_2 - \mu_1)\sigma_1 - \beta\sigma_2}{\sigma_1(\sigma_1^2 - \sigma_2^2)}, \tag{49}$$

i.e.,

$$\frac{-\beta\sigma_1^2}{\sigma_1^2 - \sigma_2^2} \geq \frac{-\beta\sigma_2^2}{\sigma_1^2 - \sigma_2^2} \implies \beta\frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2 - \sigma_2^2} \leq 0 \tag{50}$$

(50) implies that $\beta \leq 0$. However, as given in the preamble, $\beta > 0$. Thus, w_0^- does not satisfy (39), and only w_0^+ does. □

Having obtained w_0 that satisfies (39), i.e., $w_0 = w_0^+$, and substituting this into (30), we find that **w** becomes:

$$\mathbf{w} = \left[\frac{(\mu_2 - \mu_1)\sigma_2 - \beta\sigma_1}{\sigma_2(\sigma_1^2 - \sigma_2^2)}\mathbf{\Sigma}_2 - \frac{(\mu_2 - \mu_1)\sigma_1 - \beta\sigma_2}{\sigma_1(\sigma_1^2 - \sigma_2^2)}\mathbf{\Sigma}_1\right]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{51}$$

where β is as given in (36). Still, **w** has no closed-form solution from (51), since μ_1, μ_2, σ_1 and σ_2 are themselves functions of **w**.

However, if we let

$$s_2 = \frac{z_2}{\sigma_2} \quad \text{and} \quad s_1 = \frac{z_1}{\sigma_1}, \tag{52}$$

then we get the optimal **w** as given by (30) in the form of (11). Note that by multiplying **w** and w_0 by the same positive scalar c, the discrimination criterion as given by (18) is not changed. Suppose $c = (\sigma_1 z_2 - \sigma_2 z_1)/\sigma_1\sigma_2$, then **w** in (30) can also be written in the form of (15) where

$$s = \frac{\sigma_1 z_2}{\sigma_1 z_2 - \sigma_2 z_1}. \tag{53}$$

c is positive due to the inequality of (39).

Thus, (52) and (53) give the Bayes-optimal values of the parameters employed in the C-HLD and R-HLD procedures. It is worth noting that s_1 and s_2 are not free parameters, but are functions of **w** and w_0 as shown in (52). Therefore, rather than arbitrarily choosing s_1 and s_2 from systematic trial and error, one may obtain improving values of the two parameters by improving upon an initial choice of **w**. This procedure is described in Section 3.3.

3.2. *Effects of class imbalance*

As has already been indicated in Section 2, class imbalance is the scenario where the data in one class far exceeds the data in the other classes. For the two-class case, this implies that $\pi_1 \gg \pi_2$ or $\pi_1 \ll \pi_2$, since the prior probabilities π_1 and π_2 are often estimated empirically from the cardinality of the data in each class.

By defining τ as

$$\tau = \frac{\pi_2}{\pi_1}, \tag{54}$$

the limiting behaviour of LDA and the existing heteroscedastic LDA procedures can be studied as τ tends towards 0 or ∞.

3.2.1. *LDA*

From (6) and (8), as $\tau \to \infty$, the discriminating threshold w_0 approaches ∞. Similarly, as $\tau \to 0$, the discriminating threshold approaches $-\infty$. This tends to skew the decision rule in favour of the majority class.

3.2.2. *C-HLD*

To show the behaviour of C-HLD under class imbalance, consider the homoscedastic case where (6) and (8) give the Bayes-optimal linear discriminant.

1. Case 1: $\pi_1 \ll \pi_2$. Then, from the definition of τ in (54), $\tau \to \infty$. As $\tau \to \infty$, $w_0 \to \infty$ in (8). The consequence of this is that z_1 and z_2 both become positive. Note that in this scenario, the denominator in (53) is smaller than the numerator; the denominator is given to be positive due to the inequality of (39). Thus, the parameter s tends to take on values greater than 1.
2. Case 2: $\pi_1 \gg \pi_2$. Then, from the definition of τ in (54), $\tau \to 0$. As $\tau \to 0$, w_0 approaches $-\infty$ in (8). The consequence of this is that z_1 and z_2 both become negative. Since the denominator is positive due to the inequality of (39), s tends to take on negative values.

Since s tends to be greater than 1 in Case 1, as well as being negative in Case 2, clearly the C-HLD procedure that constrains s to the interval $[0, 1]$ yields solutions that are only locally optimal in the interval $[0, 1]$. Moreover, since s is constrained to this interval, the threshold w_0 computed in step 3 of the C-HLD procedure in (16), is a convex combination of the two projected means μ_1 and μ_2. Therefore, the discriminating threshold is always bounded between μ_1 and μ_2, even when it ought to approach $\pm\infty$ under class imbalance as shown in the two cases. Thus, the C-HLD heteroscedastic LDA procedure tends to be suboptimal in terms of

the misclassification rate under class imbalance, when s falls outside the interval $[0, 1]$.

3.2.3. R-HLD

1. Case 1: $\pi_1 \ll \pi_2$. Then, from the definition of τ in (54), $\tau \to \infty$. As $\tau \to \infty$, notice that in (34), $w_0 = w_0^+ \to \infty$, in which case both z_2 and z_1 approach ∞ as can be seen from (23). Therefore, if the Bayes error is to be minimised in the event of a class imbalance where $\pi_2 \gg \pi_1$, s_2 approaches ∞ and s_1 tends toward $-\infty$.
2. Case 2: $\pi_1 \gg \pi_2$. Then, from the definition of τ in (54), $\tau \to 0$. As $\tau \to 0$, the term $2(\sigma_1^2 - \sigma_2^2)\ln(\tau\sigma_1/\sigma_2)$ in (36) approaches $-\infty$. However, since the Bayes optimal threshold w_0 is supposed to be real, as is defined in the problem description in (18), β has to be non-negative in (36). Therefore, as $\tau \to 0$, $\beta \to 0$, and $w_0 = w_0^+ \to \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2}{\sigma_1^2 - \sigma_2^2}$. By substituting this value of w_0 into z_2 and z_1 as defined in (23), s_2 can be shown to approach:

$$s_2 = \frac{z_2}{\sigma_2} = \frac{w_0 - \mu_2}{\sigma_2^2} = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 - \mu_2(\sigma_1^2 - \sigma_2^2)}{\sigma_2^2(\sigma_1^2 - \sigma_2^2)} = \frac{\mu_2 - \mu_1}{\sigma_1^2 - \sigma_2^2} \tag{55}$$

while s_1 can also be shown to approach:

$$s_1 = \frac{z_1}{\sigma_1} = \frac{w_0 - \mu_1}{\sigma_1^2} = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 - \mu_1(\sigma_1^2 - \sigma_2^2)}{\sigma_1^2(\sigma_1^2 - \sigma_2^2)} = \frac{\mu_2 - \mu_1}{\sigma_1^2 - \sigma_2^2} \tag{56}$$

By considering the two cases, it will be noted that for any given dataset, s_1 is constrained in the interval $-\infty, \frac{\mu_2-\mu_1}{\sigma_1^2-\sigma_2^2}$ while s_2 is constrained in the interval $\frac{\mu_2-\mu_1}{\sigma_1^2-\sigma_2^2}, \infty$. This limiting behaviour makes it difficult to find the optimal values of s_2 and s_1 in the unbounded interval $(-\infty, \infty)$ by trial and error for an arbitrary dataset with class imbalance in the R-HLD heteroscedastic LDA procedure, unless a very large number of trials are performed. As every trial consists of a matrix inversion as shown in (11), this procedure can be prohibitive for very high-dimensional data, requiring the inversion of an equivalently high-dimensional scatter matrix.

3.3. Optimisation

As there is no closed form solution to **w** as given by (51), an iterative procedure is needed for the optimisation of **w**. However, as s has been shown to take

on negative values as well as values greater than 1, it cannot be varied between 0 and 1 with small step increments. The alternative of randomly trying different values of s_1 and s_2 in the range $(-\infty, \infty)$ is a rather unguided procedure, and has been shown to be a potentially computationally demanding task. For this reason, we take the iterative procedure described in Algorithm 1, ***known as iterative heteroscedastic linear discriminant (I-HLD)***.

Algorithm 1 Iterative HLD (I-HLD)

1: Input: $\mathcal{C}_1$ and $\mathcal{C}_2$
2: Output: $\mathbf{w}^*$
3: Obtain sample estimates of $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \pi_1, \pi_2$ from $\mathcal{C}_1$ and $\mathcal{C}_2$
4: Set a predefined tolerance ϵ
5: Randomly choose r_2, r_1 such that $r_2 > r_1$
6: Initialise $\mathbf{w}$: $\mathbf{w} = (r_2\mathbf{\Sigma}_2 - r_1\mathbf{\Sigma}_1)^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$
7: Evaluate $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ as given by (19)
8: Evaluate z_1, z_2 as given by (23)
9: Evaluate β as given by (36)
10: **while** $p_e < \epsilon$ **do**
11: **if** β is not real **then**
12: $\beta \leftarrow 0$
13: **end if**
14: Solve for $w_0 = w_0^+$ as $w_0 = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 + \sigma_1\sigma_2\beta}{\sigma_1^2 - \sigma_2^2}$
15: Evaluate z_1, z_2 as given by (23)
16: Evaluate the Bayes error $p_e = \pi_1 \; 1 - Q(z_1) \; + \pi_2 \; Q(z_2)$
17: Update $\mathbf{w}$ as $\mathbf{w} = \quad \frac{z_2}{\sigma_2}\mathbf{\Sigma}_2 - \frac{z_1}{\sigma_1}\mathbf{\Sigma}_1 \quad {}^{-1} \quad (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$
18: Evaluate $\mu_1, \mu_2, \sigma_1, \sigma_2, \beta$
19: **end while**
20: Set $\mathbf{w}^*$ as the vector $\mathbf{w}$ with the smallest Bayes error obtained so far

In Step 3 of Algorithm 1, r_1 and r_2 are chosen randomly such that $r_2 > r_1$ in order to satisfy (39). Every iteration in the proposed iterative procedure corresponds to an improved value of s_1 and s_2 or s due to the fact that they are functions of $\mathbf{w}$ and w_0.

Still, the choice of the initial solution to be used in Algorithm 1 leads one to consider the convexity or otherwise of the objective function p_e. If the objective function is convex, then we are guaranteed that for every choice of the initial

solution, we converge on the same final solution. In the following, we show that the objective function is non-convex.

Theorem 2. *The Bayes error p_e is non-convex in the convex set $\mathbb{R}^{d+1}$.*

Proof. For p_e to be non-convex, the Hessian matrix $\mathcal{H} = \nabla^2 p_e(\mathbf{w}, w_0)$ has to be positive semi-definite for all $\mathbf{w} \in \mathbb{R}^d$ and $w_0 \in \mathbb{R}$, i.e.,

$$\mathbf{v}^\top \mathcal{H} \mathbf{v} \geq 0 \tag{57}$$

for every non-zero $\mathbf{v} \in \mathbb{R}^{d+1}$. Observe that:

$$\nabla^2 p_e(\mathbf{w}, w_0) = \begin{bmatrix} \dfrac{\partial p_e^2}{\partial \mathbf{w} \partial \mathbf{w}^\top} & \dfrac{\partial p_e^2}{\partial \mathbf{w} \partial w_0} \\ \dfrac{\partial p_e^2}{\partial w_0 \partial \mathbf{w}^\top} & \dfrac{\partial p_e^2}{\partial w_0^2} \end{bmatrix} \tag{58}$$

Then, suppose that $\mathbf{v} = [\mathbf{0}_d^\top, x]^\top$, where $x \in \mathbb{R}$ and $\mathbf{0}_d$ is a d-dimensional vector of all zeros. Then,

$$\mathbf{v}^\top \mathcal{H} \mathbf{v} = x^2 \frac{\partial p_e^2}{\partial w_0^2} \tag{59}$$

It may be recalled from (38) that:

$$\frac{\partial^2 p_e}{\partial w_0^2} = \frac{\pi_1}{\sqrt{2\pi}} \quad - \frac{z_1}{\sigma_1^2} e^{-z_1^2/2} \quad + \frac{\pi_2}{\sqrt{2\pi}} \quad \frac{z_2}{\sigma_2^2} e^{-z_2^2/2}$$

Thus, the positive semi-definiteness of $\mathcal{H}$ requires that $\frac{\partial^2 p_e}{\partial w_0^2} \geq 0$.

However, recall that even when one considers the stationary points given by (33), $\frac{\partial^2 p_e}{\partial w_0^2} \geq 0$ if and only if $\frac{z_2}{\sigma_2} \geq \frac{z_1}{\sigma_1}$. Thus, as has been shown in Theorem 1, for w_0 given by:

$$w_0 = w_0^- = \frac{\mu_2 \sigma_1^2 - \mu_1 \sigma_2^2 - \sigma_1 \sigma_2 \beta}{\sigma_1^2 - \sigma_2^2}, \tag{60}$$

this condition is not satisfied.

Therefore $\frac{\partial^2 p_e}{\partial w_0^2}$ is not greater than or equal to zero for every $\mathbf{w} \in \mathbb{R}^d$ and $w_0 \in \mathbb{R}$. Hence, $\mathcal{H}$ is not positive semi-definite. This in turn implies that p_e is non-convex. □

For this reason, the Bayes error is characterised by multiple local minima, so that the iterative algorithm described has to be performed for several initial choices of $\mathbf{w}$ for R runs. As we know the general form of the optimal solution of $\mathbf{w}$ as given by (11), we may choose the initial solutions in accordance, as is done in Step 3. Each run of the iterative procedure then converges on a stationary of p_e giving us an ensemble of R classifiers. One can then find the classifier among this ensemble that minimises the probability of misclassification p_e.

After the algorithm has converged to give the optimal weight $\mathbf{w}^*$ and threshold w_0^*, one may calculate the optimal value of s, s^* as:

$$s^* = \frac{\sigma_1^* z_2^*}{\sigma_1^* z_2^* - \sigma_2^* z_1^*}. \tag{61}$$

where

$$\begin{aligned} \sigma_1^* &= \mathbf{w}^{*T}\mathbf{\Sigma}_1\mathbf{w}^*, \quad \sigma_2^* = \mathbf{w}^{*T}\mathbf{\Sigma}_2\mathbf{w}^* \\ z_1^* &= \frac{w_0^* - \mathbf{w}^{*T}\bar{\mathbf{x}}_1^*}{\sigma_1^*}, \quad z_2^* = \frac{w_0^* - \mathbf{w}^{*T}\bar{\mathbf{x}}_2^*}{\sigma_2^*} \end{aligned} \tag{62}$$

Thus, $\mathbf{w}^*$ may equivalently be calculated as:

$$\mathbf{w}^* = \ s^*\mathbf{\Sigma}_1 + (1 - s^*)\mathbf{\Sigma}_2 \ ^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{63}$$

3.4. A dynamic linear model for unbalanced data

In the event of a class imbalance, the error rate may be skewed toward the majority class. For this reason, the threshold w_0 may be varied to improve the correct classification of the minority class at the expense of the overall classification accuracy. In this manner, the true positive rate (TPR) may be plotted for different false positive rates (FPR) to obtain the receiver operating characteristics (ROC). (We have assumed that the positive class is the minority class). For a lot of classifiers such as the support vector machine (SVM) or LDA under equal covariance ((8)), the discriminating threshold w_0 is independent of the weight vector $\mathbf{w}$, and therefore varying w_0 does not change $\mathbf{w}$.

Note however that in the case of heteroscedastic LDA, the weight vector $\mathbf{w}$ is tied to the threshold w_0 through z_1 and z_2, as shown in (27). It is only for the choice of the Bayes-optimal threshold $w_0 = w_0^+$ that $\mathbf{w}$ becomes as given by (30) which can be solved for iteratively from Algorithm 1. Thus, the optimal weight vector, $\mathbf{w}^*$ is obtained for the optimal threshold w_0^*. Therefore varying the threshold to obtain the FPR and TPR necessarily causes $\mathbf{w}^*$ to be suboptimal. To

correct this, $\mathbf{w}$ ought to be optimised for every choice of w_0. As the discriminating threshold is varied, the Bayes-optimal weight vector $\mathbf{w}$ for a given non-optimal threshold w_0, denoted as $\mathbf{w}(w_0)$, can be obtained from (27) as:

$$\mathbf{w} = \left(\frac{\pi_2 z_2}{\sigma_2^2} e^{-\frac{z_2^2}{2}} \mathbf{\Sigma}_2 - \frac{\pi_1 z_1}{\sigma_1^2} e^{-\frac{z_1^2}{2}} \mathbf{\Sigma}_1 \right)^{-1} \left(\frac{\pi_1}{\sigma_1} e^{-\frac{z_1^2}{2}} \bar{\mathbf{x}}_1 - \frac{\pi_2}{\sigma_2} e^{-\frac{z_2^2}{2}} \bar{\mathbf{x}}_2 \right) \tag{64}$$

If both sides of (27) are projected onto $\mathbf{w}$, the following is obtained:

$$\mathbf{w}^\top \left(\frac{\pi_2 z_2}{\sigma_2^2} e^{-z_2^2/2} \mathbf{\Sigma}_2 - \frac{\pi_1 z_1}{\sigma_1^2} e^{-z_1^2/2} \mathbf{\Sigma}_1 \right) \mathbf{w} = \mathbf{w}^\top \left(\frac{\pi_1}{\sigma_1} e^{-z_1^2/2} \bar{\mathbf{x}}_1 - \mathbf{w}^\top \left(\frac{\pi_2}{\sigma_2} e^{-z_2^2/2} \right) \bar{\mathbf{x}}_2 \right). \tag{65}$$

The above result can be simplified as follows:

$$\pi_2 z_2 e^{-\frac{z_2^2}{2}} - \pi_1 z_1 e^{-\frac{z_1^2}{2}} = \frac{\pi_1 \mu_1}{\sigma_1} e^{-\frac{z_1^2}{2}} - \frac{\pi_2 \mu_2}{\sigma_2} e^{-\frac{z_2^2}{2}}, \tag{66}$$

$$\pi_2 e^{-\frac{z_2^2}{2}} \left(z_2 + \frac{\mu_2}{\sigma_2} \right) = \pi_1 e^{-\frac{z_1^2}{2}} \left(z_1 + \frac{\mu_1}{\sigma_1} \right), \tag{67}$$

$$\frac{\pi_2 w_0}{\sigma_2} e^{-\frac{z_2^2}{2}} = \frac{\pi_1 w_0}{\sigma_1} e^{-\frac{z_1^2}{2}}. \tag{68}$$

Substituting (68) into (64) then results in:

$$\mathbf{w} = \left(\frac{z_2}{\sigma_2} \mathbf{\Sigma}_2 - \frac{z_1}{\sigma_1} \mathbf{\Sigma}_1 \right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{69}$$

where z_1 and z_2 are functions of w_0. Thus, a dynamic model of the weight vector $\mathbf{w}$ may be obtained as a function of a given threshold value w_0 as:

$$\mathbf{w}(w_0) = \left(\frac{z_2'}{\sigma_2^*} \mathbf{\Sigma}_2 - \frac{z_1'}{\sigma_1^{*2}} \mathbf{\Sigma}_1 \right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{70}$$

where

$$\sigma_1^{*2} = \mathbf{w}^{*\top} \mathbf{\Sigma}_1 \mathbf{w}^*, \quad \sigma_2^{*2} = \mathbf{w}^{*\top} \mathbf{\Sigma}_2 \mathbf{w}^*, \quad z_1' = \frac{w_0 - \mathbf{w}^{*\top} \bar{\mathbf{x}}_1^*}{\sigma_1^*}, \quad z_2' = \frac{w_0 - \mathbf{w}^{*\top} \bar{\mathbf{x}}_2^*}{\sigma_2^*} \tag{71}$$

and $\mathbf{w}^*$ is the optimal weight vector that the iterative procedure in Algorithm 1 produces. This model, together with the iterative procedure in Algorithm 1 that yields $\mathbf{w}^*$, is referred to as dynamic heteroscedastic linear discriminant (D-HLD).

4. Experimental Validation

The effects of class imbalance is investigated experimentally in this section. While the classification accuracy may be skewed toward the majority class under class imbalance, the datasets used here are not rebalanced, since rebalancing the data results in poor estimates of the true class prior probabilities employed in LDA (see Section 1). Instead, the discriminating threshold is varied to allow for the detection of more minority samples. Thus, we have provided the AUC as the evaluation metric. The AUC provides a measure of the trade-off between the false positive rate and the true positive rate, as the discriminating threshold is varied. The proposed D-HLD model is evaluated on an artificial dataset $\mathcal{D}_a$ with the following Gaussian parameters:

$$\begin{aligned}
&\bar{\mathbf{x}}_1 = [-1.5, -0.75, 0.75, 1.5]^\top, \\
&\mathbf{\Sigma}_1 = \text{diag}(0.25, 0.75, 1.25, 1.75) \\
&\bar{\mathbf{x}}_2 = \bar{\mathbf{x}}_1 - \omega, \quad \mathbf{\Sigma}_2 = \mathbf{I}
\end{aligned} \tag{72}$$

Here, ω controls the degree of class overlap, and is set to $\omega = 0.5$ in the experiments. The choice of 0.5 was obtained empirically via cross-validation so that the classes are not so well-separated that the classification problem is trivial, nor so overlapping that a linear classifier is not appropriate. $100,000$ points are generated in class $\mathcal{C}_1$ and $100,000f$ points are generated in class $\mathcal{C}_2$ to simulate an unbalanced data. Two values of f are used, i.e., $f = 2$ and $f = 10$, representing an imbalance ratio 2 and 10 respectively. This is followed by 10 trials of 10-fold cross validation.

The prior probabilities of the classes are estimated based on the relative frequencies of the data in each class in the dataset. We run only one trial of the proposed I-HLD algorithm (Algorithm 1), employing the original LDA solution in (8) as the initial solution. The stopping criterion for the I-HLD algorithm is thus: the procedure is stopped if the Bayes error p_e is less than or equal to 10^{-6}, or else it is terminated after 20 iterations, and the solution corresponding to the minimum p_e is chosen. For the purpose of evaluating the AUC, the dynamic linear model in (70) is employed. A step size of $\Delta s = 0.001$ is used for the C-HLD algorithm, and 1000 trials are run for the R-HLD procedure. All the parameters used in the experiments are optimised via cross-validation. Note that if the sample covariance matrix is singular, the Moore-Penrose pseudo-inverse is used.

The experiment is repeated for 8 real-world UCI datasets for which the fraction of the minority class ranges between 0.77% and 42.56%. The characteristics of the UCI datasets are shown in Table 1.

Table 1: Characteristics of UCI datasets

Dataset	d	n	Minority (%)	Majority (%)
E-Coli-1	7	1484	42.56	57.44
Liver	6	345	42.03	57.97
Diabetes	8	768	34.90	65.10
WpBC	33	194	23.71	76.29
USPS-1	256	1484	16.70	83.30
Yeast-1	8	1484	16.44	83.56
Yeast-6	8	1484	3.44	96.56
Abalone-19	7	4177	0.77	99.23

d is the dimensionality of the dataset, while n is the number of samples in the dataset. Indices appended to a dataset represents the minority class, while all remaining classes form the majority class.

The average AUC and training time over all 10 folds for each of the artificial and real-world datasets for our algorithm compared with LDA, QDA, R-HLD and C-HLD are then computed. These results are shown in Tables 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11. Other metrics of interest provided in the results include the error rate (ER), and the balanced error rate (BER), which is defined as half the sum of the false positive and false negative rates, i.e., $\text{BER} = 0.5(\text{FPR} + \text{FNR})$. While the F-measure is another common evaluation metric under class imbalance scenarios, it is not included here because it only considers the precision and recall values which do not take into account the true negatives, so that the true negatives can be allowed to vary freely without significantly changing the F-measure (Powers, 2011). Additionally, the existing LDA approaches are compared in the tables with the linear SVM (Hsu & Lin, 2002) without any enhancement by rebalancing procedures such as SMOTE.

Note that in all the tables, the best average values are in bold, while the values in asterisk are those that do not differ statistically from the best values based on Wilcoxon's signed rank test at a significance level of 0.01. The p-values are provided in brackets. ***Since the Wilcoxon's test is performed by comparing the results of each algorithm with that of the best performing algorithm (shown in bold), the p-values shown against the best performing algorithms themselves are necessarily*** 1.

Table 2: Artificial dataset $\mathcal{D}_a$ (f=2): AUC, ER and BER

Algorithm	AUC	ER	BER	Time (s)
LDA	0.674 ± 0.077	0.330 ± 0.053	0.443 ± 0.062	**0.008**
	(0.008)	(0.008)	(0.001)	(1)
C-HLD	0.667 ± 0.077	0.334 ± 0.075	0.378 ± 0.084	0.153
	(0.009)	(0.003)	(0.009)	(0.000)
R-HLD	0.654 ± 0.081	0.306 ± 0.054	0.422 ± 0.067	0.133
	(0.001)	(0.005)	(0.000)	(0.000)
QDA	$\mathbf{0.760 \pm 0.080}$	$\mathbf{0.272 \pm 0.069}$	$\mathbf{0.332 \pm 0.082}$	0.012
	(1)	(1)	(1)	(0.001)
SVM	0.719 ± 0.074	0.323 ± 0.022	0.481 ± 0.029	1347.635
	(0.006)	(0.009)	(0.008)	(0.000)
Proposed	0.745 ± 0.071	0.305 ± 0.054	0.422 ± 0.066	0.013
	(0.000)	(0.009)	(0.009)	(0.001)

Best values are in bold; p-values are shown in brackets. Values with asterisks appended are those which are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at the 0.01 ***significance level.***

Table 3: Artificial dataset $\mathcal{D}_a$ (f=10): AUC, ER and BER

Algorithm	AUC	ER	BER	Time (s)
LDA	0.748 ± 0.072	$0.090^{*} \pm 0.004$	0.490 ± 0.020	**0.038**
	(0.002)	(0.796)	(0.000)	(1)
C-HLD	0.747 ± 0.074	0.195 ± 0.035	$\mathbf{0.334 \pm 0.080}$	0.174
	(0.006)	(0.001)	(1)	(0.000)
R-HLD	0.712 ± 0.087	$\mathbf{0.089 \pm 0.007}$	0.483 ± 0.027	0.154
	(0.009)	(0.998)	(0.003)	(0.000)
QDA	$\mathbf{0.817 \pm 0.055}$	$\mathbf{0.089 \pm 0.012}$	0.463 ± 0.041	0.046
	(1)	(1)	(0.008)	(0.000)
SVM	0.740 ± 0.065	$0.091^{*} \pm 0.000$	0.500 ± 0.000	7667.820
	(0.002)	(0.903)	(0.004)	(0.000)
Proposed	0.788 ± 0.072	$\mathbf{0.089 \pm 0.007}$	0.483 ± 0.027	0.046
	(0.009)	(0.968)	(0.002)	(0.005)

Best values are in bold; p-values are shown in brackets. Values with asterisks appended are those which are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at the 0.01 ***significance level.***

Table 4: E-Coli-1 dataset: AUC, ER and BER

Algorithm	AUC	ER	BER	Time (s)
LDA	0.980 ± 0.012	$0.032^{*} \pm 0.029$	$0.032^{*} \pm 0.030$	**0.022**
	(0.009)	(0.708)	(0.692)	(1)
C-HLD	0.980 ± 0.012	$0.032^{*} \pm 0.029$	$0.032^{*} \pm 0.031$	0.191
	(0.005)	(0.792)	(0.803)	(0.000)
R-HLD	0.980 ± 0.012	$0.032^{*} \pm 0.029$	$0.032^{*} \pm 0.030$	0.217
	(0.008)	(0.806)	(0.771)	(0.000)
QDA	0.971 ± 0.022	0.441 ± 0.046	0.500 ± 0.000	0.045
	(0.001)	(0.004)	(0.001)	(0.003)
SVM	0.979 ± 0.012	$\mathbf{0.031} \pm \mathbf{0.029}$	$\mathbf{0.030} \pm \mathbf{0.028}$	1.242
	(0.004)	(1)	(1)	(0.000)
Proposed	$\mathbf{0.995} \pm \mathbf{0.009}$	$0.032^{*} \pm 0.029$	$0.032^{*} \pm 0.031$	0.069
	(1)	(0.813)	(0.666)	(0.005)

Best values are in bold; p-values are shown in brackets. Values with asterisks appended are those which are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at the 0.01 ***significance level.***

Table 5: Liver disorders dataset: AUC, ER and BER

Algorithm	AUC	ER	BER	Time (s)
LDA	0.703 ± 0.097	$\mathbf{0.309} \pm \mathbf{0.074}$	$\mathbf{0.333} \pm \mathbf{0.079}$	**0.001**
	(0.008)	(1)	(1)	(1)
C-HLD	0.700 ± 0.097	0.358 ± 0.083	0.366 ± 0.085	0.155
	(0.009)	(0.005)	(0.001)	(0.000)
R-HLD	0.699 ± 0.097	0.359 ± 0.083	0.367 ± 0.085	0.132
	(0.007)	(0.004)	(0.009)	(0.000)
QDA	0.692 ± 0.085	0.401 ± 0.082	0.386 ± 0.080	0.001^{*}
	(0.000)	(0.008)	(0.06)	(0.994)
SVM	0.728 ± 0.095	0.403 ± 0.038	0.472 ± 0.040	0.014
	(0.009)	(0.006)	(0.004)	(0.008)
Proposed	$\mathbf{0.757} \pm \mathbf{0.090}$	0.359 ± 0.083	0.367 ± 0.085	0.005
	(1)	(0.006)	(0.005)	(0.005)

Best values are in bold; p-values are shown in brackets. Values with asterisks appended are those which are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at the 0.01 ***significance level.***

Table 6: Diabetes dataset: AUC, ER and BER

Algorithm	AUC	ER	BER	Time (s)
LDA	0.828 ± 0.053	0.226 ± 0.046	$\mathbf{0.274 \pm 0.057}$	0.004
	(0.007)	(0.002)	(1)	(0.004)
C-HLD	0.827 ± 0.055	0.229 ± 0.045	0.283 ± 0.055	0.167
	(0.008)	(0.003)	(0.000)	(0.000)
R-HLD	0.827 ± 0.055	0.229 ± 0.046	0.282 ± 0.056	0.144
	(0.007)	(0.008)	(0.002)	(0.000)
QDA	0.805 ± 0.055	0.258 ± 0.047	0.300 ± 0.053	**0.001**
	(0.004)	(0.008)	(0.001)	(1)
SVM	0.836 ± 0.052	$\mathbf{0.223 \pm 0.045}$	$0.275^{*} \pm 0.055$	0.031
	(0.007)	(1)	(0.682)	(0.006)
Proposed	$\mathbf{0.845 \pm 0.047}$	0.229 ± 0.045	0.283 ± 0.055	0.005
	(1)	(0.006)	(0.002)	(0.006)

Best values are in bold; p-values are shown in brackets. Values with asterisks appended are those which are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at the 0.01 ***significance level.***

Table 7: WpBC dataset: AUC, ER and BER

Algorithm	AUC	ER	BER	Time (s)
LDA	0.788 ± 0.097	$\mathbf{0.202 \pm 0.091}$	$\mathbf{0.300 \pm 0.140}$	0.043
	(0.007)	(0.995)	(1)	(0.002)
C-HLD	0.785 ± 0.091	0.212 ± 0.091	0.307 ± 0.135	0.581
	(0.000)	(0.000)	(0.000)	(0.000)
R-HLD	0.785 ± 0.092	0.212 ± 0.091	0.307 ± 0.135	0.522
	(0.003)	(0.005)	(0.009)	(0.000)
QDA	0.641 ± 0.097	0.230 ± 0.035	0.480 ± 0.046	**0.001**
	(0.000)	(0.008)	(0.009)	(1)
SVM	0.720 ± 0.124	$\mathbf{0.202 \pm 0.046}$	0.411 ± 0.077	0.031
	(0.000)	(1)	(0.005)	(0.008)
Proposed	$\mathbf{0.923 \pm 0.051}$	0.211 ± 0.090	0.306 ± 0.135	0.057
	(1)	(0.001)	(0.005)	(0.002)

Best values are in bold; p-values are shown in brackets. Values with asterisks appended are those which are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at the 0.01 ***significance level.***

Table 8: USPS-1: AUC, ER and BER

Algorithm	AUC	ER	BER	Time (s)
LDA	$0.997^{*} \pm 0.002$	0.017 ± 0.003	0.032 ± 0.008	0.091
	(0.707)	(0.006)	(0.003)	(0.002)
C-HLD	$0.997^{*} \pm 0.002$	$\mathbf{0.015} \pm \mathbf{0.003}$	$\mathbf{0.022} \pm \mathbf{0.008}$	19.722
	(0.463)	(0.925)	(0.922)	(0.000)
R-HLD	$0.997^{*} \pm 0.002$	$\mathbf{0.015} \pm \mathbf{0.003}$	$\mathbf{0.022} \pm \mathbf{0.008}$	19.732
	(0.419)	(0.900)	(0.824)	(0.000)
QDA	0.984 ± 0.009	0.167 ± 0.001	0.500 ± 0.000	**0.035**
	(0.000)	(0.003)	(0.001)	(1)
SVM	$\mathbf{0.999} \pm \mathbf{0.124}$	0.128 ± 0.003	0.384 ± 0.009	12.491
	(1)	(0.002)	(0.008)	(0.000)
Proposed	0.990 ± 0.005	$\mathbf{0.015} \pm \mathbf{0.003}$	$\mathbf{0.022} \pm \mathbf{0.008}$	0.396
	(0.004)	(1)	(1)	(0.000)

Best values are in bold; p-values are shown in brackets. Values with asterisks appended are those which are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at the 0.01 ***significance level.***

Table 9: Yeast-1 dataset: AUC, ER and BER

Algorithm	AUC	ER	BER	Time (s)
LDA	0.832 ± 0.044	0.131 ± 0.020	0.313 ± 0.042	0.002^{*}
	(0.008)	(0.003)	(0.002)	(0.504)
C-HLD	0.825 ± 0.046	0.131 ± 0.021	$\mathbf{0.273} \pm \mathbf{0.047}$	0.167
	(0.008)	(0.005)	(1)	(0.000)
R-HLD	0.820 ± 0.047	0.131 ± 0.020	0.313 ± 0.042	0.143
	(0.002)	(0.002)	(0.000)	(0.000)
QDA	0.817 ± 0.054	0.164 ± 0.028	0.500 ± 0.000	**0.001**
	(0.005)	(0.006)	(0.001)	(1)
SVM	$0.854^{*} \pm 0.046$	$\mathbf{0.117} \pm \mathbf{0.016}$	0.324 ± 0.038	0.039
	(0.995)	(1)	(0.009)	(0.004)
Proposed	$\mathbf{0.856} \pm \mathbf{0.038}$	0.131 ± 0.019	0.312 ± 0.042	0.005
	(1)	(0.007)	(0.009)	(0.001)

Best values are in bold; p-values are shown in brackets. Values with asterisks appended are those which are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at the 0.01 ***significance level.***

Table 10: Yeast-6 dataset: AUC, ER and BER

Algorithm	AUC	ER	BER	Time (s)
LDA	0.878 ± 0.091	$0.037^{*} \pm 0.011$	0.390 ± 0.090	0.001^{*}
	(0.007)	(0.055)	(0.007)	(0.633)
C-HLD	0.878 ± 0.091	0.054 ± 0.016	$\mathbf{0.243 \pm 0.110}$	0.165
	(0.008)	(0.008)	(1)	(0.000)
R-HLD	0.877 ± 0.093	$0.037^{*} \pm 0.011$	0.390 ± 0.095	0.143
	(0.003)	(0.075)	(0.302)	(0.000)
QDA	0.845 ± 0.103	$\mathbf{0.034 \pm 0.011}$	0.500 ± 0.000	**0.001**
	(0.007)	(1)	(0.994)	(1)
SVM	0.845 ± 0.101	$\mathbf{0.034 \pm 0.002}$	0.500 ± 0.000	0.028
	(0.007)	(0.992)	(1)	(0.006)
Proposed	$\mathbf{0.911 \pm 0.064}$	$0.037^{*} \pm 0.011$	0.390 ± 0.095	0.005
	(1)	(0.099)	(0.000)	(0.007)

Best values are in bold; p-values are shown in brackets. Values with asterisks appended are those which are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at the 0.01 ***significance level.***

Table 11: Abalone-19 dataset: AUC, ER and BER

Algorithm	AUC	ER	BER	Time (s)
LDA	0.847 ± 0.077	0.015 ± 0.004	0.504 ± 0.002	**0.001**
	(0.001)	(0.002)	(0.000)	(1)
C-HLD	0.848 ± 0.076	0.085 ± 0.013	$\mathbf{0.308 \pm 0.130}$	0.163
	(0.005)	(0.008)	(1)	(0.000)
R-HLD	0.724 ± 0.139	0.014 ± 0.004	0.503 ± 0.002	0.140
	(0.009)	(0.005)	(0.007)	(0.000)
QDA	0.737 ± 0.149	0.016 ± 0.005	0.504 ± 0.002	0.001^{*}
	(0.003)	(0.009)	(0.007)	(0.893)
SVM	0.662 ± 0.124	$\mathbf{0.008 \pm 0.001}$	0.500 ± 0.000	0.083
	(0.006)	(1)	(0.007)	(0.002)
Proposed	$\mathbf{0.862 \pm 0.079}$	0.014 ± 0.004	0.503 ± 0.002	0.007
	(1)	(0.004)	(0.005)	(0.005)

Best values are in bold; p-values are shown in brackets. Values with asterisks appended are those which are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at the 0.01 ***significance level.***

4.1. Results and discussions

The results in Table 2 and 3 show QDA having the largest AUC among all the classifiers compared. ***The error rate performance of QDA in Table 3 is also comparable with those of LDA, R-HLD, SVM and the proposed algorithm, which show no statistical difference from QDA.*** The robust performance of QDA is consistent with the fact that the artificial dataset is normally distributed in each class with unequal covariances. Therefore the Bayes-optimal classifier is obtained from quadratic discriminant analysis. The SVM shows a competitive performance on this dataset to QDA (below the performance of the proposed algorithm). However, since the SVM does not make any assumptions on the distribution of the data, maximising the margin between the positive and negative examples does not necessarily yield the Bayes-optimal discrimination for this dataset with a known normal distribution. Moreover, the training time of the SVM is large, taking as much as 2.1 hours in Table 3, due to the fact that as a kernel classifier, it doesn't scale well with a lot of training data.

For arbitrary non-normal distributions, however, QDA may not perform just as well, and may be prone to overfitting, due in part to the fact that it is a quadratic classifier. Linear classifiers, on the other hand, tend to be more robust to non-normality than quadratic classifiers (Raschka, 2014). Thus, LDA, as well as R-HLD and the proposed D-HLD algorithm outperform QDA in terms of the error rate on most of the real-world datasets. However, C-HLD, while also being a linear model, constrains the parameter s to $[0, 1]$ in (15) and (16). This has been shown analytically to affect the classification accuracy in Section 3.2, since s tends to fall outside the interval $[0, 1]$ under class imbalance. This accounts for why C-HLD shows the largest error rate in both Tables 2 and 3. It is for this same reason that the C-HLD achieves the best BER in both Tables 2 and 3, since by constraining s to the interval $[0, 1]$, the discriminating threshold is always bounded between the projected class means, and hence the error rate tends to be more balanced.

It will also be noted that the error rate (ER) of the LDA procedure is significantly worse than that of QDA in Table 2, but only marginally in Table 3. This is because as the degree of class imbalance increases, the majority class becomes far more probable than the minority class. Therefore, the decision rule depends less on the differences in covariance matrices, but depends more on the discriminating threshold w_0. Since the threshold obtained by LDA as given by (8) is unbounded and depends on the ratio of the prior probabilities (or equivalently the degree of class imbalance), LDA is able to track the optimal w_0 under high degrees of class imbalance and yields a satisfactory performance in terms of the error rate. This

result confirms the conclusions by Xue & Titterington (2008) that unbalanced data have no negative effect on LDA in terms of the error rate.

Unlike LDA however, R-HLD and the D-HLD account for heteroscedasticity by obtaining a linear approximation to the quadratic boundary in QDA that minimises the Bayes error. Due to this, their error rate performance is closest to QDA on the toy dataset under any degree of class imbalance as can be seen from Tables 2 and 3. Since the criterion that is minimised in the R-HLD and D-HLD procedures is the Bayes error (or the probability of misclassification) which makes use of the empirical prior probabilities, the BER is not necessarily minimised for these procedures. However, regarding the AUC, D-HLD dynamically optimises the weight vector $\mathbf{w}$ to minimise the Bayes error for any given threshold w_0, so that for the FPR corresponding to that threshold, the TPR is maximised. Therefore, D-HLD results in an improved AUC over R-HLD.

For the real-world datasets, due to the fact that they are not drawn from a normal distribution, QDA is no longer superior in terms of the error rate. For these datasets, the best error rate performance is dominated by SVM, which is a non-parametric classifier. The original LDA and heteroscedastic LDA procedures compare closely to the SVM in terms of the error rate, and consistently outperform QDA due to the fact the linear models provide robustness over QDA, even if the normal distribution assumption is not satisfied.

Still, the fact that the BER happens to be significantly larger than the ER values on most of the real-world datasets suggests that the classification is skewed toward the majority class. This is particularly so for the SVM and QDA classifiers on the USPS, Liver, WpBC, Yeast-1 and Yeast-6 datasets. The AUC is then a preferred evaluation criterion. For the same reason as indicated for the artificial datasets, the proposed D-HLD procedure yields the best AUC values over all the real-world datasets, but the USPS dataset. Moreover, D-HLD is superior to the other heteroscedastic LDA procedures (R-HLD and C-HLD) in terms of the training time, since D-HLD follows a principled optimisation procedure for minimising the Bayes error, unlike in R-HLD and C-HLD. This computational gain increases with the dimensionality d of the dataset, and is most profound on the USPS dataset, since the bulk of the computation required in the heteroscedastic LDA procedures is for the inversion of a d-sized scatter matrix.

5. Conclusion

In this paper, we have shown that existing heteroscedastic linear discriminants are either suboptimal under class imbalance or have no principled optimisation

procedure, using first and second-order optimality conditions. We have thus presented a principled iterative procedure for obtaining the Bayes optimal linear classifier for heteroscedastic LDA, which is computationally efficient. Following this, we have derived a dynamic linear model for heteroscedastic LDA under class imbalance scenarios, based on the optimality conditions for the minimisation of the Bayes error. Our approach, unlike those in the literature, has been shown to be robust against class imbalance in terms of the AUC. Experimental results based on two artificial and eight real-world datasets show that the proposed algorithm compares favourably with the existing heteroscedastic LDA procedures as well as the SVM in terms of the error rate, but at the same time, it achieves superior performance in terms of the AUC as compared to all the algorithms. Moreover, the short training time of our algorithm makes it very well-suited for large-data applications.

Our future work is focused on going beyond Gaussian families of probability distribution to obtain the Bayes error for more general distributions. Alternatively, work is on-going in obtaining a kernel transformation that implicitly maps classes of known non-normal distribution into a feature space where the classes are nearly-normally distributed.

References

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. *Machine learning: ECML 2004*, (pp. 39–50).

Anderson, T. W., & Bahadur, R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *The annals of mathematical statistics*, (pp. 420–431).

Ashkezari, A. D., Ma, H., Saha, T. K., & Ekanayake, C. (2013). Application of fuzzy support vector machine for determining the health index of the insulation system of in-service power transformers. *IEEE Transactions on Dielectrics and Electrical Insulation*, *20*, 965–973.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*, 4626–4636.

Cai, D., He, X., Zhou, K., Han, J., & Bao, H. (2007). Locality sensitive discriminant analysis. In *IJCAI* (pp. 1713–1726). volume 2007.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Decell, H. P., & Mayekar, S. M. (1977). Feature combinations and the divergence criterion. *Computers & Mathematics with Applications*, *3*, 71–76.

Decell, H. P., Odell, P. L., & Coberly, W. A. (1981). Linear dimension reduction and bayes classification. *Pattern Recognition*, *13*, 241–243.

Decell Jr, H. P., & Marani, S. K. (1976). Feature combinations and the bhattacharyya criterion. *Communications in Statistics-Theory and Methods*, *5*, 1143–1152.

Duin, R., & Loog, M. (2004). Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. *IEEE transactions on pattern analysis and machine intelligence*, *26*, 732–739.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Academic press.

Guo, Y., Hastie, T., & Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, *8*, 86–100.

Hamsici, O. C., & Martinez, A. M. (2008). Bayes optimality in linear discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, *30*, 647–657.

Hastie, T., & Tibshirani, R. (1996). Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 155–176).

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, *13*, 415–425.

Izenman, A. J. (2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Science & Business Media.

Loog, M., & Duin, R. P. (2002). Non-iterative heteroscedastic linear dimension reduction for two-class data. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (pp. 508–517). Springer.

Lyon, A. (2014). Why are normal distributions normal? *The British Journal for the Philosophy of Science*, *65*, 621–649.

Malina, W. (1981). On an extended fisher criterion for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 611–614).

Marks, S., & Dunn, O. J. (1974). Discriminant functions when covariance matrices are unequal. *Journal of the American Statistical Association*, *69*, 555–559.

McBain, J., & Timusk, M. (2011). Feature extraction for novelty detection as applied to fault detection in machinery. *Pattern Recognition Letters*, *32*, 1054–1061.

McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition* volume 544. John Wiley & Sons.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Mullers, K.-R. (1999). Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.* (pp. 41–48). IEEE.

Noury, N., Fleury, A., Rumeau, P., Bourke, A., Laighin, G., Rialle, V., & Lundy, J. (2007). Fall detection-principles and methods. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE* (pp. 1663–1666). IEEE.

Ojetola, O., Gaura, E., & Brusey, J. (2015). Data set for fall events and daily activities from inertial sensors. In *Proceedings of the 6th ACM Multimedia Systems Conference* (pp. 243–248). ACM.

Peterson, D., & Mattson, R. (1966). A method of finding linear discriminant functions for a class of performance criteria. *IEEE Transactions on Information Theory*, *12*, 380–387.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, .

Raschka, S. (2014). linear discriminant analysis bit by bit. *Blog, August*, .

Sharma, A., & Paliwal, K. K. (2008). Cancer classification by gradient lda technique using microarray gene expression data. *Data & Knowledge Engineering*, *66*, 338–347.

Xie, Y., Li, X., Ngai, E., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, *36*, 5445–5449.

Xue, J.-H., & Titterington, D. M. (2008). Do unbalanced data have a negative effect on lda? *Pattern Recognition*, *41*, 1558–1571.

Yu, H., & Yang, J. (2001). A direct lda algorithm for high-dimensional data - with application to face recognition. *Pattern Recognition*, *34*, 2067–2070.

Zhao, Z., Sun, L., Yu, S., Liu, H., & Ye, J. (2009). Multiclass probabilistic kernel discriminant analysis. In *IJCAI* (pp. 1363–1368).