Linear dimensionality reduction for classification via a sequential Bayes error minimisation with an application to flow meter diagnostics

Gyamfi, KS, Brusey, J, Hunt, A & Gaura, E

Original citation & hyperlink:
Gyamfi, KS, Brusey, J, Hunt, A & Gaura, E 2018, 'Linear dimensionality reduction for classification via a sequential Bayes error minimisation with an application to flow meter diagnostics' *Expert Systems with Applications*, vol 91, pp. 252-262
https://dx.doi.org/10.1016/j.eswa.2017.09.010

DOI 10.1016/j.eswa.2017.09.010
ISSN 0957-4174

Publisher: Elsevier

Linear dimensionality reduction for classification via a sequential Bayes error minimisation with an application to flow meter diagnostics

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

Supervised linear dimensionality reduction (LDR) performed prior to classification often improves the accuracy of classification by reducing overfitting and removing multicollinearity. If a Bayes classifier is to be used, then reduction to a dimensionality of $K-1$ is necessary and sufficient to preserve the classification information in the original feature space for the K-class problem. However, most of the existing algorithms provide no optimal dimensionality to which to reduce the data, thus classification information can be lost in the reduced space if $K-1$ dimensions are used. In this paper, we present a novel LDR technique to reduce the dimensionality of the original data to $K-1$, such that it is well-primed for Bayesian classification. This is done by sequentially constructing linear classifiers that minimise the Bayes error via a gradient descent procedure, under an assumption of normality. We experimentally validate the proposed algorithm on 10 UCI datasets. Our algorithm is shown to be superior in terms of the classification accuracy when compared to existing algorithms including LDR based on Fisher's criterion and the Chernoff criterion. The applicability of our algorithm is then demonstrated by employing it in diagnosing the health states of 2 ultrasonic flow meters. As with the UCI datasets, the proposed algorithm is found to have superior performance to the existing algorithms, achieving classification accuracies of 99.4% and 97.5% on the two flow meters. Such high classification

*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)

accuracies on the flow meters promise significant cost benefits in oil and gas operations.

Keywords: Linear dimensionality reduction, LDA, heteroscedasticity, Bayes error, flow meter diagnostics

---

Point to Point Responses to Review Comments

The authors would like to thank the anonymous reviewers for their insightful comments. Below, we address these comments point by point for each reviewer.

Responses to Reviewer 1 Comments

- *In the conclusion section, the authors need to clearly discuss their theoretical contributions in Expert and Intelligent Systems compared to those in related papers in Expert and Intelligent Systems. This MUST be added in a separate paragraph. Overall, contributions of the article are unclear and weak.*

  Response

  We have introduced the following paragraph in the conclusion to clarify the theoretical contributions of the paper to expert and intelligent systems:

  The proposed algorithm is applicable to expert and intelligent systems that require LDR to overcome overfitting in predictive classification models so that prediction accuracies may subsequently be improved. Most of the existing LDR procedures provide no optimal dimensionality to which a given dataset may be reduced, and they tend to lose class-discriminatory information in the optimal $(K-1)$*-dimensional subspace required for Bayesian classification. In contrast, the proposed algorithm provides an optimal reduction to a dimensionality of* $K-1$ *via a sequential minimisation of the Bayes error, thus guaranteeing a much better classification accuracy than the existing approaches using a Bayes classifier such as LDA or QDA.*

- *In a separate paragraph it is required to provide some including remarks to further discuss the proposed methods, for example, what are the main advantages and limitations in comparison with existing methods?*

Response

We have included a paragraph in the conclusion highlighting the strength and disadvantages of the proposed algorithm:

While the proposed algorithm has been shown to be superior to the existing procedures on the datasets tested in terms of classification accuracies, it is built on an assumption of normality of the data in each of the K-classes. Yet, since a lot of physical data tend to be nearly-normally distributed (Lyon, 2013), our algorithm is well suited for a lot of applications particularly those involving measurement errors such as machine fault diagnosis or those involving physical measurements such as accelerometer-based human activity recognition. However, for data that are radically non-normal, our procedure is expected to perform relatively poorly, as the Bayes error is not guaranteed to be minimised. Also, while the proposed LDR procedure has been derived for Bayesian classification and is thus expected to perform well on Bayesian classifiers such as LDA, QDA and the Naive Bayes classifier, it is not suitable for other discriminative classifiers such as the SVM or logistic regression. Moreover, our algorithm requires the construction of $(K^2 + K - 4)/2$ classifiers which can be rather computationally costly for a dataset having too many classes.

- *Please open a real window for future work in the conclusion section. The authors also need to clearly provide 4-5 solid future research directions in the Conclusion section. These directions should be written as at least a separate paragraph and such directions need to be insightful for most of ESWA community.*

Response

We have rewritten the last paragraph of the conclusion to clearly show our four main future research directions:

In view of the above problems, our future work is concerned with the violation of the assumption of normality employed in the proposed algorithm. This would make the procedure more robust and applicable to a wider range of problems. While the Bayes error can be analytically intractable for an arbitrary non-normal distribution, we aim to extend the proposed procedure to minimising some upper bounds on the Bayes error for a given dataset. Alternatively, future research is aimed at deriving a kernel function that implicitly transforms some data of a known non-normal distribution into a feature space where the data in each class is nearly normally distributed. Moreover, we hope to explore the use of information theoretic measures to reduce the total number of classifiers constructed in each step of the proposed algorithm. This would decrease the computational complexity of the algorithm and improve its speed. Finally, as an application to flow meter diagnostics, our future work is focused on leveraging the correct diagnosis of a flow meter in the estimation of the error associated with each flow measurement with reasonable accuracy. With knowledge of the true health state of a flow meter, the associated measurement errors can be estimated with improved accuracy. This will allow erroneous flow measurements to be self-validated, thus resulting in significant cost cuts due to incorrect flow measurements in oil and gas operations.

Responses to Reviewer 2 Comments

- *The organization and writing of the paper need to be improved significantly, including all sections.*

Response

We have reworked the paper to improve upon its organisation and the quality of the writing. In particularly, it will be noted that the short detour about flow meter diagnostics in the introduction section has been removed in order to maintain the flow of the discussion of linear dimensionality reduction. Meanwhile, the discussion of flow meter diagnostics which was originally placed under the experimental validation

section has now been moved to a standalone section. Other grammatical and typographical errors, as well as awkward phrasings, have been corrected to the best of our abilities.

- *There are various techniques or algorithms developed and employed in different fields to simplify the structure of the input data so that higher accuracy of classification may be achieved in the end. However, the efficiency of each of them actually depends on the specific data set considered. Using the algorithm proposed by the authors, the average classification accuracy regarding LDA and QDA for different natural data sets in two cases seem to be increased to certain level. This is good. If a group of statistical hypothesis tests can be conducted to show the significant difference among the results from different algorithms, it will be better.*

Response

For every test dataset in all experiments, we perform the Wilcoxon's signed rank test at a significant level of 0.01 to check for any significant difference between the classification accuracy of the best performing algorithm and those of the remaining algorithms. Based on the test results, an asterisk has been indicated against a value if that value is not statistically different from the best value in bold. This is why the captions of Table 2 and Table 3 read:

The values with asterisk (*****) are those that are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at a significance level of 0.01.

This clarification has now been included in the text in the experimental validation section.

- *LDA and QDA are two classical and traditional statistical analyzing methods, typically used for classification. There are many other classical and advanced classification approaches in statistics, such as SVM and its variants, proved to be very efficient in many application areas. A comparison between the proposed algorithm and SVM when applying*

to at least one of those real data sets can make this paper much more meaningful.

Response

Since the proposed algorithm is a linear dimensionality reduction procedure, a direct comparison with the support vector machine (SVM) is not appropriate. Instead, we have now introduced a comparison in Table 4 between the proposed LDR algorithm followed by LDA, and the linear SVM without any dimensionality reduction.

We note, however, that just as class separability of LDR methods depends on the classifier to be used (Fukunaga, 2013), our algorithm linearly projects the data onto a $(K-1)$-dimensional subspace which is optimal for Bayesian classification. It is for this reason that we have employed LDA and QDA for classification after the proposed LDR procedure, and not the SVM. This limitation of the proposed algorithm has now been highlighted in the conclusion:

Also, while the proposed LDR procedure has been derived for Bayesian classification and is thus expected to perform well on Bayesian classifiers such as LDA, QDA and the Naive Bayes classifier, it is not suitable for other discriminative classifiers such as the SVM or logistic regression.

1. Introduction

Linearly reducing the dimensionality of a dataset is an important preprocessing step in machine learning for a number of reasons. For one thing, linear dimensionality reduction (LDR) enables easy visualisation of data when the data is reduced to two or three dimensions. For another, performing LDR prior to learning can reduce model complexity while alleviating the small sample size problem in algorithms such as Fisher's linear discriminant, where a very large dimensionality and much smaller training data cause the scatter matrix to be singular (Sharma & Paliwal, 2015; Lu et al., 2003). More importantly, however, LDR often improves learning in the low-dimensional manifold in which the data is reduced to lie (Brunzell & Eriksson, 2000; Duin & Loog, 2004). This is usually due to the fact that LDR results in useful feature extraction from a dataset, thus reducing overfitting (Bermingham et al.,

2015; James et al., 2013). In algorithms such as k-Nearest Neighbours (kNN), the performance improvement obtained from LDR is also attributable to the fact that LDR mitigates the effects of the so-called curse of dimensionality (Beyer et al., 1999).

LDR has been applied to several problems such as medical diagnosis e.g. (Sharma & Paliwal, 2008; Coomans et al., 1978; Sengur, 2008; Polat et al., 2008), face and object recognition e.g. (Song et al., 2007; Chen et al., 2000; Liu et al., 2007; Yu & Yang, 2001) and credit card fraud prediction e.g. (Mahmoudi & Duman, 2015) to reduce the dimensionality of very high-dimensional feature spaces. Indeed, there are several other emerging application areas where dimensionality reduction can be employed to improve learning. One such area is flow meter diagnostics which is described in Section 4.

One of the most popular LDR techniques is Principal Components Analysis (PCA) (Barber, 2012), which works by projecting the original data onto a subspace where the variance of the data is maximised in each dimension. However, when statistical classification is desired after dimensionality reduction, PCA may lose the class-discriminatory information in the data, as the directions of maximum variance does not necessarily coincide with the most class-discriminative directions.

In order to maximise the class-discriminatory information while linearly reducing the dimensionality, LDR aimed for classification makes use of class labels to inform the choice of the transformation matrix $\mathbf{M}$. In this case, the optimum objective function to minimise is the Bayes error in the linearly reduced space (Fukunaga, 2013; Buturovic, 1994). However, as an analytic expression for the Bayes error is hard to obtain for any arbitrary probability distribution, several approximations have been made (Fukunaga, 2013; Duda et al., 2012; Buturovic, 1994), leading to several supervised dimension reduction techniques (Cunningham & Ghahramani, 2015; Duin & Loog, 2004; Brunzell & Eriksson, 2000; Barber, 2012). Notable among these techniques is Linear Discriminant Analysis (LDA) (Fisher, 1936; Izenman, 2009; Fukunaga, 2013; Barber, 2012). At its core, LDA is built on the assumption that the data is normally distributed in each class, with the covariance matrices of the classes being equal (an assumption known as homoscedasticity). Consequently, Fisher's LDA maximises Fisher's criterion (Fukunaga, 2013; Barber, 2012; Duin & Loog, 2004) as a measure of class separability, by taking only the differences in the projected class means into account, ignoring any differences in covariance matrices that might be present among the various classes in the data (Duin & Loog, 2004).

However, experimental results have shown that if one accounts for the violation of the assumptions in the original procedure, the performance of LDA can be improved (Hastie & Tibshirani, 1996; Marks & Dunn, 1974; Mika et al., 1999; Zhao et al., 2009). Along this line, our previous work describe an iterative procedure to obtain a one-dimensional subspace where the Bayes error is minimised in the two-class problem under the normality assumption in LDA, while accounting for heteroscedasticity (Gyamfi et al., 2017).

In this paper, we present a novel technique to LDR, that projects the original data onto a $(K-1)$-dimensional subspace for the K-class problem. We do this by sequentially creating linear classifiers that minimise the Bayes error under assumptions of normality and heteroscedasticity via a gradient descent procedure. This procedure is described in Section 3. Though iterative, the proposed algorithm is fast, and it remains unaffected by the number of training examples. In Section 4, we describe the applicability of LDR to flow meter diagnostics. In Section 5, we experimentally validate the proposed algorithm on 10 University of California, Irvine (UCI) datasets, as well as in the diagnosis of the health states of two ultrasonic flow meters, using datasets provided by the National Engineering Laboratory (NEL), United Kingdom.

2. Background and related work

Consider a dataset $\mathcal{D} = [\mathbf{x}_1, ..., \mathbf{x}_n]$ with n examples and a dimensionality of d. The dataset is assumed to be labelled and made up of K classes, i.e., $\mathcal{D} = [\mathcal{C}_1, ..., \mathcal{C}_K]$. We aim at finding a linear transformation $\mathbf{T}$ such that $\tilde{\mathcal{D}} = \mathbf{T}^\top \mathcal{D}$ has a dimensionality of q, i.e., $\mathbf{T} \in \mathbb{R}^{d \times q}$, where $q < d$, while maximising the class-discriminatory information.

Let $\bar{\mathbf{x}}_k$, $\mathbf{S}_k$ and $\pi_k = p(\mathcal{C}_k)$ respectively be the mean, covariance and prior probability of the kth class, for $k \in \{1, ..., K\}$. Also, let $\bar{\mathbf{x}}$ be the mean of the overall dataset $\mathcal{D}$.

2.1. Fisher's criterion

Fisher's LDA aims to maximise Fisher's criterion as given by:

$$J_F = \text{trace}((\mathbf{T}^\top \mathbf{S}_W \mathbf{T})^{-1} (\mathbf{T}^\top \mathbf{S}_B \mathbf{T})) \tag{1}$$

where $\mathbf{S}_W$, the within-class scatter matrix and $\mathbf{S}_B$, the between-class scatter matrix are both given by

$$\mathbf{S}_W = \sum_{k=1}^{K} \pi_k \mathbf{S}_k \quad \text{and} \quad \mathbf{S}_B = \sum_{k=1}^{K} \pi_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^\top. \tag{2}$$

In the two-class case, where reduction to only one dimension is possible, maximising Fisher's criterion tends to minimise the Bayes error in the one-dimensional subspace onto which the data is projected, when the normality and homoscedasticity assumptions hold (Hamsici & Martinez, 2008; Izenman, 2009).

2.2. Mahalanobis criterion

For the K-class case (where $K > 2$), however, maximisation of Fisher's criterion does not guarantee the minimisation of the Bayes error, even when the assumptions of homoscedasticity and normality are satisfied. To get around this problem, an upper bound on the Bayes error based on the Mahalanobis distance has been employed for LDR in the multi-class scenario (Brunzell & Eriksson, 2000). The Mahalanobis-based LDR seeks to preserve the separation given by

$$J_M = \prod_{1 \leq i < j \leq K} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)^\top (\mathbf{S}_i + \mathbf{S}_j)^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j) \tag{3}$$

in the linearly reduced space.

However, the Mahalanobis distance, just like Fisher's criterion, does not take the difference in covariance matrices into account.

2.3. Chernoff criterion

To account for the difference in covariance matrices among the classes, a heteroscedastic extension of the Mahalanobis distance based on the Bhattacharya distance has been proposed for LDR (Decell Jr & Marani, 1976). Following this, there has been the use of a wider class of Bregman divergences, notably, the Kullback-Leibler divergence (Decell & Mayekar, 1977) for heteroscedastic LDR. Yet, while the Bhattacharya distance provides a good enough bound on the Bayes error, it has been shown that the Chernoff bound provides a slightly tighter bound than the Bhattacharya distance (Duda et al., 2012; Nielsen, 2014). Thus, a directed distance matrix (DDM) based on the Chernoff criterion has been developed for dimensionality reduction in the two-class case (Loog & Duin, 2002), as well as in the multi-class setting (Duin & Loog, 2004). Specifically, based on this DDM, the following

Chernoff criterion is derived:

$$J_C = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \pi_i \pi_j \text{trace}\big[(\mathbf{T}^\top \mathbf{S}_W \mathbf{T})^{-1} \mathbf{T}^\top \mathbf{S}_W^{\frac{1}{2}} \big((\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}} \mathbf{S}_W^{-\frac{1}{2}}$$
$$\times (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)^\top \mathbf{S}_W^{-\frac{1}{2}} (\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}} + \frac{1}{\tau_i \tau_j} (\log \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}}$$
$$- \tau_i \log \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_i \mathbf{S}_W^{-\frac{1}{2}} - \tau_j \log \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_j \mathbf{S}_W^{-\frac{1}{2}}) \big) \mathbf{S}_W^{\frac{1}{2}} \mathbf{T} \big], \tag{4}$$

with

$$\tau_i = \frac{\pi_i}{\pi_i + \pi_j}, \quad \tau_j = \frac{\pi_j}{\pi_i + \pi_j} \quad \text{and} \quad \mathbf{S}_{ij} = \pi_i \mathbf{S}_i + \pi_j \mathbf{S}_j, \tag{5}$$

which is maximised to obtain an optimum linear transformation (Duin & Loog, 2004).

However, while the original LDA procedure provides reduction to at most $K-1$ dimensions, the LDR approaches described do not provide the optimal dimensionality to which to reduce the data. The existing procedures are often reformulated as eigenvalue decomposition or singular value decomposition (SVD) problems (Cunningham & Ghahramani, 2015), after which a desired dimensionality q is chosen by taking the first q independent vectors after the decomposition. Yet, it has been shown that if a Bayes classifier, such as quadratic discriminant analysis (QDA), is to be applied after LDR, the smallest set of independent features required is $K-1$, corresponding to a reduction to a $(K-1)$-dimensional space (Fukunaga, 2013). This is because the optimal Bayes classifier evaluates K posterior probabilities, among which the highest is chosen. Since the K probabilities must sum up to one, only $K-1$ of these K probability functions are independent. Thus, reduction to a $(K-1)$-dimensional subspace is necessary and sufficient to preserve the classification information in the original feature space (Fukunaga, 2013). In the absence of an optimal dimensionality q in the existing LDR procedures described, if q is set to $K-1$, there is no guarantee that the first $K-1$ independent vectors alone preserve the class-discriminatory information in the original space. As a result, classification information can be lost in the $(K-1)$-dimensional subspace, formed from the first $K-1$ singular vectors or eigenvectors following an SVD or eigenvalue decomposition, leading to a reduced classification accuracy using a Bayes classifier.

Our proposed algorithm, on the other hand, provides a reduction to a $(K-1)$-dimensional subspace while preserving the classification information in the original feature space.

3. Proposed algorithm

We assume that the data in each of the K classes is normally distributed with a mean of $\bar{\mathbf{x}}_k$ and a covariance matrix of $\mathbf{S}_k$ for every $k \in \{1, ..., K\}$. Our aim is to apply a Bayes classifier after LDR, thus we seek to project $\mathcal{D}$ onto a $(K-1)$-dimensional subspace. Therefore, we let the transformation matrix $\mathbf{T}$ be given as $\mathbf{T} = [\mathbf{v}_1, ..., \mathbf{v}_{K-1}]$, where $\mathbf{v}_i \in \mathbb{R}^d$ for $i \in \{1, ..., K-1\}$. The proposed algorithm is such that we find one column of $\mathbf{T}$ in each of $K-1$ steps.

3.1. K=2

In the two-class case, $\mathbf{T} = \mathbf{v}_1$, and therefore the task of finding $\mathbf{v}_1$ that preserves the classification information in the original space is equivalent to obtaining a linear discriminant that well separates the two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. That is, we obtain a linear classifier $\{\mathbf{w}_1, t_1\}$ such that for every data sample $\mathbf{x} \in \mathcal{D}$, the true class of $\mathbf{x}$, $\mathcal{C}^*(\mathbf{x})$, is decided according to the following decision rule:

$$\mathcal{C}^*(\mathbf{x}) = \begin{cases} \mathcal{C}_1 & \text{if} \quad \mathbf{w}_1^\top \mathbf{x} \geq t_1 \\ \mathcal{C}_2 & \text{if} \quad \mathbf{w}_1^\top \mathbf{x} < t_1 \end{cases} \tag{6}$$

The optimal $\mathbf{w}_1$ minimises the Bayes error given by:

$$\epsilon_1 = \pi_1 p\big(y < t_1 | \mathcal{C}_1\big) + \pi_2 p\big(y \geq t_1 | \mathcal{C}_2\big) \tag{7}$$

where $y = \mathbf{w}_1^\top \mathbf{x}$.

Since $\mathbf{x}$ is assumed to have a normal distribution in classes $\mathcal{C}_1$ and $\mathcal{C}_2$, y is expected to be normally distributed with a mean of μ_1 and a variance of σ_1^2 for class $\mathcal{C}_1$, and a mean of μ_2 and a variance of σ_2^2 for class $\mathcal{C}_2$ given as:

$$\mu_1 = \mathbf{w}_1^\top \bar{\mathbf{x}}_1 \quad \mu_2 = \mathbf{w}_1^\top \bar{\mathbf{x}}_2 \quad \sigma_1^2 = \mathbf{w}_1^\top \mathbf{S}_1 \mathbf{w}_1 \quad \sigma_2^2 = \mathbf{w}_1^\top \mathbf{S}_2 \mathbf{w}_1 \tag{8}$$

The normality assumption allows the individual misclassification probabilities in (7) to be expressed as:

$$p(y < t_1 | \mathcal{C}_1) = \int_{-\infty}^{t_1} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(r-\mu_1)^2}{2\sigma_1^2} \right] dr = 1 - Q\left(\frac{t_1 - \mu_1}{\sigma_1} \right) \tag{9}$$

and

$$p(y \geq t_1|\mathcal{C}_2) = \int_{t_1}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{(r-\mu_2)^2}{2\sigma_2^2}\right] dr = Q\left(\frac{t_1-\mu_2}{\sigma_2}\right) \tag{10}$$

where $Q(\cdot)$ is the Q-function, so that the Bayes error ϵ_1 may be rewritten as:

$$\epsilon_1 = \pi_1\big(1 - Q(z_1)\big) + \pi_2\big(Q(z_2)\big) \tag{11}$$

where

$$z_1 = \frac{t_1-\mu_1}{\sigma_1} \quad \text{and} \quad z_2 = \frac{t_1-\mu_2}{\sigma_2} \tag{12}$$

Our previous work has shown using first and second-order optimality conditions that $\mathbf{w}_1$ and t_1 that minimise ϵ_1 can be obtained by solving the following equations iteratively:

$$\mathbf{w}_1 = \left(\frac{z_2}{\sigma_2}\mathbf{S}_2 - \frac{z_1}{\sigma_1}\mathbf{S}_1\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{13}$$

and

$$t_1 = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 + \sigma_1\sigma_2\sqrt{(\mu_1-\mu_2)^2 + 2(\sigma_1^2-\sigma_2^2)\ln\left(\frac{\tau\sigma_1}{\sigma_2}\right)}}{\sigma_1^2-\sigma_2^2} \tag{14}$$

where $\tau = \pi_2/\pi_1$ (Gyamfi et al., 2017).

Though, our aim is to find only the weight vector $\mathbf{w}_1$, we note that $\mathbf{w}_1$ is not independent of t_1, as it is related to t_1 through z_1 and z_2. Therefore, the optimal choice of $\mathbf{w}_1$ is obtained only by optimising $\mathbf{w}_1$ and t_1 simultaneously.

We then set $\mathbf{v}_1$ as the optimal $\mathbf{w}_1$.

3.2. K=3

In the case where $K = 3$, the transformation matrix is $\mathbf{T} = [\mathbf{v}_1, \mathbf{v}_2]$.

3.2.1. Step 1

In the first step, we seek to find the first column of $\mathbf{T}$, i.e., $\mathbf{v}_1$. To do this, we train a linear classifier to separate one class from the remaining classes; since there are three classes, there are three different classifiers that could be constructed to this end. Our aim is to choose $\mathbf{v}_1$ to correspond to the classifier among these three that yields the smallest Bayes error.

First, we consider the possibility of training a linear classifier $\{\mathbf{w}_1, t_1\}$ to discriminate class $\mathcal{C}_1$ from classes $\mathcal{C}_2$ and $\mathcal{C}_3$. Then, for every data sample $\mathbf{x} \in \mathcal{D}$, the following decision rule applies:

$$\mathcal{C}^*(\mathbf{x}) = \begin{cases} \mathcal{C}_1 & \text{if} \quad \mathbf{w}_1^\top \mathbf{x} \geq t_1 \\ \mathcal{C}_2, \mathcal{C}_3 & \text{if} \quad \mathbf{w}_1^\top \mathbf{x} < t_1 \end{cases} \tag{15}$$

Notice that, as with the case $K = 2$, the projected data in class $\mathcal{C}_1$ is normally distributed on one side of the linear discriminant with a mean of μ_1 and a variance of σ_1^2 as given by (8), while the projected data in classes $\mathcal{C}_2$ and $\mathcal{C}_3$, on the other side of the discriminant, form a mixture of two Gaussians $\mathcal{M}_1$ given by:

$$\begin{aligned} &\mathcal{M}_1 \sim \sum_{i=2}^{3} p_i \mathcal{N}(\mu_i, \sigma_i^2), \quad \text{where} \quad p_i = \frac{\pi_i}{1 - \pi_1}, \quad \text{such that} \quad \sum_{i=2}^{3} p_i = 1 \\ &\text{and} \quad \mu_i = \mathbf{w}_1^\top \bar{\mathbf{x}}_i \quad \sigma_i^2 = \mathbf{w}_1^\top \mathbf{S}_i \mathbf{w}_1 \end{aligned} \tag{16}$$

As before, the optimal $\mathbf{w}_1$ then minimises the Bayes error. However, the Bayes error is now given by:

$$\epsilon_1 = \pi_1 p\big(y < t_1 | \mathcal{C}_1\big) + (1 - \pi_1) p\big(y \geq t_1 | \mathcal{M}_1\big) \tag{17}$$

Here, $p\big(y < t_1 | \mathcal{C}_1\big)$ is as given before in (9), while $p\big(y \geq t_1 | \mathcal{M}_1\big)$ can be expressed as:

$$\begin{aligned} p(y \geq t_1 | \mathcal{M}_1) &= \sum_{i=2}^{3} p_i \int_{t_1}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(r - \mu_i)^2}{2\sigma_i^2} \right] dr \\ &= \sum_{i=2}^{3} p_i Q\left(\frac{t_1 - \mu_i}{\sigma_i} \right). \end{aligned} \tag{18}$$

Thus, the Bayes error ϵ_1 can be evaluated as:

$$\begin{aligned} &\epsilon_1 = \pi_1 \big(1 - Q(z_{1,1})\big) + (1 - \pi_1) \sum_{i=2}^{3} p_i Q(z_{1,i}), \\ &\quad \text{where} \quad z_{1,k} = \frac{t_1 - \mathbf{w}_1^\top \bar{\mathbf{x}}_k}{(\mathbf{w}_1^\top \mathbf{S}_k \mathbf{w}_1)^{1/2}}, \quad \text{for every} \quad k \in \{1, 2, 3\}. \end{aligned} \tag{19}$$

i.e.,

$$\epsilon_1 = \pi_1\big(1 - Q(z_{1,1})\big) + \sum_{i=2}^{3} \pi_i Q(z_{1,i}) \tag{20}$$

Next, we consider the second possibility of training a linear classifier $\{\mathbf{w}_2, t_2\}$ to discriminate class $\mathcal{C}_2$ from classes $\mathcal{C}_1$ and $\mathcal{C}_3$. The optimal $\{\mathbf{w}_2, t_2\}$ minimises the Bayes error which can be shown, similar to the derivation of ϵ_1, to be given by:

$$\begin{aligned} \epsilon_2 &= \pi_2\big(1 - Q(z_{2,2})\big) + \sum_{i=1, i\neq 2}^{3} \pi_i Q(z_{2,i}), \\ &\text{where} \quad z_{2,k} = \frac{t_2 - \mathbf{w}_2^\top \bar{\mathbf{x}}_k}{(\mathbf{w}_2^\top \mathbf{S}_k \mathbf{w}_2)^{1/2}}, \quad \text{for every} \quad k \in \{1, 2, 3\}. \end{aligned} \tag{21}$$

Finally, we train a classifier $\{\mathbf{w}_3, t_3\}$ to linearly discriminate class $\mathcal{C}_3$ from classes $\mathcal{C}_1$ and $\mathcal{C}_2$. By minimising the Bayes error, ϵ_3 which can be shown to be given by:

$$\begin{aligned} \epsilon_3 &= \pi_3\big(1 - Q(z_{3,3})\big) + \sum_{i=1}^{2} \pi_i Q(z_{3,i}), \\ &\text{where} \quad z_{3,k} = \frac{t_3 - \mathbf{w}_3^\top \bar{\mathbf{x}}_k}{(\mathbf{w}_3^\top \mathbf{S}_k \mathbf{w}_3)^{1/2}}, \quad \text{for every} \quad k \in \{1, 2, 3\}, \end{aligned} \tag{22}$$

the optimal $\{\mathbf{w}_3, t_3\}$ is obtained.

We assume, without any loss of generality, that the third classifier $\{\mathbf{w}_3, t_3\}$ yields the smallest Bayes error, i.e., $\epsilon_3 < \epsilon_1, \epsilon_2$. We then set $\mathbf{v}_1$ to the optimal vector $\mathbf{w}_3$ corresponding to the minimisation of ϵ_3.

3.2.2. Step 2

The next step then is to find the second column of $\mathbf{T}$, i.e., $\mathbf{v}_2$. As we have trained a classifier to separate class $\mathcal{C}_3$ from the two remaining classes in Step 1, we remove $\mathcal{C}_3$ from the dataset $\mathcal{D}$ and proceed to construct a linear classifier $\{\mathbf{w}_1, t_1\}$, in the fashion of the case $K = 2$, to linearly discriminate classes $\mathcal{C}_1$ and $\mathcal{C}_2$. This is done by minimising the Bayes error ϵ_1 given by:

$$\epsilon_1 = \pi_1'\big(1 - Q(z_1)\big) + \pi_2'\big(Q(z_2)\big) \tag{23}$$

It will be noted that by removing $\mathcal{C}_3$ from the dataset $\mathcal{D}$, the prior probabilities of the remaining classes change. Therefore, in (23), π_1' and π_2' are the

prior probabilities of classes $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively, conditional on class $\mathcal{C}_3$ being removed from the dataset $\mathcal{D}$, and they are given by:

$$\pi_1' = \frac{\pi_1}{1-\pi_3} \quad \text{and} \quad \pi_2' = \frac{\pi_2}{1-\pi_3} \tag{24}$$

The optimal $\mathbf{w}_1$ is then assigned to $\mathbf{v}_2$.

Note that very often, the transformation matrix $\mathbf{T}$ is constrained to be orthogonal (Cunningham & Ghahramani, 2015). Thus, it would be necessary to have an orthogonality constraint in the form, $\mathbf{w}_1^\top \mathbf{w}_3 = 0$, while minimising the Bayes error ϵ_1 in the second step. However, such an orthogonality constraint is not binding, if classification is desired after dimensionality reduction (Fukunaga, 2013); it is sufficient that the component vectors of $\mathbf{T}$ be independent.

3.3. General K

Having detailed the fundamentals of the proposed LDR procedure for the special cases of $K = 2$ and $K = 3$, we proceed to describe the proposed algorithm for a general value of K.

Let $\mathcal{L} = \{1, ..., K\}$. We define a conditional prior probability $\pi_i' = p(\mathcal{C}_i|\bar{\mathcal{C}}_l)$ to be the prior probability of Class $\mathcal{C}_i$ conditional on the data in Class $\mathcal{C}_l$ being removed from the dataset $\mathcal{D}$, for all $i \in \mathcal{L}$. Then for the $k = 1$st iteration, when no class has been removed yet, $\pi_i' = \pi_i$.

3.3.1. Step 1

We construct a linear classifier $\{\mathbf{w}_i, t_i\}$ that discriminates class $\mathcal{C}_i$ from all other classes, for every $i \in \mathcal{L}$, by minimising the Bayes error ϵ_i given by:

$$\epsilon_i = \pi_i'\big(1 - Q(z_i)\big) + \sum_{j \in \mathcal{L}\setminus\{i\}} \pi_j' Q(z_j), \tag{25}$$

where

$$z_k = \frac{t_i - \mu_k}{\sigma_k}, \quad \mu_k = \mathbf{w}_i^\top \bar{\mathbf{x}}_k \quad \text{and} \quad \sigma_k^2 = \mathbf{w}_i^\top \mathbf{S}_k \mathbf{w}_i, \quad \text{for every} \quad k \in \mathcal{L}. \tag{26}$$

We then set $\mathbf{v}_k$, i.e., the kth column of $\mathbf{T}$, as the vector $\mathbf{w}_i$ corresponding to the smallest Bayes error, i.e.,

$$\mathbf{v}_k = \arg\min_{\mathbf{w}_i} \{\epsilon_1, ..., \epsilon_{|\mathcal{L}|}\} \tag{27}$$

and let

$$l = \arg\min_{i} \{\epsilon_1, ..., \epsilon_{|\mathcal{L}|}\} \tag{28}$$

3.3.2. Step 2

As the optimal classifier $\{\mathbf{w}_l, t_l\}$ linearly separates class $\mathcal{C}_l$ from all other classes, class $\mathcal{C}_l$ can be excluded from the dataset $\mathcal{D}$ to allow for the construction of other classifiers to linearly discriminate the remaining classes. Correspondingly, we remove l from the set $\mathcal{L}$. The conditional prior probabilities of the remaining classes are then updated as:

$$\pi_i' := \frac{\pi_i'}{1 - \pi_l'}, \quad \text{for all} \quad i \in \mathcal{L}. \tag{29}$$

The index k is then incremented by one.

3.3.3. Step 3

Steps 1 and 2 are repeated until all $K-1$ columns of the transformation matrix $\mathbf{T}$ have been determined.

3.4. Optimisation of the Bayes error ϵ_i

Up until this point, we have only mentioned that the classifier $\{\mathbf{w}_i, t_i\}$ ought to minimise the Bayes error given by (25). We now derive explicit expressions for the optimality conditions and propose a gradient descent procedure to minimise the error.

The first-order optimality condition for the minimisation of ϵ_i requires the gradient of ϵ_i to be zero, i.e.,

$$\nabla \epsilon_i(\mathbf{w}_i, t_i) = \left[\frac{\partial \epsilon_i}{\partial \mathbf{w}_i^\top}, \frac{\partial \epsilon_i}{\partial t_i} \right]^\top = \mathbf{0} \tag{30}$$

From (25), it can be shown that:

$$\frac{\partial \epsilon_i}{\partial \mathbf{w}_i} = \pi_i \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} \frac{\partial z_i}{\partial \mathbf{w}_i} \right) - \sum_{j \in \mathcal{L} \setminus \{i\}} \pi_j \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{z_j^2}{2}} \frac{\partial z_j}{\partial \mathbf{w}_i} \right) \tag{31}$$

where $\partial z_k / \partial \mathbf{w}_i$ can be obtained from (26) as:

$$\frac{\partial z_k}{\partial \mathbf{w}_i} = \frac{-\sigma_k \bar{\mathbf{x}}_k - z_k \mathbf{S}_k \mathbf{w}_i}{\sigma_k^2} \quad \text{for every} \quad k \in \mathcal{L}. \tag{32}$$

Therefore,

$$\frac{\partial \epsilon_i}{\partial \mathbf{w}_i} = \frac{\pi_i}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} \left(\frac{-\sigma_i \bar{\mathbf{x}}_i - z_i \mathbf{S}_i \mathbf{w}_i}{\sigma_i^2} \right) - \sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sqrt{2\pi}} e^{-\frac{z_j^2}{2}} \left(\frac{-\sigma_j \bar{\mathbf{x}}_j - z_j \mathbf{S}_j \mathbf{w}_i}{\sigma_j^2} \right) \tag{33}$$

Also,

$$\frac{\partial \epsilon_i}{\partial t_i} = \frac{\pi_i}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} \frac{\partial z_i}{\partial t_i} - \sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sqrt{2\pi}} e^{-\frac{z_j^2}{2}} \frac{\partial z_j}{\partial t_i} \tag{34}$$

where,

$$\frac{\partial z_k}{\partial t_j} = \frac{1}{\sigma_k}, \quad \text{for every} \quad k \in \mathcal{L} \tag{35}$$

which can also be obtained from (26). Therefore,

$$\frac{\partial \epsilon_i}{\partial t_i} = \frac{\pi_i}{\sqrt{2\pi}\sigma_i} e^{-\frac{z_i^2}{2}} - \sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sqrt{2\pi}\sigma_j} e^{-\frac{z_j^2}{2}} \tag{36}$$

By equating the gradient to zero, (33) yields the following:

$$\Bigg(\sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} \frac{z_j}{\sigma_j} \mathbf{S}_j - \frac{\pi_i}{\sigma_i} e^{-\frac{z_i^2}{2}} \frac{z_i}{\sigma_i} \mathbf{S}_i \Bigg) \mathbf{w}_i = \frac{\pi_i}{\sigma_i} e^{-\frac{z_i^2}{2}} \bar{\mathbf{x}}_i - \sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} \bar{\mathbf{x}}_j \tag{37}$$

while (36) results in:

$$\frac{\pi_i}{\sigma_i} e^{-\frac{z_i^2}{2}} = \sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} \tag{38}$$

Substituting (38) into (37), we obtain:

$$\begin{aligned} &\Bigg(\sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} \frac{z_j}{\sigma_j} \mathbf{S}_j - \sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} \frac{z_i}{\sigma_i} \mathbf{S}_i \Bigg) \mathbf{w}_i = \\ &\sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} \bar{\mathbf{x}}_i - \sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} \bar{\mathbf{x}}_j, \end{aligned} \tag{39}$$

i.e.,

$$\sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} \left(\frac{z_j}{\sigma_j} \mathbf{S}_j - \frac{z_i}{\sigma_i} \mathbf{S}_i \right) \mathbf{w}_i = \sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j) \tag{40}$$

$\mathbf{w}_i$ may then be obtained as:

$$\mathbf{w}_i = \Bigg[\sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} \left(\frac{z_j}{\sigma_j} \mathbf{S}_j - \frac{z_i}{\sigma_i} \mathbf{S}_i \right) \Bigg]^{-1} \sum_{j \in \mathcal{L} \setminus \{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j) \tag{41}$$

But for the fact that z_i and z_j are functions of t_i as can be seen from (26), $\mathbf{w}_i$ could have been solved for iteratively from (41) starting from an initial solution. To overcome this problem, we proceed to solve for t_i from (38), expressing it as a function of $\mathbf{w}_i$, to allow for the iterative solution of $\mathbf{w}_i$ from (41).

From (38), we derive the following:

$$\ln\left(\frac{\pi_i}{\sigma_i}\right) - \frac{z_i^2}{2} = \ln \sum_{j\in\mathcal{L}\setminus\{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}} \tag{42}$$

If the cardinality of $\mathcal{L}$, $|\mathcal{L}| > 2$, the right hand side of (42) is a logarithmic sum of exponentials, and (42) has no closed-form solution. Note, however, that (42) can be rewritten as

$$\ln\left(\frac{\pi_i}{\sigma_i}\right) - \frac{z_i^2}{2} - \ln(|\mathcal{L}|-1) = \ln\left(\frac{1}{|\mathcal{L}|-1} \sum_{j\in\mathcal{L}\setminus\{i\}} \frac{\pi_j}{\sigma_j} e^{-\frac{z_j^2}{2}}\right). \tag{43}$$

Then, as a consequence of Jensen's inequality,

$$\ln\left(\frac{\pi_i}{\sigma_i}\right) - \frac{z_i^2}{2} - \ln(|\mathcal{L}|-1) \geq \frac{1}{|\mathcal{L}|-1} \sum_{j\in\mathcal{L}\setminus\{i\}} \ln\left(\frac{\pi_j}{\sigma_j}\right) - \frac{z_j^2}{2}, \tag{44}$$

By approximating (42) using the lower bound in (44), we obtain:

$$\begin{aligned}&\frac{1}{|\mathcal{L}|-1} \sum_{j\in\mathcal{L}\setminus\{i\}} \frac{z_j^2}{2} - \frac{z_i^2}{2} + \ln\left(\frac{\pi_i}{\sigma_i}\right) - \frac{1}{|\mathcal{L}|-1} \sum_{j\in\mathcal{L}\setminus\{i\}} \ln\left(\frac{\pi_j}{\sigma_j}\right) - \ln(|\mathcal{L}|-1) \\ &= 0\end{aligned} \tag{45}$$

which can be simplified to:

$$\begin{aligned}&\left(\frac{1}{|\mathcal{L}|-1} \sum_{j\in\mathcal{L}\setminus\{i\}} \frac{1}{\sigma_j^2} - \frac{1}{\sigma_i^2}\right) t_i^2 + 2\left(\frac{\mu_i}{\sigma_i^2} - \frac{1}{|\mathcal{L}|-1} \sum_{j\in\mathcal{L}\setminus\{i\}} \frac{\mu_j}{\sigma_j^2}\right) t_i - \frac{\mu_i^2}{\sigma_i^2} \\ &+ \frac{1}{|\mathcal{L}|-1} \sum_{j\in\mathcal{L}\setminus\{i\}} \frac{\mu_j^2}{\sigma_j^2} + 2\left[\ln\left(\frac{\pi_i}{\sigma_i}\right) - \frac{1}{|\mathcal{L}|-1} \sum_{j\in\mathcal{L}\setminus\{i\}} \ln\left(\frac{\pi_j}{\sigma_j}\right) - \ln(|\mathcal{L}|-1)\right]\end{aligned} \tag{46}$$

which is a quadratic in t_i. Thus, by solving (46), t_i can be expressed as a function of $\mathbf{w}_i$ through the variables μ_i, μ_j, σ_i and σ_j. Being a quadratic,

there are two solutions to (46). Yet, by choosing the solution that yields the smaller Bayes error, (41) is then expressed solely in terms of $\mathbf{w}_i$, so that $\mathbf{w}_i$ can be solved for iteratively.

At this point, we make note of two computational issues with this procedure. First, (41) is derived from the first-order optimality condition. Therefore, there is no certainty that iteratively solving for $\mathbf{w}_i$ would converge to a local minimum of ϵ_i, as the optimality condition of (30) from which (41) is derived is also satisfied for a local maximum or a saddle point. For this reason, the iterative procedure requires the use of several different initial solutions to improve the chances of convergence to a local minimum. Moreover, there is no guarantee that (46) has any real solution, for any given dataset $\mathcal{D}$.

A better approach, which is the one used in the experimental section of this paper, is to minimise ϵ_i via a gradient descent procedure, as follows:

For $m = 0$ to some maximum number of iterations M:

$$\mathbf{w}_i^{m+1} = \mathbf{w}_i^m - \alpha \frac{\partial \epsilon_i}{\partial \mathbf{w}_i^m} \tag{47}$$

$$t_i^{m+1} = t_i^m - \alpha \frac{\partial \epsilon_i}{\partial t_i^m} \tag{48}$$

starting from an initial choice of $\mathbf{w}_i$ and t_i, where α is the learning rate. Note that the partial derivatives of ϵ_i have already been derived in (33) and (36). Since the Bayes error is known to be non-convex and is characterised by multiple local minima (Anderson & Bahadur, 1962), the gradient descent algorithm may have to be performed using different initial solutions to improve the quality of the local minima to which the algorithm converges.

Though only the optimal $\mathbf{w}_i$ is required to form the columns of $\mathbf{T}$, we observe that the optimal $\mathbf{w}_i$ is tied to the optimal threshold t_i through z_i and z_j as can be seen from (26) and (41), requiring that they both be minimised in the gradient descent procedure.

In all, $(K^2+K-4)/2$ classifiers are constructed in the proposed algorithm for a K-class problem.

4. Application to flow meter diagnostics

In this section, we demonstrate the applicability of LDR, and hence, the proposed algorithm, to flow meter diagnostics.

Flow meters are devices used to measure the volumetric or mass flow rate of a fluid. In the oil and gas industry, flow meters are subject to several problems such as transducer failure, wax deposit, as well as very harsh conditions including extremes in temperature and pressure. These problems affect the performance of the meter and cause the flow rate readings to be erroneous. This incorrect measurement is of great concern. For example, an incorrect measurement indicating a high flow rate may attract high tax liabilities.

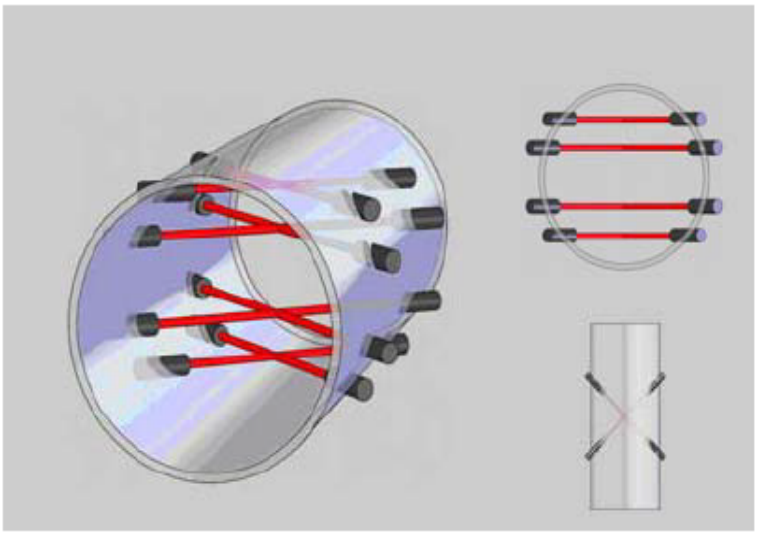
It is understood that after a period of time, the errors associated with the flow measurement may become significant and fall outside the allowable range. Thus, it is the current practice that meters are taken to accredited flow facilities to be recalibrated typically after one year in operation. However, this time-based recalibration has two main drawbacks. First, a given flow meter can encounter a problem such as a transducer failure, even before the one year schedule, and continually provide incorrect measurements until the recalibration period is up. Secondly, a given flow meter may be operating perfectly at the end of the one year period and still be taken in for recalibration, in line with regulatory requirements. However, recalibration of a flow meter can be rather expensive, especially when there is no indication the meter is not healthy. In the UK, for instance, it costs in the region of £30000 for the recalibration of an ultrasonic flow meter (TUV-NEL, 2012).

Thus, the tradeoff between accurate measurement and cutting cost due to frequent recalibration of flow meters (which may be healthy) calls for the adoption of a condition-based flow meter management. With the advent of new flow meters that provide secondary diagnostic information, it becomes possible to use machine learning to diagnose the health state of the flow meter based on the diagnostic parameters, and subsequently enable the real-time estimation of the error associated with every flow measurement. Such an expert system also becomes useful to meter operators in the field who find it difficult interpreting the wealth of diagnostic information available to them and deciding on what the true health state of the meter is. It is believed that if evidence can be provided to the regulatory body that the meter performance is within the allowable range even after the one year schedule, recalibration can be extended (TUV-NEL, 2012).

The task of correctly diagnosing the health state of the flow meter can be formulated as a classification task where it is desired to classify the meter under a number of known health states, with the diagnostic parameters forming the feature vector. For a liquid ultrasonic flow meter (USM), for instance, the most common health states of the measurement system include: waxing,

installation effects and gas injection (TUV-NEL, 2012).

Figure 1: An 8-path ultrasonic flow meter transducer configuration (TUV-NEL, 2012)



Nevertheless, the diagnostic parameters available from a given flow meter can be varied and many, so that the feature vector lives in a rather high-dimensional space. For the USM shown in Fig. 1, the diagnostic parameters include the flow profile, symmetry, crossflow, swirl angle, flow velocity (as measured by each of the eight paths), speed of sound (as measured by each of the eight paths), signal strength (as measured at both ends of each of the eight paths), turbulence (as measured by each of the eight paths), signal quality (as measured at both ends of each of the eight paths), gain (as measured at both ends of each of the eight paths) and transit time (as measured at both ends of each of the eight paths). Thus, the feature vector $\mathbf{x}$ has 92 diagnostic parameters in total.

However, some of these parameters are correlated. For example, the speed of sound, flow velocity and transit time have a known dependence (Vermeulen et al., 2012). Dimensionality reduction is therefore useful to reduce the effects of multicollinearity from the features, before the data is trained for classification.

Furthermore, some of the diagnostic parameters like the swirl angle or turbulence do not contain any classification information with respect to the most prevalent health states of the meter. Besides, it is not known if the diagnostic parameters measured from all eight paths are useful for classifi-

cation, or whether the average of all eight paths would suffice. The effect of having too many nuisance features is that the learning model can overfit the data, especially if the data is noisy. Linear dimensionality reduction alleviates this problem, and if reduction to two or three dimensions is possible, LDR makes visualisation and analysis of the diagnostics data easier for flow meter operators.

5. Experimental validation

5.1. UCI datasets

In this section, we validate the proposed LDR technique experimentally on 10 UCI datasets. The characteristics of these datasets are shown in Table 1.

Table 1: List and characteristics of datasets

Dataset	Label	d	n	K
Diabetes	(a)	8	768	2
Glass	(b)	9	214	6
Cleveland Heart	(c)	13	297	2
Vehicles	(d)	18	846	4
Image Segmentation (Statlog)	(e)	18	2310	7
Ionosphere	(f)	33	351	2
SPECTF Heart	(g)	44	267	2
Zernike Moments	(h)	47	2000	10
Optical Digits	(i)	62	5620	11
United States Postal Service	(j)	256	9298	10

This table lists the datasets used in the experimental section. K is the number of classes, d is the dimensionality of the dataset, and n is the number of data points in the dataset.

We first rescale the predictors to the range $[0, 1]$. We then perform dimensionality reduction using the proposed algorithm. This is followed by 10 independent trials of $10-$fold cross-validation. On each training set, we train two Bayes classifiers, namely QDA and LDA as used by Duin & Loog (2004). We then evaluate the average classification accuracy on the test set using the two classifiers.

For the proposed algorithm, we optimise the Bayes error using the gradient descent procedure. We use a learning rate of $\alpha = 0.1$, and we run 10000

iterations, terminating prematurely when the difference between two consecutive values of the objective function is less than 10^{-6}. We run gradient descent using only one initial solution given by:

$$\begin{aligned} w_i^{(0)} &= \mathbf{S}_L^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_L) \\ t_i^{(0)} &= \log(\tau) + \frac{1}{2}(\bar{\mathbf{x}}_i^\top \mathbf{S}_L^{-1}\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_L^\top \mathbf{S}_L^{-1}\bar{\mathbf{x}}_L) \end{aligned} \tag{49}$$

for every $i \in \mathcal{L}$, where

$$\mathbf{S}_L = \sum_{j \in \mathcal{L}\setminus\{i\}} \pi_i \mathbf{S}_j, \quad \bar{\mathbf{x}}_L = \frac{1}{|\mathcal{L}-1|} \sum_{j \in \mathcal{L}\setminus\{i\}} \bar{\mathbf{x}}_j \quad \text{and} \quad \tau = \frac{\sum_{j \in \mathcal{L}\setminus\{i\}} \pi_j}{\pi_i}. \tag{50}$$

The performance of our algorithm is then compared with the following: PCA, F-LDR, M-LDR, C-LDR, as well as the case where there is no dimensionality reduction and the full dimensionality is used (No-LDR). Note that, for PCA, F-LDR, M-LDR and C-LDR, we take the first $K-1$ independent vectors after the matrix decomposition to form the transformation matrix $\mathbf{T}$, since a dimensionality of $K-1$ is necessary and sufficient for Bayesian classification for the K-class problem (Fukunaga, 2013). The results of these experiments can be seen in Tables 2 and 3. For every test dataset, we perform the Wilcoxon's signed rank test at a significant level of 0.01 to check for any significant differences between the classification accuracy of the best performing algorithm and those of the remaining algorithms. Based on the test results, an asterisk has been indicated against a value if that value is not statistically different from the best value in bold.

To provide a more meaningful perspective on the utility of our algorithm, in Table 4, we also compare the performance of the proposed LDR algorithm followed by the LDA classifier with that of the linear Support Vector Machine (SVM) that uses no dimensionality reduction. The SVM is implemented with the MATLAB function *fitcsvm* using the default settings for a linear SVM.

5.1.1. Results and discussions

Table 2 shows that the proposed algorithm achieves the highest classification accuracy on 8 out of the 10 datasets tested using a QDA classifier, as compared to the remaining LDR procedures. This superior performance is most marked on datasets (e), (g), (h), (i) and (j). On datasets (d), using the full dimensionality results in the best classification accuracy using the LDA classifier, as LDR seems to lose useful classification information. Yet,

Table 2: Average classification accuracy (%) using QDA

Dataset	No-LDR	PCA	F-LDR	M-LDR	C-LDR	Proposed
(a)	74.20	68.57	64.55	77.69	78.09	**78.17**
(b)	55.16	55.25	54.63	**61.89**	54.09	55.21
(c)	82.00	79.56	85.06	84.82	85.22*	**85.49**
(d)	**85.27**	45.78	63.80	78.61	75.97	81.95
(e)	88.82	89.32	90.36	88.09	89.82	**93.56**
(f)	87.51	61.140	89.37	90.43	89.12	**91.12**
(g)	79.42	72.61	79.42	82.96	78.42	**84.24**
(h)	80.14	77.76	79.08	77.47	82.96	**84.93**
(i)	96.44	96.00	96.30	92.55	79.94	**97.69**
(j)	88.09	91.67	88.96	57.70	61.37	**92.94**

This table shows the average classification accuracy (%) on the test datasets using a QDA classifier for No-LDR, PCA, F-LDR, M-LDR, C-LDR and the proposed scheme. Best values are in bold. The values with asterisk (*) are those that are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at a significance level of 0.01.

Table 3: Average classification accuracy (%) using LDA

Dataset	No-LDR	PCA	F-LDR	M-LDR	C-LDR	Proposed
(a)	77.39	68.37	65.11	77.76	77.87	**78.36**
(b)	63.09	60.66	61.08	62.76	59.26	**65.09**
(c)	83.55	79.61	85.29	84.89	85.36*	**85.84**
(d)	78.19	47.18	58.31	75.02	75.65	**79.30**
(e)	**91.48**	83.68	87.83	88.24	88.84	90.33
(f)	86.72	54.50	73.73	90.13*	74.92	**90.62**
(g)	75.27	79.42	79.42	84.77	79.30	**85.55**
(h)	81.79	70.35	74.02	70.78	82.00	**83.92**
(i)	95.32	91.49	93.06	88.86	56.81	**96.09**
(j)	**91.62**	84.29	81.62	27.41	60.96	90.21

This table shows the average classification accuracy (%) on the test datasets using an LDA classifier for No-LDR, PCA, F-LDR, M-LDR, C-LDR and the proposed scheme. Best values are in bold. The values with asterisk (*) are those that are statistically indiscernible from the best values based on the Wilcoxon's signed rank test at a significance level of 0.01.

Table 4: Average classification accuracy (%): Proposed+LDA vs Linear SVM

Dataset	SVM	Proposed+LDA
(a)	76.94	**78.36**
(b)	57.79	**65.09**
(c)	83.45	**85.84**
(d)	74.11	**79.30**
(e)	**92.89**	90.33
(f)	87.73	**90.62**
(g)	79.60	**85.55**
(h)	82.94	**83.92**
(i)	**98.28**	96.09
(j)	**95.81**	90.21

This table shows the average classification accuracy (%) on the test datasets using a linear SVM without dimensionality reduction, and the proposed algorithm plus an LDA classifier. Best values are in bold. The values for both algorithms for all datasets are statistically different based on the Wilcoxon's signed rank test at a significance level of 0.01.

among the five dimensionality reduction techniques, the proposed algorithm achieves the best classification accuracy on those two datasets.

A similar performance is seen in Table 3. The proposed algorithm once again achieves superior classification accuracy on 8 out of the 10 datasets using an LDA classifier, with datasets (a), (b), (d), (g) and (h) showing the most significant performance. Again, while a classification accuracy of 91.62% is achieved on dataset (i) when no LDR is performed, the proposed algorithm achieves the best performance among all the LDR techniques tested on this dataset.

Though, the Mahalanobis distance based-LDR achieves the best performance of 61.89% on dataset (b) using the QDA classifier, we note that our algorithm achieves a superior accuracy in Table 3 on this same dataset using the LDA classifier. This is due to the fact the LDA classifier is more robust to noise, since it tends to not overfit. On the other hand, while the highest classification accuracy for dataset (j) using the LDA classifier is achieved for No-LDR, our proposed algorithm achieves a better accuracy using the QDA classifier. Thus, looking across the two tables, the proposed algorithm achieves the best classification accuracy on 9 out of the 10 datasets using either a QDA or an LDA classifier, with the exception being dataset (d).

Moreover, in Table 4, it is seen that the proposed algorithm easily outperforms the linear SVM on 7 out of the 10 datasets tested.

The poor performance of PCA in Tables 2 and 3 on most of the datasets, e.g., (b), (c), (d), (f), (h) and (i), is due to the fact that PCA reduces the dimensionality of the data without taking into account the class discriminatory information in the data.

As there is no guarantee that the choice of the first $K-1$ independent vectors are those that mostly preserve the classification information in the reduced space for PCA, M-LDR and C-LDR, a reduction to some other dimensionality $q \neq K-1$ might result in a better classification performance. Yet, these algorithms provide the optimal dimensionality q to which to reduce the data. Thus, extensive trial and error is required to obtain an optimal dimensionality in these approaches. Our algorithm, on the other hand, obtains satisfactory classification performance after a reduction to a dimensionality of $K-1$, which is the optimal dimensionality required for Bayesian classification.

We note, however, that the existing LDR algorithms, being deterministic, have faster training times than the proposed algorithm which is iterative. Our algorithm is also sequential, requiring the successive construction of $(K^2+K-4)/2$ classifiers. Thus, the time complexity is quadratic in K. Nevertheless, this is not prohibitive, as the average training time for dataset (i), which has the largest number of classes, with $K=11$ is $4.5s$.

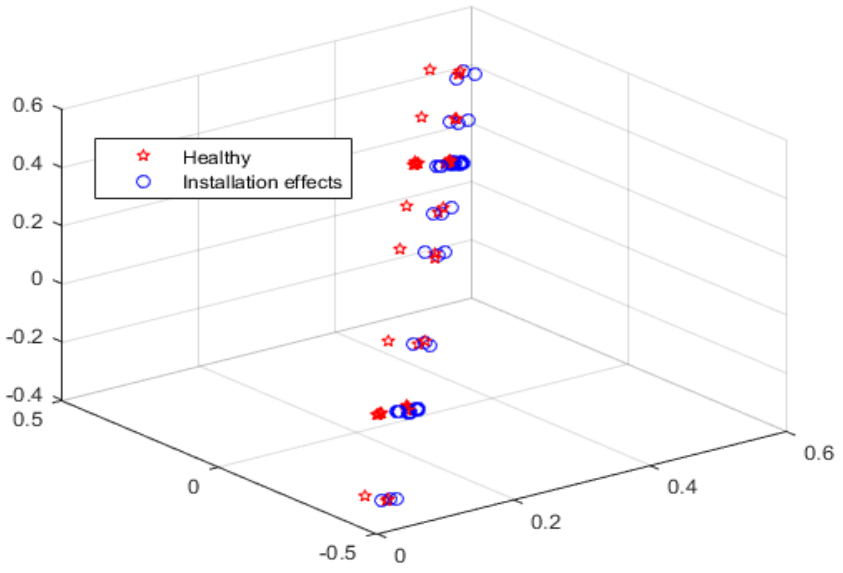
5.2. USM diagnostics datasets

We experimentally validate the proposed LDR technique on 2 different 4-path USMs denoted as Meter A and Meter B.

The two USM diagnostics datasets were obtained from experiments conducted at the National Engineering Laboratories (NEL), UK (TUV-NEL, 2012), and they can be accessed at `http://cogentee.coventry.ac.uk/~kojo/` (Marshall et al., 2012). The datasets have 4 classes or health states, namely: "Healthy", "Waxing", "Gas injection" and "Installation effects", as well as 43 diagnostic parameters. These parameters are given as: profile factor, symmetry, crossflow, flow velocity (in each of the four paths), speed of sound (in each of the four paths), signal strength (at both ends of each of the four paths), signal quality (at both ends of each of the four paths), gain (at both ends of each of the four paths), and transit time (at both ends of each of the four paths). Also, the number of examples n for Meter A is 181, while Meter B has 180 examples.

Since the different diagnostic parameters take on different ranges of values, we first normalise all features to the range $[0, 1]$. We then apply the proposed LDR technique, reducing the dimensionality of the data to $K - 1$ (where $K = 4$ in this case). In Figs. 2 to 6, we show the representation of the original diagnostics dataset for Meter A after linearly reducing its dimensionality to 3 using the proposed algorithm, PCA, F-LDR, M-LDR and C-LDR. For the sake of clarity, we show in these figures only the two classes which are particularly difficult to separate in the original feature space: "Healthy" and "Installation effects".

Figure 2: LDR performance on Meter A diagnostics data: PCA



Next, we perform 10 trials of 10−fold cross validation. For each of the training set, we train a QDA classifier, which is a special case of a Bayes classifier when the data is normally distributed in each of the classes. The classification accuracy on the test set is then evaluated using the QDA classifier.

The performance of our algorithm is then compared with the following: PCA, F-LDR, M-LDR, C-LDR, as well as the case where there is no dimensionality reduction and the full dimensionality is used (No-LDR). The same parameters used in the proposed algorithm for the experiments on the UCI

Figure 3: LDR performance on Meter A diagnostics data: F-LDR

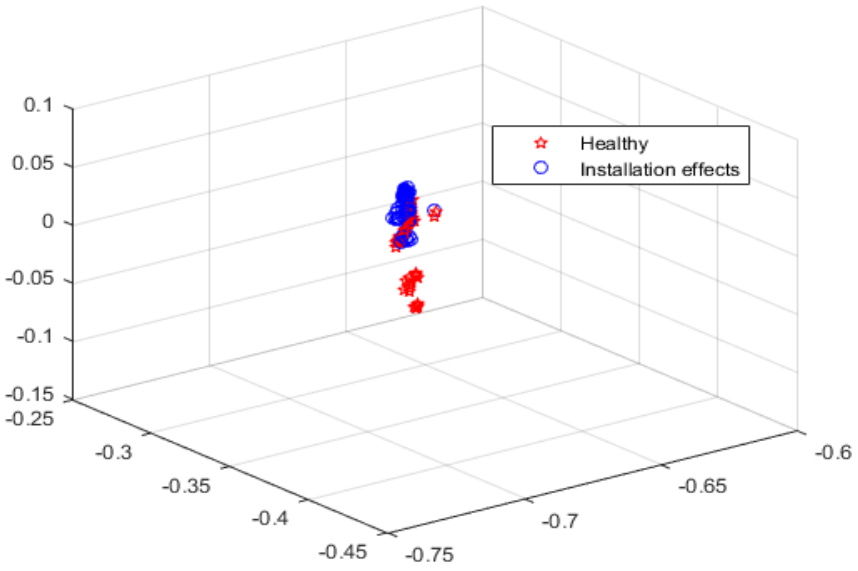


Figure 4: LDR performance on Meter A diagnostics data: M-LDR

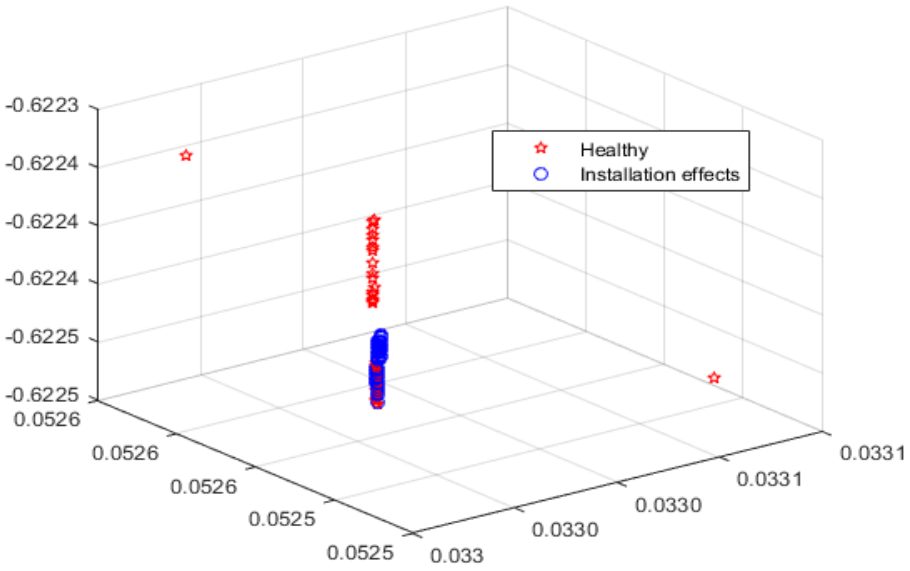


Figure 5: LDR performance on Meter A diagnostics data: C-LDR

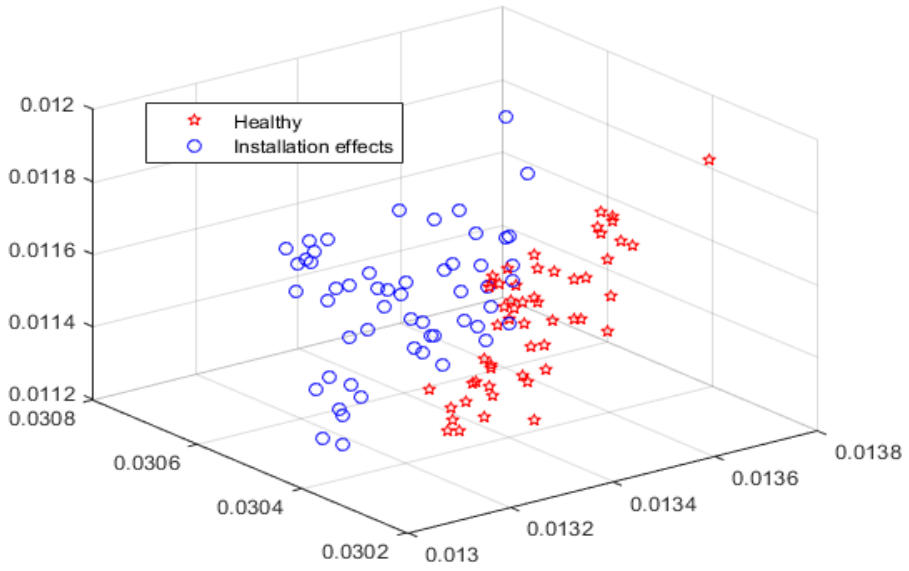
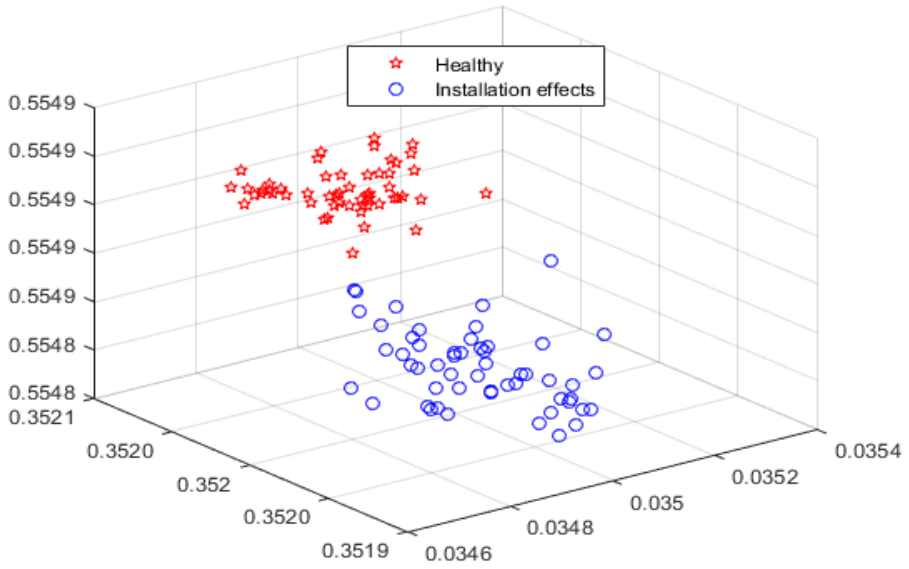
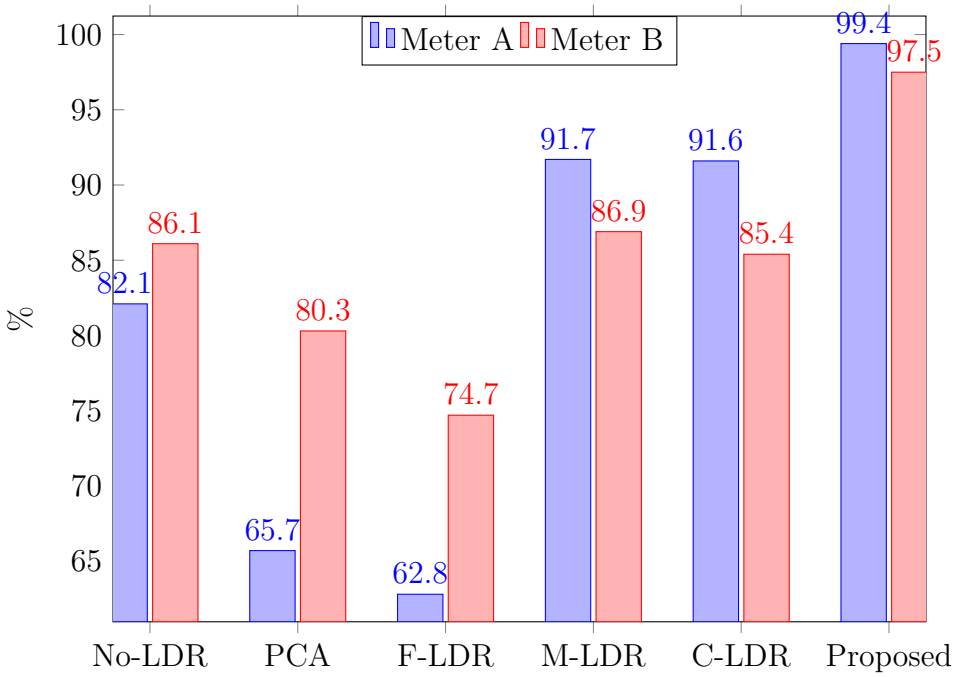


Figure 6: LDR performance on Meter A diagnostics data: Proposed



datasets are used in this experiment too. The results of these experiments can be seen in Fig. 7.

Figure 7: Average classification accuracy for Meter A and Meter B (%)



The classification accuracies for the proposed algorithm are statistically different from the remaining algorithms at the 0.01 confidence level based on the Wilcoxon's signed rank test.

5.2.1. Results and discussion

From Fig. 7, the classification performance when no LDR is performed prior to applying the QDA classifier stands at 82.1% for Meter A and 86.1% for Meter B. PCA and Fisher's LDR, however, result in reduced classification accuracies. This implies that these two algorithms tend to lose classification information when linearly reducing the dimensionality of the data to $K-1$. The Mahalanobis distance-based LDR, the Chernoff criterion-based LDR and the proposed algorithm, on the other hand, achieve an improved classification accuracy over the case where the full dimensionality is used. These results also indicate that Meter A has better diagnostic capabilities than Meter B.

Moreover, Fig. 7 shows that the proposed algorithm achieves the highest classification accuracy: 99.4% on Meter A and 97.5% on Meter B. This is followed by the Mahalanobis distance based LDR at 91.6% and 86.9% respectively for Meter A and Meter B. The reasons for the relative performance of the various LDR algorithms on the flow meter datasets in Fig. 7 are as given by the analysis of the results of the experiments on the UCI datasets in Section 4.1.1.

6. Conclusion

In this paper, we have presented a novel technique for supervised linear dimensionality reduction (LDR). The proposed algorithm projects the original data onto a $(K-1)$-dimensional subspace, where K is the number of classes. This is done by sequentially creating linear classifiers that separate one class from the remaining classes under a normality assumption, while minimising the Bayes error through a gradient descent procedure.

The proposed algorithm is applicable to expert and intelligent systems that require LDR to overcome overfitting in predictive classification models so that prediction accuracies may subsequently be improved. Most of the existing LDR procedures provide no optimal dimensionality to which a given dataset may be reduced, and they tend to lose class-discriminatory information in the optimal $(K-1)$-dimensional subspace required for Bayesian classification. In contrast, the proposed algorithm provides an optimal reduction to a dimensionality of $K-1$ via a sequential minimisation of the Bayes error, thus guaranteeing a much better classification accuracy than the existing approaches using a Bayes classifier such as LDA or QDA.

To demonstrate this, we have validated the proposed algorithm experimentally on 10 UCI datasets that cut across a wide range of application areas including medical diagnosis, handwriting recognition and object detection. The application of the proposed algorithm to flow meter diagnostics has also been discussed. This is followed by employing the proposed algorithm in the diagnosis of the health state of two ultrasonic flow meters, achieving classification accuracies of 99.4% and 97.5%. On both the ultrasonic flow meter datasets and the UCI datasets, our algorithm is shown to achieve superior performance in terms of the classification accuracy, as compared to the existing linear dimensionality reduction techniques.

While the proposed algorithm has been shown to be superior to the existing procedures on the datasets tested in terms of classification accuracies,

it is built on an assumption of normality of the data in each of the K-classes. Yet, since a lot of physical data tend to be nearly-normally distributed (Lyon, 2013), our algorithm is well suited for a lot of applications particularly those involving measurement errors such as machine fault diagnosis or those involving physical measurements such as accelerometer-based human activity recognition. However, for data that are radically non-normal, our procedure is expected to perform relatively poorly, as the Bayes error is not guaranteed to be minimised. Also, while the proposed LDR procedure has been derived for Bayesian classification and is thus expected to perform well on Bayesian classifiers such as LDA, QDA and the Naive Bayes classifier, it is not suitable for other discriminative classifiers such as the SVM or logistic regression. Moreover, our algorithm requires the construction of $(K^2 + K - 4)/2$ classifiers which can be rather computationally costly for a dataset having too many classes.

In view of the above problems, our future work is concerned with the violation of the assumption of normality employed in the proposed algorithm. This would make the procedure more robust and applicable to a wider range of problems. While the Bayes error can be analytically intractable for an arbitrary non-normal distribution, we aim to extend the proposed procedure to minimising some upper bounds on the Bayes error for a given dataset. Alternatively, future research is aimed at deriving a kernel function that implicitly transforms some data of a known non-normal distribution into a feature space where the data in each class is nearly normally distributed. Moreover, we hope to explore the use of information theoretic measures to reduce the total number of classifiers constructed in each step of the proposed algorithm. This would decrease the computational complexity of the algorithm and improve its speed. Finally, as an application to flow meter diagnostics, our future work is focused on leveraging the correct diagnosis of a flow meter in the estimation of the error associated with each flow measurement with reasonable accuracy. With knowledge of the true health state of a flow meter, the associated measurement errors can be estimated with improved accuracy. This will allow erroneous flow measurements to be self-validated, thus resulting in significant cost cuts due to incorrect flow measurements in oil and gas operations.

References

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

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

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A. F., Wilson, J. F., Agakov, F., Navarro, P. et al. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific reports*, *5*.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is nearest neighbor meaningful? In *International conference on database theory* (pp. 217–235). Springer.

Brunzell, H., & Eriksson, J. (2000). Feature reduction for classification of multidimensional data. *Pattern Recognition*, *33*, 1741–1748.

Buturovic, L. J. (1994). Toward Bayes-optimal linear dimension reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*, 420–424.

Chen, L.-F., Liao, H.-Y. M., Ko, M.-T., Lin, J.-C., & Yu, G.-J. (2000). A new LDA-based face recognition system which can solve the small sample size problem. *Pattern recognition*, *33*, 1713–1726.

Coomans, D., Jonckheer, M., Massart, D. L., Broeckaert, I., & Blockx, P. (1978). The application of linear discriminant analysis in the diagnosis of thyroid diseases. *Analytica chimica acta*, *103*, 409–415.

Cunningham, J. P., & Ghahramani, Z. (2015). Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, *16*, 2859–2900.

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

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

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

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.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*, 179–188.

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

Gyamfi, K. S., Brusey, J., Hunt, A., & Gaura, E. (2017). Linear classifier design under heteroscedasticity in linear discriminant analysis. *Expert Systems with Applications*, *79*, 44–52.

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).

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* volume 6. Springer.

Liu, J., Chen, S., Tan, X., & Zhang, D. (2007). Efficient pseudoinverse linear discriminant analysis and its nonlinear form for face recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, *21*, 1265–1278.

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.

Lu, J., Plataniotis, K. N., & Venetsanopoulos, A. N. (2003). Regularized discriminant analysis for the small sample size problem in face recognition. *Pattern Recognition Letters*, *24*, 3079–3087.

Lyon, A. (2013). Why are normal distributions normal? *The British Journal for the Philosophy of Science*, *65*, 621–649.

Mahmoudi, N., & Duman, E. (2015). Detecting credit card fraud by modified Fisher discriminant analysis. *Expert Systems with Applications*, *42*, 2510–2516.

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

Marshall, C., Mills, C., & Gyamfi, K. S. (2012). *USM diagnostic dataset, 2012*. URL `http://cogentee.coventry.ac.uk/~kojo/`.

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.

Nielsen, F. (2014). Generalized bhattacharyya and chernoff upper bounds on bayes error using quasi-arithmetic means. *Pattern Recognition Letters*, *42*, 25–34.

Polat, K., Güneş, S., & Arslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert systems with applications*, *34*, 482–487.

Sengur, A. (2008). An expert system based on linear discriminant analysis and adaptive neuro-fuzzy inference system to diagnosis heart valve diseases. *Expert Systems with Applications*, *35*, 214–222.

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

Sharma, A., & Paliwal, K. K. (2015). Linear discriminant analysis for the small sample size problem: an overview. *International Journal of Machine Learning and Cybernetics*, *6*, 443–454.

Song, F., Zhang, D., Wang, J., Liu, H., & Tao, Q. (2007). A parameterized direct LDA and its application to face recognition. *Neurocomputing*, *71*, 191–196.

TUV-NEL (2012). *Testing the diagnostic capabilities of liquid ultrasonic flow meters*. National Measurement System.

Vermeulen, M. J. M., Drenthen, J. G., & Hollander, d. H. (2012). *Understanding diagnostic and expert systems in ultrasonic flow meters*. KROHNE Oil and Gas, CT Products.

Yu, H., & Yang, J. (2001). A direct LDA algorithm for high-dimensional datawith 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 *Proceedings of the 21st international jont conference on Artifical intelligence* (pp. 1363–1368). Morgan Kaufmann Publishers Inc.