Abstract—Gaussian classifiers make the modelling assumption that each class follows a multivariate Gaussian distribution, and include the Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). In high dimensional low sample size settings, a general drawback of these methods is that the sample covariance matrix may be singular, so their inverse is not available. Among many regularisation methods that attempt to get around this problem, the random projection ensemble approach for LDA has been both practically successful and theoretically well justified. The aim of this work is to investigate random projection ensembles for QDA, and to experimentally determine the relative strengths of LDA vs QDA in this context. We identify two different ways to construct averaging ensembles for QDA. The first is to aggregate low dimensional matrices to construct a smoothed estimate of the inverse covariance, which is then plugged into QDA. The second approach is to aggregate the predictions of low dimensional QDAs instead. By implementing experiments in eight different kinds of data sets, we conduct a thorough evaluation. From it we conclude that both QDA ensembles are able to improve on vanilla QDA, however the LDA ensemble tends to be more stable with respect to the choice of target dimension of the random projections.

Keywords—high dimensional data, random projection, Gaussian classifier, linear discriminant analysis, quadratic discriminant analysis

I. INTRODUCTION

Gaussian classifiers model each class as a multivariate Gaussian. Their simplest version, known as the Linear Discriminant Analysis (LDA) or Fisher’s Discriminant Analysis assumes that each class has the same shape. That is, the class covariance matrix is shared among all classes. As its name indicates, it produces a linear decision boundary. A more flexible model is the Quadratic Discriminant Analysis, which allows each class to have its own covariance shape. This results in a quadratic decision boundary.

The estimation of high dimensional multivariate Gaussians is prone to the curse of dimensionality. In the absence of a sufficiently large sample size, the covariance estimates become ill-conditioned. Dimensionality reduction is one popular approach to tackle this problem, and Random Projection (RP) is a computationally attractive candidate. LDA with RP is guaranteed to retain good predictive performance under mild conditions, see e.g. [1], [2] and references therein. Similarly, QDA with RP has good theoretical guarantees [3]. Furthermore, an averaging ensemble of RP-ed LDA classifiers was shown to enjoy state-of-the art performance [4]. It is then natural to wonder if this idea could be extended to devise QDA ensembles?

This work is motivated by the above question. We investigate random projection ensembles for QDA, and conduct experiments to assess its potential relative to the analogous LDA ensemble. The main idea is to use the random projections to reduce the dimensionality of the covariance matrix, so the smaller dimensional covariance matrix will not be singular.

We identify two different ways to construct averaging ensembles for QDA, which will be referred to as $\Sigma_{RP}$-QDA and $RP_{Ensemble}$-QDA respectively. The $\Sigma_{RP}$-QDA approach will be to approximate the inverse of the covariance matrix with an ensemble of the inverses of its random projections. Then, when applying classification, the new approximate inverse covariance matrix will be used. Our second approach, $RP_{Ensemble}$-QDA, is to reduce the dimension of the data set multiple times by independent random projections, feed each to the classifier, and combine the predictions. In the case of LDA, these two approaches both coincide with the existing RP-Ensemble-LDA of [4].

We implement our proposed methods and devise experiments in eight data sets of differing characteristics to conduct a thorough evaluation. By deploying experimentation and subsequent analysis of the results, the following questions will be addressed: Are any of the two QDA ensembles able to improve on vanilla QDA? Of the two QDA ensemble methods, $\Sigma_{RP}$-QDA and RP Ensemble, which one is better? Of the LDA and QDA ensembles, which one performs better?

The next section will give some background about Gaussian classifiers and Random Projections, then related work will be discussed. In Section III, the details of our QDA ensemble approaches will be presented. In Section IV, our experiment design, results, analysis, and discussions will be provided. Finally, we conclude and discuss directions for future work.

II. BACKGROUND AND RELATED WORK

A. Supervised Learning

Supervised learning is a branch of machine learning, concerned with establishing a rule from a labelled data set to deal with new data points. In supervised learning, there is a
teacher or supervisor to guide us to categorise the training set into different labels or calculate the cost of all the classes. The aim of the learner is to decrease the cost of the category [5, p. 16]. We call this learning method supervised learning because the existing information or results will point out the way for the learning process [6, p. 2].

B. Gaussian Classifier

The rules used to implement classification are called classifiers. Gaussian classifiers represent one approach, which we will focus on in this paper.

A Gaussian classifier is based on the data set’s density function, posterior probability, and maximum likelihood [7, p.20 21]. Each class has one model to represent it, that is each class has one discriminant function. The new point \( x \) will be classified into the class that has the largest value among all the discriminant function values about \( x \) [6, p. 101]. The Gaussian classifier assumes that for each pre-defined class \( t \), the data points are independent and follow a multivariate Gaussian distribution. When there is a new point \( x \) to be classified, calculating its posterior probability for all the classes using density functions, and selecting the largest value yields the prediction.

Let \( x = (x_1, x_2, \cdots , x_d) \) be the vector of attributes for a point \( x \). So \( x \) has \( d \) attributes. Suppose this data set has \( T \) pre-defined classes. The posterior function for class \( t \) is:

\[
p(x|h = t) = \frac{1}{(2\pi)^{d/2} |\Sigma_t|} \exp \left( -\frac{1}{2} (x - m_t)^T \Sigma_t^{-1} (x - m_t) \right)
\]

\( p(x|h = t) \) is the density function of class \( t \), it means the likelihood that the distribution of class \( t \) could have generated the point \( x \). \( \Sigma_t \) is the covariance matrix of this class. \( m_t \) is the mean vector that is the center of the class. The density function of each class is represented by its covariance matrix and mean vector – that is using the mean and covariance to represent the Gaussian distribution [6, p.125 126].

Estimation of parameters from data via maximising the likelihood amounts to setting \( m_t \) to the average of points in class \( t \), and \( \Sigma_t \) to the sample covariance of class \( t \). As we already mentioned the latter is problematic in high dimensions as its inverse may not exist. Instead, in Section III we will propose some novel approaches to get around this problem, which we will use later in our experiments.

To make classification with this model, for each point \( x \) we calculate \( p(h = t|x) \), which we call the posterior probability. With class prior probabilities \( p(h = t) \), the class posterior is computed by Bayes rule \( p(h = t|x) = p(x|h = t)p(h = t)/p(x) \), and the class that maximises this value is predicted.

The Gaussian classifier has some major advantages in terms of interpretability. From the covariance matrix we can learn about the correlation structure of the attributes or features in the data set. It captures relationships between the features. Furthermore, the entries of the inverse covariance are conditional correlations between the features.

According to whether the class-covariances coincide or not, the Gaussian classifiers can be divided in two categories, one is called Linear Discriminant Analysis (LDA), and the other one is called Quadratic Discriminant analysis (QDA).

1) Linear Discriminant Analysis: Linear Discriminant Analysis (LDA) assumes that the classes follow multivariate Gaussian distributions with different means but all have the same covariance matrix [7, p.22 23]. That means, in the density function Equation (1), the value of \( \Sigma_t = \Sigma, \forall t = 1, ..., T \). We call it shared covariance. The equation of shared covariance is:

\[
\Sigma = \frac{1}{N} \sum_{p=1}^{N}(x_p - m_{class(x_p)})(x_p - m_{class(x_p)})^T
\]

In Equation (2), \( N \) is the number of points, \( x_p \) where \( p \in \{1, ..., N\} \) are the points in the data set. All classes share one covariance matrix – that means, the shape of all classes is the same.

2) Quadratic Discriminant Analysis: Quadratic Discriminant Analysis (QDA) only assumes that each class obeys the Gaussian distribution with its own parameter settings. So, each class has its own covariance matrix, and the shapes of the classes can be different.

Figure 1 shows the differences between Quadratic Discriminant Analysis and Linear Discriminant Analysis. From the diagram we can see, for Linear Discriminant Analysis, the shapes of all the classes stay the same and they can be separated by a straight line. But for Quadratic Discriminant Analysis, the shapes of different classes are different, and the separating boundary is a quadratic curve. Quadratic Discriminant Analysis is more powerful than Linear Discriminant Analysis, for Quadratic Discriminant Analysis does not require all the classes have identical covariance.

C. Random Projections

A theorem by Johnson and Lindenstrauss [8] indicates that a set of points in \( d \) dimensions can be mapped into a \( k \) dimensional space, where \( k \) is smaller than \( d \), such that the distortion of the distances between points are under some bounds. Dasgupta and Gupta [9] give a simplified proof of this theorem, showing that a linear map with standard Gaussian entries can be used. This gives a method to reduce the high dimensional data into lower dimension without too much distortion in the Euclidean space. In this paper we use Gaussians random projections, that is, a dimensionally reducing random matrix with i.i.d. standard Gaussian entries.
Random projection (RP) consists of pre-multiplying the data points with a $k \times d, k < d$ matrix having entries drawn i.i.d. from a standard Gaussian. Since $k < d$, this reduces dimension:

$$Y_{k \times n} = R_{k \times d} \cdot X_{d \times n}$$

(3)

where $X_{d \times n}$ is the original data matrix, which has $n$ points and $d$ attributes, $R_{k \times d}$ is the random projection matrix where $k$ is smaller than $d$. After the projection, the matrix $Y_{k \times n}$ is the new matrix with $n$ points in $k$ dimensions.

From Equation (3) we can see that the computational cost of Random Projection is relatively low: $O(dkn)$ [10]. More importantly, it can preserve the distances between points to a good approximation. Achlioptas [11] also gives computationally cheaper variants.

D. Related Work

A number of previous works have exploited random projections (RP) to deal with high dimensional problems.

Bingham and Mannila [12] conduct experiments on images and text data, showing the measured distortions of pointwise distances can be much lower in practice than the theory predicts.

Dasgupta [13] used RP combined with maximum-likelihood estimation by the EM algorithm, and found improvements compared with regular maximum-likelihood in handwritten digit recognition. The main advantage of the classifier is that it performs quickly.

Deegalla and Boström [10] implement experiments to compare Principal Component Analysis (PCA) against Random Projection for the nearest neighbour classifier. They apply these methods to classify images and micro arrays. Their results indicate that PCA performs better than RP but RP is faster and easier to compute. These results agree with the experimental study of Fradkin and Madigan [14]. Goel et al. [15] present experiments on face recognition, again using PCA and RP to reduce dimension. Their results conclude comparable performance despite the speed-up of RP.

Lin and Gunopulos [16] and Lagus et al. [17] have studied the performance of Latent Semantic Indexing while reducing the dimension of data by RP, and found this combination very effective for image retrieval.

Fern and Brodley [18] combine RP with clustering, and construct ensembles that improve the stability of clustering.

Liu et al. [19] apply multiplicative random projection matrices for privacy-preserving data mining. They find this can preserve the statistical properties of the data while also offering privacy-preservation. They also conclude that this method can be deployed in many different types of data mining tasks, like inner product or Euclidean distance estimation, correlation matrix computation, clustering, outlier detection, linear classification.

Several works considered Gaussian classifiers with small sample size, and devised methods to deal with the ill-conditioning of covariance estimates. Among these, Yau and Manry [20] map the Gaussian classifier to a specific type of neural network, Guo [21] devise a regularisation for LDA. Durrant and Kabán [4] use the Random Projection ensemble while doing classification. They focus on the Fisher Linear Discriminant classifier, give some theory, and state of the art experiment results. It is therefore very tempting to extend this further for QDA.

III. METHODS

The approach of Durrant and Kabán [4] is to average an ensemble of LDA classifiers, each fed by an independently RP-ed version of the data. Therefore the component covariance matrices are all low-dimensional, but as the paper proves, the ensemble can be viewed as a single LDA classifier having an aggregate covariance of the form studied in Marzetta et al. [22]. Here we extend the idea to include Quadratic Discriminant Analysis. However, we find that, for QDA the two views mentioned above lead to two different methods. We now describe these.

A. Method 1

Marzetta et al. [22] gives a method to deal with singular covariance matrices based on random dimensionality reduction. The main idea is, first project the covariance to lower dimension several times using an independent random matrices which will make these non-singular in the lower dimensional space, so they can be inverted. Then project these back to inverse matrices into the high dimensional space and average them. As the number of terms in the average increases, the resulting aggregate high dimensional matrix will no longer be singular, while it will have the same eigenvectors as the original. For details see [4], [22].

Applying this to the sample covariance of the $t$-th class in QDA, we get the aggregated inverse covariance:

$$invcov_k(\Sigma_t) = E(R^T(R\Sigma_tR^T)^{-1}R)$$

(4)

Since we have finitely many terms in the average, in practice we work with the following instead:

$$\Sigma_{t,RP}^{-1} = \frac{1}{M} \sum_{i=1}^{M} R_i^T(\Sigma_t R_i R_i^T)^{-1} R_i$$

(5)

where $M$ is the number of terms in the average, that is the size of the ensemble.

This matrix $\Sigma_{t,RP}^{-1}$ will be plugged in the place of the required inverse covariance in the Gaussian classifiers density function. The mean vector estimation remains the same as that of the original QDA. We call this method $\Sigma_{RP}$-QDA. We also know that $M$ needs to be of order $O(d)$ for the finite average in eq. (5) to approach the expectation in eq. (4) [23].

B. Method 2

Our second method is to use Random Projections to project the data set into $k < d$ dimensions $M$ times in parallel, use a classifier on each, and combine the outputs of these classifiers using majority voting. The whole idea of this method is that, after the reduction of dimensionality, the $k \times k$
covariance estimates will not be singular, so the individual QDA classifiers should have no problems.

The difference between the methods used in this paper with [10], [14] and [3] is that we deploy the random dimensionality reduction and classification process \( M \) times, whereas these previous studies used only one random projection. We call this method the RP-Ensemble-QDA.

IV. EXPERIMENT DESIGN

A. Data Sets

Eight data sets were chosen for the forthcoming experiments. Table I shows the characteristics of the data sets used. Five of them – Ionosphere, Colon, Leukemia, Spam and Ads – are chosen from Fradkin and Madigans paper [14], who have been among the first to conduct a detailed experimental study on classification of randomly projected data. The data sets MIAS and COIL100 are chosen from Deegalla and Boströms paper [10], where these data sets were used to compare dimensionality reduction methods specifically for nearest neighbour classification. Finally, the Cancer data set is chosen from Guillemot et al. s work [24], which was used in a computational study on functional genomics. Taken together, these data sets represent different types of data: images, text documents, and micro arrays or gene expressions.

<table>
<thead>
<tr>
<th>Date Sets</th>
<th># Train</th>
<th>Attributes</th>
<th>Classes</th>
<th># Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad</td>
<td>3279</td>
<td>1554</td>
<td>2</td>
<td>1079</td>
<td>Internet Ads</td>
</tr>
<tr>
<td>Cancer</td>
<td>102</td>
<td>282</td>
<td>2</td>
<td>25</td>
<td>Micro Array</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>351</td>
<td>34</td>
<td>2</td>
<td>51</td>
<td>Radar Data</td>
</tr>
<tr>
<td>MIAS</td>
<td>322</td>
<td>1024</td>
<td>7</td>
<td>64</td>
<td>Images</td>
</tr>
<tr>
<td>Spam</td>
<td>4601</td>
<td>57</td>
<td>2</td>
<td>1601</td>
<td>Emails</td>
</tr>
<tr>
<td>Leukemia</td>
<td>72</td>
<td>3571</td>
<td>3</td>
<td>12</td>
<td>Micro Array</td>
</tr>
<tr>
<td>Colon</td>
<td>62</td>
<td>2000</td>
<td>2</td>
<td>12</td>
<td>Micro Array</td>
</tr>
<tr>
<td>COIL100</td>
<td>7200</td>
<td>1024</td>
<td>100</td>
<td>1440</td>
<td>Images of Objects</td>
</tr>
</tbody>
</table>

From Table I we can see that, some of these chosen data sets have more instances than attributes, others have more attributes than instances. Some are two class problems, others multi or many class problems. In this way, we have a variety of data characteristics in our test suite.

All data sets are standardised to zero mean and unit variance for all features for the reported results. Though we also experimented with the raw data, and interestingly found very little variation.

We implemented the methods described in MatLab, making use of the in-built function classify with appropriate options for LDA and QDA model estimation and prediction.

B. Experiments

The purpose of our experiments is to find out the following.

- Will our two methods of random projection ensemble improve on Quadratic Discriminant Analysis? We expect this to be the case unless a massive sample size would be available.
- To what extent our two RP-aggregation methods differ in terms of behaviour and performance, and is any of them better than the other?
- How do the QDA ensemble variants compare with the simpler and more restrictive LDA ensemble? We find this question interesting, since QDA allows for different shapes in each cluster whereas LDA assumes the same shape. This would suggest an advantage for the former. On the other hand, there is a price to pay for the added flexibility as QDA will need more data to estimate those class-specific shapes. In addition to this trade-off, there is the effect that random projection to low dimension makes covariance shapes less elongated. It was shown theoretically in previous work [2] that this diminishes the ill-effect of covariance mis-specification in LDA. It will be interesting to see from our experiments the joint effect of these different tendencies in practice.

In order to answer the questions raised above, we implemented QDA with \( \Sigma_{RP} \)-smoothing (\( \Sigma_{RP} \)-QDA), QDA with RP-Ensemble, vanilla QDA, that is using and pseudo-inverting the sample estimates of the class-covariance matrices; we also implemented a multi-class extension of the previously proposed RP-Ensemble LDA. By construction, the latter is equivalent to the \( \Sigma_{RP} \)-smoothing of LDA (while for QDA they are different).

The evaluation criterion of the behaviour of these various approaches are the out-of-sample accuracy of the classification results. Each data set is divided into two disjoint subsets: a training set, and a testing set. The training set is used to estimate the parameters (class means and covariances). These are then used to calculate the posterior probability of each class for points of the test set. The predicted class is the one with higher posterior probability. Finally the percentages of the points that are correctly classified when compared with their true labels represents the accuracy measurement for the classifier under evaluation.

For each data set, the test set size is fixed for the split during experiments. We will vary the target dimension (\( k \)) of the random projection. In practice the target dimension is a parameter that can be set for instance by cross-validation. Though here we are interested to learn about the behaviour of the above approaches comparatively. We note that the range of target dimensions we consider for Linear Discriminant Analysis based and Quadratic Discriminant Analysis based methods are different, in order to make sure the covariance is not singular. For Linear Discriminant Analysis, the values of \( k \) cannot be greater than the number of points minus one. For Quadratic Discriminant Analysis, the values of \( k \) cannot be greater than the smallest number of points among all the classes minus one. This is because, if \( \Sigma \) denotes a \( d \)-dimensional sample estimate of a covariance, estimated from \( N \) points, then:

\[
\text{rank}(\Sigma) = \begin{cases} 
  d, & d < N \\
  N - 1, & d \geq N
\end{cases}
\] (6)

In Quadratic Discriminant Analysis we need to invert the
covariance matrix for each class, so the rank of this matrix must equal the working dimension, namely \( k \) – hence the number of points in each class must be at least \( k + 1 \) (\( k \) is the target dimension). The target dimension is increased up to \( d \). If \( d \) is large, then \( k \) will start from 5 and increase in steps of 5 or 10. If \( d \) is small, then \( k \) will start from 1 and increase in steps of 2 or 3.

We ran some preliminary experiments to find out the effects of the ensemble size, \( M \), and observed the larger the \( M \) the better. We also noted that convergence tends to be quick at the start, so we chose the ensemble size of \( M = 100 \) for the reported experiments.

We run each approach at least 25 times (RP Ensemble was run for 100 times). Each time, the testing set size and training set size is the same, but the points are randomly shuffled among training and testing sets. In order to avoid potential bias from the random projection, the random matrices were generated independently anew for each of the repetitions.

V. RESULTS AND ANALYSIS

We first present a snapshot of evaluation results in Table II, followed by a more comprehensive analysis in Figure 3. Table II presents the two RP-aggregation methods for Quadratic Discriminant Analysis when the best \( k \) is chosen, against the original QDA. From Table II we can see the following.

- The RP-aggregation approaches increase the accuracy of QDA classification except for one data set. Out of the data sets tested, the MIAS data set is the only one where the original QDA wins – the reasons for this will be discussed later in this section.
- The data space results can be very bad due to ill conditioning of the covariance estimates.
- The two aggregation approaches for QDA that we devised are indeed different in terms of behaviour. For instance, on Ionosphere the RP-Ensemble QDA significantly outperforms the covariance-aggregation \( \Sigma_{RP} \)-QDA, whereas on COIL100, \( \Sigma_{RP} \)-QDA significantly outperforms RP-Ensemble QDA. We have not been able to find out what makes these differences, this seems like a hard question and may require theoretical analysis which was outside of the scope of this work. We conjecture that RP-Ensemble QDA has a better ability to take advantage of diversity within the ensemble.

Figure 3 presents the results of our experiments. The data sets are arranged according to their properties, that is the relationship between number of points and number of features. The data sets which have more points than features are arranged on the left column, and the data sets in the opposite situation are arranged on the right column. These diagrams can provide answers to the question about which discriminant analysis performs better (linear or quadratic), and which aggregation method works better for QDA (\( \Sigma_{RP} \) or RP Ensemble).

In the experiments in Figure 3 we vary the projection dimension \( k \), and display the accuracy curves comparatively for our RP-Ensemble QDA and \( \Sigma_{RP} \)-QDA versus QDA. In addition we also show superimposed on these figures the performances of Fisher Linear Discriminant Analysis (LDA) and RP-Ensemble LDA. Note, for the LDA model, this is the same as \( \Sigma_{RP} \)-LDA would be. The error bars span one standard error about the mean, and these are most of the time so small that are hardly visible on the plots.

From Figure 3, we can see the performances of \( \Sigma_{RP} \)-QDA, seen as red curves in the diagrams, are first rising and then dropping. That is because the Quadratic Discriminant Analysis has many covariance matrices, so more points need to be used when the target dimension \( k \) increases. There are not enough points with the increasing dimension, so the accuracy will drop. The trends of RP-Ensemble QDA have similar behaviour. We can conclude from the figures that the good range of \( k \) for both QDA ensemble variants tends to be quite narrow.

The trends of performances of the RP-ensemble-LDA, as dimension \( k \) increases, present two types of behaviour, which we can see from the diagrams. The first one is the case of data sets with more points than features. That is the figures on the left column. In these cases we see a steady rise. That is because the Linear Discriminant Analysis has only one covariance matrix, so the data sets that have more points than features can provide enough points to calculate the covariance.

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>Original, QDA (%)</th>
<th>( \Sigma_{RP} )-QDA</th>
<th>RP Ensemble QDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad</td>
<td>0.00±0.00</td>
<td>95.80±0.01 (k=50)</td>
<td>95.12±0.01 (k=50)</td>
</tr>
<tr>
<td>Cancer</td>
<td>49.76±0.11</td>
<td>78.08±0.02 (k=12)</td>
<td>80.32±0.01 (k=10)</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>0.00±0.00</td>
<td>62.20±0.06 (k=25)</td>
<td>91.94±0.04 (k=5)</td>
</tr>
<tr>
<td>MIAS</td>
<td>59.06±0.18</td>
<td>56.13±0.06 (k=5)</td>
<td>55.22±0.05 (k=12)</td>
</tr>
<tr>
<td>Spam</td>
<td>73.56±0.15</td>
<td>90.31±0.01 (k=50)</td>
<td>77.65±0.01 (k=30)</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.00±0.00</td>
<td>70.00±0.15 (k=5)</td>
<td>89.58±0.09 (k=4)</td>
</tr>
<tr>
<td>Colon</td>
<td>34.33±0.30</td>
<td>61.33±0.12 (k=1)</td>
<td>63.75±0.12 (k=10)</td>
</tr>
<tr>
<td>COIL100</td>
<td>0.56±0.01</td>
<td>90.93±0.12 (k=40)</td>
<td>6.02±0.08 (k=20)</td>
</tr>
</tbody>
</table>

Fig. 2. Fast decaying eigen-spectrum of the data set MIAS.
Fig. 3. The test-set accuracy results of $\Sigma_{RP}$-QDA and RP-Ensemble QDA versus QDA, and versus the linear restriction of these, i.e. LDA and RP-Ensemble LDA. The error bars depict one standard deviation.
matrix. For the other four data sets which have more features than points, only one data set, Leukemia, enjoys a stable good performance. Other three data sets present a rise then drop. The latter happens for the same reason as we have seen in QDA – though LDA only has one covariance matrix to estimate, the number of points is still not enough in the case of these high dimensional data sets.

When comparing Quadratic Discriminant Analysis ensembles with the Linear Discriminant Analysis ensemble, the latter performs steadier, it displays more stability than the former. On the contrary, the performance of the QDA based methods is sensitive to the value of \( k \) that is the Random Projections dimension. Except for the data sets Ionosphere and Spam, all other results have narrow peaks of performance – for instance, the data set COIL100, Cancer, MIAS and Leukemia. Second, the accuracies rely on the structure of the data sets. If the data set follows identical distribution for all the classes – that is all the classes have the same covariance shape – then the results of Linear Discriminant Analysis will be better than Quadratic Discriminant Analysis. This is the case of the data set Colon. The Quadratic Discriminant Analysis outperforms Linear Discriminant Analysis when the data set has much more points than attributes, for instances COIL100.

The next question is the performance of \( \Sigma_{RP} \) versus RP Ensemble for QDA. When comparing these, we see significant differences in some of the data sets, and comparable performances in others. However, we find it difficult to explain when and why this is the case. The results on the MIAS data set deserve some further investigation. We see, it is the only data set for which the vanilla QDA performs the best. Considering that it has 7 classes (so QDA estimates 7 covariance matrices), 1024 dimensions, and only 322 points, this looks very strange at first. A Principal Component Analysis was deployed to gain further insight. Figure 2 shows the eigenvalues. We can see, about 10 features account for the classification. From Figure 3, we see that the peak point is around 10 dimensions. In other words, this data set has a low intrinsic dimension, far lower than the sample size. That is the reason why QDA was able to perform so well (although we are unsure at this point why the QDA ensembles drop so sharply). So we see, it is not the representation dimension that matters, but the intrinsic structure, which is invisible to the naked eye. Though the original data space dimension is 1024, which is much larger than the number of instances, its core is only around 10 features. In this way, its actual dimension is very low.

To summarise, from Figure 3 we can say the following.

- The peaks of accuracy for both variants of RP-aggregation QDA tend to be rather narrow, as QDA requires more and more points to estimate the class-covariances as \( k \) increases, while the available sample size is fixed.
- In turn, the LDA ensemble is stable because it only needs to estimate one covariance matrix. In most cases this makes it succeed better than QDA ensembles despite its clearly restrictive and unrealistic model assumption of identical covariance in each class.
- The above effects are more pronounced in the very small sample size settings (right column plots). When the small sample condition is less severe (left column plots) then at least one of the RP-aggregation QDA variants performs no worse – better on the COIL100 data set – in comparison with RP-Ensemble LDA.

From these results we can conclude that, although RP-aggregation is beneficial to address the covariance singularity issue of QDA, and improves on vanilla QDA, the added flexibility of distinct class covariances has not been observed to bring much benefits in comparison to Fisher’s Linear Discriminant Analysis. One may wonder if this might still show up on data sets that have a much larger sample size. However, as we see on the Spam data set that, vanilla LDA works very well, and if the sample size is really large then there is no need for the regularisation effects of RP-aggregation as far as accuracy is concerned. Its computational benefits remain of course, since all it takes is low dimensional computations, which can be run in parallel.

VI. Conclusions and future work

We extended previous work on random projection ensemble LDA, to include QDA classifiers for high dimensional small sample problems. In the case of QDA, the idea of RP-aggregation leads to two different approaches, one is to aggregate the predictions, the other is to plug in an aggregate of covariance projections. We experimentally examined these methods in conjunction with their linear counterpart on eight data sets chosen to span a range of characteristics. The most prominent conclusions from our experiments is that RP-aggregation improves the performance of plain QDA by its regularisation effect, whenever the sample size is small relative to the intrinsic dimension of the data. However, for reasons discussed above, the seemingly restrictive linear LDA based RP-ensemble enjoys a better stability and better overall performance, while QDA displays rather narrow ranges of target dimensions that work well.

A worthwhile aim for future work is to gain a better understanding of the reasons for our empirical findings. Furthermore, in applications, it would be interesting to interpret the inverse covariance that results from RP-aggregations as a graph, and use it to infer conditional correlations between the features.

Acknowledgements

Most of this work was done during the first author’s MSc project at the University of Birmingham. The work of AK is supported by EPSRC under Fellowship grant EP/P004245/1, and a Turing Fellowship (grant EP/N510129/1).

References


