

GO-LDA: Generalised Optimal Linear Discriminant Analysis

Jiahui Liu, Xiaohao Cai, and Mahesan Niranjan

Abstract—Linear discriminant analysis (LDA) has been a useful tool in pattern recognition and data analysis research and practice. While linearity of class boundaries cannot always be expected, nonlinear projections through pre-trained deep neural networks have served to map complex data onto feature spaces in which linear discrimination has served well. Unlike principal component analysis which is variance preserving, LDA maximises the separation between classes, simultaneously minimising the projected scatter of each class on a subspace. The solution to binary LDA is obtained by eigenvalue analysis of within-class and between-class scatter matrices. It is well known that the multiclass LDA is solved by an extension to the binary LDA, a generalised eigenvalue problem, from which the largest subspace that can be extracted is of dimension one lower than the number of classes in the given problem. In this paper, we show that, apart from the first of the discriminant directions, the generalised eigenanalysis solution to multiclass LDA does neither yield orthogonal discriminant directions nor maximise discrimination of projected data along them. Surprisingly, to the best of our knowledge, this has not been noted in decades of literature on LDA. To overcome this drawback, we present a derivation with a strict theoretical support for sequentially obtaining discriminant directions that are orthogonal to previously computed ones and maximise in each step the Fisher criterion. We show distributions of projections along these axes and demonstrate that discrimination of data projected onto these discriminant directions has optimal separation, which is much higher than those from the generalised eigenvectors of the multiclass LDA. Using a wide range of benchmark tasks, we present a comprehensive empirical demonstration that on a number of pattern recognition and classification problems, the optimal discriminant subspaces obtained by the proposed method, referred to as GO-LDA (Generalised Optimal LDA), can offer superior accuracy.

Index Terms—LDA; PCA; dimensionality reduction; machine learning; Fisher criterion; multiclass; pattern recognition; classification.

1 INTRODUCTION

Linear discriminant analysis (LDA) has been a widely used technique for pattern recognition over several decades starting from the seminal work of Fisher [1]. While recent large-scale problems, such as computer vision, are solved by deep neural networks, where do exist several problems of industrial and societal importance posed on relatively small data sets for which LDA is still the effective tool of choice. LDA, in a vast majority of problems such as screening for a medical condition solving binary classification problems, seeks to find a direction in the space of features such that Fisher criterion of separation between projected means is maximised and simultaneously the within-class scatter of projections is kept low. Foley and Sammon in [2] extended LDA in the binary regime by constructing a discriminant subspace following the derivation of discriminant directions. It begins with the first discriminant direction given by maximising the Fisher criterion and sequentially constructing discriminant directions which also maximise discrimination information but are constructed to be orthogonal to those previously derived.

Fisher formulation has a multiclass extension for which the solution is arrived at by solving a generalised eigenvalue problem involving the sum of within-class scatter matrices and between-class matrices obtained on the sum of outer products of vectors linking the pairwise means [3]. A consequence of this is that the dimension of the obtained

discriminant subspace is limited to one less than the number of classes of the given problem [4].

An aspect of the multiclass LDA, formed by generalised eigenvectors, is that apart from the first of the directions, the others neither maximise discrimination nor are they orthogonal to each other. To the best of our knowledge, this particular fact/drawback has been overlooked in the LDA literature for decades. In this paper, we show this to be the case and, building on Foley and Sammon's work [2] on sequential construction, derive orthogonal discriminant directions for multiclass LDA. This result is an optimal subspace on which to project data while maintaining their separation. We present the necessary algebraic derivations, illustrative examples showing the distributions of projected data and a comprehensive set of experiments to demonstrate how successive directions computed carry discriminant information and that combination offers subspace in which simple classifiers (e.g., linear, quadratic and k -nearest neighbour) would be applied. Another important result of the work is that in the case of multiclass problems, we are no longer limited to the discriminant subspace being of a smaller (one fewer) size than the number of classes of the given problem. We refer to our work as GO-LDA – Generalised Optimal LDA; see Fig. 1 for a brief illustration of the difference between the multiclass LDA (i.e., classic-LDA) and GO-LDA on a three-class problem.

Our derivation, supported by empirical work, shows that a sequential construction of mutually orthogonal discriminant directions finds a linear subspace in which greater accuracy of classification may often be obtained for multiclass problems. Surprisingly, this construction goes beyond

The authors are with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK.
E-mail: jl4f19@soton.ac.uk; x.cai@soton.ac.uk; mn@ecs.soton.ac.uk

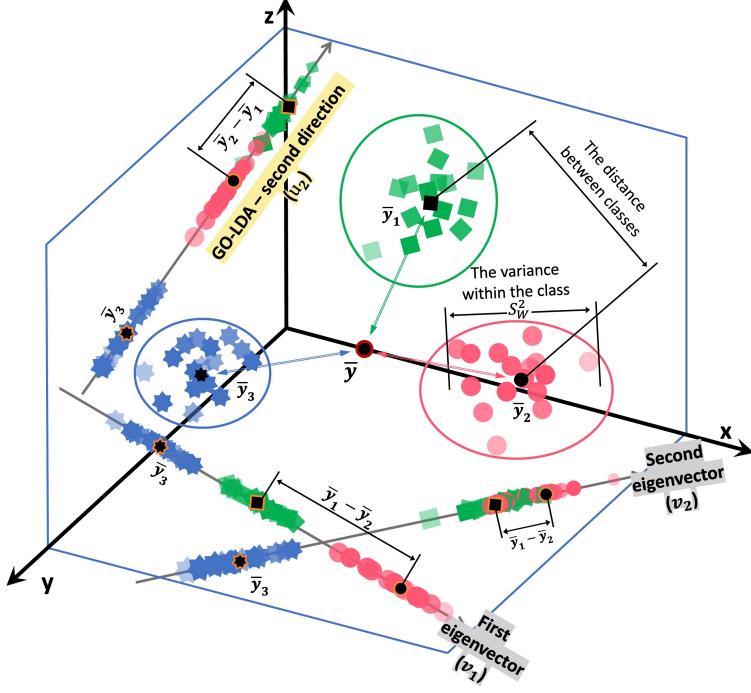


Fig. 1: Schematic diagram for a three-class problem showing the difference between the classic-LDA (generalized eigenanalysis solution) and the proposed GO-LDA (sequential computation of discrimination maximizing orthogonal projections). Along the direction of the first eigenvector v_1 , the three classes are separated well, but the second direction (i.e., v_2) of the classic-LDA is not orthogonal to the first and fails to separate two of the three classes, whereas a better solution exists, i.e., the direction (i.e., u_2) of our GO-LDA. For ease of reference, the symbols in the diagram for the class means, i.e., $\{\bar{y}_j\}_{j=1}^3$, are reused for the mean of the projected samples in each class.

the limit set by the number of classes of the given problem owing to the rank deficiency of the between-class scatter matrix encountered in the classic-LDA formulation. We note, however, consistent with the *no free lunch theorem*, that accuracy gains cannot always be expected for two reasons. Some problems may be sufficiently easy that a small number of discriminant directions (as few as one) might carry all useful information. At the other extreme, some problems might require non-linear classification boundaries that cannot be modelled by linear projections. Nevertheless, it is an intriguing finding and of paramount importance that for several multiclass problems there still remains unextracted information even with linear projections. Using a wide range of benchmark tasks, we show that GO-LDA outperforms the classic-LDA and other related methods by a large margin. It can also be exploited with deep learning techniques and can achieve state-of-the-art performance.

The remainder of this paper is organized as follows. Section 2 defines the necessary notation. A brief survey of discriminant analysis is given in Section 3. Section 4 recalls the mathematical aspect regarding LDA. In Section 5, we present the proposed methodology GO-LDA, including mathematical derivations, computational complexity analysis and some illustrations of the discriminant ability and projections of data onto the computed discriminant directions. Extensive experimental results and comparisons validating the superior performance of GO-LDA are given in Section 6. We finally conclude with a discussion in Section 7. Further related derivation and illustrations are given in Appendix.

2 NOTATIONS

Given N samples $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iM})^\top \in \mathbb{R}^M$, $1 \leq i \leq N$, we form a data matrix $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)^\top \in \mathbb{R}^{N \times M}$,

where M is the number of features of every sample. Suppose that these N samples belong to C different classes, namely Λ_j , and their cardinality $|\Lambda_j| = N_j$, $1 \leq j \leq C$. Let \mathbf{y}_k^j represent the k -th sample in class Λ_j . Clearly, $N = \sum_{j=1}^C N_j$, $\Lambda_j = \{\mathbf{y}_k^j\}_{k=1}^{N_j}$ and $\{\mathbf{y}_i\}_{i=1}^N = \bigcup_{j=1}^C \{\mathbf{y}_k^j\}_{k=1}^{N_j}$. Let $\bar{\mathbf{y}}$ and $\bar{\mathbf{y}}_j$ respectively be the mean of the whole samples and the samples in class j , i.e., $\bar{\mathbf{y}} = \frac{1}{N} \sum_{i=1}^N \mathbf{y}_i$, $\bar{\mathbf{y}}_j = \frac{1}{N_j} \sum_{\mathbf{y} \in \Lambda_j} \mathbf{y}$, $1 \leq j \leq C$.

Denote S_B and S_W , respectively, as the inter- and intra-class (also known as between- and within-class) scatters, i.e.,

$$S_B = \sum_{j=1}^C (\bar{\mathbf{y}}_j - \bar{\mathbf{y}})(\bar{\mathbf{y}}_j - \bar{\mathbf{y}})^\top, \quad S_W = \sum_{j=1}^C S_W^j, \quad (1)$$

where S_W^j represents the intra-class scatter for class j , i.e.,

$$S_W^j = \sum_{k=1}^{N_j} (\mathbf{y}_k^j - \bar{\mathbf{y}}_j)(\mathbf{y}_k^j - \bar{\mathbf{y}}_j)^\top, \quad 1 \leq j \leq C. \quad (2)$$

Specifically, for the case of $C = 2$, we also name \tilde{S}_B as the inter-class scatter, i.e.,

$$\tilde{S}_B = \mathbf{s}_b \mathbf{s}_b^\top, \quad (3)$$

where $\mathbf{s}_b = \bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2$.

3 BRIEF SURVEY OF DISCRIMINANT ANALYSIS

This section conducts a brief survey of discriminant analysis in terms of different challenges it has been tackling. We begin with its successes in discriminant dimension reduction, followed by its variants in tackling nonlinear data, data with outliers and/or noises, data with small size, and multimodel data. Finally, we point out the multiclass LDA's limited representation ability for multiclass problems, which is also the main focus of this paper. Readers familiar with discriminant analysis may prefer to skip this section and continue reading from Section 4.

3.1 Discriminant dimension reduction

In the fields of pattern recognition and machine learning, for example, feature extraction has proven effective in mitigating complexity, improving efficiency, and enhancing classification performance [5]. It can also be considered as subspace learning since it seeks to identify a low-dimensional representation of high-dimensional data, i.e., discovering a projection matrix using which to transform the original high-dimensional data into a low-dimensional subspace.

Fisher discriminant analysis (FDA) [1] is one of the most commonly used techniques for linear discriminant supervised dimensionality reduction. It requires an embedding transformation that maximises the between-class scatter and minimises the within-class scatter, tackling binary problems. In 1948, Rao et al. [3] extended it to multiclass problems by solving a generalised eigenvalue problem, which becomes the most popular (classic) LDA in use presently. To handle multiclass problems with nonlinear challenges, variants with kernelised strategies have been developed [6], [7]. The performance of the kernelised techniques is highly dependent on the choice of the kernel function family and parameters for supervised dimensionality reduction.

3.2 Data with outliers and noises

Classic-LDA utilises the ℓ_2 -norm distance, which may be sensitive to outliers and noises. By weighing the relative contributions of pairwise terms, weighted pairwise Fisher criteria minimising the effect of some dominant terms on the final criterion were proposed [8]. They are more robust compared to the classic-LDA and are capable of limiting the impact of outlier classes on the final linear dimension reduction transformation, although there is no guarantee that they would always lead to a classification rate increase since numerous estimates were required to obtain a computationally simple solution.

Li et al. [9] proposed to utilise a rotationally invariant ℓ_1 -norm to measure the two scatter matrices for discriminant projection learning. Using a weighting parameter, the relative importance of the two scatter matrices is balanced. Its application is limited due to its difficulty in identifying the optimal weighting value for varied tasks. There are many LDA variants based on the ℓ_1 -norm, e.g. [10], [11], whereas obtaining the global solution is challenging. For example, an iterative technique was used in [10] to obtain a local solution with each projection vector obtained repeatedly; a set of local optimal projection vectors was learned by a more robust version proposed in [11] using a greedy strategy. Note that the local optimal projection vectors obtained do not always optimise their objective function. To improve the performance, a non-greedy variant of the iterative technique was developed in [12].

3.3 Data with small size

For data with a small sample size, occurring when the number of data samples is much less than the number of the features, i.e., $N \ll M$, the performance of LDA methods may be degraded. The work in [13] showed that principal component analysis (PCA) outperforms LDA when the amount of training data is limited. A great deal of discriminant analysis research has focused on developing methods

that can better address the small sample size challenge, e.g., the methods utilising regularisation terms in the scatter matrix [14], [15], applying PCA for LDA (where PCA was used to obtain an intermediary subspace) [16], [17], and using null space projections (to eliminate the null space of the within-class matrix) [18]–[20].

3.4 Multimodal data

For multimodal data, each data class may contain several different modalities and each modality may form a separate cluster. LDA might produce unreliable results for multimodal data since it attempts to separate class means as much as possible [21]. Many LDA variants were proposed to address the multimodal data challenge.

An eigenvector-based linear dimensionality reduction approach (a multiclass extension of the technique in [22]) for multiclass data with heteroskedasticity was proposed in [23]. Attention was paid to heteroskedasticity data by generalising the scatter between classes using a Chernoff distance metric and using the separation information from the class mean and class covariance matrix. Moreover, the mixture discriminant analysis (MDA) was proposed in [24] and a combination of Gaussians to represent MDA subclasses was proposed in [25]. Due to the difficulty in determining the number of mixing components, an iterative process was proposed by Gkalelis et al. [26] to estimate the number of mixing components in a Gaussian mixture model.

Many other LDA variants tackled multimodal data by examining manifold (or Laplacian graph) that depicts local structures (which are often more relevant than global structures when there are insufficient training samples for discriminant analysis). For example, the local FDA (LFDA) and local sensitive discriminant analysis (LSDA) were proposed in [27] and [28], respectively. LFDA combines FDA with a locality preserving projection that can efficiently handle multimodal data; LSDA determines a projection that optimises the distance between data points belonging to various classes in each local region. It was highlighted in [29] that Laplacian-based approaches only take pairwise differences into account and disregard regional consistency, and suggested manifold partition discriminant analysis. It seeks to identify a linear embedding space in which intra-class similarity is obtained along a direction consistent with local variations in the structure of the data stream and nearby data belonging to different classes are kept separate.

Non-parametric discriminant analysis (NDA) is another strategy for multimodal data. The work in [30] suggested a non-parametric extension based on the frequently used scatter matrix for the binary scenario. It is also regarded as a k -nearest neighbour version of the non-parametric valley finding technique. A non-parametric feature analysis (NFA) was presented in [31] based on the null space of the intra-class scatter matrix. The work in [32] focused on locally averaged nearest neighbour discriminant analysis and proposed a method by combining with classifiers. A different local discriminator was created in [33] to directly specify the scattering matrix across the neighbourhood. The discriminator does not require an independently and identically distributed assumption, and the neighbourhood can be viewed naturally as the smallest subclass, which is easier to acquire than subclasses without a clustering technique.

3.5 Limitation

It is understood that LDA methods for the multiclass case are only able to obtain $(C - 1)$ discriminant directions (recall that C is the number of classes for a given problem). Foley and Sammon in [2] addressed this limitation for the binary (i.e., two-class) problem in 1975. They proposed an algorithm to derive a collection of orthogonal discriminant directions for the binary problem, in which the criteria used for selecting each discriminant direction are directly related to the discriminating potential of each direction. The number of obtained orthogonal discriminant directions can be as large as M , i.e., the maximum number of orthogonal directions in \mathbb{R}^M . In [34], folded LDA was proposed in order to extract more than $(C - 1)$ features from hyperspectral images. However, the number of the obtained discriminant directions (which are in a different dimensional space from the given samples and are not optimal) by the folded LDA is still restricted by the rank of its between-class variance matrix as what the multiclass LDA suffers.

In this paper, we focus on the very challenging/general problem, i.e., the multiclass problem, and propose our GO-LDA, which is capable of deriving a collection and maximum number of orthogonal and optimal discriminant directions in \mathbb{R}^M .

4 LINEAR DISCRIMINANT ANALYSIS

This section briefly presents the mathematical foundations of LDA for both the two-class and multiclass cases and the rationale for their limited discriminant ability.

Let $\mathbf{v} \in \mathbb{R}^M$ be an LDA projection vector. Following the Fisher criterion, \mathbf{v} maximises

$$\mathcal{R}(\mathbf{v}) = \frac{\mathbf{v}^\top \mathbf{S}_B \mathbf{v}}{\mathbf{v}^\top \mathbf{S}_W \mathbf{v}}, \quad (4)$$

which is also called the Fisher ratio. Note that \mathbf{S}_B and \mathbf{S}_W defined in Eq. (1) are the between-class scatter and within-class scatter, respectively. Unlike PCA which is variance preserving of the whole data, LDA seeks to find directions along which discrimination of projected data is maximised. The insight from Fisher was to find a direction along which the projected means of individual classes are as far apart as possible while simultaneously data from each class are projected with minimum scatter. This is achieved by defining an objective given by Eq. (4).

4.1 Two-class case

We start with the binary classification case ($C = 2$) for which Foley and Sammon [2] derived a sequential construction of mutually orthogonal vectors that maximise the Fisher criterion (*cf.* Eq. (4)), i.e.,

$$\tilde{\mathcal{R}}(\mathbf{d}) = \frac{\mathbf{d}^\top \tilde{\mathbf{S}}_B \mathbf{d}}{\mathbf{d}^\top \mathbf{S}_W \mathbf{d}}. \quad (5)$$

Note that $\tilde{\mathcal{R}}(\mathbf{d})$ is independent of the magnitude of \mathbf{d} .

The first discriminant direction, say \mathbf{d}_1 , is founded by maximising $\tilde{\mathcal{R}}(\cdot)$, and then we have

$$\mathbf{d}_1 = \alpha_1 \mathbf{S}_W^{-1} \mathbf{s}_b, \quad (6)$$

where α_1 (i.e., $\alpha_1^2 = (\mathbf{s}_b^\top [\mathbf{S}_W^{-1}]^2 \mathbf{s}_b)^{-1}$) is the normalising constant such that $\|\mathbf{d}_1\|_2 = 1$ and recall that \mathbf{s}_b defined in Eq. (3) is the difference of the means of the two classes. Note that the discriminant direction obtained in Eq. (6) is also the classic-LDA for the binary case.

The second discriminant direction \mathbf{d}_2 is required to maximise $\tilde{\mathcal{R}}(\cdot)$ in Eq. (5) and be orthogonal to \mathbf{d}_1 , which can be derived by using the Lagrange multipliers, i.e., finding the stationary points from the Lagrangian function

$$\tilde{\mathcal{R}}(\mathbf{d}_2) - \lambda [\mathbf{d}_2^\top \mathbf{d}_1], \quad (7)$$

where λ is the Lagrange multiplier. We can then obtain

$$\mathbf{d}_2 = \alpha_2 \left(\mathbf{S}_W^{-1} - \frac{\mathbf{s}_b^\top (\mathbf{S}_W^{-1})^2 \mathbf{s}_b}{\mathbf{s}_b^\top (\mathbf{S}_W^{-1})^3 \mathbf{s}_b} (\mathbf{S}_W^{-1})^2 \right) \mathbf{s}_b, \quad (8)$$

where α_2 is the normalising constant such that $\|\mathbf{d}_2\|_2 = 1$; see Appendix A for the detailed derivation.

The above procedure can be extended to derive any number of discriminant directions recursively as follows. The n -th discriminant direction \mathbf{d}_n is required to maximise $\tilde{\mathcal{R}}(\cdot)$ in Eq. (5) and be orthogonal to $\mathbf{d}_k, k = 1, 2, \dots, n-1$. It can be shown that

$$\mathbf{d}_n = \alpha_n \mathbf{S}_W^{-1} \left\{ \mathbf{s}_b - [\mathbf{d}_1 \cdots \mathbf{d}_{n-1}] \mathbf{S}_{n-1}^{-1} \begin{bmatrix} 1/\alpha_1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \right\}, \quad (9)$$

where α_n is the normalising constant such that $\|\mathbf{d}_n\|_2 = 1$ and matrix $\mathbf{S}_{n-1} \in \mathbb{R}^{(n-1) \times (n-1)}$ whose (i, j) entries are defined as

$$\mathbf{d}_i^\top \mathbf{S}_W^{-1} \mathbf{d}_j, \quad 1 \leq i, j \leq n-1. \quad (10)$$

It is worth highlighting that the way of obtaining \mathbf{d}_n in Eq. (9) only works for the binary case rather than the multiclass case, since the representation of \mathbf{s}_b defined in Eq. (3) is only for the binary case and no such a representation for the multiclass case.

After obtaining a collection of n ($n \leq M$) mutually orthogonal discriminant directions, i.e., $\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n \in \mathbb{R}^M$, a transformation matrix $\mathbf{W} = (\mathbf{d}_1, \dots, \mathbf{d}_n) \in \mathbb{R}^{M \times n}$ can be formed, which will be used to transform $\forall \mathbf{y} \in \mathbb{R}^M$ to a vector in the low-dimensional space, i.e., $\mathbf{W}^\top \mathbf{y} \in \mathbb{R}^n$.

4.2 Multiclass case

Multiclass LDA seeks discriminant directions maximising $\mathcal{R}(\cdot)$ in Eq. (4), which, utilising the Rayleigh-Ritz quotient approach [35], can be simplified as

$$\max_{\mathbf{v}} \mathbf{v}^\top \mathbf{S}_B \mathbf{v}, \quad \text{s.t. } \mathbf{v}^\top \mathbf{S}_W \mathbf{v} = 1. \quad (11)$$

The Lagrangian reads

$$\mathcal{L}(\mathbf{v}, \lambda) = \mathbf{v}^\top \mathbf{S}_B \mathbf{v} - \lambda (\mathbf{v}^\top \mathbf{S}_W \mathbf{v} - 1), \quad (12)$$

where λ is the Lagrange multiplier. Finding the stationary points of \mathcal{L} regarding \mathbf{v} yields

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}} = 2\mathbf{S}_B \mathbf{v} - 2\lambda \mathbf{S}_W \mathbf{v} \stackrel{\text{set}}{=} \mathbf{0}, \quad (13)$$

i.e.,

$$\mathbf{S}_B \mathbf{v} = \lambda \mathbf{S}_W \mathbf{v}, \quad (14)$$

which is known as the generalised eigenvalue problem regarding \mathbf{S}_B and \mathbf{S}_W . Therefore, the eigenvector say \mathbf{v}_1 corresponding to the largest non-zero eigenvalue say λ_1 of the generalised eigenvalue problem (14) gives the first discriminant direction for the multiclass problem. This \mathbf{v}_1 also maximises the Fisher criterion $\mathcal{R}(\cdot)$ in Eq. (4).

Since the ranks of \mathbf{S}_W and \mathbf{S}_B are $(M - C)$ and $(C - 1)$, respectively, the generalised eigenvalue problem (14) has maximum of $(C - 1)$ non-zero eigenvalues. After obtaining the remaining $(C - 2)$ eigenvectors, i.e., $\mathbf{v}_i, i = 2, \dots, C - 1$, corresponding to the remaining $(C - 2)$ eigenvalues, i.e., $\lambda_i, i = 2, \dots, C - 1$, together with \mathbf{v}_1 forming the $(C - 1)$ discriminant directions, that is the so-called classic-LDA for the multiclass problem.

For the classic-LDA, the $(C - 1)$ discriminant directions can also be obtained by conducting eigendecomposition of $(\mathbf{S}_W^{-1} \mathbf{S}_B)$. A small perturbation can be adopted to deal with the singularity of \mathbf{S}_W [36], which is replaced by e.g.,

$$\mathbf{S}_W \leftarrow \mathbf{S}_W + \delta \mathbf{I}, \quad (15)$$

where \mathbf{I} is the identity matrix and δ (e.g., $\delta = 5 \times 10^{-3}$) is a relatively small value such that \mathbf{S}_W is non-singular and therefore invertible.

Note that both \mathbf{S}_B and \mathbf{S}_W^{-1} are symmetric matrices. This does not imply that $(\mathbf{S}_W^{-1} \mathbf{S}_B)$ is still symmetric. From the formation of \mathbf{S}_B and \mathbf{S}_W in Eq. (1), $(\mathbf{S}_W^{-1} \mathbf{S}_B)$ in practice is asymmetric, implying that the eigenvectors (i.e., the discriminant directions of the classic-LDA) obtained by eigendecomposition of $(\mathbf{S}_W^{-1} \mathbf{S}_B)$ may not be mutually orthogonal, see Theorem 4.1 below.

Theorem 4.1. Let $\mathbf{v}_i, i = 1, \dots, C - 1$ be the $(C - 1)$ discriminant directions obtained by solving the generalised eigenvalue problem in Eq. (14), where \mathbf{v}_i is the i -th eigenvector corresponding to the i -th largest eigenvalue λ_i for Eq. (14). $\forall i, j \in \{1, \dots, C - 1\}, i \neq j$, if

$$\mathbf{v}_j^\top \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{v}_i \neq \mathbf{v}_j^\top (\mathbf{S}_W^{-1} \mathbf{S}_B)^\top \mathbf{v}_i \quad (16)$$

then

$$\mathbf{v}_i \not\perp \mathbf{v}_j. \quad (17)$$

Proof: For $(\mathbf{v}_i, \lambda_i), 1 \leq i \leq C - 1$, we have

$$\mathbf{S}_B \mathbf{v}_i = \lambda_i \mathbf{S}_W \mathbf{v}_i, \quad (18)$$

which yields

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{v}_i = \lambda_i \mathbf{v}_i. \quad (19)$$

$\forall i, j \in \{1, \dots, C - 1\}, i \neq j$, satisfying Eq. (16), we have

$$\begin{aligned} & (\lambda_i - \lambda_j) \mathbf{v}_j^\top \mathbf{v}_i \\ &= \lambda_i \mathbf{v}_j^\top \mathbf{v}_i - \lambda_j \mathbf{v}_i^\top \mathbf{v}_j \\ &= \mathbf{v}_j^\top \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{v}_i - \mathbf{v}_i^\top \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{v}_j \\ &= \mathbf{v}_j^\top \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{v}_i - \mathbf{v}_j^\top (\mathbf{S}_W^{-1} \mathbf{S}_B)^\top \mathbf{v}_i \\ &\neq 0 \quad (\text{using Eq. (16)}), \end{aligned} \quad (20)$$

which implies

$$\mathbf{v}_i \not\perp \mathbf{v}_j. \quad (21)$$

This completes the proof. \square

It is obvious that condition in Eq. (16) is quite common to be satisfied given $(\mathbf{S}_W^{-1} \mathbf{S}_B)$ is asymmetric. Therefore, the discriminant directions, i.e., $\mathbf{v}_i, i = 1, \dots, C - 1$, obtained by the classic-LDA are generally not mutually orthogonal. Most importantly, we notice that only the first discriminant direction \mathbf{v}_1 maximises the Fisher criterion $\mathcal{R}(\mathbf{v})$ in Eq. (4). The remainder of the discriminant directions, i.e., $\mathbf{v}_i, i = 2, \dots, C - 1$, are just the eigenvectors of the generalised eigenvalue problem (14), rather than meeting/maximising the Fisher criterion $\mathcal{R}(\cdot)$ in Eq. (4), and therefore, are non-optimal discriminant directions. The discriminant ability of the discriminant directions by the classic-LDA actually dropped dramatically from the second direction \mathbf{v}_2 . The reason is that the Fisher ratio $\mathcal{R}(\cdot)$ in Eq. (4) is equal to the generalised eigenvalue λ_i of Eq. (14) for the i -th discriminant direction \mathbf{v}_i ; however, the eigenvalue λ_i becomes very small when i ($\leq C - 1$) is growing, indicating the dramatic drop of \mathbf{v}_i 's discriminant ability when i is large. Our illustrated experiments in Sections 5 and 6 also show that the discriminant directions by the classic-LDA, i.e., $\mathbf{v}_i, i \leq C - 1$, indeed carry more and more limited (non-optimal) discriminant information. Our proposed GO-LDA in Section 5 below overcomes all of these aforementioned limitations in the classic-LDA for the multiclass problem.

5 PROPOSED METHOD

We first present a straightforward trial of improving the classic-LDA's performance, i.e., orthogonalising the classic-LDA's non-orthogonal discriminant directions using the Gram-Schmidt process [37]. Then we propose our GO-LDA, which can produce the maximum number of optimal and also mutually orthogonal discriminant directions in space \mathbb{R}^M . Afterwards, we conduct a computational complexity analysis and provide some illustrations of the discriminant ability and projections of data by GO-LDA.

5.1 Gram-Schmidt process

The Gram-Schmidt process is a way of orthonormalizing a set of non-orthogonal vectors in an inner product space; e.g., the most commonly used Euclidean space \mathbb{R}^M equipped with the standard inner product defined as $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^M, \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i^\top \mathbf{x}_j$. For the mutually non-orthogonal discriminant directions, i.e., $\mathbf{v}_i, i = 1, \dots, C - 1$, obtained by the classic-LDA, the Gram-Schmidt process with $\tilde{\mathbf{v}}_1 = \mathbf{v}_1$, for $i = 2, \dots, C - 1$, reads,

$$\tilde{\mathbf{v}}_i = \mathbf{v}_i - \sum_{k=1}^{i-1} \langle \mathbf{v}_i, \tilde{\mathbf{v}}_k \rangle \frac{\tilde{\mathbf{v}}_k}{\|\tilde{\mathbf{v}}_k\|_2}. \quad (22)$$

Then $\tilde{\mathbf{v}}_1 \perp \tilde{\mathbf{v}}_2 \perp \dots \perp \tilde{\mathbf{v}}_{C-1}$ (normalised) will be used as discriminant directions replacing the mutually non-orthogonal ones from the classic-LDA.

We found that orthogonalisation improves the performance of the classic-LDA to some extent, but will not change the number of discriminant directions. Moreover, the orthogonalised discriminant directions are still non-optimal (since the Fisher criterion is not maximised by $\tilde{\mathbf{v}}_i, i = 2, \dots, C - 1$). Without loss of generality, in the rest of the paper, we assume all the directions are normalised.

5.2 Generalised optimal LDA

This section presents our GO-LDA method, deriving as many as M optimal mutually orthogonal discriminant directions in \mathbb{R}^M . Note that in \mathbb{R}^M , it is understood that only M mutually orthogonal directions exist.

Let $\mathbf{u}_n, n = 1, \dots, M$, be the discriminant directions of GO-LDA for the multiclass problem. The first GO-LDA's discriminant direction \mathbf{u}_1 is defined as

$$\mathbf{u}_1 = \mathbf{v}_1 \quad (23)$$

(i.e., the same as that of the classic-LDA), which is the eigenvector corresponding to the largest eigenvalue of the generalised eigenvalue problem (*cf.* Eq. (14)), maximising the Fisher criterion $\mathcal{R}(\cdot)$ in Eq. (4). We require the second GO-LDA's discriminant direction \mathbf{u}_2 to maximise the Fisher criterion $\mathcal{R}(\cdot)$ in Eq. (4) and to be orthogonal to the first discriminant direction \mathbf{u}_1 . Then we derive that \mathbf{u}_2 can be found by solving the generalised eigenvalue problem given in Eq. (25), see Theorem 5.1 below.

Theorem 5.1. Let $\mathbf{u}_2 \in \mathbb{R}^M$ maximise the Fisher criterion $\mathcal{R}(\cdot)$ in Eq. (4) and be orthogonal to \mathbf{u}_1 in Eq. (23), i.e., \mathbf{u}_2 is the solution of the following problem

$$\max_{\mathbf{u}} \mathcal{R}(\mathbf{u}) = \frac{\mathbf{u}^\top \mathbf{S}_B \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_W \mathbf{u}}, \quad \text{s.t. } \mathbf{u} \perp \mathbf{u}_1. \quad (24)$$

Then \mathbf{u}_2 is the eigenvector corresponding to the largest eigenvalue of the generalised eigenvalue problem

$$(\mathbf{S}_B - \mathbf{k}_1) \mathbf{u} = \mu \mathbf{S}_W \mathbf{u}, \quad (25)$$

where

$$\mathbf{k}_1 = \frac{\mathbf{u}_1^\top \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{u}_1}{\mathbf{u}_1^\top \mathbf{S}_W^{-1} \mathbf{u}_1}. \quad (26)$$

Proof: The maximisation problem (24) can be derived by finding the stationary points from the Lagrangian function

$$\hat{\mathcal{L}}(\mathbf{u}, \beta) = \frac{\mathbf{u}^\top \mathbf{S}_B \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_W \mathbf{u}} - \beta \mathbf{u}^\top \mathbf{u}_1, \quad (27)$$

where β is the Lagrange multiplier. This means its partial derivation with respect to \mathbf{u} should be zero, i.e.,

$$\frac{2\mathbf{S}_B \mathbf{u} \mathbf{u}^\top \mathbf{S}_W \mathbf{u} - 2\mathbf{u}^\top \mathbf{S}_B \mathbf{u} \mathbf{S}_W \mathbf{u}}{(\mathbf{u}^\top \mathbf{S}_W \mathbf{u})^2} - \beta \mathbf{u}_1 = 0. \quad (28)$$

Since $\mathbf{u}_1^\top \mathbf{u} = 0$ and note that $\mathbf{u}^\top \mathbf{S}_B \mathbf{u}$ is a scalar, we have

$$(\mathbf{u}_1^\top \mathbf{S}_W^{-1})(2\mathbf{u}^\top \mathbf{S}_B \mathbf{u} \mathbf{S}_W \mathbf{u}) = 2\mathbf{u}^\top \mathbf{S}_B \mathbf{u} (\mathbf{u}_1^\top \mathbf{u}) = 0. \quad (29)$$

Multiplying $\mathbf{u}_1^\top \mathbf{S}_W^{-1}$ on both sides of Eq. (28) and using Eq. (29) yield

$$\frac{2\mathbf{u}_1^\top \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_W \mathbf{u}} - \beta \mathbf{u}_1^\top \mathbf{S}_W^{-1} \mathbf{u}_1 = 0. \quad (30)$$

Then we have

$$\beta = \frac{2\mathbf{u}_1^\top \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{u}}{(\mathbf{u}^\top \mathbf{S}_W \mathbf{u}) \mathbf{u}_1^\top \mathbf{S}_W^{-1} \mathbf{u}_1}. \quad (31)$$

Substituting the representation of β above into Eq. (28) and then multiplying $\mathbf{u}^\top \mathbf{S}_W \mathbf{u}/2$ on both sides, we have

$$\mathbf{S}_B \mathbf{u} - \frac{\mathbf{u}^\top \mathbf{S}_B \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_W \mathbf{u}} \mathbf{S}_W \mathbf{u} - \frac{\mathbf{u}_1^\top \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{u}_1}{\mathbf{u}_1^\top \mathbf{S}_W^{-1} \mathbf{u}_1} \mathbf{u} = 0, \quad (32)$$

which can be rewritten as

$$\left(\mathbf{S}_B - \frac{\mathbf{u}_1^\top \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{u}_1}{\mathbf{u}_1^\top \mathbf{S}_W^{-1} \mathbf{u}_1} \right) \mathbf{u} = \frac{\mathbf{u}^\top \mathbf{S}_B \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_W \mathbf{u}} \mathbf{S}_W \mathbf{u}, \quad (33)$$

i.e. (using the definition of \mathbf{k}_1 in Eq. (26)),

$$(\mathbf{S}_B - \mathbf{k}_1) \mathbf{u} = \frac{\mathbf{u}^\top \mathbf{S}_B \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_W \mathbf{u}} \mathbf{S}_W \mathbf{u}. \quad (34)$$

Since \mathbf{u}_2 satisfies the problem in (24), it is also a solution of Eq. (34) above, i.e.,

$$(\mathbf{S}_B - \mathbf{k}_1) \mathbf{u}_2 = \frac{\mathbf{u}_2^\top \mathbf{S}_B \mathbf{u}_2}{\mathbf{u}_2^\top \mathbf{S}_W \mathbf{u}_2} \mathbf{S}_W \mathbf{u}_2. \quad (35)$$

This means \mathbf{u}_2 is an eigenvector of the eigenvalue

$$\mu_1^{(2)} = \frac{\mathbf{u}_2^\top \mathbf{S}_B \mathbf{u}_2}{\mathbf{u}_2^\top \mathbf{S}_W \mathbf{u}_2} \quad (36)$$

for the generalised eigenvalue problem in (25), with \mathbf{k}_1 defined in (26). Since \mathbf{u}_2 maximises $\mathcal{R}(\mathbf{u})$ in (24), indicating $\mu_1^{(2)}$ is the largest eigenvalue of the generalised eigenvalue problem in (25). This completes the proof. \square

Theorem 5.1 shows that the second GO-LDA's discriminant direction \mathbf{u}_2 (also orthogonal to \mathbf{u}_1) is the eigenvector corresponding to the largest eigenvalue of the generalised eigenvalue problem given in Eq. (25), which is different from the generalised eigenvalue problem in Eq. (14) used in the classic-LDA. For the difference between our GO-LDA and the classic-LDA in deriving their second discriminant direction, except for solving different generalised eigenvalue problems, GO-LDA takes the eigenvector corresponding to the largest eigenvalue while the classic-LDA takes the eigenvector corresponding to the second largest eigenvalue of their generalised eigenvalue problems. It also indicates that the second discriminant direction of GO-LDA is optimal whereas that of the classic-LDA is not.

Now we give the derivation of all the discriminant directions of GO-LDA for the multiclass problem, i.e., $\mathbf{u}_n, n = 2, \dots, M$, where each $\mathbf{u}_n \in \mathbb{R}^M$ maximises the Fisher criterion $\mathcal{R}(\cdot)$ in Eq. (4) and is orthogonal to the discriminant directions $\mathbf{u}_i, i = 1, \dots, n-1$. Theorem 5.2 below shows that each $\mathbf{u}_n, 2 \leq n \leq M$, can be derived by solving a different generalised eigenvalue problem given in Eq. (38).

Theorem 5.2. Let $\mathbf{u}_n \in \mathbb{R}^M, 2 \leq n \leq M$ maximise the Fisher criterion $\mathcal{R}(\cdot)$ in Eq. (4) and be orthogonal to the GO-LDA's discriminant directions $\mathbf{u}_i, i = 1, \dots, n-1$, i.e., \mathbf{u}_n is the solution of the following problem

$$\begin{aligned} \max_{\mathbf{u}} \mathcal{R}(\mathbf{u}) &= \frac{\mathbf{u}^\top \mathbf{S}_B \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_W \mathbf{u}}, \\ \text{s.t. } \mathbf{u} &\perp \mathbf{u}_1 \perp \dots \perp \mathbf{u}_{n-1}. \end{aligned} \quad (37)$$

Then \mathbf{u}_n is the eigenvector corresponding to the largest eigenvalue of the generalised eigenvalue problem

$$(\mathbf{S}_B - \mathbf{U}_{n-1} \mathbf{T}_{n-1}^{-1} \mathbf{B}_{n-1}) \mathbf{u} = \mu \mathbf{S}_W \mathbf{u}, \quad (38)$$

where matrices

$$\mathbf{U}_{n-1} = (\mathbf{u}_1 \cdots \mathbf{u}_{n-1}), \quad (39)$$

Algorithm 1 GO-LDA: Generalised Optimal LDA

Input: Data $\mathbf{Y} \in \mathbb{R}^{N \times M}$, number of classes C , and $K \leq M$ number of discriminant directions.

Output: Discriminant directions $\{\mathbf{u}_n\}_{n=1}^K$.

Compute \mathbf{S}_W and \mathbf{S}_B in Eq. (1);

Compute \mathbf{u}_1 in Eq. (23) and normalise it;

$n = 2$;

for $n \leq K$ **do**

Form the matrices \mathbf{U}_{n-1} , \mathbf{B}_{n-1} and \mathbf{T}_{n-1} using the definitions in (39), (40) and (41), respectively.

Form the generalised eigenvalue problem (38).

Compute the eigenvector \mathbf{u}_n corresponding to the largest eigenvalue of the problem (38) and normalise it;

$n = n + 1$;

end for

Return $\{\mathbf{u}_n\}_{n=1}^K$

$$\mathbf{B}_{n-1} = \begin{pmatrix} \mathbf{u}_1^\top \mathbf{S}_W^{-1} \mathbf{S}_B \\ \mathbf{u}_2^\top \mathbf{S}_W^{-1} \mathbf{S}_B \\ \vdots \\ \mathbf{u}_{n-1}^\top \mathbf{S}_W^{-1} \mathbf{S}_B \end{pmatrix}, \quad (40)$$

and $\mathbf{T}_{n-1} \in \mathbb{R}^{(n-1) \times (n-1)}$ whose (i, j) entries are defined as

$$\mathbf{u}_i^\top \mathbf{S}_W^{-1} \mathbf{u}_j, \quad 1 \leq i, j \leq n-1. \quad (41)$$

Proof: The maximisation problem (37) can be derived by finding the stationary points from the Lagrangian function

$$\bar{\mathcal{L}}(\mathbf{u}, \boldsymbol{\beta}) = \frac{\mathbf{u}^\top \mathbf{S}_B \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_W \mathbf{u}} - \sum_{i=1}^{n-1} \beta_i \mathbf{u}^\top \mathbf{u}_i, \quad (42)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{n-1})^\top$ is the Lagrange multiplier. This means its partial derivation regarding \mathbf{u} should be zero, i.e.,

$$\frac{2\mathbf{S}_B \mathbf{u} \mathbf{u}^\top \mathbf{S}_W \mathbf{u}}{(\mathbf{u}^\top \mathbf{S}_W \mathbf{u})^2} - \frac{2\mathbf{u}^\top \mathbf{S}_B \mathbf{u} \mathbf{S}_W \mathbf{u}}{(\mathbf{u}^\top \mathbf{S}_W \mathbf{u})^2} - \sum_{i=1}^{n-1} \beta_i \mathbf{u}_i = 0, \quad (43)$$

which is equivalent to (using \mathbf{U}_{n-1} defined in Eq. (39))

$$\frac{2\mathbf{S}_B \mathbf{u} \mathbf{u}^\top \mathbf{S}_W \mathbf{u}}{(\mathbf{u}^\top \mathbf{S}_W \mathbf{u})^2} - \frac{2\mathbf{u}^\top \mathbf{S}_B \mathbf{u} \mathbf{S}_W \mathbf{u}}{(\mathbf{u}^\top \mathbf{S}_W \mathbf{u})^2} - \mathbf{U}_{n-1} \boldsymbol{\beta} = 0. \quad (44)$$

Since $\mathbf{u}_k^\top \mathbf{u} = 0$, $k = 1, \dots, n-1$, multiplying $\mathbf{u}_k^\top \mathbf{S}_W^{-1}$, $k = 1, \dots, n-1$, individually on both sides of Eq. (44) and using Eq. (29) yield

$$\alpha \mathbf{u}_k^\top \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{u} - \sum_{i=1}^{n-1} \beta_i \mathbf{u}_k^\top \mathbf{S}_W^{-1} \mathbf{u}_i = 0, \quad k = 1, \dots, n-1, \quad (45)$$

i.e.,

$$\sum_{i=1}^{n-1} \beta_i \mathbf{u}_k^\top \mathbf{S}_W^{-1} \mathbf{u}_i = \alpha (\mathbf{u}_k^\top \mathbf{S}_W^{-1} \mathbf{S}_B) \mathbf{u}, \quad k = 1, \dots, n-1, \quad (46)$$

where

$$\alpha = 2 / (\mathbf{u}^\top \mathbf{S}_W \mathbf{u}). \quad (47)$$

Using the definition of \mathbf{B}_{n-1} and \mathbf{T}_{n-1} in Eq. (40) and Eq. (41), respectively, Eq. (46) can be rewritten as

$$\mathbf{T}_{n-1} \boldsymbol{\beta} = \alpha \mathbf{B}_{n-1} \mathbf{u}. \quad (48)$$

We then have

$$\boldsymbol{\beta} = \alpha \mathbf{T}_{n-1}^{-1} \mathbf{B}_{n-1} \mathbf{u}. \quad (49)$$

Substituting the representation of $\boldsymbol{\beta}$ above into Eq. (44) and then multiplying $\mathbf{u}^\top \mathbf{S}_W \mathbf{u} / 2$ on both sides, we have

$$\mathbf{S}_B \mathbf{u} - \frac{\mathbf{u}^\top \mathbf{S}_B \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_W \mathbf{u}} \mathbf{S}_W \mathbf{u} - \mathbf{U}_{n-1} \mathbf{T}_{n-1}^{-1} \mathbf{B}_{n-1} \mathbf{u} = 0, \quad (50)$$

i.e.,

$$(\mathbf{S}_B - \mathbf{U}_{n-1} \mathbf{T}_{n-1}^{-1} \mathbf{B}_{n-1}) \mathbf{u} = \frac{\mathbf{u}^\top \mathbf{S}_B \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_W \mathbf{u}} \mathbf{S}_W \mathbf{u}. \quad (51)$$

Since \mathbf{u}_n satisfies the problem in (37), it is also a solution of Eq. (51) above, i.e.,

$$(\mathbf{S}_B - \mathbf{U}_{n-1} \mathbf{T}_{n-1}^{-1} \mathbf{B}_{n-1}) \mathbf{u}_n = \frac{\mathbf{u}_n^\top \mathbf{S}_B \mathbf{u}_n}{\mathbf{u}_n^\top \mathbf{S}_W \mathbf{u}_n} \mathbf{S}_W \mathbf{u}_n. \quad (52)$$

This means \mathbf{u}_n is an eigenvector of the eigenvalue say

$$\mu_1^{(n)} = \frac{\mathbf{u}_2^\top \mathbf{S}_B \mathbf{u}_2}{\mathbf{u}_2^\top \mathbf{S}_W \mathbf{u}_2} \quad (53)$$

for the generalised eigenvalue problem in Eq. (38). Since \mathbf{u}_n maximises $\mathcal{R}(\mathbf{u})$ in Eq. (37), indicating $\mu_1^{(n)}$ is the largest eigenvalue of the generalised eigenvalue problem in Eq. (38). This completes the proof. \square

Theorem 5.2 tells us that GO-LDA's discriminant directions, i.e., \mathbf{u}_n , $n = 1, \dots, M$, can be derived recursively with \mathbf{u}_1 computed in Eq. (23), by computing the eigenvector corresponding to the largest eigenvalue of the generalised eigenvalue problem given in Eq. (38) each time. Obviously, the GO-LDA's discriminant directions are optimal since they are all maximising the Fisher criterion $\mathcal{R}(\cdot)$ in Eq. (4) with the mutually orthogonal constraint. We finally remark that the ideas proposed in Theorems 5.1 and 5.2 could be exploited to enhance the performance of most, if not all, of the LDA variants (e.g. the ones surveyed in Section 3).

The whole procedure of finding GO-LDA's discriminant directions, i.e., $\{\mathbf{u}_n\}_{n=1}^M$, is summarised in Algorithm 1.

5.3 Computational complexity

We now analyse GO-LDA's computational complexity and make comparison to the classic-LDA.

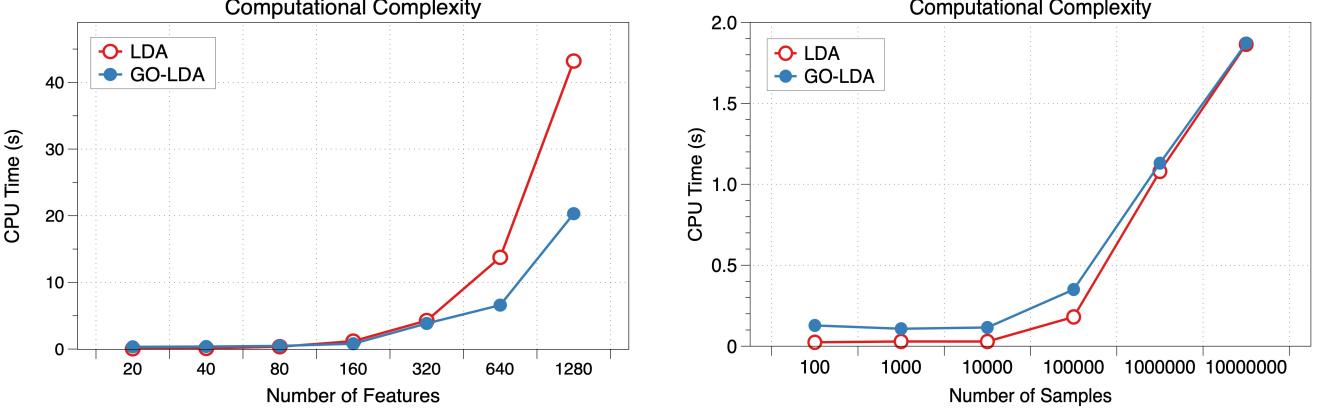


Fig. 2: Computation cost comparison between the classic-LDA (referred to LDA in the plots/tables for simplicity) and GO-LDA. The left panel shows the CPU time for synthetic data sets with a fixed number of samples and an increased number of data features. The right panel shows the CPU time for synthetic data sets with an increased number of samples and a fixed number of data features. The synthetic data sets are created by using the Python built-in function `make_blobs` to generate isotropic Gaussian samples. The number of classes is $C = 5$, and four discriminant directions are derived for both the classic-LDA and GO-LDA. In particular, for the left panel, $N = 1,000$ with the number of data features M increased from 20 to over 1,000; and for the right panel, $M = 10$ with the number of samples N increased from 100 to over 10^7 .

5.3.1 Computation cost

Recall that N is the number of samples of the given data set and M is the number of features of every sample. The classic-LDA (presented in Section 4.2) consists of the following two key steps. The first is to form the inter-class matrix S_B and the intra-class matrix S_W , as specified in Eq. (1), and the second is to solve the generalised eigenvalue problem (14). The computational complexity of forming S_B and S_W is dominated by S_W , which is $\mathcal{O}(NM^2)$. Solving the generalised eigenvalue problem (14) takes $\mathcal{O}(M^3)$ if e.g. we calculate S_W^{-1} first (taking $\mathcal{O}(M^3)$) and then conduct eigendecomposition. The computation cost of the classic-LDA is therefore

$$\mathcal{O}(NM^2) + 2\mathcal{O}(M^3). \quad (54)$$

Note that the above representation is informal since our main purpose here is for ease of comparison (*cf.* Eq. (55)).

Analogously, GO-LDA also firstly computes matrices S_B and S_W , taking $\mathcal{O}(NM^2)$, see Algorithm 1. Forming matrices U_{n-1}, B_{n-1} and T_{n-1} using the definitions in Eq. (39), (40) and (41), respectively, takes $\mathcal{O}(M^3)$, since the computation cost is dominated by previously forming S_W^{-1} , which is $\mathcal{O}(M^3)$. Then, to compute K , $K \leq M$, number of GO-LDA discriminate directions, K number of generalised eigenvalue problems given in Eq. (38) need to be solved. Note that, for each generalised eigenvalue problem in Eq. (38), GO-LDA only computes one eigenvector corresponding to the largest eigenvalue (rather than the complete eigendecomposition), taking $\mathcal{O}(M^2)$ [38]. We then obtain the computation cost of GO-LDA, i.e.,

$$\mathcal{O}(NM^2) + \mathcal{O}(M^3) + K\mathcal{O}(M^2). \quad (55)$$

From the computation cost representations of the classic-LDA and GO-LDA respectively in (54) and (55), we see that if $N \gg M$, then $2\mathcal{O}(M^3)$ in Eq. (54) and $(\mathcal{O}(M^3) + K\mathcal{O}(M^2))$ in Eq. (55) will be dominated by the term $\mathcal{O}(NM^2)$. Moreover, $K \leq M$ implies that $\mathcal{O}(M^3)$ and

$K\mathcal{O}(M^2)$ are comparable. Therefore, the computation cost representations in (54) and (55) indicate that both the classic-LDA and GO-LDA have comparable computation cost, even though our GO-LDA needs to solve K different generalised eigenvalue problems.

5.3.2 Experimental demonstration

Below we experimentally demonstrate the GO-LDA's computation cost and make comparison to the classic-LDA. The Python package `np.linalg.eig` is used to perform eigen-decomposition; in particular, the Python built-in function `scipy.sparse.linalg.eigs` is used to calculate GO-LDA's individual discriminant directions, i.e., calculating one eigenvector corresponding to the largest eigenvalue of each generalised eigenvalue problem in Eq. (38).

Fig. 2 gives the CPU time of the classic-LDA and GO-LDA on synthetic data sets with different number of features M and different number of samples N , see the left and right panels of Fig. 2. In particular, the left panel of Fig. 2 shows that our GO-LDA is surprisingly even more economical than the classic-LDA when the number of features M increases. This could be explained by the computation cost representations in (54) and (55), i.e., $K\mathcal{O}(M^2)$ is less than $\mathcal{O}(M^3)$ when $K \ll M$. The right panel of Fig. 2 shows that, when both M and N are small, the CPU cost of our GO-LDA is slightly higher than that of the classic-LDA, but the difference vanishes when the number of samples N increases. This could again be explained by the computation cost representations in (54) and (55), i.e., the term $\mathcal{O}(NM^2)$ dominates both representations when N is large. On the whole, Fig. 2 experimentally demonstrates that the computation cost of our GO-LDA is quite economical and is comparable to the classic-LDA.

5.4 Discriminant ability

We finally illustrate the discriminant ability of the discriminant directions derived by our proposed GO-LDA and make

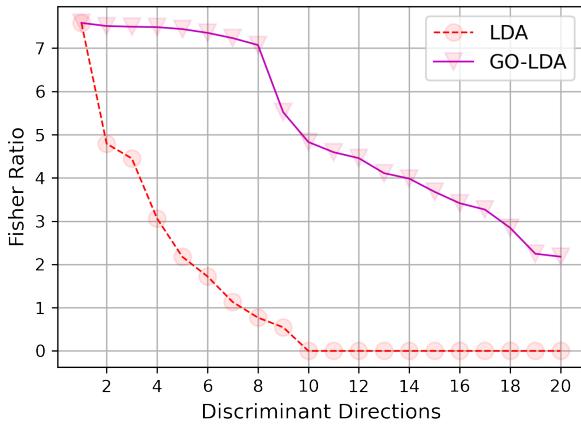


Fig. 3: Comparison between the classic-LDA and GO-LDA in terms of Fisher ratio corresponding to their individual discriminant directions on the Handwritten Digits data set. The performance of the classic-LDA and GO-LDA is represented by the dashed and solid lines, respectively. It demonstrates that GO-LDA achieves a much higher level of discriminant ability compared to the classic-LDA even beyond the limit of the discriminant directions (i.e., eight directions) of the classic-LDA.

comparison to the classic-LDA in terms of the Fisher ratio and projections on discriminant directions.

5.4.1 Fisher ratio on discriminant directions

The Fisher criterion is the objective function of LDA methods. Below we compare the Fisher ratio achieved by the discriminant directions of the classic-LDA and GO-LDA, see Fig. 3. The Handwritten Digits classification data set [39] containing 10 classes in 64-dimensional feature space (i.e., pixel values of 8×8 images) is used in Fig. 3. Hence the classic-LDA only has nine discriminant directions available; in contrast, our GO-LDA is not restricted by the number of classes and can derive as many discriminant directions as the dimension of the feature space. Fig. 3 clearly shows that the classic-LDA yields discriminant directions along which the Fisher ratio decreases rapidly; in other words, the discriminant ability of the classic-LDA is indeed quite limited. In contrast, our GO-LDA can derive more (optimal according to Theorem 5.2) discriminant directions and retain much higher levels of discriminant ability even beyond the limit of the nine discriminant directions of the classic-LDA.

5.4.2 Projections on discriminant directions

To visually validate the optimal discriminant ability of GO-LDA, we investigate the projections on GO-LDA's discriminant directions and make comparison to the classic-LDA. The Wine data set, a benchmark classification problem containing 3 classes in 13-dimensional feature space taken from the UCI ML repository [39], is used here. Hence the classic-LDA will only be able to give two meaningful discriminant directions, whereas GO-LDA can derive 13 discriminant directions (i.e., the dimension of the feature space).

Projections of the data onto the discriminant directions obtained by the classic-LDA and GO-LDA are shown in Fig.

4. From Fig. 4 (a), we see that the projections of the three classes of the Wine data set can be separated by the first discriminant direction (i.e., v_1 or u_1) of the classic-LDA and GO-LDA. However, the projections of two classes overlap along the classic-LDA's second discriminant direction v_2 . Fig. 4 (a) (see the right two plots) also shows the projections of the Wine data set on two more directions obtained by solving the generalised eigenvalue problem of the classic-LDA, indicating that all of these directions are not useful and carry no discriminant information. In contrast, the GO-LDA's performance shown in Fig. 4 (b) clearly demonstrates that all the GO-LDA's discriminant directions carry important discriminant information. In detail, the projections of the Wine data set can be well separated by the first three GO-LDA's discriminant directions (cf. only the first discriminant direction of the classic-LDA can do so). Moreover, the following five GO-LDA's discriminant directions, i.e., u_4 to u_8 , as shown in Fig. 4 (b), can also achieve high separability after projecting the Wine data set on them (cf. the directions derived by solving the generalised eigenvalue problem of the classic-LDA deliver no separability after v_1 and v_2). In particular, two classes of the Wine data set slightly overlap after projecting onto the GO-LDA's discriminant directions u_4 and u_5 for example, but the overlap classes are different, indicating that each of the GO-LDA's discriminant directions carries different discriminant information, which benefits from the orthogonality between the GO-LDA's discriminant directions. Hence, interestingly, a combination of e.g. u_4 and u_5 in a two-dimensional projection can also separate all the three classes of the Wine data set, see Appendix B.

The above illustrations showed the great performance of GO-LDA and the surprisingly limited performance of the classic-LDA in terms of discriminant ability. In the next section, we conduct a comprehensive set of experiments and comparisons to further validate GO-LDA's great discriminant ability and its importance in various applications.

6 EXPERIMENTAL RESULTS

To showcase the effectiveness and importance of GO-LDA, we illustrate classification problems using both individual and multiple discriminant directions by carrying out extensive empirical work on a total of twenty benchmark data sets spanning different data types, numbers of classes and class imbalances. Throughout we compare the performance of our GO-LDA against the classic-LDA. Further, for completeness, we also compare with PCA by projecting the data onto the same number of principal components as that of the discriminant directions. Classifiers are built on those subspaces formed by principle/discriminant directions.

Data. Among the twenty benchmark data sets, eleven are taken from the widely used UCI ML repository [39], i.e., IrisPlants, BreastTissue, ForestTypeMapping, Glass, Handwritten Digits, Landsat, Nursery, ThyroidGland, UrbanLandCover, Vowel and Wine; five are taken from the KEEL repository [40] with high imbalance across classes, i.e., contraceptive, Ecoli, Hayes-Roth, New-Thyroid and Yeast; two are face recognition data sets, i.e., LFW [41] and ORL [42]; and the rest two are medical data sets, i.e., BrainTumor [43] and

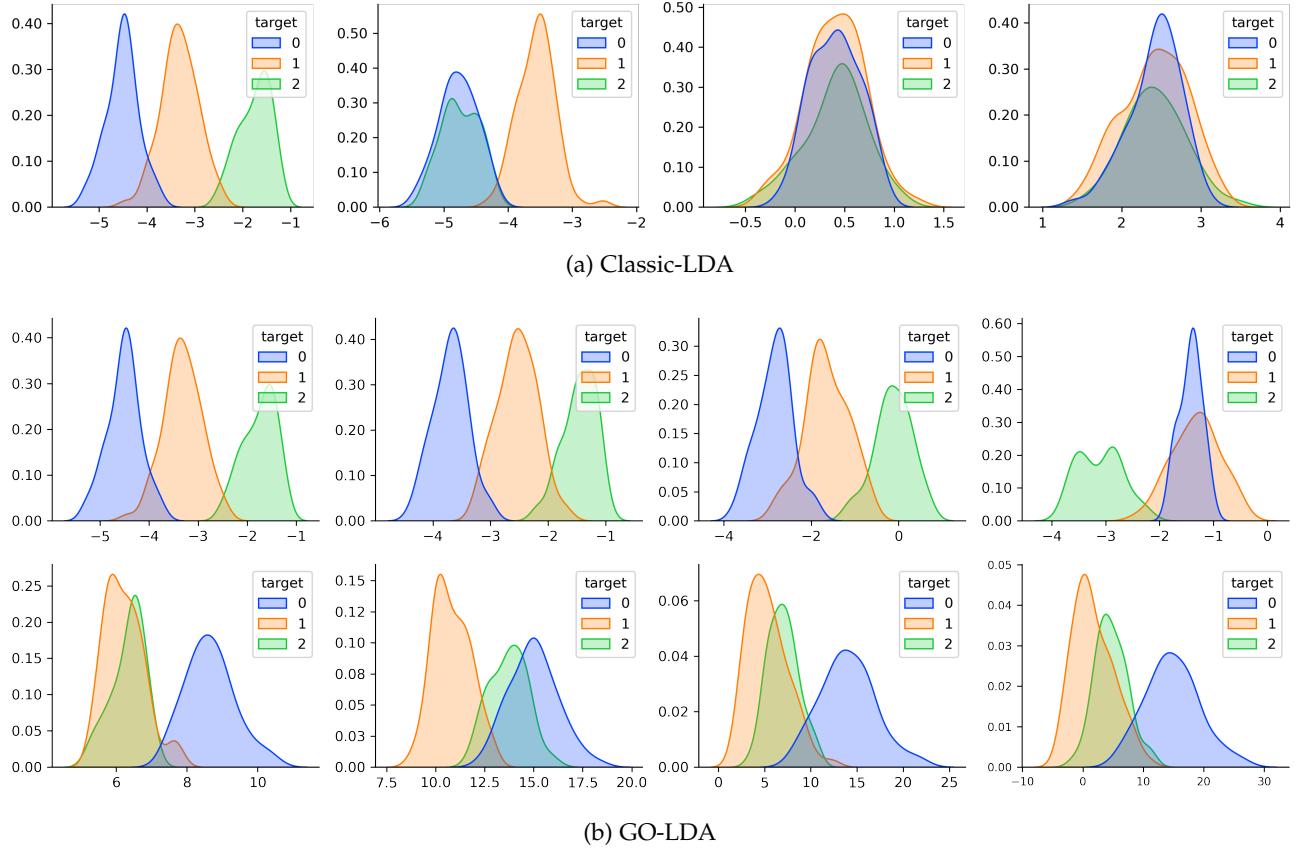


Fig. 4: Comparison between the classic-LDA and GO-LDA in terms of projections on their discriminant directions. The three-class Wine data set with 13 features is used in this illustration. Panel (a) gives the projections of the Wine data set on the classic-LDA's discriminant directions, i.e., v_1 and v_2 , plus two more directions from solving the same generalised eigenvalue problem of the classic-LDA, respectively. Panel (b) gives the projections of the Wine data set on the first eight GO-LDA's discriminant directions, i.e., u_1 to u_8 , respectively. The projections clearly show that GO-LDA delivers significantly better discriminant ability compared to the classic-LDA. In particular, the optimality of discrimination in GO-LDA is preserved for all the eight shown discriminant directions (*cf.* the classic-LDA can only find two discriminant directions for the three-class Wine data set, with its second direction not optimal). A detailed description is given in the main text. A scatter plot in the two dimensional subspace consisting of u_4 and u_5 is shown in Appendix B.

DeepDrid [44]. Some necessary characteristics (e.g. the values of C , N and M) of the data sets are given along with the corresponding results in Tables 1–4.

Setting. To make the comparisons more convincing, we choose different classifiers (i.e., k -nearest neighbour, linear and quadratic) to act on the subspaces. To illustrate our work on inference from medical images, we first transfer the image data through a pre-trained deep neural network (i.e., ResNet18) into a fixed vector representation. The medical problems we consider here are characterised by the data scarcity scenario and thus only training deep image analysis models is generally insufficient. The projection techniques with discriminant abilities like PCA, classic-LDA and GO-LDA are essential. In the transferred space, images are represented in a 512-dimensional feature space. The two medical problems are related to brain imaging (i.e., the BrainTumor data set) [43] and diabetic retinopathy (i.e., the DeepDrid data set) [44]. In the former, 592 images are used for training and 148 for test; and with the latter, 529 images are used for training and 133 for test. Their figures are chosen to give a training set slightly higher than the

image embedding dimensions (i.e., 512).

6.1 Results on individual discriminant directions

Table 1 gives the comparison between PCA, classic-LDA and GO-LDA on six data sets from the UCI ML repository [39], using quadratic classifiers on projections of the data onto different principle/discriminant directions taken one at a time. Note again that the number of the discriminant directions of the classic-LDA is limited by the number of classes C , which is significantly less than the number of the principle/discriminant directions of PCA and GO-LDA (which can go as large as the number of the features M).

The results in Table 1 show that GO-LDA outperforms the classic-LDA and PCA by a large margin. In detail, for the first principle/discriminant direction, both GO-LDA and classic-LDA achieve much higher classification accuracy than PCA, indicating the dramatic discriminant ability of GO-LDA and classic-LDA against PCA (which is variance preserving of the whole data but lack of discriminant ability). After that, all the GO-LDA's discriminant directions can

TABLE 1: Classification performance comparison between PCA, classic-LDA and GO-LDA using their individual principle/discriminant directions. Six data sets all from the UCI ML repository [39] are tested, with a quadratic classifier. Recall that C , N and M respectively represent the number of classes, the number of samples and the number of features of every sample in each data set. Up to $K = 15$ principle/discriminant directions are used for PCA and GO-LDA.

Data	$C/N/M$	Method	Accuracy on Different Principle/Discriminant Direction												
			1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	\dots	15th	
IrisPlants	3/150/4	PCA	0.90	0.93	0.40	0.27									
		LDA	1.0	0.50	N/A	N/A									
		GO-LDA	1.0	0.8	0.90	0.80									
ThyroidGland	3/215/5	PCA	0.79	0.97	0.82	0.70	0.72								
		LDA	0.95	0.79	N/A	N/A	N/A								
		GO-LDA	0.95	0.88	0.86	0.74	0.86								
Glass	6/214/9	PCA	0.46	0.51	0.42	0.46	0.37	0.39	0.37	0.35	0.36				
		LDA	0.65	0.39	0.51	0.30	0.42	N/A	N/A	N/A	N/A				
		GO-LDA	0.65	0.69	0.69	0.58	0.51	0.49	0.47	0.40	0.40				
Wine	3/178/13	PCA	0.64	0.55	0.39	0.80	0.47	0.39	0.36	0.36	0.39	0.42			
		LDA	0.89	0.69	N/A										
		GO-LDA	0.89	0.86	0.88	0.81	0.72	0.67	0.67	0.69	0.64	0.67			
Landsat	6/6435/36	PCA	0.47	0.64	0.57	0.22	0.33	0.25	0.29	0.24	0.26	0.23	\dots	0.26	
		LDA	0.55	0.66	0.47	0.38	0.22	N/A	N/A	N/A	N/A	N/A	\dots	N/A	
		GO-LDA	0.55	0.73	0.64	0.62	0.63	0.62	0.53	0.59	0.52	0.45	\dots	0.46	
Handwritten Digits	10/1797/64	PCA	0.17	0.40	0.35	0.34	0.26	0.32	0.24	0.22	0.25	0.22	\dots	0.15	
		LDA	0.46	0.41	0.34	0.29	0.26	0.28	0.26	0.22	0.20	N/A	\dots	N/A	
		GO-LDA	0.46	0.46	0.47	0.48	0.45	0.46	0.46	0.36	0.39	0.42	\dots	0.32	

achieve significantly better results than that of the classic-LDA, showing GO-LDA’s discriminant optimality. In particular, after the third direction, all the GO-LDA’s discriminant directions also achieve significantly better results than the principle directions of PCA. Another great advantage of GO-LDA shown from the results is that all the discriminant directions carry important discriminant information, i.e., their discriminant ability slightly decreases (which is what we expect according to the theory in Theorem 5.2) when going to higher directions but in a quite slow manner (*cf.* the classic-LDA’s discriminant ability drops sharply to not applicable after the $(C - 1)$ direction).

6.2 Results on discriminant subspaces

Let Ω_l denote the discriminant subspace formed by the first l principle/discriminant directions of PCA, classic-LDA or GO-LDA. Therefore, for PCA and GO-LDA, l is the dimensionality of Ω_l and can be as large as M , while for the classic-LDA, l can only be as large as $(C - 1)$, which is generally much smaller than M . All the quantitative results with uncertainties in the tables below are estimated by ten-fold cross-validation.

Table 2 gives the comparison between PCA, classic-LDA and GO-LDA using the k -nearest neighbour classifier (with k set to 1 for simplicity) on projections of the data onto discriminant subspaces Ω_l in terms of the mean classification accuracy (MCA). The table shows results on five different UCI problems, two face recognition and two medical problems. Consistent results are obtained in Table 2. For the results on the discriminant subspace Ω_{C-1} , apart from a single outlier (i.e., results on the ForestTypeMapping data set with discriminant subspace Ω_3), the performance of the classic-LDA and GO-LDA outperforms PCA dramatically, and GO-LDA outperforms the classic-LDA. For the results on the discriminant subspace Ω_l where $l > C - 1$, i.e., the

subspace formed by using more principle/discriminant directions than the limit of the classic-LDA, GO-LDA achieves better results than PCA for all the cases including the case of the ForestTypeMapping data set, again demonstrating GO-LDA’s optimality of discriminant ability. In particular, we find that when more discriminant directions are used, GO-LDA’s discriminant ability increases in six of the nine problems, sometimes substantially e.g. see the case of the DeepDrid problem. This indicates that the data may have useful information over and beyond the first $(C - 1)$ directions that GO-LDA can discover but the classic-LDA is constrained to find.

Table 3 shows the results obtained similar to the way in Table 2 but with a linear classifier on different discriminant subspaces and comparing with previously published results by Duin et al. [23]. Three data sets from the UCI ML repository are tested. Similar and consistent results are obtained. In particular, our GO-LDA achieves the best results compared to PCA and classic-LDA for all the cases by a large margin. We also note that going from a discriminant subspace Ω_l formed by a small number of discriminant directions, e.g., $l \leq 3$, to one formed by a large number of discriminant directions, e.g., $l = C - 1$, GO-LDA can already achieve higher accuracies than that reported in [23].

Finally, we make similar comparisons on the highly imbalanced data sets from the KEEL repository [40]. The results are given in Table 4 using the k -nearest neighbour classifier. Again, our GO-LDA outperforms the classic-LDA for all the cases and outperforms PCA for most of the cases (i.e., eight out of the ten cases). Interestingly, compared to the previous results, the performance difference between PCA, classic-LDA and GO-LDA is not that significant on the highly imbalanced data sets, and the performance gains are minor when considering subspaces formed by more principle/discriminant directions. This might be because

TABLE 2: Classification performance comparison between PCA, classic-LDA and GO-LDA using their discriminant subspaces (Ω_l) in terms of the MCA (mean classification accuracy). Nine data sets (containing five different UCI problems, two face recognition and two medical problems) are tested, with the k -nearest neighbour classifier where k is set to 1 for simplicity. Uncertainties are estimated by ten-fold cross-validation. The results on the ForestTypeMapping data set have no uncertainty since its training and test data sets are fixed by default. Recall that Ω_l represents the discriminant subspace formed by the first l principle/discriminant directions of PCA, classic-LDA and GO-LDA individually.

Data	$C/N/M$	Method	MCA on Discriminant Subspaces	
			Ω_{C-1}	Ω_l
IrisPlants	3/150/4	PCA	0.94 ± 0.05 (on Ω_2)	0.96 ± 0.05 (on Ω_4)
		LDA	0.96 ± 0.05 (on Ω_2)	N/A
		GO-LDA	0.98 ± 0.03 (on Ω_2)	0.96 ± 0.04 (on Ω_4)
UrbanLandCover	9/168/147	PCA	0.21 ± 0.12 (on Ω_8)	0.42 ± 0.10 (on Ω_{20})
		LDA	0.31 ± 0.11 (on Ω_8)	N/A
		GO-LDA	0.43 ± 0.12 (on Ω_8)	0.56 ± 0.08 (on Ω_{20})
BreastTissue	6/106/9	PCA	0.52 ± 0.19 (on Ω_5)	0.51 ± 0.10 (on Ω_9)
		LDA	0.55 ± 0.14 (on Ω_5)	N/A
		GO-LDA	0.61 ± 0.16 (on Ω_5)	0.52 ± 0.13 (on Ω_9)
ForestTypeMapping	4/523/27	PCA	0.82 (on Ω_3)	0.83 (on Ω_{10})
		LDA	0.79 (on Ω_3)	N/A
		GO-LDA	0.76 (on Ω_3)	0.84 (on Ω_{10})
Nursery	4/12960/8	PCA	0.51 ± 0.08 (on Ω_3)	0.89 ± 0.03 (on Ω_8)
		LDA	0.86 ± 0.02 (on Ω_3)	N/A
		GO-LDA	0.90 \pm 0.01 (on Ω_3)	0.95 \pm 0.02 (on Ω_8)
LFW	5/1140/1850	PCA	0.30 ± 0.03 (on Ω_4)	0.46 ± 0.04 (on Ω_{10})
		LDA	0.66 \pm 0.02 (on Ω_4)	N/A
		GO-LDA	0.66 \pm 0.02 (on Ω_4)	0.74 \pm 0.04 (on Ω_{10})
ORL	40/400/10304	PCA	0.78 ± 0.02 (on Ω_{39})	0.97 ± 0.02 (on Ω_{50})
		LDA	0.98 ± 0.01 (on Ω_{39})	N/A
		GO-LDA	0.99 \pm 0.02 (on Ω_{39})	0.99 \pm 0.01 (on Ω_{50})
BrainTumor	4/800/512	PCA	0.38 ± 0.01 (on Ω_3)	0.39 ± 0.01 (on Ω_{10})
		LDA	0.53 ± 0.02 (on Ω_3)	N/A
		GO-LDA	0.57 \pm 0.01 (on Ω_3)	0.59 \pm 0.02 (on Ω_{10})
DeepDrid	5/662/512	PCA	0.27 ± 0.01 (on Ω_4)	0.44 ± 0.05 (on Ω_{20})
		LDA	0.27 ± 0.04 (on Ω_4)	N/A
		GO-LDA	0.29 \pm 0.04 (on Ω_4)	0.50 \pm 0.05 (on Ω_{20})

the dominated classes could be predicted well by just using the first few number of principle/discriminant directions, and the classes being dominated are insignificant for the MCA due to their small size. Further investigation of highly imbalanced data is of great interest for future work.

7 CONCLUSION

We in this paper proposed GO-LDA, deriving how discriminant directions can be sequentially extracted for multiclass data analysis and pattern classification problems through maximising Fisher criterion and retaining orthogonality to previously computed ones. The textbook solution to computing such a discriminant subspace has been by solving a generalized eigenvalue problem, i.e., the classic-LDA. Surprisingly, this solution, which has been in the literature for several decades, does not preserve the mutual orthogonality of resulting directions; nor do the resulting individual directions hold high discrimination. Moreover, the classic solution is restricted to finding a subspace, the dimensionality of which is limited by the number of classes in the problem. Our derivation of GO-LDA in this paper sequentially optimizes a set of discriminant directions that, while preserving discrimination and mutual orthogonality, can naturally go beyond the limit imposed by the rank of the

between-class scatter matrix. The excellent performance of GO-LDA was supported by illustrative examples showing the objective function (i.e., Fisher ratio) and distributions of projections as well as an extensive set of multiclass classification experiments taken from machine learning benchmark data sets with thorough comparisons.

Problems we consider for discriminant analysis may be seen as relatively small by standards of the very large-scale problems such as computer vision arising in modern machine learning. However, they are quite pertinent to validate the property and power of the proposed GO-LDA in both theoretical and practical manners. It is also well established that small data-size problems are still of paramount interest in applications such as medical diagnostics, where either due to the prevalence of a complex disease or due to restrictions arising from privacy issues that limit the amount of data available for training and validating models. Even with settings that demand non-linear classification boundaries, it is possible to have fixed nonlinear transformations using for example pre-trained deep neural networks followed by linear discrimination models acting on their feature spaces. The two medical problems we used for illustration in this paper fall precisely in this space, emphasising the importance of the study we report. In the present work,

TABLE 3: Classification performance comparison between PCA, classic-LDA and GO-LDA using their discriminant subspaces (Ω_l) in terms of the MCA. Three data sets all from the UCI ML repository [39] are tested, with a linear classifier. Please refer to the work reported in [23] for comparison against published results.

Data	$C/N/M$	Method	MCA on Discriminant Subspaces		
			$\Omega_l, l \leq 3$	Ω_{C-1}	$\Omega_l, l = \min\{M, 10\}$
Glass	6/214/9	PCA	0.45 ± 0.09 (on Ω_3)	0.54 ± 0.10 (on Ω_5)	0.61 ± 0.07 (on Ω_9)
		LDA	0.42 ± 0.11 (on Ω_3)	0.51 ± 0.09 (on Ω_5)	N/A
		GO-LDA	0.53 ± 0.09 (on Ω_3)	0.57 ± 0.06 (on Ω_5)	0.63 ± 0.07 (on Ω_9)
Landsat	6/6435/36	PCA	0.55 ± 0.09 (on Ω_3)	0.49 ± 0.10 (on Ω_5)	0.49 ± 0.09 (on Ω_{10})
		LDA	0.71 ± 0.07 (on Ω_3)	0.69 ± 0.07 (on Ω_5)	N/A
		GO-LDA	0.75 ± 0.07 (on Ω_3)	0.77 ± 0.05 (on Ω_5)	0.74 ± 0.06 (on Ω_{10})
Vowel	11/990/10	PCA	0.41 ± 0.06 (on Ω_2)	0.51 ± 0.03 (on Ω_{10})	0.51 ± 0.03 (on Ω_{10})
		LDA	0.49 ± 0.05 (on Ω_2)	0.53 ± 0.05 (on Ω_{10})	0.53 ± 0.05 (on Ω_{10})
		GO-LDA	0.50 ± 0.04 (on Ω_2)	0.54 ± 0.05 (on Ω_{10})	0.54 ± 0.05 (on Ω_{10})

TABLE 4: Classification performance comparison between PCA, classic-LDA and GO-LDA using their discriminant subspaces (Ω_l) in terms of the MCA. Five highly imbalanced data sets (indicated by the imbalanced ratio in the third column of the table below) from the KEEL repository [40] are tested, with the k -nearest neighbour classifier.

Data	$C/N/M$	Imbalanced Ratio	Method	MCA on Discriminant Subspaces	
				Ω_{C-1}	Ω_M
contraceptive	3/1473/9	1.89	PCA	0.46 ± 0.04 (on Ω_2)	0.45 ± 0.03 (on Ω_9)
			LDA	0.42 ± 0.05 (on Ω_2)	N/A
			GO-LDA	0.42 ± 0.04 (on Ω_2)	0.45 ± 0.02 (on Ω_9)
Hayes-Roth	3/132/4	1.7	PCA	0.69 ± 0.11 (on Ω_2)	0.75 ± 0.09 (on Ω_4)
			LDA	0.70 ± 0.13 (on Ω_2)	N/A
			GO-LDA	0.70 ± 0.09 (on Ω_2)	0.73 ± 0.09 (on Ω_4)
New-Thyroid	3/215/5	4.84	PCA	0.91 ± 0.05 (on Ω_2)	0.90 ± 0.06 (on Ω_5)
			LDA	0.95 ± 0.04 (on Ω_2)	N/A
			GO-LDA	0.95 ± 0.03 (on Ω_2)	0.96 ± 0.03 (on Ω_5)
Ecoli	8/336/7	71.5	PCA	0.81 ± 0.05 (on Ω_7)	0.81 ± 0.05 (on Ω_7)
			LDA	0.79 ± 0.05 (on Ω_7)	N/A
			GO-LDA	0.83 ± 0.05 (on Ω_7)	0.83 ± 0.05 (on Ω_7)
Yeast	10/1484/8	23.15	PCA	0.52 ± 0.04 (on Ω_9)	0.51 ± 0.04 (on Ω_8)
			LDA	0.52 ± 0.02 (on Ω_9)	N/A
			GO-LDA	0.52 ± 0.02 (on Ω_9)	0.52 ± 0.02 (on Ω_8)

we continue in this area of medical inference, incorporating uncertainty via probabilistic modelling in the derivation of discriminant subspaces.

Ubiquitous applications of GO-LDA are evident. For future avenues, it is of great interest to investigate highly imbalanced data and transfer the essence of GO-LDA to other LDA variants.

APPENDIX A DERIVATION OF d_2 FOR THE BINARY PROBLEM

This Appendix derives an important step in the derivation of the second discriminant direction d_2 for the binary problem, which is not shown in [2], [45].

Maximising the objective function in Eq. (7) in Section 4 with respect to d_2 can be addressed by solving

$$\frac{2\tilde{S}_B d_2}{d_2^\top S_W d_2} - \frac{2d_2^\top \tilde{S}_B d_2 S_W d_2}{(d_2^\top S_W d_2)^2} - \lambda d_1 = 0. \quad (56)$$

Note that $\tilde{S}_B = s_b s_b^\top$. Substituting it into Eq. (56) yields

$$\frac{2s_b s_b^\top d_2}{d_2^\top S_W d_2} - \frac{2d_2^\top s_b s_b^\top d_2 S_W d_2}{(d_2^\top S_W d_2)^2} - \lambda d_1 = 0. \quad (57)$$

Let $\kappa = s_b^\top d_2 / d_2^\top S_W d_2$, which is a scalar. Eq. (57) can be rewritten as

$$2\kappa s_b - 2\kappa^2 S_W d_2 - \lambda d_1 = 0. \quad (58)$$

Then we have

$$d_2 = \frac{1}{\kappa} S_W^{-1} \left(s_b - \frac{\lambda}{2\kappa} d_1 \right). \quad (59)$$

Since $d_1 = S_W^{-1} s_b$, we get

$$d_2 = \frac{1}{\kappa} \left(S_W^{-1} - \frac{\lambda}{2\kappa} (S_W^{-1})^2 \right) s_b. \quad (60)$$

Let $S_{11} = d_1^\top S_W^{-1} d_1$. Since $d_1^\top d_2 = 0$, we have

$$d_1^\top d_2 = \frac{1}{\kappa} d_1^\top S_W^{-1} s_b - \frac{\lambda}{2\kappa^2} d_1^\top S_W^{-1} d_1 = 0, \quad (61)$$

which gives

$$\frac{1}{\kappa} d_1^\top d_1 - \frac{\lambda}{2\kappa^2} S_{11} = 0. \quad (62)$$

Therefore,

$$\frac{\lambda}{2\kappa} = \frac{d_1^\top d_1}{S_{11}}. \quad (63)$$

Since

$$\begin{aligned} S_{11} &= \mathbf{d}_1^\top \mathbf{S}_W^{-1} \mathbf{d}_1 \\ &= \mathbf{s}_b^\top \mathbf{S}_W^{-1} \mathbf{S}_W^{-1} \mathbf{S}_W^{-1} \mathbf{s}_b, \\ &= \mathbf{s}_b^\top (\mathbf{S}_W^{-1})^3 \mathbf{s}_b, \end{aligned} \quad (64)$$

and $\mathbf{d}_1^\top \mathbf{d}_1 = \mathbf{s}_b^\top (\mathbf{S}_W^{-1})^2 \mathbf{s}_b$, we have

$$\frac{\lambda}{2\kappa} = \frac{\mathbf{s}_b^\top (\mathbf{S}_W^{-1})^2 \mathbf{s}_b}{\mathbf{s}_b^\top (\mathbf{S}_W^{-1})^3 \mathbf{s}_b}. \quad (65)$$

Substituting it into Eq. (60), we have

$$\mathbf{d}_2 = \frac{1}{\kappa} \left(\mathbf{S}_W^{-1} - \frac{\mathbf{s}_b^\top (\mathbf{S}_W^{-1})^2 \mathbf{s}_b}{\mathbf{s}_b^\top (\mathbf{S}_W^{-1})^3 \mathbf{s}_b} (\mathbf{S}_W^{-1})^2 \right) \mathbf{s}_b. \quad (66)$$

Normalising \mathbf{d}_2 above, we then complete the derivation.

APPENDIX B COMBINATION OF GO-LDA's DISCRIMINANT DIRECTIONS

When extracting multiple mutually orthogonal discriminant directions with GO-LDA, an intriguing observation is that even when individual directions do not necessarily separate all the classes, their combinations do as illustrated in Fig. 5. The data set used here is from the Wine problem, a benchmark classification problem sourced from the UCI ML repository [39], consisting of 178 13-dimensional samples that are categorised into three distinct classes.

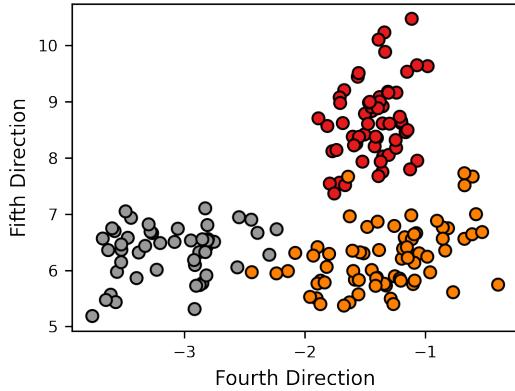


Fig. 5: Projection of the three-class Wine data set on two of GO-LDA's discriminant directions. It shows that, when projected along the two directions \mathbf{u}_4 and \mathbf{u}_5 , this three-class data set continues to show separation between classes. This would not have been possible with the classic-LDA's solution from which only two discriminant directions could have been extracted; therefore, individuals seeking to acquire additional discriminant directions for further investigation would find the fourth and fifth directions of limited applicability in the classic-LDA.

ACKNOWLEDGMENTS

MN's contribution to the work was partially funded by Engineering and Physical Sciences Research Council (EPSRC) grant "Early detection of contact distress for enhanced performance monitoring and predictive inspection of machines" (EP/S005463/1).

REFERENCES

- [1] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
- [2] D. H. Foley and J. W. Sammon, "An Optimal Set of Discriminant Vectors," *IEEE Transactions on Computers*, vol. 100, no. 3, pp. 281–289, 1975.
- [3] C. R. Rao, "The Utilization of Multiple Measurements in Problems of Biological Classification," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 10, no. 2, pp. 159–203, 1948.
- [4] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006, vol. 4, no. 4.
- [5] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," *The Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.
- [6] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Muller, "Fisher Discriminant Analysis with Kernels," in *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (cat. no. 98th8468)*. Ieee, 1999, pp. 41–48.
- [7] G. Baudat and F. Anouar, "Generalized Discriminant Analysis using a Kernel Approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.
- [8] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, "Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 762–766, 2001.
- [9] X. Li, W. Hu, H. Wang, and Z. Zhang, "Linear Discriminant Analysis using Rotational Invariant L1 Norm," *Neurocomputing*, vol. 73, no. 13–15, pp. 2571–2579, 2010.
- [10] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher Discriminant Analysis with L1-Norm," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 828–842, 2013.
- [11] F. Zhong and J. Zhang, "Linear Discriminant Analysis Based on L1-Norm Maximization," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3018–3027, 2013.
- [12] Y. Liu, Q. Gao, S. Miao, X. Gao, F. Nie, and Y. Li, "A Non-Greedy Algorithm for L1-Norm lda," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 684–695, 2016.
- [13] A. M. Martinez and A. C. Kak, "PCA Versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.
- [14] J. H. Friedman, "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.
- [15] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization Studies of Linear Discriminant Analysis in Small Sample Size Scenarios with Application to Face Recognition," *Pattern Recognition Letters*, vol. 26, no. 2, pp. 181–191, 2005.
- [16] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
- [17] K. Liu, Y.-Q. Cheng, J.-Y. Yang, and X. Liu, "An Efficient Algorithm for Foley-Sammon Optimal set of Discriminant Vectors by Algebraic Method," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 6, no. 05, pp. 817–829, 1992.
- [18] W. Liu, Y. Wang, S. Z. Li, and T. Tan, "Null Space-based Kernel Fisher Discriminant Analysis for Face Recognition," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings*. IEEE, 2004, pp. 369–374.
- [19] L. Yang, W. Gong, X. Gu, W. Li, and Y. Liang, "Null Space Discriminant Locality Preserving Projections for Face Recognition," *Neurocomputing*, vol. 71, no. 16–18, pp. 3644–3649, 2008.
- [20] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A New LDA-Based Face Recognition System which can Solve the Small Sample Size Problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.
- [21] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Elsevier, 2013.

- [22] M. Loog and R. P. Duin, "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)*. Springer, 2002, pp. 508–517.
- [23] R. P. Duin and M. Loog, "Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: the Chernoff Criterion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 732–739, 2004.
- [24] T. Hastie and R. Tibshirani, "Discriminant Analysis by Gaussian Mixtures," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 155–176, 1996.
- [25] M. Zhu and A. M. Martinez, "Subclass Discriminant Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1274–1286, 2006.
- [26] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki, "Mixture Subclass Discriminant Analysis Link to Restricted Gaussian Model and Other Generalizations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 1, pp. 8–21, 2012.
- [27] M. Sugiyama, "Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis." *Journal of Machine Learning Research*, vol. 8, no. 5, 2007.
- [28] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality Sensitive Discriminant Analysis." in *IJCAI*, vol. 2007, 2007, pp. 1713–1726.
- [29] Y. Zhou and S. Sun, "Manifold Partition Discriminant Analysis," *IEEE Transactions on Cybernetics*, vol. 47, no. 4, pp. 830–840, 2016.
- [30] K. Fukunaga and J. Mantock, "Nonparametric Discriminant Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 671–678, 1983.
- [31] Z. Li, D. Lin, and X. Tang, "Nonparametric Discriminant Analysis for Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 755–761, 2009.
- [32] J. Yang, L. Zhang, J.-y. Yang, and D. Zhang, "From Classifiers to Discriminators: A Nearest Neighbor Rule induced Discriminant Analysis," *Pattern Recognition*, vol. 44, no. 7, pp. 1387–1402, 2011.
- [33] F. Zhu, J. Gao, J. Yang, and N. Ye, "Neighborhood Linear Discriminant Analysis," *Pattern Recognition*, vol. 123, p. 108422, 2022.
- [34] S. D. Fabiyi, P. Murray, J. Zabalza, and J. Ren, "Folded LDA: Extending the Linear Discriminant Analysis Algorithm for Feature Extraction and Data Reduction in Hyperspectral Remote Sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 12312–12331, 2021.
- [35] L. Meirovitch and M. K. Kwak, "Convergence of the Classical Rayleigh-ritz Method and the Finite Element Method," *AIAA Journal*, vol. 28, no. 8, pp. 1509–1516, 1990.
- [36] T. Li, S. Zhu, and M. Oghihara, "Using Discriminant Analysis for Multi-class Classification: an Experimental Investigation," *Knowledge and Information Systems*, vol. 10, pp. 453–472, 2006.
- [37] J. W. Daniel, W. B. Gragg, L. Kaufman, and G. W. Stewart, "Re-orthogonalization and Stable Algorithms for Updating the Gram-schmidt Factorization," *Mathematics of Computation*, vol. 30, no. 136, pp. 772–795, 1976.
- [38] G. Strang, *Linear Algebra and its Applications*. Belmont, CA: Thomson, Brooks/Cole, 2006.
- [39] D. Dua and C. Graff, "UCI Machine Learning Repository," 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [40] J. Derrac, S. Garcia, L. Sanchez, and F. Herrera, "Keel Data-mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework," *J. Mult. Valued Logic Soft Comput*, vol. 17, 2015.
- [41] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.
- [42] "ORL Face Database AT&T laboratories, cambridge, U.K," <http://www.cam-orl.co.uk/facedatabase.html>, accessed: 2010-09-30.
- [43] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. Abou Elwafa, and H. Kurdi, "Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain," *Applied Sciences*, vol. 11, no. 2, p. 796, 2021.
- [44] R. Liu, X. Wang, Q. Wu, L. Dai, X. Fang, T. Yan, J. Son, S. Tang, J. Li, Z. Gao, A. Galdran, J. Poorneshwaran, H. Liu, J. Wang, Y. Chen, P. Porwal, G. S. Wei Tan, X. Yang, C. Dai, H. Song, M. Chen, H. Li, W. Jia, D. Shen, B. Sheng, and P. Zhang, "Deepdrid: Diabetic Retinopathy—Grading and Image Quality Estimation Challenge," *Patterns*, p. 100512, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666389922001040>
- [45] J. W. Sammon, "An Optimal Discriminant Plane," *IEEE Transactions on Computers*, vol. 100, no. 9, pp. 826–829, 1970.