processes

MDPI

Review

A Comprehensive Review on Discriminant Analysis for Addressing Challenges of Class-Level Limitations, Small Sample Size, and Robustness

Lingxiao Qu [1] **and Yan Pei** [2,*]

[1] Graduate School of Computer Science and Engineering, University of Aizu, Aizuwakamatsu 965-8580, Japan; d8232113@u-aizu.ac.jp
[2] Computer Science Division, University of Aizu, Aizuwakamatsu 965-8580, Japan
* Correspondence: peiyan@u-aizu.ac.jp; Tel.: +81-(242)-37-2765

Abstract: The classical linear discriminant analysis (LDA) algorithm has three primary drawbacks, i.e., small sample size problem, sensitivity to noise and outliers, and inability to deal with multi-modal-class data. This paper reviews LDA technology and its variants, covering the taxonomy and characteristics of these technologies and comparing their innovations and developments in addressing these three shortcomings. Additionally, we describe the application areas and emphasize the kernel extensions of these technologies to solve nonlinear problems. Most importantly, this paper presents perspectives on future research directions and potential research areas in this field.

Keywords: discriminant analysis; small sample size; singularity; multi-modality; kernel method

check for updates

Citation: Qu, L.; Pei, Y. A Comprehensive Review on Discriminant Analysis for Addressing Challenges of Class-Level Limitations, Small Sample Size, and Robustness. *Processes* **2024**, *12*, 1382. https://doi.org/10.3390/pr12071382

Academic Editor: Gorazd Karer

Received: 21 May 2024
Revised: 14 June 2024
Accepted: 26 June 2024
Published: 2 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Linear discriminant analysis (LDA) is a classical linear learning method, being first proposed in 1936 by R.A. Fisher, also known as Fisher discriminant analysis (FDA) [1] for binary classification. In a strict sense, LDA grounds the assumption that the covariance matrix of all classes of sample data is the same and full-rank, which slightly differs from FDA [2]. LDA is a supervised data analysis method using an orthogonal transformation that makes the process and analysis of data more efficient and convenient.

The definition for the term "perpendicular" by Euclid, a synonym for "orthogonal", is depicted in [3]. Mathematically and specifically, a linear transformation is a mapping from one vector space to another that preserves the operations of vector additions and scalar multiplications. Especially in linear algebra, an orthogonal transformation $T: V \to V$ is a linear transformation conducted in an inner product space V, which preserves the inner product of two vectors $u, v \in V$: $< u, v >=< Tu, Tv >$; accordingly, it preserves the norms of vectors and angles between vectors. Orthogonality is a significant property, meaning a certain kind of non-dependence of things, that makes components separated for clearer and easier observations, analysis and manipulation. From the viewpoint of mathematics, in a vector space, any signal is a vector that can be represented by a set of orthogonal bases, being decomposed into uncorrelated components along different axes as much as possible. This meets the need for processing and analyzing data more effectively and conveniently. Orthogonal transform rotates signals with orthogonal components, from one set of orthogonal bases to another one but more approximately or accurately, or more proper for a favorable aim, preserving the equivalent representations of identical inner products between vectors before and after the transform [4]. This addresses quite a lot of necessary needs for varied data processing and analysis across a wide range of various areas and fields.

Notably, it is necessary to talk about the difference between statistical analysis and data analysis in the introduction of this paper, both being always mentioned along with each

other, which was the first question that bothered us at the beginning of our investigation. Are the terms statistical analysis and data analysis, which often appear together, equivalent? Are they interchangeable? Can the term statistical analysis be an alternative for data analysis in this paper? The answer is no. Data analysis is to collect, clean, and learn the original data set aiming to gain an insight into hidden and potential information under the data. Statistical analysis draws inferences about the larger population using part quantitative samples. Statistical analysis can infer what is beyond the data analysis. Given a brief description of the differences between the two terms from some perspectives, it is indicated that this paper limits the research scope under the data analysis. Using different methods to compare the same data with the same aim has become a big benefit for data analysis, said by Tukey [5], and what is taken to include by data analysis is analyzing procedures, results interpreting techniques, data collection methods for easier, more precise or more accurate analysis, and mathematical statistics tools applied to analyzing data.

In statistical discriminant analysis, the betweenclass scatter matrix S_b as shown in Equation (1), the within-class scatter matrix S_w as shown in Equation (2), and the total (or mixture) scatter matrix S_t as shown in Equation (3) are utilized for the formulations of the class separability criterions. For a L-classes problem of N samples $x \in \mathbb{R}^d$ in the d-dimensional original space, the scatter matrices are defined as follows [1].

$$S_b = \frac{1}{N}\sum_{j=1}^{L} N_j (m_j - m)(m_j - m)^T, \tag{1}$$

$$S_w = \frac{1}{N}\sum_{j=1}^{L} N_j \sum_{x \in L_j} (x - m_j)(x - m_j)^T, \tag{2}$$

$$S_t = S_b + S_w = \frac{1}{N}\sum_{x} (x - m)(x - m)^T, \tag{3}$$

where N_j denotes the size of samples in class L_j, $j = 1, \cdots, L$, $N = \sum_{j=1}^{L} N_j$, m_j and m denote the class mean of L_j and the total mean, respectively. S_t is the total scatter matrix of all samples regardless of their class assignments [6]. The target of LDA is to obtain the transform vector v that satisfies Equation (4).

$$v = \arg\max_{v} \frac{v^T S_b v}{v^T S_w v}. \tag{4}$$

The objective transform vector is the most discriminative projection direction of the maximum distance between classes and the minimum variance within each class, as shown in Figure 1.

In Equation (1), the rank of vector $m_j - m$ is 1, so that the rank of S_b is at most L after summing up all vectors of L classes. Due to the nonlinear independence between all m_j of L classes and m, that is the m_j of $(L-1)$ and m can linearly represent the L-th m_j, it is inferred that $\max(rank(S_b)) = L - 1$. So as to that of $\max(rank(S_w^{-1} S_b))$ because of $rank(AB) \leq \min\{rank(A), rank(B)\}$. From Equation (4), consequently, there are at most $L - 1$ non-zero eigenvalues and valid eigenvectors, respectively. In consequence, the reduced space by LDA is of at most $L - 1$ dimensions.

Nonetheless, LDA has three main drawbacks: inapplicability for multi-modal datasets; the singularity of the within-class scatter matrix; and insufficient robustness against outliers and noises [7,8]. We present the technical causes, the existing cases and the resulting bad influence of three drawbacks of conventional LDA in Table 1. In the last few decades, the purpose of mitigating the three drawbacks has been motivating the rush towards many extensions to LDA around a wide variety of disciplines and areas.

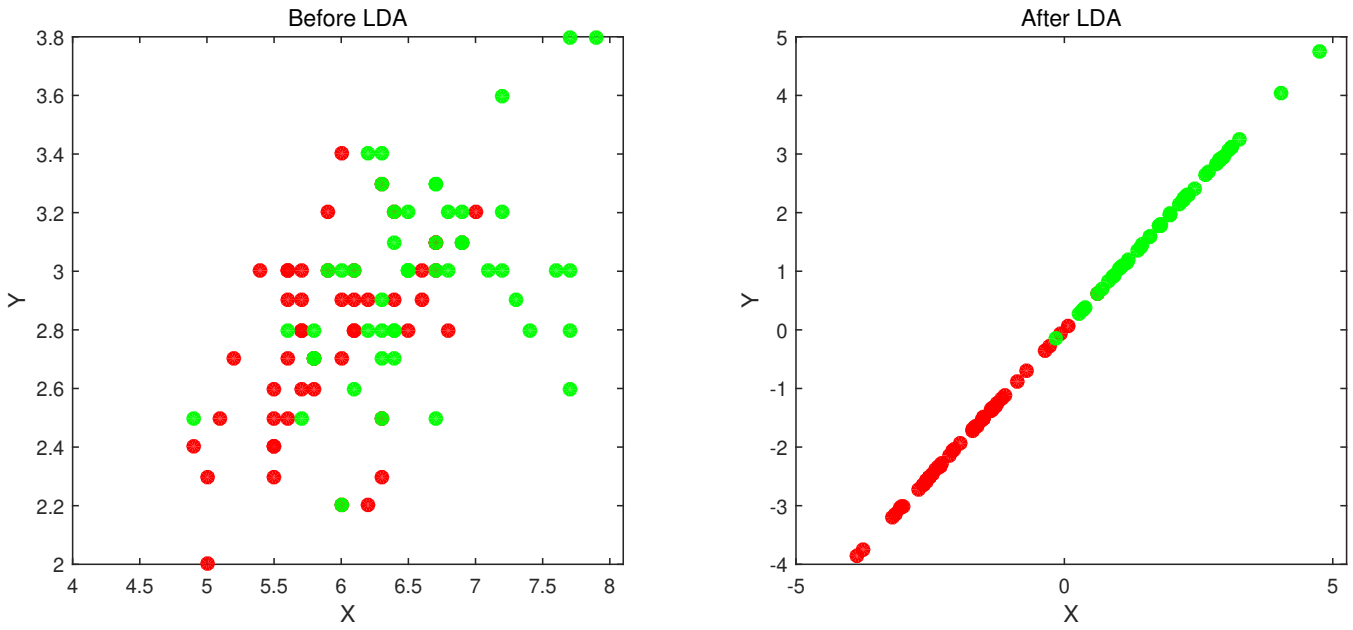

Figure 1. A two-class case before and after using LDA.

Table 1. The types, causes, situation of existence, and negative impacts of the three drawbacks of LDA.

Types of Drawbacks	Cause	Existence	Negative Impacts
Inapplicability of multi-modality	The i.i.d. assumption fails.	Classes are multi-modal containing some sub-classes or clusters.	LDA may make the information in classes be distorted and inseparable.
Small Sample Size (SSS)	S_w is (almost) singular.	Training samples are of high dimension but small size.	The singularity of S_w leads to severe instability and over-fitting.
Unrobustness	L_2 norm in LDA is unrobust.	Outliers exist in training samples.	Projection vectors may drift from the target directions.

In this paper, we intend to review the last two to three decades of articles working on suppressing the affection of three drawbacks of LDA, summarize the taxonomy of techniques and their applications, and compare the characteristics and innovation of those methods. For clarity, we build Table 2 for the name abbreviations of the reviewed methods regarding their references.

Our objective is to provide the readers with comprehensive knowledge about the existence, the primary philosophy, the solutions, and the corresponding applications of the three drawbacks of conventional discriminant analysis. This knowledge can guide the readers on how these methods play their roles in the machine learning area based on various benefits and usage, to identify the real knowledge differences from past to present, which motivates us to make this article serve as the foundation of developing theory and predicting the heading directions in discriminative data analysis area. How best to guide readers from the underlying mathematical theories to the technical comparisons, to the application for the realistic situations, and finally to the future development is the originality of this paper. Motivation has been the guiding principle throughout the writing of this review article.

Table 2. The name abbreviations of the reviewed methods from their references.

Method Names and References	Abbreviations
Linear discriminant analysis [1]	LDA
Fisher discriminant analysis [1]	FDA
Mixture discriminant analysis [9]	MDA
Subclass discriminant analysis [10]	SDA
Mixture subclass discriminant analysis [11]	MSDA
Separability oriented subclass discriminant analysis [12]	SSDA
Resilient SDA [13]	RSDA
One MSDA extension [14]	EM-MSDA
Fractional step MSDA [14]	FSMSDA
Local Fisher discriminant analysis [15,16]	LFDA
Locality-preserving projection [17]	LPP
Locality Sensitive discriminant analysis [18]	LSDA
Manifold partition discriminant analysis [19]	MP-DA
Adaptive and fuzzy locality discriminant analysis [20]	AFLDA
Approximate pairwise accuracy criterion [21,22]	aPAC
Penalized discriminant analysis [23]	PDA
Heteroscedastic linear dimension reduction [24]	HLDR
Local mean based nearest neighbor discriminant analysis [25]	LM-NNDA
Nonparametric discriminant analysis [26]	NDA
Neighborhood linear discriminant analysis [27]	nLDA
Eigenspectrum regularization reverse neighborhood discriminative learning [28]	ERRNDL
An alternative of null-space-based LDA methods [29]	Fast NLDA
Direct LDA [30]	DLDA
Orthogonal LDA [31]	OLDA
Rotational invariant L_1-norm based LDA [32]	LDA-R_1
L_1-norm based LDA [33]	LDA-L_1
Two-dimensional LDA with L_1-norm [34]	L_1-2DLDA
L_1-norm based LDA [35]	L_1-LDA
Kernel extension of L_1-norm based LDA [35]	L_1-KDA
Kernel discriminant analysis [36]	KDA
A robust LDA measured by $L_{2,1}$-norm [37]	$L_{2,1}$-LDA
L_p-norm based LDA [38]	LDA-L_p
A bilateral two-dimensional LDA using the L_p-norm [39]	BL_p2DLDA
LDA measured by L_s-and L_p-norm [40]	FLDA-L_{sp}

The rest of this paper is organized as follows. Section 2 summarizes and compares the methods and techniques of LDA variants addressing three drawbacks especially. Section 3 summarizes the main applications of LDA variants from solving each drawback. Section 4 summarizes the review methods that have been extended to the kernel version and discusses the ones that can be extended to the kernel version. Finally, Section 5 is the conclusion of this paper.

2. LDA Extensions: Variations in Principle

2.1. LDA Variants for Multi-Modal Classes

LDA relies on the assumption that all data samples of the same class are independently and identically distributed (i.i.d.). It can be described as a maximum likelihood estimation for Gaussian distributions for each class with common covariance and distinct means for different classes. In case the assumption fails, the original LDA with class-level scatter matrices cannot deal with the cases that the class is multi-modal containing some independent sub-classes or clusters [6,41]. Moreover, in practice, due to complex nonlinear distributions, outliers and any possible real factors, segmenting classes into different sub-classes is in favor of making them more separable, which preserves the information involved in the multi-modal structure.

2.1.1. Using a Mixture of Gaussians

Approximating the fundamental distribution in every class as the mixture of Gaussians is a good way of describing a big size of data distributions, no matter if they correspond with compact sets [42,43].

To perform classification effectively, the proposal mixture discriminant analysis (MDA) [9] fits Gaussian mixtures to each class especially when there are sub-classes. MDA has the feature that the classes are structured as mixtures of Gaussian distributions, instead of only Gaussian distributions as in traditional LDA. To be specific, using Gaussian mixtures for modelling the class densities of the predictors $P(X|G)$. Suppose N training data with label set $\mathcal{L}$: $(x_i, g_i) \in \mathbb{R}^d \times \mathcal{L}, i = 1, 2, \cdots, N$, and g_i means the label of x_i. Dividing each class L_j into $R_j, j = 1, 2, \cdots, L$ artificial sub-classes presented as $l_{jr}, r = 1, 2, \cdots, R_j$. The model supposes that every subclass follows a multivariate normal distribution as well as the mean vector u_{jr} of itself and a shared common covariance matrix Σ. Let $\prod_j$ be the prior probability of class L_j, and make π_{jr} be the mixing probability of the r-th subclass inside the class L_j. The class L_j has its mixture density, as shown in Equation (5),

$$\begin{aligned} m_j(x) &= P(X = x|G = j) \\ &= |2\pi\Sigma|^{-1/2} \sum_{r=1}^{R_j} \pi_{jr} \exp\{-D(x, u_{jr})/2\}, \end{aligned} \tag{5}$$

and the conditional log-likelihood of the data is shown in Equation (6).

$$l^{mix}(u_{jr}, \Sigma, \pi_{jr}) = \sum_{i=1}^{N} m_{g_i}(x_i). \tag{6}$$

To maximize $l^{mix}(\theta)$, executing expectation-maximization (EM) algorithm [44] iteratively via Bayes theorem, obtaining the posterior class probabilities $P(G = j|X = x)$ and maximizing it for optimal classification.

From the above perspective, MDA applies the EM algorithm firstly for estimating the real underlying distribution inside every class ahead of utilizing LDA. It has to be mentioned that the EM algorithm can match a mixture of Gaussians efficiently only when the sample number is very large. Besides the standard MDA, Bashir et al. [45] consider the case that in Gaussian mixture models, the estimators for those unknown parameters inside the EM algorithm are affected due to outliers, which results in the non-robustness. They substitute those unrobust estimators of the M-step in the EM algorithm for the robust S-estimators, defined as having higher breakdown points, of the unknown parameters, where the compared results show that the average probability of misclassification reduces slightly than the standard mixture discriminant analysis. This proposal is called high breakdown mixture discriminant analysis.

There is a different method using a mixture of Gaussian, addressing the goal of optimizing classification, rather than recovering the real underlying but unknown data distribution, that is, subclass discriminant analysis (SDA) [10]. It defines criteria to ensure the best amount of Gaussians in each class, which is the number of sub-classes, and uses nearest neighbor-based clustering to divide classes into sub-classes. The target is solving the generalized eigenvalue decomposition problem to find the optimal discriminant vectors for the classification, as shown in Equation (7).

$$S_b V = \Sigma V \Lambda, \tag{7}$$

where Σ represents the covariance matrix of the data, V denotes the eigenvector matrix, Λ is the relevant diagonal eigenvalue matrix. S_b is the between-subclass scatter matrix represented as Equation (8).

$$S_b = \sum_{j=1}^{L-1} \sum_{r=1}^{R_j} \sum_{k=j+1}^{L} \sum_{l=1}^{R_k} p_{jr} p_{kl} (u_{jr} - u_{kl})(u_{jr} - u_{kl})^T, \quad (8)$$

where L is classes number, R_j is the size of sub-classes divisions in class L_j, p_{jr} is the prior of the r-th subclass of class L_j, and u_{jr} is the mean of the r-th subclass in class L_j.

Notably, Gkalelis et al. mention two shortcomings of SDA. One is SDA does not guarantee the minimum Bayes error, and another one is the covariance matrix does not work in the minimization of the discriminant analysis stability [11]. To this end, in [11], the authors propose a modified SDA using a partitioning procedure that alleviates the aforementioned two shortcomings, referred to as mixture subclass discriminant analysis (MSDA). SDA and MSDA divide each class into sub-classes and reformulate the within-subclass and between-subclass matrices.

Wan et al. make a study on obtaining the sub-classes more efficiently to realize higher separation of classes, named separability oriented subclass discriminant analysis (SSDA) [12]. In particular, to lessen the overlap between models of sub-classes, the authors utilize clustering to separate each class into sub-classes based on the criterion of separability-oriented and then redefine scatter matrices for discriminant analysis. The experimental results show that, compared to LDA, SDA, and MSDA, SSDA performs better and has higher class separation in most cases. SSDA differs from LDA due to the existence of a subclass. SSDA differs from SDA/MSDA due to different criteria for separating a class into sub-classes and redefining the subclass-based scatter matrices. Specifically, SSDA aims to divide each class into distinct sub-classes and further separate the distinct classes.

It is pointed out by Wu et al. that in Gaussian mixture models, the EM algorithm normally needs a big size of samples to estimate the mixture parameters accurately, so it may be unstable for small dataset problems. Wu et al. propose the resilient SDA (RSDA) [13] with a modified EM algorithm by first projecting the data into the space of much lower dimensionality of highest class separation and clustering the mapped data to the novel space. In comparison with the conventional EM algorithm, RSDA improves the robustness of clustering the mixtures of Gaussians regardless of the sample size and the modified subclass-based covariance matrices are quite smaller to be easier for inversion, and also, lower the computational cost because the most costly step of assigning samples to each subclass is conducted on a much lower dimensional space. Additionally, RSDA uses a stepwise cross-validation procedure to determine the optimal number of subclasses, rather than an exhaustive search, significantly reducing computational cost.

SDA works well with higher dimensionality sub-spaces as the dimensionality of the learned feature space is limited by the between-subclass matrix that is limited under the entire amount of sub-classes. One of the major disadvantages of low speed in the case of large numbers and large dimensionality of a dataset of SDA is presented by Chumachenko et al. [46]. To this end, the authors propose the speed-up SDA method that overcomes both the low speed and the limited dimension of the subspace. Specifically, this method is based on graph embedding and spectral regression approaches, where the exploitation of the between-class Laplacian matrix makes the eigendecomposition process quite faster. The authors formulate a multi-view SDA criterion, allowing the method to be used for the multi-view data.

The conventional MDA and SDA separate subclass within each isolated class before addressing the generalized eigenvalue issue, which ignores the relations between each class and may not keep the locality in the original data space leading to the unguaranteed classification performance. A novel iterative subclass separation method of the EM-like framework is presented to solve such questions [47]. The authors seek the eigenvectors and operate subclass division by k-means clustering class by class in the projected space iteratively under an EM-alike framework. Compared to conventional MDA and SDA,

the experimental results demonstrate this method has a better performance and costs a bit more time.

The authors in [14] extend MSDA in three ways. The first one is called the EM-MSDA. It estimates the optimal amount of mixed components for every Gaussian mixture density iteratively, where during every iteration a novel Gaussian model is determined. The second one is the fractional step MSDA (FSMSDA) solving the subclass division problem. Specifically, the dimension of the learned subspace is completely lower than the rank of the subspace-based scatter matrix. It is solved by a proper weighting strategy with an iterative algorithm. The third one is the kernel extension for nonlinear problems.

2.1.2. Using Manifold (Laplacian Graph)

In this part, we will summarize the extensions of LDA methods of using manifold to exploit the local data structure, to a degree, which can be regarded as the joint study of LDA and manifold learning. As is well known, graphs are usually taken as the proxy of a manifold. More specifically, these methods all fall into graph Laplacian-based framework, applying the Laplacian matrix on specific graphs to depict the local data structure, and thus projecting the nearby data of the same labels to the reduced space as closely as possible, whereas those nearby data of distinct labels are projected as far as possible.

There is an essential limitation of FDA that it only works in the case that the dimensionality of embedding space is smaller than class numbers due to the rank deficiency of the between-class scatter matrix [6]. Here is another essential problem for multi-modal dimensionality reduction to protect the local structure of the data from being changed. The proposed local Fisher discriminant analysis (LFDA) [15,16] makes a combination of the FDA and locality-preserving projection (LPP) [17] without losing the local structure. It is one of the most typical LDA extensions and could be regarded as a supervised modification of LPP. This paper's proposal overcomes the limitation of rank deficiency in S_b by reducing dimensionality into an arbitrary dimensional space.

LPP is a linear dimensionality reduction technique projecting the data along the directions of maximal variances and optimally preserving the neighborhood structure of the data. LPP uses one graph to model the geometrical structure in the data. The high dimensional samples are located on a low dimensional manifold, where LPP is found by searching the best linear approximations for the eigenfunctions of the Laplace–Beltrami operator.

Given N samples, $\{x_i | x_i \in \mathbb{R}^d, i = 1, \cdots, N\}$. Find a transformation matrix T mapping these N points to $\{y_i | y_i \in \mathbb{R}^l, l \ll d, i = 1, \cdots, N\}$, where $y_i = T^T x_i, T = (t_0, t_1, \cdots, t_{l-1})$ denotes the transformed sample from x_i by T. The objective of LPP is to seek the transformation matrix T that meets Equation (9).

$$\min \frac{1}{2} \sum_{ij} (y_i - y_j)^2 A_{ij}. \tag{9}$$

The first step of the LPP algorithm is constructing the adjacency graph with n nodes by putting an edge connecting the nodes i and j if x_i and x_j are close, defined under certain criteria. It chooses the *affinity matrix* A_{ij} for the edge joining vertices i and j by certain variations. One common variant is by heat kernel: $A_{ij} = e^{-\frac{||x_i - x_j||}{t}}, t \in \mathbb{R}$. Based on the objective Equation (9), when x_i and x_j are far apart, the corresponding A_{ij} will be small. When x_i and x_j are close, A_{ij} will be large, so y_i and y_j will consequently also be close to meet the minimizing requirement of Equation (9).

After a simple introduction of LPP, it goes back to LFDA. Suppose $x_i \in \mathbb{R}^d$, $i = 1, 2, \cdots, N$ is the original training data of d-dimension, N is the amount of training data and N_l is the sample amount of class L_l. The local within-class scatter matrix $\widetilde{S}^{(w)}$

and the local between-class scatter matrix $\widetilde{S}^{(b)}$ of LFDA are defined in a pairwise expression as shown in Equations (10)–(13).

$$\widetilde{S}^{(w)} = \frac{1}{2}\sum_{i,j=1}^{N} \widetilde{W}_{i,j}^{(w)}(x_i - x_j)(x_i - x_j)^T, \tag{10}$$

$$\widetilde{S}^{(b)} = \frac{1}{2}\sum_{i,j=1}^{N} \widetilde{W}_{i,j}^{(b)}(x_i - x_j)(x_i - x_j)^T, \tag{11}$$

where

$$\widetilde{W}_{i,j}^{(w)} = \begin{cases} A_{i,j}/N_l & if\ y_i = y_j = l, \\ 0 & if\ y_i \neq y_j, \end{cases} \tag{12}$$

$$\widetilde{W}_{i,j}^{(b)} = \begin{cases} A_{i,j}(1/N - 1/N_l) & if\ y_i = y_j = l, \\ 1/N & if\ y_i \neq y_j. \end{cases} \tag{13}$$

The affinity $A_{i,j}$ weights for the data pairs of the same class, from which the far apart samples of the same class make few effects on $\widetilde{S}^{(w)}$ and $\widetilde{S}^{(b)}$. Moreover, the samples in different classes are irrespective of affinity since they are expected to be separated from each other. The objective function of LFDA is to obtain the transformation matrix T_{LFDA} shown as Equation (14).

$$T_{LFDA} = \arg\max_{T\in\mathbb{R}^{d\times r}} tr\left(\frac{T^T\widetilde{S}^{(b)}T}{T^T\widetilde{S}^{(w)}T}\right), \tag{14}$$

It makes the neighbored data pairs in the same class close; far apart ones are not imposed, while the data pairs in distinct classes are apart.

The authors in [18] propose the locality sensitive discriminant analysis (LSDA) algorithm. It preserves the locality and discriminant properties of the data. Specifically, they model the local geometry of the underlying manifold by constructing a nearest-neighbor graph. Assume N data points $\{x_i \in \mathbb{R}^d | i = 1, \cdots, N\}$ sampled from the underlying submanifold $\mathcal{M}$. To model the local geometrical structure of $\mathcal{M}$, the authors construct the nearest neighbor graph G by finding the k nearest neighbors set $N(x_i) = \{x_i^1, \cdots, x_i^k\}$ of x_i and putting edges between x_i and its neighbors. The nearest neighbor graph G with its weight matrix W depicts the local geometric structure of $\mathcal{M}$. Next, splitting the graph G into within-class graph G_w and between-class graph G_b. The $N(x_i)$ can be split into $N_b(x_i)$ and $N_w(x_i)$ shown as Equations (15) and (16), containing the neighbors with the same and distinct labels with x_i.

$$N_w(x_i) = \{x_i^j | x_i^j\ has\ same\ label\ to\ x_i, 1 \leq j \leq k\}. \tag{15}$$

$$N_b(x_i) = \{x_i^j | x_i^j\ has\ distinct\ label\ to\ x_i, 1 \leq j \leq k\}. \tag{16}$$

Accordingly, the weight matrix W is split into W_w and W_b, as shown in Equations (17) and (18), respectively, corresponding to G_w and G_b.

$$W_{w,ij} = \begin{cases} 1, & if\ x_i \in N_w(x_j)\ or\ x_j \in N_w(x_i) \\ 0, & otherwise. \end{cases} \tag{17}$$

$$W_{b,ij} = \begin{cases} 1, & if\ x_i \in N_b(x_j)\ or\ x_j \in N_b(x_i) \\ 0, & otherwise. \end{cases} \tag{18}$$

It identifies a linear transformation matrix to project the data into a reduced space, ensuring that closely related samples with the same label remain near each other, while closely

related samples with different labels are separated by a greater distance. The objective is to optimize the following functions by the eigendecomposition, shown as Equation (19).

$$\max \sum_{ij} \frac{(y_i - y_j)^2 W_{b,ij}}{(y_i - y_j)^2 W_{w,ij}}, \tag{19}$$

where $y_i = v^T x_i, i = 1, \cdots, N$ is the projected value mapped from x_i into reduced space by the projection vector v. The LSDA has been extended into reproducing kernel Hilbert space (RKHS) by kernel method in this paper.

Besides the pairwise difference being considered, the proposal in [19] establishes a manifold representation that also characterizes piecewise regional consistency of potential manifold, called manifold partition discriminant analysis (MPDA). It splits the manifold into some regional ones in a piecewise manner and represents the partitioned manifold using the first-Taylor expansion based on both pairwise differences as well as piecewise regional consistency for the manifold. Thus, MPDA can obtain the projection matching the local change in the underlying manifold.

There is a more robust proposal that eliminates the interference of noise and redundancy by Wang et al., named adaptive and fuzzy locality discriminant analysis (AFLDA) [20]. The potential submanifold structures are learned through the subclass partition. An adaptively updated fuzzy membership matrix is designed to learn the multi-modal data, promising an optimized subspace to alleviate the impact of noise and redundant information.

2.1.3. Setting Weights for LDA

Incorporating weights into the estimation of matrices is another strategy to flexibly reduce or penalize the effects of unstable distributed data. It allows a slight escape from the Gaussian distribution assumption, which is an advantage over LDA, whose data follow the normal distribution.

In addition to LFDA and NDA depicted previously, two alternatives also using a weight version of the original LDA will be introduced here. The first one is the approximate pairwise accuracy criterion (aPAC) [21,22]. It modifies by redefining the matrix S_b shown in Equation (20),

$$S_b = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \omega(\Delta_{ij})(m_i - m_j)(m_i - m_j)^T, \tag{20}$$

where L is the number of classes, m_i is the mean of class L_i, and Δ_{ij} is the Mahalanobis distance between classes L_i and L_j. $\omega(\Delta_{ij})$ is a weighting function depending on Δ_{ij}, which contributes to every class pair being equivalent to the accuracy of the classification.

Another method is penalized discriminant analysis (PDA) [23] by redefining the matrix S_w. It introduces a penalizing matrix Ω onto S_w, rewritten as Equation (21).

$$S_w = \Sigma_w + \Omega, \tag{21}$$

where Σ_w is the unpenalized within-class scatter matrix. By weighting the features according to their proportion, the noise eigenvectors can be effectively penalized.

An extension of LDA using a heteroscedastic two-class technique that follows the Chernoff criterion is proposed, called heteroscedastic linear dimension reduction (HLDR) [24]. Specifically, the authors use the Chernoff distance to evaluate the class similarity with means and covariances. Consequently, S_b is modified as shown in Equation (22).

$$S_b = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \Big[\Sigma_{ij}^{-1/2} (m_i - m_j)(m_i - m_j)^T \Sigma_{ij}^{-1/2} + 4\Big(\log \Sigma_{ij} - \frac{1}{2} \log \Sigma_i - \frac{1}{2} \log \Sigma_j \Big) \Big], \tag{22}$$

where Σ_i is the covariance matrix of data in class L_i, Σ_{ij} is the average between Σ_i and Σ_j, and equal priors are assumed. This method extends the two-class case into a multiclass version of the Chernoff criterion.

2.1.4. Using k-Nearest Neighbors

The proposal in [25] defines the scatter matrices on k-nearest neighbors of each sample, called local mean-based nearest neighbor discriminant analysis (LM-NNDA). Given N training samples of L classes $\{x_{ij} | i = 1, \cdots, L; j = 1, \cdots, N_i\}$, where N_i is the number of samples in class i. For each sample x_{ij}, search its k-nearest neighbors in every class. Let m_{ij}^s be the local mean vector of k-nearest neighbors of x_{ij} in class s.

The local within-class scatter matrix of LM-NNDA is defined as Equation (23).

$$S_w^{LMNNDA} = \frac{1}{N} \sum_{i,j} (x_{ij} - m_{ij}^i)(x_{ij} - m_{ij}^i)^T. \tag{23}$$

The local between-class scatter matrix of LM-NNDA is defined as Equation (24).

$$S_b^{LMNNDA} = \frac{1}{N(L-1)} \sum_{i,j} \sum_{s \neq i} (x_{ij} - m_{ij}^s)(x_{ij} - m_{ij}^s)^T. \tag{24}$$

A non-parametric form of discriminant analysis is first presented in [48] to overcome two problems. One is in parametric discriminant analysis, only at most $L - 1$ features (L: # of classes) are extracted due to the rank-deficient between-neighborhood scatter matrix while the non-parametric matrices are full-rank. Another one is that non-Gaussian datasets are allowed in non-parametric matrices. It redefines S_b using kNN techniques, focusing on two-classes cases. The proposal in [26] gives an extension of S_b shown in Equations (25) and (26) to multiclass classification under a face recognition scenario, referred to as multiclass non-parametric discriminant analysis (NDA).

$$S_b^{NDA} = \sum_{i=1}^{L} \sum_{\substack{j=1 \\ j \neq i}}^{L} \sum_{l=1}^{N_i} w(i,j,l)(x_l^i - m_j(x_l^i))(x_l^i - m_j(x_l^i))^T, \tag{25}$$

where $w(i,j,l)$ is the value of the weighting function depicted as

$$w(i,j,l) = \frac{\min\{d^\alpha(x_l^i, NN_k(x_l^i, i)), d^\alpha(x_l^i, NN_k(x_l^i, j))\}}{d^\alpha(x_l^i, NN_k(x_l^i, i)) + d^\alpha(x_l^i, NN_k(x_l^i, j))}. \tag{26}$$

The x_l^i is the l-th sample in class L_i, $NN_k(x_l^i, j)$ is the k-th nearest neighbor from class L_j to the sample x_l^i, $m_j(x_l^i)$ is the local KNN mean of $NN_k(x_l^i, j)$, $\alpha \in (0, +inf)$ is the parameter controlling the weight, and $d(\cdot, \cdot)$ is the Euclidean distance of two vectors. The weighting function explicitly emphasizes the data points around the boundary.

2.1.5. Neighborhood Linear Discriminant Analysis

Differing from the strategies above whose scatter matrices are defined on k-NN sets, neighborhood linear discriminant analysis (nLDA) [27] proposes a discriminator oriented to multi-modal classes where the scatter matrices are based on other types of

the neighborhood. It is motivated by the neighborhood can be taken as the smallest subclass and there is no need for any prior knowledge of the inner structure inside a class, avoiding the difficulty of determining the number of sub-classes inside a class. The scatter matrices are based on reverse k-nearest neighbor sets [49], shown as Equation (27). Given a training set $X = \{x_i \in \mathbb{R}^d | i \in \{1, \cdots, N\}\}$ and its label set $\mathcal{L} = \{g_i | g_i \in \{1, \cdots, L\}\}$. $X_j, |X_j| = N_j, \sum_{j=1}^{L} N_j = N$ consists all samples in class L_j. Given a dataset D and a sample $x_p \in D$, the reverse k-nearest neighbor set of x_p is defined as Equation (27).

$$\mathrm{RNN_k}(x_p, D) = \{x_q | x_q \in D \setminus \{x_p\}, x_p \in \mathrm{NN_k}(x_q, D)\}, \tag{27}$$

where $\mathrm{NN_k}(x_q, D)$ is the k-nearest neighbor set of $x_q \in D$. The within-neighborhood scatter matrix is depicted as Equation (28).

$$S_w^{nLDA} = \sum_{\substack{i=1 \\ |\mathrm{RNN_k}(x_i, X_{g_i})| \geq t}}^{N} \sum_{x_j \in \mathrm{RNN_k}(x_i, X_{g_i})} (x_j - \tilde{m}_i)^T (x_j - \tilde{m}_i), \tag{28}$$

where $\tilde{m}_i$ is the mean of the data in $\mathrm{RNN_k}(x_i, X_{g_i})$. Here is a threshold for nLDA that $|\mathrm{RNN_k}(x_i, X_{g_i})| \geq t$. There are $\mathcal{O}(kN)$ times of computing the outer product. The between-neighborhood scatter matrix is presented as Equation (29).

$$\mathbf{S}_b^{nLDA} = \sum_{|\mathrm{RNN_k}(x_i, X_{g_i})| \geq t}^{N} \sum_{\substack{j=1 \\ |\mathrm{RNN_k}(x_j, X_{g_j})| \geq t \\ g_i \neq g_j}}^{N} (\tilde{m}_i - \tilde{m}_j)^T (\tilde{m}_i - \tilde{m}_j). \tag{29}$$

From a point view of calculating times of outer product between vectors, the complexity of $\mathbf{S}_b^{nLDA}$ is $\mathcal{O}(N^2)$, which is too large with a large dataset. Here is an approximate alternative $\mathbf{S}_b^{nLDA^{app}}$ shown in Equation (30).

$$\mathbf{S}_b^{nLDA^{app}} = \sum_{|\mathrm{RNN_k}(x_i, X_{g_i})| \geq t}^{N} \sum_{\substack{x_j \in \mathrm{NN_k}(x_i, X - X_{g_i}) \\ |\mathrm{RNN_k}(x_j, X_{g_j})| \geq t \\ g_i \neq g_j}}^{N} (\tilde{m}_i - \tilde{m}_j)^T (\tilde{m}_i - \tilde{m}_j). \tag{30}$$

This reduces the complexity of the between-neighbor scatter matrix to $\mathcal{O}(kN)$. The target function is to find the projected directions v satisfying Equation (31).

$$v = \arg\max \left| \frac{v^T S_b^{nLDA^{app}} v}{v S_w^{nLDA} v} \right|. \tag{31}$$

The cost of nLDA contains two parts. One part is finding reverse nearest neighbors by computing $\mathcal{O}(N^2)$ times of distance between samples. Another part is computing scatter matrices and solving an eigenvalue problem. The latter is the same as that in LDA. The former is $\mathcal{O}(kN)$. So the cost of calculating Equation (31) is $\mathcal{O}(N^2)$ times of distance and $\mathcal{O}(kN)$ times of vector product. The empirical results demonstrate that nLDA outperforms greatly than LDA and some other discriminators. Notably, a proposal to solve the unstable and poor general issue resulting from the SSS problem of nLDA is presented by Xie et al. [28], where the singularity of within-neighborhood scatter matrix is evaded by the eigenspectrum regularisation techniques so that the method is called eigenspectrum regularisation reverse neighborhood discriminative learning (ERRNDL).

The conceptual comparisons of methods of LDA variants for multi-modal classes are shown in Table 3. Compared to the four discriminant analysis methods nLDA, LM-NNDA, LFDA and NDA which are all oriented to multi-modal class, we conclude several main connections and distinctions here.

Table 3. Conceptual comparison of LDA variants for multi-modal classes.

Type	Goal	Technique	References	Advantages
Mixture of Gaussians	Estimate underlying distribution of every class as mixture of Gaussians	Optimize mixture density by EM alg.	MDA [9] [45]	More robust EM than MDA
		Optimize # of sub-classes	SDA [10] MSDA [11] [47]	Overcome shortcomings of SDA Preserve original locality
			SSDA [12]	Better performance and better class separability than LDA, SDA and MSDA
			speed-up SDA [46]	
		Both of above	RSDA [13]	Improved robustness and lower computation cost
			extended MSDA [14]	
Applied problems	Classification and recognition algorithms for data with outliers			
Manifold	Depict local structure using manifold	Combine LDA and LPP	LFDA [15,16]	Overcome rank deficiency in S_b and protect local structure
		Use NN graph	LSDA [18]	
		Characterize piecewise regional consistency	MP-DA [19]	
		Subclass partition	AFLDA [20]	
Applied problems	Dimension reduction method for multimodal-labeled data and face recognition			
Setting weights	Import weights to penalize unstable data	Redefine S_b	aPAC [21,22] HLDR [24]	
		Redefine S_w	aPAC [23]	
Applied problems	Classification algorithms for real data			
KNN	k-Nearest Neighbor set based scatter matrices	Redefine S_b	NDA [26]	
		Redefine S_b and S_w	LM-NNDA [25]	
Applied problems	Classification and feature extraction methods for face databases			
RNN_k	Reverse k-nearest neighbors (RNN_k) set based scatter matrices	Redefine S_b and S_w	nLDA [27]	RNN_k can be regarded as the smallest subclass
		Three eigen-spectrum regularisation models	ERRNDL [28]	Overcome SSS in nLDA
Applied problems	Recognition and discriminative algorithm for multimodal-class data			

The nLDA uses the reverse k-nearest neighbor set to describe the multi-modality in classes, while LM-NNDA uses the k-nearest neighbors. LFDA depicts the local structure of

multi-modal class by combining with LPP, and NDA used k-nearest neighbors to rebuild scatter matrices, where they share the commonality of inheriting Fisher's criterion and differ from each other on the definitions of S_w and S_b.

As shown in Equation (28), the S_w^{nLDA} is defined on the set of $RNN_k(x_i, X_{g_i})$ that focuses on the scatter between samples x_i with the mean of its RNN_k set. Similarly, in Equation (23), the $S_w^{LMNNLDA}$ depicts the scatter between each sample and the k-nearest neighbors' mean inside the associated class, while in Equation (10), the within-scatter matrix S_w^{LFDA} depicts not the NN_k set but the scatter between the samples x_i to its neighbors by manifold.

In Equation (29), the S_b^{nLDA} is defined on the neighborhood of each sample that is found within its RNN_k set. However, as shown in Equation (25), the between-scatter matrix S_b^{NDA} is defined on the k-nearest neighbors for each sample that are found around all remaining classes. This is similar to $S_b^{LM-NNDA}$ in Equation (24) depicting the scatter between each sample and the mean of its k-nearest neighbors searched from other classes.

2.2. LDA Variants Solving the Small Sample Size (SSS) Problem

There is another main drawback of LDA. If the training samples are of high dimensionality but the the size of training samples is limited, S_w may have rank deficiency, that is, it almost becomes a singular matrix resulting in severe instability and over-fitting [50]. This is commonly considered as the small sample size (SSS) problem [6], and it always happens in pattern recognition which makes it a widely researched problem in related areas.

From Equation (3), as well as $\max(rank(S_b)) = L - 1$ as mentioned previously, it can be easily proved that $\max(rank(S_t)) = N - 1$ and $\max(rank(S_w) = rank(S_t) - rank(S_b)) = N - L$. Namely, the rank of S_b, S_w and S_t have the upper bounds of $L - 1$, $N - L$ and $N - 1$, respectively, and all of them are quite smaller than d under the scenario of high-dimensional but limited-sized samples. That is to say, S_b, S_w and S_t are all of singularity, resulting in the unsolvable for the objective Equation (4). We summarize and analyze different methods proposed to solve the SSS problem.

2.2.1. Fisherface Method (PCA + LDA)

The Fisherface method [51] is used in a wide variety of disciplines and areas that applies PCA initially such that the original d-dimensional features are reduced to a medium dimensionality d_1 under the guarantee of $d_1 \leq rank(S_w) = N - L$ aiming to make the consequent within-class scatter matrix full-rank. Then applying the standard LDA technique for further reducing the dimensionality to d_2 that has to be guaranteed $d_2 \leq L - 1$ because of the $\max(rank(S_b)) = L - 1$. Consequently, the SSS problem is overcome. In [52], we introduced in Section 2.2.2 a regularization procedure for the SSS problem; the author also applies PCA first to obtain full-rank S_w.

However, there exists a drawback that the PCA application of the first dimensionality reduction process leads to the loss of some useful discriminant information.

2.2.2. Regularization Method

For face recognition problems, it is common that the samples' dimensionality is very large resulting in S_w being singular. In [52], the authors slightly modify matrix S_w to $S_w + \kappa I$, where κ is a quite small positive number making $S_w + \kappa I$ absolutely positive definite. This is a regularization procedure by adding a small diagonal positive matrix to S_w. The same technique is used in references [53,54] to solve the SSS problem. However, the drawbacks are also obvious. Firstly, the computational complexity is quite high to deal with S_w of such a high dimension. Secondly, adding κ is just used for performing the inverse operation feasibly without any physical meaning. It is not able to evaluate κ and its poor choice may degrade the generalization performance of the method.

Besides regularizing matrix S_w directly, Jiang et al. [55] present an approach of eigenfeature regularization for face recognition. Using eigenvectors of S_w to span to image space and decomposing it into three sub-spaces, that is a null subspace, an unstable subspace

because of noise and limited sample size, and a reliable subspace spanned mostly by the facial variation, where eigen features are regularized in different ways within the three sub-spaces. This proposed approach remits the issue of limited sample size and noise leading to uncertain small and zero eigenvalues and is verified to be more stable, less over-fitted, or better generalized.

Another eigenfeature regularization method is proposed in [55] that regularises S_w by extrapolating its eigenvalues of the range space into the null space where the extrapolation is made by using exponential functions.

2.2.3. Null Space Method

The null space (or kernel) of a matrix $A \in \mathbb{R}^{m\times d}$ [56]: $ker(A) = \{x \in \mathbb{R}^d | Ax = \mathbf{0}\}$. The range space of A: $range(A) = \{y | y = Ax, x \in \mathbb{R}^d\}$.

The Fisher's criterion function [57] is shown in Equation (32).

$$F(v) = \frac{v^T S_b v}{v^T S_w v}, \tag{32}$$

where v denotes the projected vector. The authors in [58] introduce a revised Fisher's criterion $\widehat{F}(v)$ shown in Equation (33),

$$\widehat{F}(v) = \frac{v^T S_b v}{v^T S_b v + v^T S_w v}, \tag{33}$$

and have proved Equation (34), that is, $F(v)$ and $\widehat{F}(v)$ obtain the same optimal v.

$$\arg\max \widehat{F}(v) = \arg\max F(v). \tag{34}$$

Based on Equations (33) and (34), the authors in [59] introduce a different LDA technique to compute the best mapping directions using $\widehat{F}(v)$. If S_w is non-singular, then the $S_t = S_w + S_b$ is also non-singular. For the circumstances oriented towards the SSS problem, the process utilizes the null space of S_w. Suppose the original feature space $\mathbb{R}^d$, and the rank of S_w is denoted as r_w and $r_w < d$, that is S_w is singular. Thus, there exists the null space of S_w: $null(S_w) \subset \mathbb{R}^d$ such that $null(S_w) = span\{\alpha_i | S_w\alpha_i = 0, i = 1, \cdots, d - r_w\}$. Let all samples in $\mathbb{R}^d$ be projected into $null(S_w)$ via the transformation matrix T^T, where $T = (\alpha_1, \cdots, \alpha_{N-r})$. The within-class scatter matrix $\widetilde{S}_w$ of the mapped data in $null(S_w)$ is proved a complete zero matrix. That is to say, $\widetilde{S}_t = \widetilde{S}_w + \widetilde{S}_b = \widetilde{S}_b$. So maximizing the between-class scatter matrix $\widetilde{S}_b$ in $null(S_w)$ is the same as maximizing the total scatter in $null(S_w)$. In such cases, the author applies the PCA method to calculate the eigenvectors related to the largest eigenvalues of $\widetilde{S}_b$ that are the vectors of optimal discrimination fulfilling the requirements of LDA. However, projecting all data to the useful null subspace of S_w displays its strong clustering ability to achieve nice generalization performance, which seems that it achieves optimal discriminant ability but leads to over-fitting. The step of the diagonalization of S_b needs to be eliminated for the aim of avoiding possible over-fitting, which is mentioned by Liu et al. in [60].

There is quite a high computational complexity in the process of identifying the null space of S_w because of its high dimension. To escape the high calculating complexity, in [59], the pixel grouping technique is applied beforehand for artificial feature extraction and dimension reduction of the original data, and after that, the null space LDA is realized in reduced feature space $null(S_w)$ rather than the original space.

Due to the computation complexity problem of the original null space LDA method we introduced hereinabove, the authors in [61] propose a more efficient null space approach to solve that. If there are $v^T S_w v = 0$ and $v^T S_b v \neq 0$, then the eigen vector v is valuable for discriminating, whereas if $v^T S_w v = 0$ and $v^T S_b v = 0$, v is useless. Consequently, they remove the null space of S_t without losing valuable discrimination. Suppose U is the matrix

with columns being all the eigenvectors of S_t that correspond to the nonzero eigenvalues, and it is able to obtain the S'_w shown in Equation (35).

$$S'_w = U^T S_w U, \; S'_b = U^T S_b U. \tag{35}$$

Next a reduced but equally useful subspace of the null space of S_w is calculated, and projecting the data onto it, and then deriving the most discriminative vectors. Let Q be the null space of S'_w, thus $Q^T S'_w Q = 0$, then there are Equations (36) and (37).

$$S''_w = Q^T S'_w Q = Q^T U^T S_w U Q = (UQ)^T S_w (UQ), \tag{36}$$

$$S''_b = Q^T S'_b Q = Q^T U^T S_b U Q = (UQ)^T S_b (UQ), \tag{37}$$

where UQ is a subspace of all the null space of S_w which is reduced but of quite use to derive the most discriminative vectors. It is notable that if there is a null space of S''_b, it is necessary to remove it. Due to $\max(rank(S_t)) = N - 1$, the dimensionality of S'_w is bounded at $N - 1$; and due to $\max(rank(S_w)) = N - L$ and $rank(S'_w) = rank(S_w)$, the dimensionality of the null space of S'_w is $L - 1$. While S''_b is always full-rank so the amount of the optimal discriminant vectors is $L - 1$. This method improves the computational problem of the null space by removing redundant information without decreasing the discriminant efficiency.

Liu et al. [60] present the most appropriate condition for the null-space-based LDA method: $N = d - 1$, that is S_t, is full-rank, where N is the amount of all data and d is the dimensionality of original space. The procedure of null-space-based LDA under this most suitable situation removes the null space of S_t first and extracts the null space of reduced S_w. It is most straightforward with just one time of eigen-analysis. It not only saves a lot of computational costs but also keeps the performance simultaneously. Moreover, the authors incorporate the kernel technique into the null space method by using the Cosine kernel function. They discovered that in kernel space, S_t is full-rank, so the process of the null space method is extremely faster and more stable during calculation.

A faster null space method than [59] is proposed in [62] by only carrying out QR factorizations to implement LDA without carrying out eigendecomposition and SVD, of which computational complexity is approximately $4dN^2 + 2dNL$.

An alternative method of null-space-based LDA methods named Fast NLDA [29] modifies a fast process for the null space technique based on random matrix multiplication with scatter matrices. It is based on the assumption that the vectors are linearly independent. The oriented transformation matrix is gotten by $T = S_t^+ S_b Y$, where S_t^+ is the pseudo inverse of S_t, and Y is a random matrix of rank $L - 1$. This approach requires $dN^2 + 2dNL$ computations. The pseudoinverse LDA method of pseudo inverse S_w is studied in [63] for image classification.

2.2.4. Direct LDA

Based on the drawbacks of the techniques we introduced above: discarded dimensions which carry key discriminative information in Section 2.2.1, falling short of using information out of null space of S_w and computational problems related to large scatter matrices in Section 2.2.3. We introduce a direct LDA algorithm that permits data of high dimensionality and optimizes Fisher's criterion without any feature extraction or dimension reduction steps in advance, referred to as direct LDA (DLDA) by Yu et al. [30].

DLDA discards the null space of S_b firstly, where there is no discriminative information, but abandoning the null space of S_w where there is of the best discrimination. This is achieved in reverse order of traditional procedure by performing a simultaneous diagonalization procedure on S_b first by the found matrix W, as shown in Equation (38),

$$W^T S_b W = I, \tag{38}$$

and then on S_w, which keeps its null space to find the most discriminative vectors, as shown in Equation (39), and D_w means the diagonalizable S_w.

$$W^T S_w W = D_w. \tag{39}$$

It is worth mentioning that DLDA seems to reserve the null space of S_w, from which the optimal discriminant vectors of LDA can be deduced [51,59]. But it cannot substantially avoid it because removing the null space of S_b could lead to the portion loss of null space of S_w. S_b has a smaller rank than S_w in most instances, so the subspace guaranteeing the full rank of S_b is also guaranteeing that of S_w. DLDA does not take full advantage of the null space of S_w, by abandoning the null space of S_b via reducing dimension and indirectly leading to the loss of the null space of S_w. Additionally, in this paper, calculating skills are introduced to deal with large scatter matrices along with an accurate solution to Fisher's criterion being given.

2.2.5. Orthogonal LDA

Ye et al. propose an orthogonal LDA (OLDA) method against SSS problem [31] defining a new criterion that does not require the non-singularity of the scatter matrices. It has presented to be the same as those null-space-based LDA methods [59,61] limited in a soft condition of the data are linearly independent [64]. The null-space-based method [61] and OLDA all lead to the orthogonal transformations, while the former performs the orthogonal transformation in the null space of S_w and the latter executes that via diagonalizing S_b, S_w and S_t simultaneously. The calculation cost of OLDA is smaller than that of the null space method [59] and it is measured as $14dN^2 + 4dNL + 2dL^2$ flops.

2.2.6. Against Over-Fitting

Another serious issue of the SSS problem is the over-fitting problem. It is mainly because the between- and within-class scatter matrices calculated from the limited number of data drift extremely from the underlying ones. Pang et al. introduce a regularization term via clustering to solve this problem [65], specifically regularizing the within-class and between-class scatter matrices with within-cluster and between-cluster scatter matrices, respectively, and simultaneously.

We compare the methods of solving the SSS problem of LDA variants conceptually in Table 4 on metrics of method types, method goals, specific techniques, advantages and disadvantages.

Table 4. Conceptual comparison of LDA variants for SSS problem.

Type	Goal	Technique	References	Advantages	Disadvantages
Fisherface method	Reduce dimension of S_w	Use PCA initially to reduce dimension of S_w	[51]		lose useful information
Applied problems	Face recognition algorithms				
Regularization	Regularize S_w	$S_w + \kappa I$	[52] [53] [54]		high computational complexity; uncontrollable parameter
		Regularize eigenfeatures of S_w	[55]	more stability less over-fitting or better generalization	

Table 4. *Cont.*

Type	Goal	Technique	References	Advantages	Disadvantages
Applied problems	Face recognition algorithms				
Null space	Utilize null space of S_w that fulfills LDA criterion	Utilize full null space of S_w	[59]		high computational complexity
		Remove useless null space part of S_w and use the reduced part	[61]	more efficiency than [59]	
		Same as above but under the most suitable situation $N = d - 1$	[60]		
		QR factorizations only	[62]	faster than [59]	
		Random matrix multiplies scatter matrices	[29]	faster than [62]	
Applied problems	Face recognition algorithms				
Direct LDA	Indirect dimension reduction of S_w	Remove null space of S_b firstly	[30]		
Applied problems	Face recognition algorithms				
Orthogonal LDA	Orthogonal transformation in three scatter matrices simultaneously	New criterion no need non-singularity	[31]	lower complexity than [59]	
Applied problems	Classification algorithm for real-world data				
Against over-fitting	Solve over-fitting caused by SSS	Cluster-based scatter matrices	[65]		
Applied problems	Face recognition algorithms				

2.3. LDA Variants with Robustness

The conventional LDA method is based on L_2-norm [1] that is sensitive to outliers [66]. The outliers may lead to the projection vectors drifting from the objective directions. It is advisable to think about the robust modelling of the classical L_2-norm LDA to suppress the affection of outliers.

2.3.1. L_1 Norm

It is known to all that the L_1-norm is of better robustness than L_2-norm because the L_1-norm does not heighten the impact of outliers related to many errors as the L_2-norm does [66–68]. Li et al. [32] present a rotational invariant L_1-norm (i.e., R_1-norm) based LDA, denoted as LDA-R_1. It uses the gradient ascending iterative algorithm upon eigenvalues decomposition that leads to much time costs to perform convergence in input space of high dimensionality.

Wang et al. [33] introduce a technique, named LDA-L_1. It maximizes the proportion of the between-class dispersion to that of the within-class. They are defined by the L_1-norm

rather than the L_2-norm. Recall that the number of classes L, the N_i is the number of samples in class L_i, the class mean m_i, the total mean m, the projected vector v, and x_j^i is the j-th sample of class L_i. The Fisher-like criterion of L_1-norm is presented as shown in Equation (40).

$$F(v) = \frac{\sum_{i=1}^{L} N_i |v^T(m_i - m)|}{\sum_{i=1}^{L} \sum_{j=1}^{N_i} |v^T(x_j^i - m_i)|}. \tag{40}$$

The criterion (40) is termed as LDA-L_1 by authors used to maximize the proportion of between-class dispersion to that of the within-class. However, it is intractable to optimize the objective function (40) and obtain the global solution of the LDA-L_1. The authors give a gradient ascending (GA) iterative algorithm in order to seek a local solution v of L_1-norm LDA that maximizes the objective function. It is worth mentioning that the LDA-L_1 does not suffer from the problem of SSS or rank limit because the criterion is not based on the conventional matrices S_b and S_w anymore. Similar work of LDA-L_1 was published by Zhong et al. in the same year [69] which obtains a single locally optimal solution realized iteratively and obtains multiple locally optimum solutions via a greedy search method, as well as solving the singularity of S_w.

On the contrary, Liu et al. [70] propose a non-greedy iterative algorithm to address the objective function Equation (41) and obtain a closed-form solution for all projections.

$$\begin{aligned} F(V, \lambda) = arg \max_{V^T V = I} & \left(\sum_{i=1}^{L} N_i |V^T(m_i - m)| \right) \\ & - \lambda \left(\sum_{i=1}^{L} \sum_{j=1}^{N_i} |V^T(x_j^i - m_i)| \right), \end{aligned} \tag{41}$$

where V is the projection matrix, λ is related to V that are optimized iteratively, L and N_i are the numbers of total classes and the class L_i, respectively, m and m_i are the means of total samples and samples in L_i, respectively, and x_j^i is the j-th sample of class L_i.

When it comes to matrix-input issues, a matrix must be converted into a vector before applying the LDA methods, which are vector-based. This conversion can lead to high-dimensional data and the loss of some fundamental local information. Besides the matrix-based methods such as the matrix-based PCA [71–73], the matrix-based SVM [74,75], and the matrix-based LPP [76–79], the first proposal of L_2-norm-based 2-dimensional LDA (2DLDA) appears in [80] and afterwards many extensions are raised [81–86]. However, the 2DLDA may suffer from the robustness due to the effects of outliers and noise although it remits the SSS problem based on a weak assumption and preserves its original structural information. Li et al. [34] extend conventional 2-dimensional LDA with L_2-norm into 2-dimensional LDA with L_1-norm, termed as L_1-2DLDA, where the optimization problem is solved by the greedy iterative algorithm with its convergence being guaranteed. The authors in [87] further solve L_1-2DLDA through a nongreedy algorithm.

The iterative algorithms in the above L_1-norm based literature unfortunately mostly require selecting a suitable stepsize by iteratively modifying discriminant vectors. Due to the nonconvexity in the updating process, an unsuitable selection of the stepsize will impact the deduction of an optimum greatly. To handle the LDA-L_1 optimization problem, Zheng et al. [35] present an iterative algorithm that uses a new surrogate convex function for the optimization objective inside every iteration which only solves a convex problem and guarantees a closed-form result, referred to as L_1-LDA.

Furthermore, same as the equivalence relation sharing between the kernel discriminant analysis (KDA) [36] and the kernel principal component analysis (KPCA) plus LDA, which is found by Yang et al. [88], the authors generalize the proposed L_1-LDA method into

reproducing kernel Hilbert space (RKHS) to handle the nonlinear robust feature extraction through the kernel techniques, which hence termed as the L_1-norm kernel discriminant analysis (L_1-KDA = KPCA + L_1-LDA) method. Even though there is no need for choice of stepsize through this new efficient iterative algorithm, it has been indicated that its utilization results in L_1-LDA being easy to get in trouble with a lot of serious problems [89]: the existence of singularity problem, insufficient robustness against outliers because of the updated weighting mechanism, and unguaranteed Bayes solution optimality of discriminative criterion of L_1-LDA.

To this end, Ye et al. [89] present an efficient iterative method to deal with a general L_1-norm min–max issue and perform conceptual insight into its convergency, which overcomes the above problems that exist in both LDA-L_1 and L_1-LDA.

2.3.2. $L_{2,1}$ Norm

The utilization of L_1-norm in the above works is of limited robustness and mostly based on the greedy search strategy to seek the projections each by each where the process is time-consuming and may be trapped in local optimality. The $L_{2,1}$-norm-based loss function is firstly proposed by Nie et al. [90] to overcome the outliers and used as a regularization to fulfill feature selection. Inspired by this work, Nie et al. propose a novel robust LDA measured by $L_{2,1}$-norm, named as $L_{2,1}$-LDA [37].

The $L_{2,1}$-norm of a matrix $A \in \mathbb{R}^{d\times m}$ with its elements $a_{ij}, i = 1, \cdots, d, j = 1, \cdots, m$ is shown in Equation (42).

$$||A||_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{m} a_{ij}^2}. \tag{42}$$

The $L_{2,1}$-norm can measure the distances of spatial dims in L_2-norm, specifically, which is designed to enforce sparsity over the row-by-row data points to improve the robustness resisting outliers in L_1-norm. In comparison with the $L_{2,1}$-norm, the L_1-norm only focuses on inhibiting anomaly overall values without keeping an eye on the distinction between row and column, leading to insufficient robustness against outliers. To this end, the authors design the $L_{2,1}$-norm criterion to min–max within-class scatter and total data scatter at the same time to enhance the robustness and discriminability of the formulation as in Equation (43).

$$\begin{cases} \min \sum\limits_{j=1}^{L} \sum\limits_{i=1}^{N_j} ||V^T(x_i^j - m_j)||_2, \\ \max \frac{1}{N} \sum\limits_{k=1}^{N} ||V^T x_k||_2 = \frac{1}{N}||X^T V||_{2,1}, \end{cases} \tag{43}$$

where V is the projection matrix, and X is the data matrix. This $L_{2,1}$-norm criterion suppresses the anomaly of outliers from the learned sparsity structure by capturing the distinction between spatial dimensions and sample vectors and promoting the sparsity at the data points level. This improved the robustness. The authors propose a min–max iterative re-weighted optimization algorithm to deal with (43) which is a big challenge to be solved perfectly.

2.3.3. L_p Norm

Contrasting to the L_2-norm, the robust analysis of L_p-norm is investigated widely in data mining, for example, the robust locality preserving projections [91], L_p-norm based principal component analysis [92–94]. An L_p-norm based LDA is proposed in [38], termed as LDA-L_p by the authors. In this scheme, arbitrary values of p can be used to acquire a robust and rotation-invariant extension of LDA, for which the optimal solution is found

using the steepest gradient method. The objective of the maximization problem is presented as Equation (44).

$$F_p(v) = \frac{\sum_{i=1}^{L} N_i |v^T(m_i - m)|^p}{\sum_{i=1}^{L} \sum_{j=1}^{N_i} |v^T(x_j^i - m_i)|^p}, \tag{44}$$

which can be worked out by computing the gradient of $F_p(v)$ regarding v. However, there is a problem that the gradient of $F_p(v)$ is not well defined on some singular points due to the absolute value operator in this formula. The author introduces a sign function to escape the technical difficulty shown in Equation (45).

$$sgn(a) = \begin{cases} 1, & if\ a > 0, \\ 0, & if\ a = 0, \\ -1, & if\ a < 0. \end{cases} \tag{45}$$

Hence, the above objective formula can be rewritten as presented in Equation (46).

$$F_p(v) = \frac{\sum_{i=1}^{L} N_i [sgn(v^T(m_i - m))v^T(m_i - m)]^p}{\sum_{i=1}^{L} \sum_{j=1}^{N_i} [sgn(v^T(x_j^i - m_i))v^T(x_j^i - m_i)]^p}. \tag{46}$$

The optimal v that maximizes (46) can be obtained by taking a gradient of $F_p(v)$ regarding v: $\nabla_v = \frac{dF_p(v)}{dv}$, as depicted in a steepest gradient iterative algorithm with singular check and convergence check steps in detail.

To handle the matrix-input problem, contrasting to the L_1-2DLDA we introduced above that is still sensitive to outliers and noise, the L_p-norm is of much more robustness for $0 < p \leq 1$. Li et al. [39] introduce a bilateral two-dimensional LDA using the L_p-norm, named BL_p2DLDA. The criterion of BL_p2DLDA shares an equivalence relation with an upper bound of the theoretical principal of the optimal Bayes which theoretically guarantees the reasonability of its optimization via the Bayes error bound. The objective is solved by a modified ascent iterative technique.

2.3.4. L_{sp} Norm

Inspired by successful PCA-L_p algorithms [93–97], Ye et al. [40] propose a robust LDA, referred to as FLDA-L_{sp}. It maximizes L_s norm distance and minimizes L_p norm distance simultaneously via L_s-and L_p-norm measuring the between- and within-class distances, respectively, which differs from LDA-L_p [38] by a more effective iterative algorithm to obtain the target solution. The objective function of FLDA-L_{sp} is presented in Equation (47).

$$F(v) = \max_{v^T v = 1} \frac{\sum_{i=1}^{L} N_i |v^T(m_i - m)|^s}{\sum_{i=1}^{L} \sum_{j=1}^{N_i} |v^T(x_j^i - m_i)|^p}. \tag{47}$$

It is obvious that when $0 < s < 2$ and $0 < p < 2$, the objective is conferred with robustness. Moreover, the LDA-L_1 and LDA-L_2 become the special cases by setting specific values of s and p.

Compared with the gradient ascending iterative algorithm [33,69], the iterative algorithm used in LDA-L_{sp} does not require to apply the non-convex surrogate function, and it overcomes the challenge of choosing stepsize. Compared to the alternative algorithm

addressing the drawbacks in the gradient ascending iterative algorithm [33,69], the iterative algorithm used in LDA-L_{sp} avoids transforming the original objective during each iteration.

The norm types of LDA extensions, with comparisons of corresponding optimization methods, advantages as well as disadvantages are shown in Table 5.

Table 5. Conceptual comparison of LDA variants for robustness.

Type of Norm	References	Optimization Method	Advantages	Disadvantages
L_1 norm	[32] LDA-R_1	GA iterative algorithm		High computational complexity
	[33] LDA-L_1	Local solution by GA iterative algorithm	No SSS or rank limit	
	[69] LDA-L_1	Single local solution by iteration algorithm; Multiple local solutions by greedy searching	No SSS problem	
	[70]	A non-greedy iterative algorithm; A closed-form solution for all projections		
	[34] L_1-2DLDA matrix input	Greedy iterative algorithm; Convergence being guaranteed		
	[87] L_1-2DLDA	A nongreedy algorithm		Bad selection of stepsize may impact the optimality
	[35] L_1-LDA	Iterative algorithm; A closed-form solution during every iteration	No stepsize	Easy singularity; Insufficient robustness; Unguaranteed Bayes optimality
	[89]	An effective iterative framework	Overcome problems in LDA-L_1 and L_1-LDA	
Applied problems	Robust classification and recognition algorithms for suppressing outliers			
$L_{2,1}$ norm	[37]	Minmax iterative re-weighted optimization algorithm	More robust than L_1-norm	Hard to solve objective
Applied problems	Robust classification and visualization algorithms for synthetic data and image datasets			
L_p norm	[38] LDA-L_p	Steepest gradient iterative algorithm	Arbitrary p can obtain robust and other LDA versions	Technique difficulty in optimization
	[39] BL_p2DLDA matrix input	Modified ascent iterative technique	More robust than L_1-2DLDA [34] for $0 < p \leq 1$	
Applied problems	Robust discriminant analysis methods for contaminated databases			
L_{sp} norm	[40] FLDA-L_{sp}	A more effective iterative algorithm	Robustness at $0 < s, p < 2$; LDA-L_1, LDA-L_2 are special cases; Needless of non-convex surrogate function and stepsize compared to [33,69]; No transforming original objective iteratively	
Applied problems	Robust discriminant analysis methods for image data in suppressing the noise			

3. Applications of LDA Variants

In this section, we focus on the usage of discriminant analysis in addressing the drawbacks of small sample size problems, being sensitive to noise and outliers, and being unable to deal with multi-modal-class data. We summarize the application fields across face recognition, fault detection and diagnosis, system condition monitoring, process monitoring et al., which belong to the main areas of computer vision, pattern recognition, and automation control systems. The objective of this section is to guide the readers on the benefits of suppressing three drawbacks in discriminant analysis and how to optimally utilize suitable techniques in certain cases.

3.1. Applications of LDA Variants for Multi-Modal Data

3.1.1. Mixture of Gaussians

The mixture of Gaussian-based discriminant analysis, depicting the mixtures of the multi-modal density models in each class, has been acting as an excellent technique to offer a better estimation and description of multi-modal data distributions. This technique motivates extensive applications aiming to address the multi-modal problem in many scenarios. MDA is applied to face detection [98], human–robot interaction [99], remote sensing [100], process monitoring [101], drug distribution in humans [102], digit recognition [103], and speaker verification [104]. In addition, MDA is used as a per-field classification method [105] and a curve classification method [106]. The subclass-based mixture of Gaussian, such as SDA, is applied for face recognition [107,108], disease diagnosis [109], behavior recognition [110], and bug prediction [111].

3.1.2. Manifolds

LFDA is applied in pedestrian re-identification [112], diagnose prediction [113], facial expression recognition [114,115], fault diagnosis [116,117], spoken language identification [118], industrial process fault classification [119], process monitoring [120], the physical load prediction [121], and the spoken emotion recognition [122]. Additionally, there are various of LFDA extensions being of much usage. For examples, sparse LFDA for facial expression recognition [123] and status monitoring [124], maximum LFDA for face recognition [125], complete LFDA for face recognition [126], uncorrelated LFDA for ear recognition [127], geometric preserving LFDA for person re-identification [128], wavelet LFDA based bearing defect classification [129], orthogonal LFDA for fault diagnosis [130] and facial expression recognition [131], projection-optimal LFDA for feature extraction [132] and palmprint recognition [133], self-adaptive LFDA based semi-supervised image recognition [134], the unsupervised image-adapted LFDA [135], the fault diagnosis based on local centroid mean LFDA [136], fault diagnosis for blast furnace ironmaking process using randomized LFDA [137], and online soft measurement method based on improved LFDA [138].

Additionally, the studies focusing on semi-supervised LFDA are welcome in many scenarios, such as enhanced semi-supervised LFDA for face recognition [139], for sparse dimensionality reduction of the hyperspectral image [140], and gene expression data classification [141].

LSDA is applied as another discriminant approach based on manifold learning in kinds of fields, for example, stable LSDA-based image recognition [142], improved LSDA-based feature extraction [143], orthogonal LSDA-based face recognition [144,145], identification of breast cancer [146], hyperspectral imagery classification [147], fault diagnosis [148], image recognition [149], face recognition [150,151], and video semantic detection [152].

3.1.3. k-Nearest Neighbors

NDA has been applied into various areas, for example, face recognition [26], face detection [153], feature extraction [154,155], image recognition [156,157], imagery classification [158], image retrieval [157,159], incremental subspace learning and recognition [160], and 3-D model classification [161].

3.1.4. Setting Weights

Penalized discriminant analysis has been used in conifer species recognition [162], tumor classification [163,164], classification of bladder cancer patients [165], image-based morphometry [166], detection of wild-grown and cultivated Ganoderma lucidum [167], noise removal [168], brain images [169], microarrays [170], and predicting choice [171].

3.2. Applications of LDA Variants for Solving SSS

The SSS problem happens in the case of the larger feature dimension but the limited data size which arouses great concern within the face recognition community to solve the poor generalization, instability or the over-fitting problems when performing face recognition on a larger face dataset but with very few available training face images [30,51–53,55,59–61]. The applications of LDA variants against the SSS problems promote a feature presentation of more discrimination and stability in low-dimensional space for the face images to perform extracting features, classifying, and reducing dimension issues in pattern recognition. Tian et al. discussed the image classification issue in the case where the total amount is smaller than the dimensionality of training samples to be classified and provided a good classification performance with a low error rate [63].

Additionally, the null space discriminant analysis for the SSS problem has been applied for novelty detection [172,173], and person re-identification [174,175].

The Fisherface method for the SSS problem is used for a face recognition problem which slowly reacts to big changes in the cases of lighting direction and facial expression [51]. This technique is also applied for image retrieval [176].

Pang et al. applied the proposed enhanced LDA into pattern recognition systems of face and ear recognition to solve the over-fitting problem [65].

3.3. Applications of LDA Variants with Robustness

LDA variants based on different norms have been applied widely by reducing the influence of outliers, for example, L_1-norm LDA for robust feature extraction [35,177], human activity recognition [178,179], $L_{2,1}$-norm LDA for face recognition [180] and image recognition [181].

3.4. Discussions on the Applications of LDA Variants

Based on the above summaries of the application fields of LDA variants for addressing three drawbacks, we can conduct some analysis as follows.

Firstly, the applications of the methods for multi-modal data are mainly distributed in fault detection and diagnosis, process and status monitoring, recognition and identification of information, and the classification, which is coming from the fact that the multi-modality exists in the samples of complicated distributions, such as outliers or noise. The techniques of different theoretical philosophies oriented to multi-modality may help to guide the applications to the different detection or recognition scenarios with complex samples.

Secondly, the application fields of LDA variants for solving the SSS mostly gather in face recognition, because facial information is of high features that quite easily limit to the rank-deficiency of the within scatter matrix. This can direct us to deal with other realistic problems with large features by discriminant analysis methods.

Thirdly, the application fields of LDA variants with robustness are similar to those of multi-modal data impairing the influence of outliers. There is an optimization problem in solving the eigendecomposition of LDA variants with other norms, which motivates us to optimize and apply discriminant analysis with L_1, $L_{2,1}$, L_p and L_{sp} norm.

4. Kernelization

The kernel method involves performing a projection from the original low-dimensional space into a higher-dimensional feature space, specifically the reproducing kernel Hilbert space (RKHS) [182]. This transformation changes the data from being linearly inseparable to linearly separable.

We provide a simplified introduction to RKHS here. We start with the vector space, defined as a set of vectors equipped with addition and scalar multiplication. The inner product space is a vector space equipped with an inner product operation. The Hilbert space is a complete inner product space, where all Cauchy sequences converge within this space. An RKHS is a Hilbert space of functions where the inner product of the mapping functions is equivalent to the kernel function when data is mapped into this space.

The process of feature mapping, depicted in Figure 2, clearly shows how a nonlinear problem is transformed into a linear problem. This projection is achieved by the feature map utilizing the kernel function.

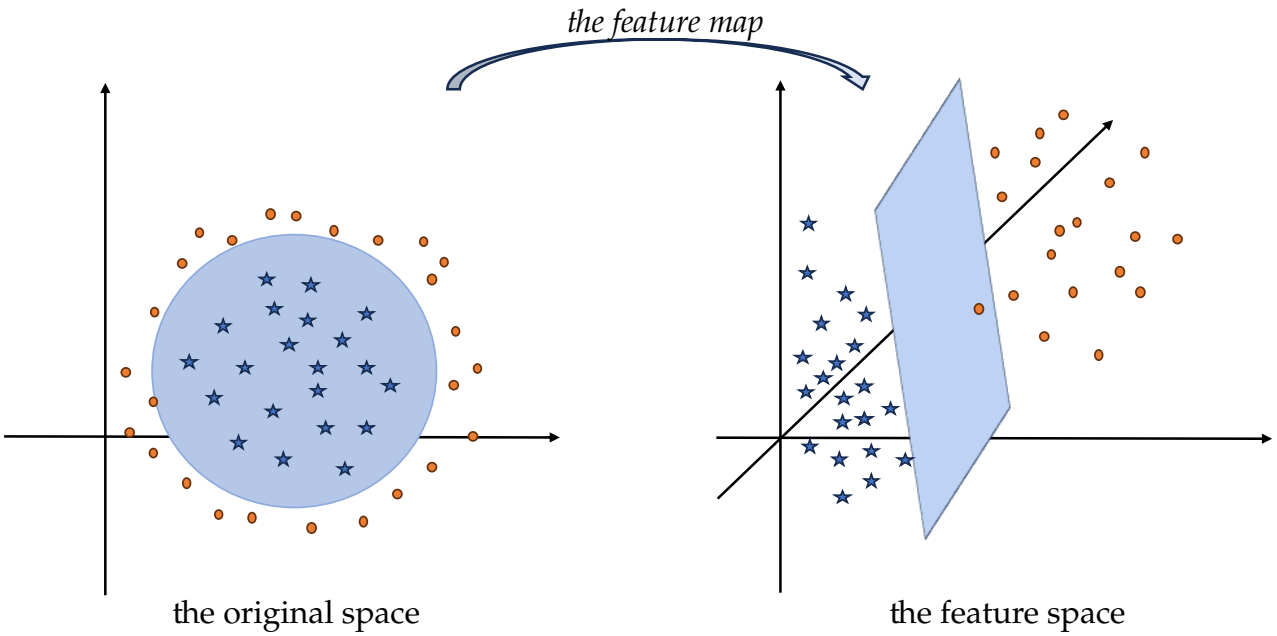

Figure 2. The feature map of kernel method.

Here, we summarise the methods above which have been extended to the kernel version for solving nonlinear problems. First of all, we have a proposal of a kernelized data analysis method using an orthogonal transformation that combines the three objectives of kernel PCA and kernel LDA [183]. Our proposal possesses both feature extraction and discriminative properties to solve nonlinear problems in reproducing kernel Hilbert space. SDA is extended into kernel version in [184], and the kernel extension of SDA is used for yielding the optimal recognition rates [185]. The speed-up and multi-view SDA and its kernelized form are proposed in [46]. The kernel MSDA (KMSDA) is proposed in [14]. LFDA has been extended to non-linear dimensionality reduction cases with kernel trick, called KLFDA [15,16], making so many applications: sparse kernel LFDA for fault diagnosis [186], multiple kernel LFDA based face recognition [187] and fault diagnosis [188], financial distress predictions [189], wavelet kernel LFDA for bearing defect classification [190], manifold adaptive kernel LFDA for face recognition [191], individual geographic origin prediction [192], hyperspectral image classification [193], and semi-supervised kernel LFDA for bearing defect classification [194].

The NDA is extended into kernel version in [185] and is used for data classification by Diaf et al. [195] and 3-D model classification [161]. The kernel technique is incorporated into the null-space-based LDA effectively solving the SSS problem [60] and is used for novelty detection [173]. The L_1-norm LDA has been kernelized in [35], and the L_p LDA has been kernelized in [196].

5. Conclusions and Study Perspectives in LDA

We searched the Scopus and Elsevier databases, as well as the Web of Science Core Collection (WoSCC), using the relevant keywords "multi-modal", "Small Sample Size", "robust" and "discriminant analysis". This search yielded more than 300 papers. We applied a priority selection criterion based on stronger relevance, a higher number of citations, more recent publication years, and a higher ranking of journals according to the Journal Citation Reports (JCR) and conference papers according to the China Computer Federation (CCF) recommendations, resulting in a review of 197 articles in total.

In general, 175 articles within our references are sourced on WoSCC across ten fields, as illustrated in Figure 3. The two most covered fields are computer science artificial intelligence, and engineering electrical electronics. We summarized and compared the extensions on techniques, innovations and main applications of discriminant analysis-based algorithms, focusing on addressing the three main drawbacks of conventional LDA: its inability to handle multi-modal data, small sample sizes (SSS), and lack of robustness. Finally, we summarized the kernel-extended algorithms designed for nonlinear problems.

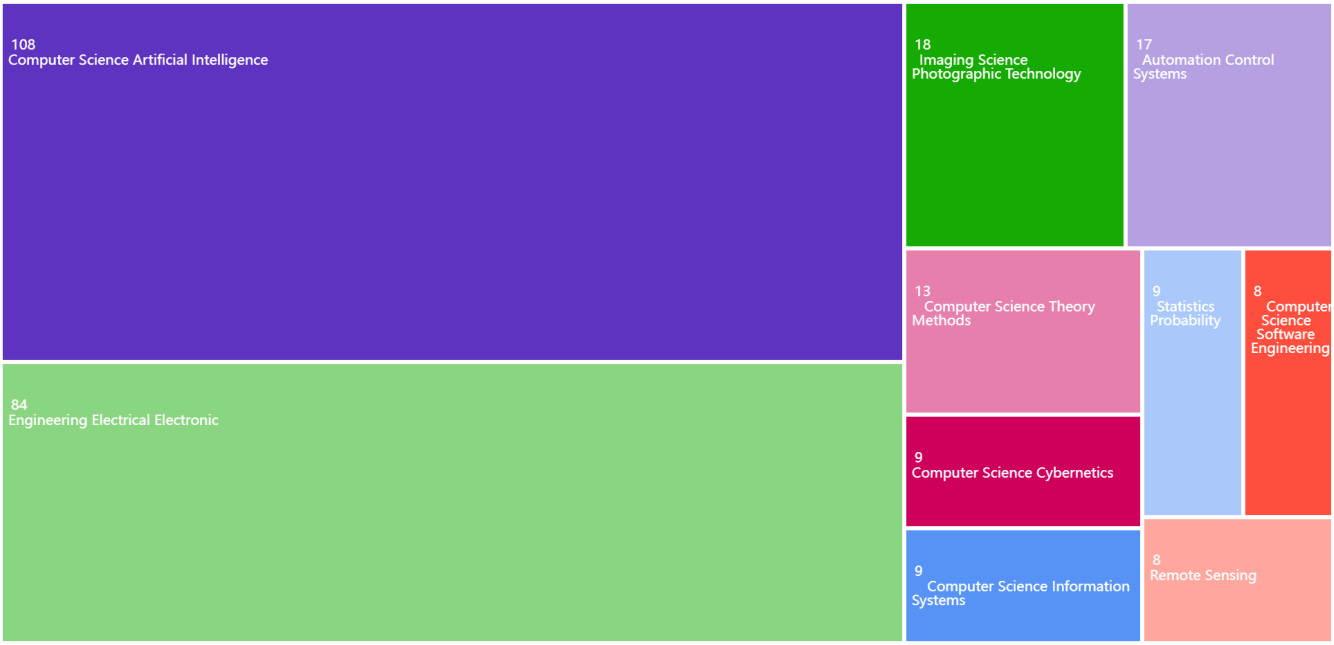

Figure 3. The analysis results on Web of Science categories of the references from Web of Science Core Collection.

As part of future research, three drawbacks of LDA can be considered common issues in classification, clustering, and regression problems using discriminative analysis. The application areas of the reviewed variants may provide optimization-oriented guidance on how to apply these methods better to appropriate real-world problems for optimal performance. This constitutes the first future direction for applications.

Additionally, by examining how these LDA variants address the three drawbacks, we can gain insights into the underlying relationships between data distributions, structures, and algorithms. This generates another open question regarding the potential designs of the fusions of more robust, stable, and general algorithms with discriminative properties, which should be explored further.

There are two types of fusion methods. The first involves combining the objectives of different algorithms. Reference [197] proposes a novel framework of seven data analysis methods that combines the objectives of PCA and LDA. Based on this framework, we extend the first method of the framework into RKHS with a kernel method [183]. The second type of fusion is the staged usage of different algorithms. The methods [51,88]

reviewed above that utilize PCA or KPCA ahead of LDA are two-phase fusions. The potential designs for fusions of different objective functions and staged methods, aimed at enhancing robustness, stability, and generality, represent a significant future research topic.

Furthermore, the kernel extension is an important research topic for data analysis methods to address nonlinear problems. Building on techniques that resolve the drawbacks of LDA, our future research will focus on three subjects.

1. Improving robustness for nonlinear problems;
2. Handling multi-modal-class data with complicated nonlinear distributions or outliers;
3. Addressing small sample size problems in reproducing kernel Hilbert space.

It is a promising research direction that investigates extending discriminative analysis methods that have already overcome these drawbacks into their kernel versions.

Author Contributions: Literature search and reading, conceptual analysis and comparisons, application investigation, writing—original draft preparation, L.Q.; supervision, project administration, review and revision, Y.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
2. Zhou, Z.H. Linear Models. In *Machine Learning*; Springer: Singapore, 2021; pp. 57–77.
3. Byrne, O. Book 1, Defination 10. In *The First Six Books of the Elements of Euclid*; Taschen America LLC: Beverly Hills, CA, USA, 2010.
4. Wang, R. *Introduction to Orthogonal Transforms: With Applications in Data Processing and Analysis*; Cambridge University Press: Cambridge, UK, 2012.
5. Tukey, J.W. The future of data analysis. *Ann. Math. Stat.* **1962**, *33*, 1–67. [CrossRef]
6. Keinosuke, F. *Introduction to Statistical Pattern Recognition*, 2nd ed.; Academic Press: Cambridge, MA, USA, 1990.
7. Zhu, F.; Yang, J.; Gao, J.; Xu, C. Extended nearest neighbor chain induced instance-weights for SVMs. *Pattern Recognit.* **2016**, *60*, 863–874. [CrossRef]
8. Zhu, F.; Yang, J.; Gao, C.; Xu, S.; Ye, N.; Yin, T. A weighted one-class support vector machine. *Neurocomputing* **2016**, *189*, 1–10. [CrossRef]
9. Hastie, T.; Tibshirani, R. Discriminant analysis by Gaussian mixtures. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 155–176. [CrossRef]
10. Zhu, M.; Martinez, A.M. Subclass discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1274–1286. [PubMed]
11. Gkalelis, N.; Mezaris, V.; Kompatsiaris, I. Mixture subclass discriminant analysis. *IEEE Signal Process. Lett.* **2011**, *18*, 319–322. [CrossRef]
12. Wan, H.; Wang, H.; Guo, G.; Wei, X. Separability-oriented subclass discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 409–422. [CrossRef] [PubMed]
13. Wu, D.; Boyer, K.L. Resilient subclass discriminant analysis. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009 ; pp. 389–396.
14. Gkalelis, N.; Mezaris, V.; Kompatsiaris, I.; Stathaki, T. Mixture subclass discriminant analysis link to restricted Gaussian model and other generalizations. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *24*, 8–21. [CrossRef]
15. Sugiyama, M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J. Mach. Learn. Res.* **2007**, *8*, 1027–1061.
16. Sugiyama, M. Local fisher discriminant analysis for supervised dimensionality reduction. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 905–912.
17. He, X.; Niyogi, P. Locality preserving projections. *Adv. Neural Inf. Process. Syst.* **2003**, *16*, 186-197.
18. Cai, D.; He, X.; Zhou, K.; Han, J.; Bao, H. Locality sensitive discriminant analysis. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, 6–12 January 2007; Volume 2007, pp. 1713–1726.
19. Zhou, Y.; Sun, S. Manifold partition discriminant analysis. *IEEE Trans. Cybern.* **2016**, *47*, 830–840. [CrossRef] [PubMed]
20. Wang, J.; Yin, H.; Nie, F.; Li, X. Adaptive and fuzzy locality discriminant analysis for dimensionality reduction. *Pattern Recognit.* **2024**, *151*, 110382. [CrossRef]
21. Loog, M.; Duin, R.P.W.; Haeb-Umbach, R. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 762–766. [CrossRef]
22. Loog, M. *Approximate Pairwise Accuracy Criteria for Multiclass Linear Dimension in Reduction*; Delft University Press: Delft, The Netherlands, 1999.
23. Hastie, T.; Buja, A.; Tibshirani, R. Penalized discriminant analysis. *Ann. Stat.* **1995**, *23*, 73–102. [CrossRef]

24. Duin, R.P.; Loog, M. Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 732–739. [CrossRef] [PubMed]
25. Yang, J.; Zhang, L.; Yang, J.y.; Zhang, D. From classifiers to discriminators: A nearest neighbor rule induced discriminant analysis. *Pattern Recognit.* **2011**, *44*, 1387–1402. [CrossRef]
26. Li, Z.; Lin, D.; Tang, X. Nonparametric discriminant analysis for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 755–761. [PubMed]
27. Zhu, F.; Gao, J.; Yang, J.; Ye, N. Neighborhood linear discriminant analysis. *Pattern Recognit.* **2022**, *123*, 108422. [CrossRef]
28. Xie, M.; Tan, H.; Du, J.; Yang, S.; Yan, G.; Li, W.; Feng, J. Eigenspectrum regularisation reverse neighbourhood discriminative learning. *IET Comput. Vis.* **2024**. [CrossRef]
29. Sharma, A.; Paliwal, K.K. A new perspective to null linear discriminant analysis method and its fast implementation using random matrix multiplication with scatter matrices. *Pattern Recognit.* **2012**, *45*, 2205–2213. [CrossRef]
30. Yu, H.; Yang, J. A direct LDA algorithm for high-dimensional data-with application to face recognition. *Pattern Recognit.* **2001**, *34*, 2067–2070. [CrossRef]
31. Ye, J.; Yu, B. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *J. Mach. Learn. Res.* **2005**, *6*, 483–502.
32. Li, X.; Hu, W.; Wang, H.; Zhang, Z. Linear discriminant analysis using rotational invariant L1 norm. *Neurocomputing* **2010**, *73*, 2571–2579. [CrossRef]
33. Wang, H.; Lu, X.; Hu, Z.; Zheng, W. Fisher discriminant analysis with L1-norm. *IEEE Trans. Cybern.* **2013**, *44*, 828–842. [CrossRef] [PubMed]
34. Li, C.N.; Shao, Y.H.; Deng, N.Y. Robust L1-norm two-dimensional linear discriminant analysis. *Neural Netw.* **2015**, *65*, 92–104. [CrossRef] [PubMed]
35. Zheng, W.; Lin, Z.; Wang, H. L1-norm kernel discriminant analysis via Bayes error bound optimization for robust feature extraction. *IEEE Trans. Neural Netw. Learn. Syst.* **2013**, *25*, 793–805. [CrossRef]
36. Yang, M. Kernel Eigenfaces vs. Kernel Fisherfaces: Face recognition using Kernel methods. In Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, 20–21 May 2002; IEEE Comp Soc TC PAMI; pp. 215–220.
37. Nie, F.; Wang, Z.; Wang, R.; Wang, Z.; Li, X. Towards Robust Discriminative Projections Learning via Non-Greedy $l_{2,1}$-Norm MinMax. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 2086–2100. [CrossRef] [PubMed]
38. Oh, J.H.; Kwak, N. Generalization of linear discriminant analysis using Lp-norm. *Pattern Recognit. Lett.* **2013**, *34*, 679–685. [CrossRef]
39. Li, C.N.; Shao, Y.H.; Wang, Z.; Deng, N.Y. Robust bilateral Lp-norm two-dimensional linear discriminant analysis. *Inf. Sci.* **2019**, *500*, 274–297. [CrossRef]
40. Ye, Q.; Fu, L.; Zhang, Z.; Zhao, H.; Naiem, M. Lp-and Ls-norm distance based robust linear discriminant analysis. *Neural Netw.* **2018**, *105*, 393–404. [CrossRef] [PubMed]
41. Zollanvari, A.; Dougherty, E.R. Generalized consistent error estimator of linear discriminant analysis. *IEEE Trans. Signal Process.* **2015**, *63*, 2804–2814. [CrossRef]
42. Jain, A.K.; Duin, R.P.W.; Mao, J. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 4–37. [CrossRef]
43. McLachlan, G.J.; Basford, K.E. *Mixture Models: Inference and Applications to Clustering*; M. Dekker: New York, NY, USA 1988; Volume 38.
44. Moon, T.K. The expectation-maximization algorithm. *IEEE Signal Process. Mag.* **1996**, *13*, 47–60. [CrossRef]
45. Bashir, S.; Carter, E. High breakdown mixture discriminant analysis. *J. Multivar. Anal.* **2005**, *93*, 102–111. [CrossRef]
46. Chumachenko, K.; Raitoharju, J.; Iosifidis, A.; Gabbouj, M. Speed-up and multi-view extensions to subclass discriminant analysis. *Pattern Recognit.* **2021**, *111*, 107660. [CrossRef]
47. Tao, Y.; Yang, J.; Chang, H. Enhanced iterative projection for subclass discriminant analysis under EM-alike framework. *Pattern Recognit.* **2014**, *47*, 1113–1125. [CrossRef]
48. Fukunaga, K.; Mantock, J. Nonparametric discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1983**, *PAMI-5*, 671–678. [CrossRef]
49. Korn, F.; Muthukrishnan, S. Influence sets based on reverse nearest neighbor queries. *ACM Sigmod Rec.* **2000**, *29*, 201–212. [CrossRef]
50. Sharma, A.; Paliwal, K.K. Linear discriminant analysis for the small sample size problem: An overview. *Int. J. Mach. Learn. Cybern.* **2015**, *6*, 443–454. [CrossRef]
51. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 711–720. [CrossRef]
52. Zhao, W.; Chellappa, R.; Phillips, P.J. *Subspace Linear Discriminant Analysis for Face Recognition*; Technical Report CAR-TR-914; Center for Automation Research, Univ. of Maryland: College Park, MD, USA, 1999.
53. Dai, D.Q.; Yuen, P.C. Face recognition by regularized discriminant analysis. *IEEE Trans. Syst. Man, Cybern. Part B (Cybern.)* **2007**, *37*, 1080–1085.
54. Friedman, J.H. Regularized discriminant analysis. *J. Am. Stat. Assoc.* **1989**, *84*, 165–175. [CrossRef]

55. Jiang, X.; Mandal, B.; Kot, A. Eigenfeature regularization and extraction in face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 383–394. [CrossRef] [PubMed]
56. Farjoun, E.D. *Cellular Spaces, Null Spaces and Homotopy Localization*; Springer: Berlin, Heidelberg, German, 1996.
57. Fisher, A. *The Mathematical Theory of Probabilities and Its Application to Frequency Curves and Statistical Methods*; Macmillan: New York, NY, USA, 1922; Volume 1.
58. Liu, K.; Cheng, Y.Q.; Yang, J.Y.; Liu, X. An efficient algorithm for Foley-Sammon optimal set of discriminant vectors by algebraic method. *Int. J. Pattern Recognit. Artif. Intell.* **1992**, *6*, 817–829. [CrossRef]
59. Chen, L.F.; Liao, H.Y.M.; Ko, M.T.; Lin, J.C.; Yu, G.J. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognit.* **2000**, *33*, 1713–1726. [CrossRef]
60. Liu, W.; Wang, Y.; Li, S.Z.; Tan, T. Null space-based kernel fisher discriminant analysis for face recognition. In Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, Republic of Korea, 17–19 May 2004; pp. 369–374.
61. Huang, R.; Liu, Q.; Lu, H.; Ma, S. Solving the small sample size problem of LDA. In Proceedings of the 2002 International Conference on Pattern Recognition, Quebec, QC, Canada, 11–15 August 2002; Volume 3, pp. 29–32.
62. Chu, D.; Thye, G.S. A new and fast implementation for null space based linear discriminant analysis. *Pattern Recognit.* **2010**, *43*, 1373–1379. [CrossRef]
63. Tian, Q.; Barbero, M.; Gu, Z.H.; Lee, S.H. Image classification by the Foley-Sammon transform. *Opt. Eng.* **1986**, *25*, 834–840. [CrossRef]
64. Ye, J.; Xiong, T.; Madigan, D. Computational and Theoretical Analysis of Null Space and Orthogonal Linear Discriminant Analysis. *J. Mach. Learn. Res.* **2006**, *7*, 1183–1204.
65. Pang, Y.; Wang, S.; Yuan, Y. Learning regularized LDA by clustering. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *25*, 2191–2201. [CrossRef]
66. Ke, Q.; Kanade, T. Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 739–746.
67. Xu, L.; Yuille, A.L. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Trans. Neural Netw.* **1995**, *6*, 131–143. [PubMed]
68. Pang, Y.; Li, X.; Yuan, Y. Robust tensor analysis with L1-norm. *IEEE Trans. Circuits Syst. Video Technol.* **2009**, *20*, 172–178. [CrossRef]
69. Zhong, F.; Zhang, J. Linear discriminant analysis based on L1-norm maximization. *IEEE Trans. Image Process.* **2013**, *22*, 3018–3027. [CrossRef] [PubMed]
70. Liu, Y.; Gao, Q.; Miao, S.; Gao, X.; Nie, F.; Li, Y. A non-greedy algorithm for L1-norm LDA. *IEEE Trans. Image Process.* **2016**, *26*, 684–695. [CrossRef] [PubMed]
71. Kong, H.; Wang, L.; Teoh, E.K.; Li, X.; Wang, J.G.; Venkateswarlu, R. Generalized 2D principal component analysis for face image representation and recognition. *Neural Netw.* **2005**, *18*, 585–594. [CrossRef]
72. Yang, J.; Yang, J.y. From image vector to matrix: A straightforward image projection technique-IMPCA vs. PCA. *Pattern Recognit.* **2002**, *35*, 1997–1999. [CrossRef]
73. Yang, J.; Zhang, D.; Frangi, A.F.; Yang, J.y. Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 131–137. [CrossRef]
74. Wang, Z.; Chen, S. New least squares support vector machines based on matrix patterns. *Neural Process. Lett.* **2007**, *26*, 41–56. [CrossRef]
75. Nurvitadhi, E.; Mishra, A.; Marr, D. A sparse matrix vector multiply accelerator for support vector machine. In Proceedings of the 2015 International Conference on Compilers, Architecture and Synthesis for Embedded Systems (CASES), Amsterdam, The Netherlands, 4–9 October 2015; pp. 109–116.
76. Hu, D.; Feng, G.; Zhou, Z. Two-dimensional locality preserving projections (2DLPP) with its application to palmprint recognition. *Pattern Recognit.* **2007**, *40*, 339–342. [CrossRef]
77. Chen, S.; Zhao, H.; Kong, M.; Luo, B. 2D-LPP: A two-dimensional extension of locality preserving projections. *Neurocomputing* **2007**, *70*, 912–921. [CrossRef]
78. Chen, S.B.; Luo, B.; Hu, G.P.; Wang, R.H. Bilateral two-dimensional locality preserving projections. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, HI, USA, 15–20 April 2007; Volume 2, pp. II–601–II–604.
79. Xu, Y.; Feng, G.; Zhao, Y. One improvement to two-dimensional locality preserving projection method for use with face recognition. *Neurocomputing* **2009**, *73*, 245–249. [CrossRef]
80. Liu, K.; Cheng, Y.Q.; Yang, J.Y. Algebraic feature extraction for image recognition based on an optimal discriminant criterion. *Pattern Recognit.* **1993**, *26*, 903–911. [CrossRef]
81. Jing, X.Y.; Wong, H.S.; Zhang, D. Face recognition based on 2D Fisherface approach. *Pattern Recognit.* **2006**, *39*, 707–710. [CrossRef]
82. Kong, H.; Wang, L.; Teoh, E.K.; Wang, J.G.; Venkateswarlu, R. A framework of 2D Fisher discriminant analysis: Application to face recognition with small number of training samples. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 1083–1088.
83. Li, M.; Yuan, B. 2D-LDA: A statistical linear discriminant analysis for image matrix. *Pattern Recognit. Lett.* **2005**, *26*, 527–532. [CrossRef]

84. Xiong, H.; Swamy, M.; Ahmad, M.O. Two-dimensional FLD for face recognition. *Pattern Recognit.* **2005**, *38*, 1121–1124. [CrossRef]
85. Yang, J.; Zhang, D.; Yong, X.; Yang, J.y. Two-dimensional discriminant transform for face recognition. *Pattern Recognit.* **2005**, *38*, 1125–1129. [CrossRef]
86. Xu, X.; Deng, J.; Cummins, N.; Zhang, Z.; Wu, C.; Zhao, L.; Schuller, B. A two-dimensional framework of multiple kernel subspace learning for recognizing emotion in speech. *IEEE/ACM Trans. Audio Speech, Lang. Process.* **2017**, *25*, 1436–1449. [CrossRef]
87. Li, M.; Wang, J.; Wang, Q.; Gao, Q. Trace ratio 2DLDA with L1-norm optimization. *Neurocomputing* **2017**, *266*, 216–225. [CrossRef]
88. Yang, J.; Frangi, A.F.; Yang, J.y.; Zhang, D.; Jin, Z. KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 230–244. [CrossRef]
89. Ye, Q.; Yang, J.; Liu, F.; Zhao, C.; Ye, N.; Yin, T. L1-norm distance linear discriminant analysis based on an effective iterative algorithm. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *28*, 114–129. [CrossRef]
90. Nie, F.; Huang, H.; Cai, X.; Ding, C. Efficient and robust feature selection via joint l2, 1-norms minimization. Advances in Neural Information Processing Systems 23. In Proceedings of the 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010, Vancouver, BC, Canada, 6–9 December 2010; Volume 23, pp.1813–1821.
91. Wang, H.; Nie, F.; Huang, H. Learning robust locality preserving projection via p-order minimization. In Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 4, pp. 3059–3065.
92. Liang, Z.; Xia, S.; Zhou, Y.; Zhang, L.; Li, Y. Feature extraction based on Lp-norm generalized principal component analysis. *Pattern Recognit. Lett.* **2013**, *34*, 1037–1045. [CrossRef]
93. Wang, J. Generalized 2-D principal component analysis by Lp-norm for image analysis. *IEEE Trans. Cybern.* **2015**, *46*, 792–803. [CrossRef] [PubMed]
94. Kwak, N. Principal component analysis by L_p-norm maximization. *IEEE Trans. Cybern.* **2013**, *44*, 594–609. [CrossRef]
95. Tao, H.; Hou, C.; Nie, F.; Jiao, Y.; Yi, D. Effective discriminative feature selection with nontrivial solution. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *27*, 796–808. [CrossRef]
96. Wang, H.; Nie, F.; Cai, W.; Huang, H. Semi-supervised robust dictionary learning via efficient $l_{2,0+}$-norms minimization. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1145–1152.
97. Nie, F.; Huang, Y.; Wang, X.; Huang, H. New primal SVM solver with linear computational cost for big data classifications. In Proceedings of the 31st International Conference on International Conference on Machine Learning, Beijing, China, 21–26 June 2014.; Volume 32, pp. II–505–II–513.
98. Yang, M.H.; Kriegman, D.; Ahuja, N. Face detection using multimodal density models. *Comput. Vis. Image Underst.* **2001**, *84*, 264–284. [CrossRef]
99. Martínez, A.M.; Vitria, J. Clustering in image space for place recognition and visual annotations for human–robot interaction. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2001**, *31*, 669–682. [CrossRef] [PubMed]
100. Ju, J.; Kolaczyk, E.D.; Gopal, S. Gaussian mixture discriminant analysis and sub-pixel land cover characterization in remote sensing. *Remote Sens. Environ.* **2003**, *84*, 550–560. [CrossRef]
101. Choi, S.W.; Park, J.H.; Lee, I.B. Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. *Comput. Chem. Eng.* **2004**, *28*, 1377–1387. [CrossRef]
102. Lombardo, F.; Obach, R.S.; DiCapua, F.M.; Bakken, G.A.; Lu, J.; Potter, D.M.; Gao, F.; Miller, M.D.; Zhang, Y. A hybrid mixture discriminant analysis-random forest computational model for the prediction of volume of distribution of drugs in human. *J. Med. Chem.* **2006**, *49*, 2262–2267. [CrossRef]
103. Haeb-Umbach, R.; Geller, D.; Ney, H. Improvements in connected digit recognition using linear discriminant analysis and mixture densities. In Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, MN, USA, 27–30 April 1993; Volume 2, pp. 239–242.
104. Pekhovsky, T.; Sizov, A. Comparison between supervised and unsupervised learning of probabilistic linear discriminant analysis mixture models for speaker verification. *Pattern Recognit. Lett.* **2013**, *34*, 1307–1313. [CrossRef]
105. Calis, N.; Erol, H. A new per-field classification method using mixture discriminant analysis. *J. Appl. Stat.* **2012**, *39*, 2129–2140. [CrossRef]
106. Chamroukhi, F.; Glotin, H.; Sam, A. Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing* **2013**, *112*, 153–163. [CrossRef]
107. Pnevmatikakis, A.; Polymenakos, L. Subclass linear discriminant analysis for video-based face recognition. *J. Vis. Commun. Image Represent.* **2009**, *20*, 543–551. [CrossRef]
108. Nikitidis, S.; Tefas, A.; Nikolaidis, N.; Pitas, I. Subclass discriminant nonnegative matrix factorization for facial image analysis. *Pattern Recognit.* **2012**, *45*, 4080–4091. [CrossRef]
109. Di Cataldo, S.; Bottino, A.; Islam, I.U.; Vieira, T.F.; Ficarra, E. Subclass discriminant analysis of morphological and textural features for hep-2 staining pattern classification. *Pattern Recognit.* **2014**, *47*, 2389–2399. [CrossRef]
110. Mandal, B.; Fajtl, J.; Argyriou, V.; Monekosso, D.; Remagnino, P. Deep residual network with subclass discriminant analysis for crowd behavior recognition. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 938–942.
111. Xu, B.; Zhao, D.; Jia, K.; Zhou, J.; Tian, J.; Xiang, J. Cross-project aging-related bug prediction based on joint distribution adaptation and improved subclass discriminant analysis. In Proceedings of the 2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE), Coimbra, Portugal, 2–5 November 2020; pp. 325–334.

112. Pedagadi, S.; Orwell, J.; Velastin, S.; Boghossian, B. Local Fisher Discriminant Analysis for Pedestrian Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 3318–3325.
113. Prakash, P.S.; Rajkumar, N. Improved local fisher discriminant analysis based dimensionality reduction for cancer disease prediction. *J. Ambient Intell. Humaniz. Comput.* **2021**, *12*, 8083–8098. [CrossRef]
114. Zhang, S.; Zhao, X.; Lei, B. Facial expression recognition based on local binary patterns and local fisher discriminant analysis. *WSEAS Trans. Signal Process.* **2012**, *8*, 21–31.
115. Rahulamathavan, Y.; Phan, R.C.W.; Chambers, J.A.; Parish, D.J. Facial expression recognition in the encrypted domain based on local fisher discriminant analysis. *IEEE Trans. Affect. Comput.* **2012**, *4*, 83–92. [CrossRef]
116. Feng, J.; Wang, J.; Zhang, H.; Han, Z. Fault diagnosis method of joint fisher discriminant analysis based on the local and global manifold learning and its kernel version. *IEEE Trans. Autom. Sci. Eng.* **2015**, *13*, 122–133. [CrossRef]
117. Liu, H.; Yang, C.; Kim, M.J.; Yoo, C.K. Fault Diagnosis of Subway Indoor Air Quality Based on Local Fisher Discriminant Analysis. *Environ. Eng. Sci.* **2018**, *35*, 1206–1215. [CrossRef]
118. Shen, P.; Lu, X.; Liu, L.; Kawai, H. Local fisher discriminant analysis for spoken language identification. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5825–5829.
119. Deng, X.; Tian, X.; Chen, S.; Harris, C. Statistics local fisher discriminant analysis for industrial process fault classification. In Proceedings of the 2016 UKACC 11th International Conference on Control (CONTROL), Belfast, UK, 31 August–2 September 2016; pp. 1–6.
120. Yu, J. Localized Fisher discriminant analysis based complex chemical process monitoring. *AIChE J.* **2011**, *57*, 1817–1828. [CrossRef]
121. Kaya, H.; Özkaptan, T.; Salah, A.A.; Gürgen, S.F. Canonical correlation analysis and local fisher discriminant analysis based multi-view acoustic feature reduction for physical load prediction. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 442–446.
122. Zhang, S.; Lei, B.; Chen, A.; Chen, C.; Chen, Y. Spoken emotion recognition using local fisher discriminant analysis. In Proceedings of the IEEE 10th International Conference on Signal Processing, Beijing, China, 24–28 October 2010; pp. 538–540.
123. Wang, Z.; Ruan, Q.; An, G. Facial expression recognition using sparse local Fisher discriminant analysis. *Neurocomputing* **2016**, *174*, 756–766. [CrossRef]
124. Wu, W.; Tan, C.; Zhang, S.; Dong, F. Sparse local fisher discriminant analysis for gas-water two-phase flow status monitoring with multisensor signals. *IEEE Trans. Ind. Inform.* **2022**, *19*, 2886–2898. [CrossRef]
125. Wang, L.; Ji, H.; Shi, Y. Face recognition using maximum local fisher discriminant analysis. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 1737-1740.
126. Huang, H.; Feng, H.; Peng, C. Complete local Fisher discriminant analysis with Laplacian score ranking for face recognition. *Neurocomputing* **2012**, *89*, 64–77. [CrossRef]
127. Huang, H.; Liu, J.; Feng, H.; He, T. Ear recognition based on uncorrelated local Fisher discriminant analysis. *Neurocomputing* **2011**, *74*, 3103–3113. [CrossRef]
128. Jia, J.; Ruan, Q.; Jin, Y. Geometric preserving local fisher discriminant analysis for person re-identification. *Neurocomputing* **2016**, *205*, 92–105. [CrossRef]
129. Van, M.; Kang, H.J. Bearing defect classification based on individual wavelet local fisher discriminant analysis with particle swarm optimization. *IEEE Trans. Ind. Inform.* **2015**, *12*, 124–135. [CrossRef]
130. Li, F.; Wang, J.; Chyu, M.K.; Tang, B. Weak fault diagnosis of rotating machinery based on feature reduction with Supervised Orthogonal Local Fisher Discriminant Analysis. *Neurocomputing* **2015**, *168*, 505–519. [CrossRef]
131. Wang, Z.; Ruan, Q. Facial expression recognition based orthogonal local fisher discriminant analysis. In Proceedings of the IEEE 10th International Conference on Signal Processing, Beijing, China, 24–28 October 2010; pp. 1358-1361.
132. Wang, Z.; Ruan, Q.; An, G. Projection-optimal local Fisher discriminant analysis for feature extraction. *Neural Comput. Appl.* **2015**, *26*, 589–601. [CrossRef]
133. Guo, J.; Chen, H.; Li, Y. Palmprint Recognition Based on Local Fisher Discriminant Analysis. *J. Softw.* **2014**, *9*, 287–292. [CrossRef]
134. Liu, Z.; Wang, J.; Man, J.; Li, Y.; You, X.; Wang, C. Self-adaptive Local Fisher Discriminant Analysis for semi-supervised image recognition. *Int. J. Biomed.* **2012**, *4*, 338–356. [CrossRef]
135. Zaatour, R.; Bouzidi, S.; Zagrouba, E. Unsupervised image-adapted local fisher discriminant analysis to reduce hyperspectral images without ground truth. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7931–7941. [CrossRef]
136. Sun, Z.; Wang, Y.; Sun, G. Fault diagnosis of rotating machinery based on local centroid mean local fisher discriminant analysis. *J. Vib. Eng. Technol.* **2023**, *11*, 1417–1441. [CrossRef]
137. Zhou, J.; Wu, P.; Ye, H.; Song, Y.; Wu, X.; He, Y.; Pan, H. Fault diagnosis for blast furnace ironmaking process based on randomized local fisher discriminant analysis. *Can. J. Chem. Eng.* **2024**. [CrossRef]
138. Peng, J.; Zhao, L.; Gao, Y.; Yang, J. ILFDA Model: An Online Soft Measurement Method Using Improved Local Fisher Discriminant Analysis. *J. Adv. Comput. Intell. Intell. Inform.* **2024**, *28*, 284–295. [CrossRef]
139. Huang, H.; Li, J.; Liu, J. Enhanced semi-supervised local Fisher discriminant analysis for face recognition. *Future Gener. Comput. Syst.* **2012**, *28*, 244–253. [CrossRef]
140. Shao, Z.; Zhang, L. Sparse dimensionality reduction of hyperspectral image based on semi-supervised local Fisher discriminant analysis. *Int. J. Appl. Earth Obs. Geoinf.* **2014**, *31*, 122–129. [CrossRef]

141. Huang, H.; Li, J.; Feng, H.; Xiang, R. Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis. *Expert Syst. Appl.* **2012**, *39*, 2314–2320. [CrossRef]
142. Gao, Q.; Liu, J.; Cui, K.; Zhang, H.; Wang, X. Stable locality sensitive discriminant analysis for image recognition. *Neural Netw.* **2014**, *54*, 49–56. [CrossRef]
143. Yi, Y.; Zhang, B.; Kong, J.; Wang, J. An improved locality sensitive discriminant analysis approach for feature extraction. *Multimed. Tools Appl.* **2015**, *74*, 85–104. [CrossRef]
144. Jin, Y.; Ruan, Q.Q. Orthogonal Locality Sensitive Discriminant Analysis for Face Recognition. *J. Inf. Sci. Eng.* **2009**, *25*, 419–433.
145. Ding, Z.; Du, Y. Fusion of Log-Gabor wavelet and orthogonal locality sensitive discriminant analysis for face recognition. In Proceedings of the 2011 International Conference on Image Analysis and Signal Processing, Ravenna, Italy, 14–16 September 2011; pp. 177–180.
146. Raghavendra, U.; Acharya, U.R.; Fujita, H.; Gudigar, A.; Tan, J.H.; Chokkadi, S. Application of Gabor wavelet and Locality Sensitive Discriminant Analysis for automated identification of breast cancer using digitized mammogram images. *Appl. Soft Comput.* **2016**, *46*, 151–161. [CrossRef]
147. Yu, H.; Gao, L.; Li, W.; Du, Q.; Zhang, B. Locality sensitive discriminant analysis for group sparse representation-based hyperspectral imagery classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1358–1362. [CrossRef]
148. Zhang, X.; Zhang, Q.; Li, H.; Sun, Y.; Qin, X. Fault diagnosis using locality sensitive discriminant analysis for feature extraction. In Proceedings of the 2017 Prognostics and System Health Management Conference, Harbin, China, 9–12 July 2017; pp. 1–6.
149. Lu, J. Enhanced locality sensitive discriminant analysis for image recognition. *Electron. Lett.* **2010**, *46*, 213–214. [CrossRef]
150. Bala, A.; Rani, A.; Kumar, S. An Illumination Insensitive Normalization Approach to Face Recognition Using Locality Sensitive Discriminant Analysis. *Trait. Signal* **2020**, *37*, 451. [CrossRef]
151. Wei, Y.K.; Jin, C. Locality sensitive discriminant projection for feature extraction and face recognition. *J. Electron. Imaging* **2019**, *28*, 043028. [CrossRef]
152. Zhan, Y.; Liu, J.; Gou, J.; Wang, M. A video semantic detection method based on locality-sensitive discriminant sparse representation and weighted KNN. *J. Vis. Commun. Image Represent.* **2016**, *41*, 65–73. [CrossRef]
153. Prins, D.; Gool, V. Svm-based nonparametric discriminant analysis, an application to face detection. In Proceedings of the Proceedings Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; pp. 1289–1296.
154. Zhu, M.; Hastie, T.J. Feature extraction for nonparametric discriminant analysis. *J. Comput. Graph. Stat.* **2003**, *12*, 101–120. [CrossRef]
155. Zheng, Y.J.; Yang, J.Y.; Yang, J.; Wu, X.J.; Jin, Z. Nearest neighbour line nonparametric discriminant analysis for feature extraction. *Electron. Lett.* **2006**, *42*, 679–680. [CrossRef]
156. Li, Z.; Zhao, F.; Liu, J.; Qiao, Y. Pairwise nonparametric discriminant analysis for binary plankton image recognition. *IEEE J. Ocean. Eng.* **2013**, *39*, 695–701. [CrossRef]
157. Cao, G.; Iosifidis, A.; Gabbouj, M. Multi-view nonparametric discriminant analysis for image retrieval and recognition. *IEEE Signal Process. Lett.* **2017**, *24*, 1537–1541. [CrossRef]
158. Knick, S.T.; Rotenberry, J.T.; Zarriello, T.J. Supervised classification of Landsat Thematic Mapper imagery in a semi-arid rangeland by nonparametric discriminant analysis. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 79–86.
159. Tao, D.; Tang, X. Nonparametric discriminant analysis in relevance feedback for content-based image retrieval. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 23–26 August 2004; Volume 2, pp. 1013–1016.
160. Raducanu, B.; Vitrià, J. Online nonparametric discriminant analysis for incremental subspace learning and recognition. *Pattern Anal. Appl.* **2008**, *11*, 259–268. [CrossRef]
161. Li, J.B.; Sun, W.H.; Wang, Y.H.; Tang, L.L. 3D model classification based on nonparametric discriminant analysis with kernels. *Neural Comput. Appl.* **2013**, *22*, 771–781. [CrossRef]
162. Yu, B.; Ostland, M.; Gong, P.; Pu, R. Penalized discriminant analysis of in situ hyperspectral data for conifer species recognition. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 2569–2577. [CrossRef]
163. Huang, D.S.; Zheng, C.H. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* **2006**, *22*, 1855–1862. [CrossRef] [PubMed]
164. Ghosh, D. Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics* **2003**, *59*, 992–1000. [CrossRef] [PubMed]
165. Shahraki, H.R.; Bemani, P.; Jalali, M. Classification of bladder cancer patients via penalized linear discriminant analysis. *Asian Pac. J. Cancer Prev. APJCP* **2017**, *18*, 1453.
166. Wang, W.; Mo, Y.; Ozol. Penalized fisher discriminant analysis and its application to image-based morphometry. *Pattern Recognit. Lett.* **2011**. *32*, 2128–2135. [CrossRef]
167. Zhu, Y.; Tan, T.L. Penalized discriminant analysis for the detection of wild-grown and cultivated Ganoderma lucidum using Fourier transform infrared spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2016**, *159*, 68–77. [CrossRef]
168. Lu, M.; Hu, L.; Yue, T.; Chen, Z.; Chen, B.; Lu, X.; Xu, B. Penalized linear discriminant analysis of hyperspectral imagery for noise removal. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 359–363. [CrossRef]
169. Kustra, R.; Strother, S. Penalized discriminant analysis of $[^{15}O]$-water PET brain images with prediction error selection of smoothness and regularization hyperparameters. *IEEE Trans. Med. Imaging* **2001**, *20*, 376–387. [CrossRef] [PubMed]
170. Guo, Y.; Hastie, T.; Tibshirani, R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **2007**, *8*, 86–100. [CrossRef] [PubMed]

171. Grosenick, L.; Klingenberg, B.; Greer, S.; Taylor, J.; Knutson, B. Whole-brain sparse penalized discriminant analysis for predicting choice. *NeuroImage* **2009**, *47*, S58. [CrossRef]
172. Bodesheim, P.; Freytag, A.; Rodner, E.; Kemmler, M.; Denzler, J. Kernel null space methods for novelty detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3374–3381.
173. Liu, J.; Lian, Z.; Wang, Y.; Xiao, J. Incremental kernel null space discriminant analysis for novelty detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4123–4131.
174. Zhang, L.; Xiang, T.; Gong, S. Learning a discriminative null space for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1239–1248.
175. Ali, T.; Chaudhuri, S. Maximum margin metric learning over discriminative nullspace for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 122–138.
176. Swets, D.L.; Weng, J.J. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 831–836. [CrossRef]
177. Chen, X.; Yang, J.; Jin, Z. An improved linear discriminant analysis with L1-norm for robust feature extraction. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 1585–1590.
178. Markopoulos, P.P.; Zlotnikov, S.; Ahmad, F. Adaptive radar-based human activity recognition with L1-norm linear discriminant analysis. *IEEE J. Electromagn. RF Microw. Med. Biol.* **2019**, *3*, 120–126. [CrossRef]
179. Zhou, W.; Kamata, S.i. L1-norm based linear discriminant analysis: An application to face recognition. *IEICE Trans. Inf. Syst.* **2013**, *96*, 550–558. [CrossRef]
180. Shi, X.; Yang, Y.; Guo, Z.; Lai, Z. Face recognition by sparse discriminant analysis via joint L2, 1-norm minimization. *Pattern Recognit.* **2014**, *47*, 2447–2453. [CrossRef]
181. Li, C.N.; Qi, Y.F.; Shao, Y.H.; Guo, Y.R.; Ye, Y.F. Robust two-dimensional capped l2, 1-norm linear discriminant analysis with regularization and its applications on image recognition. *Eng. Appl. Artif. Intell.* **2021**, *104*, 104367. [CrossRef]
182. Paulsen, V.I.; Raghupathi, M. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*; Cambridge University Press: Cambridge, UK, 2016; Volume 152.
183. Qu, L.; Pei, Y.; Li, J. A Data Analysis Method Using Orthogonal Transformation in a Reproducing Kernel Hilbert Space. In Proceedings of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Honolulu, HI, USA, 1–4 October 2023; pp. 887–892.
184. Chen, B.; Yuan, L.; Liu, H.; Bao, Z. Kernel subclass discriminant analysis. *Neurocomputing* **2007**, *71*, 455–458. [CrossRef]
185. You, D.; Hamsici, O.C.; Martinez, A.M. Kernel optimization in discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 631–638. [CrossRef] [PubMed]
186. Zhong, K.; Han, M.; Qiu, T.; Han, B. Fault diagnosis of complex processes using sparse kernel local Fisher discriminant analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 1581–1591. [CrossRef]
187. Wang, Z.; Sun, X. Multiple kernel local Fisher discriminant analysis for face recognition. *Signal Process.* **2013**, *93*, 1496–1509. [CrossRef]
188. Zhang, Q.; Li, H.; Zhang, X.; Wang, H. Optimal multi-kernel local fisher discriminant analysis for feature dimensionality reduction and fault diagnosis. *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.* **2021**, *235*, 1041–1056. [CrossRef]
189. Huang, S.C.; Tang, Y.C.; Lee, C.W.; Chang, M.J. Kernel local Fisher discriminant analysis based manifold-regularized SVM model for financial distress predictions. *Expert Syst. Appl.* **2012**, *39*, 3855–3861. [CrossRef]
190. Van, M.; Kang, H.J. Wavelet kernel local fisher discriminant analysis with particle swarm optimization algorithm for bearing defect classification. *IEEE Trans. Instrum. Meas.* **2015**, *64*, 3588–3600. [CrossRef]
191. Wang, Z.; Sun, X. Manifold Adaptive Kernel Local Fisher Discriminant Analysis for Face Recognition. *J. Multimed.* **2012**, *7*, 387. [CrossRef]
192. Qin, X.; Chiang, C.W.; Gaggiotti, O.E. Kernel local fisher discriminant analysis of principal components (KLFDAPC) significantly improves the accuracy of predicting geographic origin of individuals. *bioRxiv* **2021**. [CrossRef]
193. Li, W.; Prasad, S.; Fowler, J.E.; Bruce, L.M. Locality-preserving discriminant analysis in kernel-induced feature spaces for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 894–898. [CrossRef]
194. Tao, X.; Ren, C.; Li, Q.; Guo, W.; Liu, R.; He, Q.; Zou, J. Bearing defect diagnosis based on semi-supervised kernel Local Fisher Discriminant Analysis using pseudo labels. *ISA Trans.* **2021**, *110*, 394–412. [CrossRef]
195. Diaf, A.; Boufama, B.; Benlamri, R. Non-parametric Fishers discriminant analysis with kernels for data classification. *Pattern Recognit. Lett.* **2013**, *34*, 552–558. [CrossRef]
196. Yan, F.; Mikolajczyk, K.; Barnard, M.; Cai, H.; Kittler, J. Lp norm multiple kernel fisher discriminant analysis for object and image categorisation. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3626–3632.
197. Pei, Y. Linear principal component discriminant analysis. In Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics, Kowloon, Hong Kong, 9–12 October 2015; pp. 2108–2113.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.