

A Relationship between LDA and the Generalized Minimum Squared Error Solution*

Cheong Hee Park¹ and Haesun Park² †

Dept. of Computer Science and Engineering^{1,2}

University of Minnesota

Minneapolis, MN 55455, USA

{chpark,hpark}@cs.umn.edu

The National Science Foundation²

4201 Wilson Boulevard, Arlington, Virginia 22230, USA

Abstract

In this paper, a relationship between Linear Discriminant Analysis (LDA) and the generalized Minimum Squared Error (MSE) solution is presented. The generalized MSE solution is shown to be equivalent to applying a certain classification rule in the space defined by LDA. The relationship between the MSE solution and Fisher Discriminant Analysis (FDA) is extended to multi-class problems and also to undersampled problems for which the classical LDA is not applicable due to singularity of the scatter matrices. In addition, an efficient algorithm for LDA is proposed exploiting its relationship with the MSE procedure. Extensive experiments verify the theoretical results.

Keywords. Dimension reduction method, Discriminant function, Generalized singular value decomposition, Linear Discriminant Analysis, Minimum Squared Error Solution, Pseudo-inverse, Undersampled problems.

1 Introduction

The extensive utilization of linear discriminant functions for pattern recognition is attributed to their simple concept and ease of computation. In linear discriminant function-based methods, the existence of a hyperplane which can optimally separate two classes is assumed. The optimality of the separating hyperplanes can be measured by various criteria, and numerous methods have

*This material is based upon work supported in part by the National Science Foundation Grants CCR-0204109 and ACI-0305543. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

†The work of Haesun Park has been performed while at the NSF (National Science Foundation) and was partly supported by IR/D from the NSF.

been proposed originating from the work by R.A. Fisher [1]. In the Perceptron method, a linear discriminant function is obtained by iterative procedures to reduce the amount of misclassified training data [2]. Support Vector Machines (SVM) search for a linear function which maximizes the margin between classes either in the original data space or nonlinearly transformed feature spaces [3]. The Minimum Squared Error (MSE) solution seeks a linear discriminant function that minimizes the squared error [4, 5]. Bayesian classifiers are based on estimation of distribution functions by which data is generated. From the estimated class density function, the class posterior probability is computed using Bayes Theorem and a new data item is assigned to the class having the maximum posterior probability. In particular, assuming normal density functions for each class, where different means but a common covariance are to be estimated, Bayesian classification rules turn into a set of linear discriminant functions [6, 7].

While being similar in that linear functions are utilized, Linear Discriminant Analysis (LDA) is different from the above mentioned methods since LDA is a dimension reduction method rather than a discriminant classifier. LDA finds a linear transformation that can maximize the class separability in the reduced dimensional space. The criterion used in LDA is to find a dimension reducing transformation that maximizes the between-class scatter and minimizes the within-class scatter [7]. When the dataset has two classes, the relationship between the MSE solution and Fisher Discriminant Analysis (FDA) has been studied [8, 4], where FDA is a special case of LDA for two-class cases. Since both FDA and the MSE solution in two-class cases deals with *one linear function*, the relationship between them follows naturally. The MSE procedure is generalized for multi-class problems by setting up multiple two-class problems each of which is constructed by one class and the remaining data points forming the other class [4, 6]. In this paper, we develop the relationship between LDA and the generalized MSE procedure for multi-class problems and also for undersampled problems. Utilizing the developed relationships, it is shown that the MSE solution can be obtained by applying a certain classification rule in the reduced dimensional space obtained by LDA, and conversely LDA can be performed through the MSE procedure without solving the eigenvalue problem explicitly.

The term *Linear Discriminant Analysis (LDA)* has also been used to denote Bayesian linear classifiers resulting from the assumption of normal density functions with a common covariance. We note that in the rest of paper LDA refers to a dimension reduction method. Many problems including generalization of Bayesian linear classifiers have been studied [9, 10], and under certain restrictions the relationships between Bayesian linear classifiers and LDA were investigated [11, 10]. While the relationships developed in this paper are applicable for both oversampled and undersampled problems, the results in [11, 10] are restricted to oversampled problems. For the singular or ill-conditioned covariance which occurs in undersampled problems, regularization methods can be applied for computation of the eigenvalue decomposition [12, 13]. However, the estimation of the regularization parameters can be expensive and generalization errors by overfitting, especially in undersampled problems, should be taken care of.

The rest of the paper is organized as follows. In Sections 2 and 3, LDA and the MSE procedures are reviewed. In Section 4, we generalize the relation between the MSE solution and FDA for undersampled problems for which the classical FDA fails due to the singularity of the scatter matrices. We also derive the relationship between LDA and the generalized MSE solution for multi-class problems. In Section 5, we propose an efficient algorithm for LDA which utilizes the relationship with the MSE solution and does not require the solution of eigenvalue problems. The experimental results in Section 6 verify the theoretical results.

2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a linear dimension reduction method which can be used as a preprocessing step for data analysis. Based on the information from the given data, LDA finds a linear transformation that maximizes the between-class distances and minimizes the within-class scatter so that the class separability can be optimized in the transformed space. Throughout the paper, we assume the vector space representation of a data set A

$$A = [a_1, \dots, a_n] = [A_1, A_2, \dots, A_r] \in \mathbb{R}^{m \times n} \quad (1)$$

where each data item in the m -dimensional space is represented as a column vector a_i and a collection of data items in the i -th class as a block matrix $A_i \in \mathbb{R}^{m \times n_i}$. Each class i ($1 \leq i \leq r$) has n_i elements and the total number of data items is $n = \sum_{i=1}^r n_i$. Let N_i ($1 \leq i \leq r$) be the index set of data items in the class i . The data set A can be considered a training set on which the modelling of data analysis algorithms is based. For example, searching for a linear transformation for LDA and discriminant functions for the MSE procedure.

Given a data set A , the between-class scatter matrix S_b , within-class scatter matrix S_w , and total scatter matrix S_t are defined as

$$\begin{aligned} S_b &= \sum_{i=1}^r n_i (c_i - c)(c_i - c)^T, & S_w &= \sum_{i=1}^r \sum_{j \in N_i} (a_j - c_i)(a_j - c_i)^T, \\ S_t &= S_b + S_w = \sum_{j=1}^n (a_j - c)(a_j - c)^T, \end{aligned} \quad (2)$$

where $c_i = \frac{1}{n_i} \sum_{j \in N_i} a_j$ and $c = \frac{1}{n} \sum_{j=1}^n a_j$ are class centroids and the global centroid, respectively. The *traces* of the scatter matrices can be used to measure the quality of the cluster structure in the data set as

$$\text{trace}(S_b) = \sum_{i=1}^r n_i \|c_i - c\|_2^2, \quad \text{trace}(S_w) = \sum_{i=1}^r \sum_{j \in N_i} \|a_j - c_i\|_2^2.$$

The distance between classes is quantified by $\text{trace}(S_b)$, and $\text{trace}(S_w)$ measures the scatter within classes. The optimal dimension reducing transformation $G^T \in \mathbb{R}^{l \times m}$ for LDA is the one that maximizes

$$J(G) = \text{trace}((G^T S_w G)^{-1} (G^T S_b G)) \quad (3)$$

where $G^T S_b G$ and $G^T S_w G$ are scatter matrices in the transformed space. It is well known [7] that the criterion in (3) is maximized when the columns of $G \in \mathbb{R}^{m \times (r-1)}$ are the eigenvectors x corresponding to the $r - 1$ largest eigenvalue λ of

$$S_b x = \lambda S_w x. \quad (4)$$

When S_w is nonsingular, one can solve the eigenvalue problem

$$S_w^{-1} S_b x = \lambda x, \quad (5)$$

referred to as the classical LDA.

In order to overcome some limitations in the classical LDA, several generalization methods have been proposed. The problems caused by the singularity of the scatter matrices on undersampled problems are circumvented by two-stage decompositions of the scatter matrices [14, 15, 16], and the criterion itself of LDA is criticized in [17]. Howland et al. [18, 19] applied the Generalized Singular Value Decomposition (GSVD) due to Paige and Saunders [20] which is applicable for undersampled problems. We briefly review the method in [18] and give a new approach to it, which will be used in deriving the relationship between LDA and the generalized MSE solution.

When the GSVD [20] is applied to two matrices Z_b and Z_w with the same number of columns, p , we have

$$U_b^T Z_b X = \left[\underbrace{\Gamma_b}_s \quad \underbrace{0}_{p-s} \right] \quad \text{and} \quad U_w^T Z_w X = \left[\underbrace{\Gamma_w}_s \quad \underbrace{0}_{p-s} \right] \quad \text{for} \quad s = \text{rank} \left(\begin{bmatrix} Z_b \\ Z_w \end{bmatrix} \right) \quad (6)$$

where U_b and U_w are orthogonal and X is nonsingular,

$$\Gamma_b^T \Gamma_b + \Gamma_w^T \Gamma_w = I_s$$

and $\Gamma_b^T \Gamma_b$ and $\Gamma_w^T \Gamma_w$ are diagonal matrices with nonincreasing and nondecreasing diagonal components respectively. The method due to Howland et al. [18] utilizes the fact that the scatter matrices can be expressed as

$$\begin{aligned} S_b &= H_b H_b^T, & H_b &= [\sqrt{n_1}(c_1 - c), \dots, \sqrt{n_r}(c_r - c)] \in \mathbb{R}^{m \times r}, \\ S_w &= H_w H_w^T, & H_w &= [A_1 - c_1 e_1^T, \dots, A_r - c_r e_r^T] \in \mathbb{R}^{m \times n}, \\ S_t &= H_t H_t^T, & H_t &= [a_1 - c, \dots, a_n - c] \in \mathbb{R}^{m \times n}, \end{aligned}$$

where $e_i = [1, \dots, 1]^T \in \mathbb{R}^{n_i \times 1}$. Suppose the GSVD is applied to the matrix pair (H_b^T, H_w^T) , and we obtain

$$U_b^T H_b^T X = [\Gamma_b \quad 0] \quad \text{and} \quad U_w^T H_w^T X = [\Gamma_w \quad 0]. \quad (7)$$

From (7) and $\Gamma_b^T \Gamma_b + \Gamma_w^T \Gamma_w = I_s$,

$$X^T S_b X = \begin{bmatrix} \Gamma_b^T \Gamma_b & \\ & 0_{m-s} \end{bmatrix} \equiv \begin{bmatrix} I_\mu & & & \\ & D_\tau & & \\ & & 0_{s-\mu-\tau} & \\ & & & 0_{m-s} \end{bmatrix} \quad \text{and} \quad (8)$$

$$X^T S_w X = \begin{bmatrix} \Gamma_w^T \Gamma_w & \\ & 0_{m-s} \end{bmatrix} \equiv \begin{bmatrix} 0_\mu & & & \\ & E_\tau & & \\ & & I_{s-\mu-\tau} & \\ & & & 0_{m-s} \end{bmatrix}, \quad (9)$$

where the subscripts on I and 0 denote the order of square identity and zero matrices. Denoting the diagonal elements in (8) as η_i 's and the diagonal elements in (9) as ζ_i 's, we have

$$\zeta_i S_b x_i = \eta_i S_w x_i \quad i = 1, \dots, m, \quad (10)$$

where x_i are the column vectors of X . Note that $x_i, i = s+1, \dots, m$, belong to $\text{null}(S_b) \cap \text{null}(S_w)$ and therefore they do not convey any discriminant information. Since

$$\eta_1 \geq \dots \geq \eta_s \quad \text{and} \quad \zeta_1 \leq \dots \leq \zeta_s,$$

the $r - 1$ leftmost columns of X give an optimal transformation for LDA. This method is called LDA/GSVD [18, 19].

The algorithm to compute the GSVD for the pair (H_b^T, H_w^T) was presented in [18] as follows.

1. Compute the SVD of $Z = \begin{bmatrix} H_b^T \\ H_w^T \end{bmatrix} \in \mathbb{R}^{(r+n) \times m}$: $Z = P \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} U^T$ where $s = \text{rank}(Z)$ and $P \in \mathbb{R}^{(r+n) \times (r+n)}$ and $U \in \mathbb{R}^{m \times m}$ are orthogonal and the diagonal components of $\Sigma \in \mathbb{R}^{s \times s}$ are nonincreasing.
2. Compute V from the SVD of $P(1 : r, 1 : s)$, which is $P(1 : r, 1 : s) = W \Gamma_b V^T$.
3. Compute the first $r - 1$ columns of $X = U \begin{bmatrix} \Sigma^{-1} V & 0 \\ 0 & I \end{bmatrix}$, and assign them to the transformation matrix G .

Since $\Gamma_b^T \Gamma_b + \Gamma_w^T \Gamma_w = I_s$, from (8-9), we have

$$X^T S_t X = X^T S_b X + X^T S_w X = \begin{bmatrix} I_s & 0 \\ 0 & 0 \end{bmatrix} \quad (11)$$

where $s = \text{rank}(Z)$. Eq. (11) implies $s = \text{rank}(S_t)$ and from the step 3 in the LDA/GSVD algorithm

$$S_t = X^{-T} \begin{bmatrix} I_s & 0 \\ 0 & 0 \end{bmatrix} X^{-1} = U \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} U^T \quad (12)$$

which results in the EVD of S_t . Partitioning U as

$$U = \left[\underbrace{U_1}_s \quad \underbrace{U_2}_{m-s} \right],$$

we have

$$X = U \begin{bmatrix} \Sigma^{-1} V & 0 \\ 0 & I \end{bmatrix} = \left[U_1 \Sigma^{-1} V \quad U_2 \right]. \quad (13)$$

By substituting X in (8) with Eq. (13),

$$\Sigma^{-1} U_1^T S_b U_1 \Sigma^{-1} = V \Gamma_b^T \Gamma_b V^T. \quad (14)$$

Note that the optimal transformation matrix G by LDA/GSVD is obtained by the leftmost $r - 1$ columns of X , which are the leftmost $r - 1$ columns of $U_1 \Sigma^{-1} V$. Hence (12) and (14) show that the solution to LDA/GSVD can be obtained as

1. Compute the EVD of S_t :

$$S_t = [U_1 \quad U_2] \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix}. \quad (15)$$

2. Compute V from the EVD of $\tilde{S}_b \equiv \Sigma^{-1}U_1^T S_b U_1 \Sigma^{-1}$: $\tilde{S}_b = V\Gamma_b^T \Gamma_b V^T$.

3. Assign the first $r - 1$ columns of $U_1 \Sigma^{-1}V$ to G .

In the step 2 of the new approach, denoting

$$F = U_1 \Sigma^{-1} \in \mathbb{R}^{m \times s}, \quad (16)$$

the EVD of \tilde{S}_b can be computed from the SVD of $F^T H_b$ as

$$F^T H_b = V\Gamma_b^T S^T = \begin{bmatrix} \underbrace{V_1}_{r-1} & \underbrace{V_2}_{s-r+1} \end{bmatrix} \begin{bmatrix} \Gamma_{b1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} S_1^T \\ S_2^T \end{bmatrix} \quad (17)$$

where $V \in \mathbb{R}^{s \times s}$, $S \in \mathbb{R}^{r \times r}$ are orthogonal and $\Gamma_{b1} \in \mathbb{R}^{(r-1) \times (r-1)}$ is a diagonal matrix with nonincreasing diagonal elements. Hence

$$X = [U_1 \Sigma^{-1}V \quad U_2] = [FV_1 \quad FV_2 \quad U_2] \quad (18)$$

and the transformation matrix G is given as

$$G = FV_1 = U_1 \Sigma^{-1}V_1. \quad (19)$$

For any $x \in \text{null}(S_b) \cap \text{null}(S_w)$,

$$\begin{aligned} 0 &= x^T S_b x = (x^T H_b)(H_b^T x) = \|x^T H_b\|^2 = \sum_{i=1}^r n_i |x^T c_i - x^T c|^2 \quad \text{and} \\ 0 &= x^T S_w x = \sum_{j=1}^n |x^T a_j - x^T c_i|^2 \quad \text{where } a_j \text{ belongs to the } i\text{-th class.} \end{aligned}$$

Hence

$$\begin{cases} x^T c_i = x^T c & \text{for } i = 1, \dots, r \\ x^T a_j = x^T c_i & \text{for all } j \text{ in } N_i \text{ and } i = 1, \dots, r, \end{cases}$$

therefore

$$U_2^T z = U_2^T c \quad (20)$$

for any given data item z . This implies that the vectors x_i , $i = s + 1, \dots, m$, belonging to $\text{null}(S_b) \cap \text{null}(S_w)$ do not convey discriminative information among the classes, even though the corresponding eigenvalues are not necessarily zero. Since $\text{rank}(S_b) \leq r - 1$, from Eqs. (8-9)

$$x_i^T S_b x_i = 0 \quad \text{and} \quad x_i^T S_w x_i = 1 \quad \text{for } r \leq i \leq s,$$

and the between-class scatter becomes zero by the projection onto the vector x_i . Hence it is justifiable that the linear transformation G^T for LDA can be formed by taking the first $r - 1$ columns from $[FV \quad U_2] = [FV_1 \quad FV_2 \quad U_2]$.

3 Minimum Squared Error (MSE) Solution

The Minimum Squared Error (MSE) solution in a two-class problem (i.e., $r = 2$) seeks a linear discriminant function

$$g(z) = w_0 + w^T z$$

for which

$$g(z) = w_0 + w^T z = \begin{cases} \beta_1, & \text{if } z \in \text{class 1} \\ \beta_2, & \text{if } z \in \text{class 2} \end{cases}, \quad (21)$$

where β_i is the prespecified number for each class. For the data set A given in (1), the problem (21) can be reformulated to minimize the squared error

$$\left\| \begin{bmatrix} 1 & a_1^T \\ \vdots & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \right\|_2^2 \quad (22)$$

where $y_i = \beta_1$ if a_i is in the class 1 and $y_i = \beta_2$ if a_i is in the class 2. Denoting

$$\mathcal{P} = \begin{bmatrix} 1 & a_1^T \\ \vdots & \vdots \\ 1 & a_n^T \end{bmatrix}, \quad (23)$$

a solution which minimizes the squared error (22) can be computed using the pseudo-inverse \mathcal{P}^+ of \mathcal{P} as

$$\begin{bmatrix} w_0 \\ w \end{bmatrix} = \mathcal{P}^+ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \quad (24)$$

When the number of columns of \mathcal{P} is greater than the number of rows, i.e., $m + 1 > n$, the least squares problem of (22) is underdetermined and there may exist infinitely many solutions. The one given in (24) is one of many possible solutions.

Different choices of $\beta = [\beta_1, \beta_2]^T$ would give different discriminant functions. In particular, when $\beta_1 = n/n_1$ and $\beta_2 = -n/n_2$, the MSE solution is related to the Fisher Discriminant Analysis (FDA) [4]. The vector w in (24) is the same as the solution x of FDA except for some scaling factor α as

$$w = \alpha S_w^{-1}(c_1 - c_2) \equiv \alpha x \quad \text{and} \quad w_0 = -w^T c \quad (25)$$

where c and c_i are the global and class centroids, respectively. A new data item is assigned to class 1 if

$$w^T z + w_0 = w^T(z - c) = \alpha x^T(z - c) > 0, \quad (26)$$

otherwise it is assigned to class 2.

The MSE procedure is generalized to multi-class cases as a set of multiple two-class problems [4]. For each class i ($1 \leq i \leq r$), the MSE solution to the problem

$$g_i(z) = w_{0i} + w_i^T z = \begin{cases} \beta_i, & \text{if } z \in \text{class } i \\ 0, & \text{otherwise} \end{cases} \quad (27)$$

is to be found. The solution of the multi-class problem (27) in contrast to the problem (21) will be referred to as the *generalized* MSE solution whenever the distinction is needed. As in [4], one choice for β_i would be assigning $\beta_i = 1$ for $i = 1, \dots, r$. The squared error function in the multi-class problem is expressed using the Frobenius norm as

$$\left\| \begin{bmatrix} 1 & a_1^T \\ \vdots & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_{01} & \cdots & w_{0r} \\ w_1 & \cdots & w_r \end{bmatrix} - \begin{bmatrix} y_{11} & \cdots & y_{1r} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nr} \end{bmatrix} \right\|_F^2, \quad (28)$$

where $y_{ji} = \beta_i$ if a_j belongs to the class i , and 0 otherwise. Denoting

$$\mathcal{W} = \begin{bmatrix} w_{01} & \cdots & w_{0r} \\ w_1 & \cdots & w_r \end{bmatrix} \quad \text{and} \quad \mathcal{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1r} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nr} \end{bmatrix} \quad (29)$$

and with \mathcal{P} defined as in (23), the MSE solution of the problem (28) can be obtained by

$$\mathcal{W} = \mathcal{P}^+ \mathcal{Y}, \quad (30)$$

and a new data item z is assigned to the class i if for all $j \neq i$

$$g_i(z) > g_j(z). \quad (31)$$

Let us consider the mapping defined by the discriminant functions of the MSE solution (30) as

$$x \longrightarrow [g_1(x), \dots, g_r(x)]^T \in \mathbb{R}^{r \times 1}. \quad (32)$$

Then Eq. (28) can be represented as

$$\|\mathcal{P}\mathcal{W} - \mathcal{Y}\|_F^2 = \sum_{1 \leq i \leq r} \sum_{j \in \mathcal{N}_i} \left\| \begin{bmatrix} g_1(a_j) \\ \vdots \\ g_r(a_j) \end{bmatrix} - \beta_i b_i \right\|_2^2$$

where $b_i \in \mathbb{R}^{r \times 1}$ ($1 \leq i \leq r$) is the column vector with 1 in the i -th position and 0 elsewhere. Hence in the space transformed by the mapping (32), the i -th class centroid will be mapped close to the point $\beta_i b_i$. Figure 1 visualizes the transformed spaces by LDA and the mapping (32) where β_i was set to 1, $1 \leq i \leq 3$, for a problem with three classes. The figures on the first row were obtained by the mapping (32) resulting in the dimension which is the same as the number of classes, while the second row figures show the reduced dimensional space by LDA for which the dimension is one less than the number of classes. The first two figures on the top were obtained by randomly

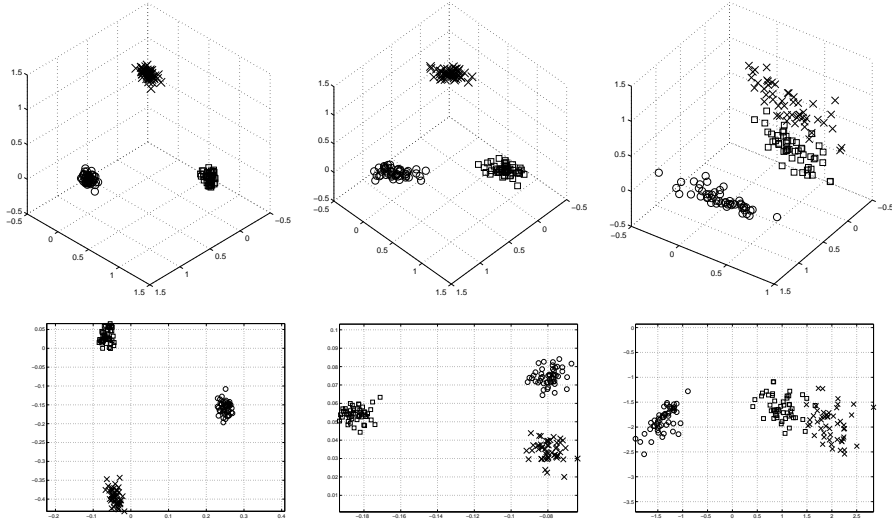


Figure 1: The visualization of the spaces transformed by the mapping (32) induced by the generalized MSE solution [the first row] and LDA [the second row].

taking three subclasses in the Isolet dataset from the UCI Machine Learning Repository ¹. The Isolet dataset has 26 classes and a detailed explanation of the dataset will be given in Section 6. The third figure on the top which was obtained by the Iris dataset illustrates that two classes among three classes are not well separable. The figures in the second row show the transformed space by LDA which corresponds to the figures on the top. The corresponding figures look quite similar.

What is the mathematical relationship between the two methods? If there is any relationship, is it possible to take advantage of the merits from each method and combine them? In the next section, we answer these questions by studying the relationship between LDA and the generalized MSE solution for multi-class problems.

4 Relationships between LDA and the generalized MSE solution

The relationship of the MSE solution and FDA given in (25) holds when the within-class scatter matrix S_w is nonsingular. Now we show that the relationship (25) can be generalized for multi-class and undersampled problems by using the algorithm discussed in Section 2.

4.1 FDA and the MSE Solution on Undersampled Problems

Let

$$g(z) = w_0 + w^T z$$

¹<http://www.ics.uci.edu/~mlern/MLRepository.html>

be the MSE solution to the problem

$$g(z) = w_0 + w^T z = \begin{cases} n/n_1, & \text{if } z \in \text{class 1} \\ -n/n_2, & \text{if } z \in \text{class 2} \end{cases} . \quad (33)$$

The normal equations for the problem in Eq. (22) are

$$\begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} 1 & a_1^T \\ \vdots & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} \frac{n}{n_1} e_{n_1} \\ -\frac{n}{n_2} e_{n_2} \end{bmatrix} \quad (34)$$

where e_{n_i} is the $n_i \times 1$ column vector with elements 1. From Eq. (34), we obtain

$$\begin{cases} nw_0 + nc^T w = 0 \\ ncw_0 + (\sum_{1 \leq j \leq n} a_j a_j^T) w = \frac{n}{n_1} \sum_{j \in N_1} a_j - \frac{n}{n_2} \sum_{j \in N_2} a_j. \end{cases} \quad (35)$$

From the first equation in (35) we have

$$w_0 = -c^T w. \quad (36)$$

By substituting Eq. (36) in the second equation of (35) and using the expressions of S_b and S_w for two-class problems

$$S_b = \sum_{1 \leq i \leq 2} n_i c_i c_i^T - ncc^T \quad \text{and} \quad S_w = \sum_{1 \leq j \leq n} a_j a_j^T - \sum_{1 \leq i \leq 2} n_i c_i c_i^T,$$

we obtain

$$(S_b + S_w)w = n(c_1 - c_2). \quad (37)$$

Let x_1 be the first column vector of $[FV_1 \ FV_2 \ U_2]$ in Eq. (18). From the discussion given in Section 2 and the fact that $\text{rank}(S_b) = 1$, we have

$$\zeta_i S_b x_i = \eta_i S_w x_i \quad \text{for} \quad \eta_1 > \eta_2 = \cdots = \eta_m = 0.$$

Since $\eta_1 + \zeta_1 = 1$,

$$\eta_1 (S_b + S_w) x_1 = (\eta_1 + \zeta_1) S_b x_1 = S_b x_1 = \frac{n_1 n_2}{n} (c_1 - c_2)(c_1 - c_2)^T x_1. \quad (38)$$

Denoting

$$\mu = \eta_1 \frac{n^2}{n_1 n_2 (c_1 - c_2)^T x_1},$$

Eq. (38) becomes

$$(S_b + S_w) \mu x_1 = n(c_1 - c_2). \quad (39)$$

Then by Eqs. (37) and (39), we have

$$S_t w = (S_b + S_w)w = (S_b + S_w)\mu x_1 = S_t \mu x_1. \quad (40)$$

From (40) and the EVD of S_t in Eq. (15),

$$U_1 \Sigma^2 U_1^T w = U_1 \Sigma^2 U_1^T \mu x_1 \quad \text{and} \quad U_1^T w = U_1^T \mu x_1$$

and from (20) and (36) we obtain

$$\begin{aligned} w^T z + w_0 &= w^T (z - c) = w^T (U_1 U_1^T + U_2 U_2^T)(z - c) \\ &= w^T U_1 U_1^T (z - c) = \mu x_1^T U_1 U_1^T (z - c) \\ &= \mu x_1^T (U_1 U_1^T + U_2 U_2^T)(z - c) = \mu x_1^T (z - c). \end{aligned} \quad (41)$$

Eq. (41) gives the relation between the MSE solution and the generalized solution of FDA, which holds regardless of the singularity of the scatter matrices.

While FDA gives a 1-dimensional reduced representation and the MSE solution produces one discriminant function, the generalized MSE solution works with r linear discriminant functions and LDA gives an $r - 1$ dimensional representation of the original data space. Now we show the relationship between LDA and the generalized MSE solution.

4.2 LDA and the Generalized MSE Solution

The generalized MSE solution to the problem

$$g_i(z) = w_{0i} + w_i^T z = \begin{cases} \beta_i, & \text{if } z \in \text{class } i \\ 0, & \text{otherwise.} \end{cases} \quad \text{for } i = 1, \dots, r$$

can be solved by the normal equations

$$\mathcal{P}^T \mathcal{P} \mathcal{W} = \mathcal{P}^T \mathcal{Y} \quad (42)$$

where \mathcal{P} , \mathcal{Y} and \mathcal{W} were defined in (23) and (29). From Eq. (42), we obtain

$$\begin{bmatrix} n & \sum_{j=1}^n a_j^T \\ \sum_{j=1}^n a_j & \sum_{j=1}^n a_j a_j^T \end{bmatrix} \begin{bmatrix} w_{01} & \cdots & w_{0r} \\ w_1 & \cdots & w_r \end{bmatrix} = \begin{bmatrix} n_1 \beta_1 & \cdots & n_r \beta_r \\ (\sum_{j \in N_1} a_j) \beta_1 & \cdots & (\sum_{j \in N_r} a_j) \beta_r \end{bmatrix},$$

resulting in a linear system

$$\begin{cases} n w_{0i} + n c^T w_i = n_i \beta_i \\ n c w_{0i} + (\sum_{j=1}^n a_j a_j^T) w_i = n_i \beta_i c_i \end{cases} \quad \text{for } i = 1, \dots, r. \quad (43)$$

By substituting w_{0i} of the second equation with w_{0i} of the first equation in (43), (43) becomes

$$(n_i \beta_i - n c^T w_i) c + \left(\sum_{j=1}^n a_j a_j^T \right) w_i = n_i \beta_i c_i, \quad i = 1, \dots, r. \quad (44)$$

From (44) and

$$S_b = \sum_{1 \leq i \leq r} n_i c_i c_i^T - n c c^T \quad \text{and} \quad S_w = \sum_{1 \leq j \leq n} a_j a_j^T - \sum_{1 \leq i \leq r} n_i c_i c_i^T,$$

we have

$$S_t w_i = (S_b + S_w) w_i = n_i \beta_i (c_i - c), \quad i = 1, \dots, r. \quad (45)$$

Recall that according to Eqs. (16), (17) and (19), the transformation matrix G for LDA was obtained by

$$G = F V_1 = U_1 \Sigma^{-1} V_1 \quad (46)$$

where

$$S_t = U_1 \Sigma^2 U_1^T \quad \text{and} \quad F^T H_b = \Sigma^{-1} U_1^T H_b = V_1 \Gamma_{b1} S_1^T. \quad (47)$$

The following Theorem gives the relation between the MSE solution and the matrix G for LDA.

THEOREM 1 *Let G be the dimension reducing transformation matrix from LDA given in (46) and*

$$\{g_i(z) = w_{0i} + w_i^T z\}_{1 \leq i \leq r}$$

be the discriminant functions for the MSE problem (27). Then

$$\begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_1 U_1^T = \begin{bmatrix} n_1 \beta_1 (c_1 - c)^T \\ \vdots \\ n_r \beta_r (c_r - c)^T \end{bmatrix} G G^T. \quad (48)$$

PROOF. *From (45), we have*

$$\begin{aligned} S_t w_i = n_i \beta_i (c_i - c) &\rightarrow U_1 \Sigma^2 U_1^T w_i = n_i \beta_i (c_i - c) \\ &\rightarrow w_i^T U_1 = n_i \beta_i (c_i - c)^T U_1 \Sigma^{-2}. \end{aligned} \quad (49)$$

Then by (49),

$$\begin{aligned} \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_1 U_1^T &= \begin{bmatrix} n_1 \beta_1 (c_1 - c)^T \\ \vdots \\ n_r \beta_r (c_r - c)^T \end{bmatrix} U_1 \Sigma^{-2} U_1^T \\ &= \text{diag}(\sqrt{n_1} \beta_1, \dots, \sqrt{n_r} \beta_r) H_b^T U_1 \Sigma^{-1} (V_1 V_1^T + V_2 V_2^T) \Sigma^{-1} U_1^T \\ &= \text{diag}(\sqrt{n_1} \beta_1, \dots, \sqrt{n_r} \beta_r) H_b^T F V_1 V_1^T F^T \\ &= \begin{bmatrix} n_1 \beta_1 (c_1 - c)^T \\ \vdots \\ n_r \beta_r (c_r - c)^T \end{bmatrix} G G^T. \end{aligned} \quad (50)$$

The third equality in Eq. (50) holds, since

$$\text{span}(V_2) \subset \text{null}(F^T S_b F) = \text{null}(H_b^T F)$$

from Eqs. (14) and (17). \square

Let us denote the reduced dimensional representation obtained by the linear transformation G^T from LDA as

$$\tilde{z} = G^T z \text{ for any data item } z.$$

First we consider the case that S_w is nonsingular and therefore S_t is nonsingular. In this case, $U = U_1$ is orthogonal and U_2 does not appear in the EVD of S_t in (15). Then by Theorem 1 and Eqs. in (43), for any data item z

$$\begin{aligned} \begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} &= \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} z = \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} (z - c) \\ &= \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_1 U_1^T (z - c) \\ &= \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} n_1\beta_1(\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r\beta_r(\tilde{c}_r - \tilde{c})^T \end{bmatrix} (\tilde{z} - \tilde{c}). \end{aligned} \quad (51)$$

Eq. (51) shows that the decision rule in the generalized MSE solution

$$\arg \max_{1 \leq i \leq r} \{g_i(z)\} \quad (52)$$

is equivalent to

$$\arg \max_{1 \leq i \leq r} \{n_i\beta_i/n + n_i\beta_i(\tilde{c}_i - \tilde{c})^T(\tilde{z} - \tilde{c})\} \quad (53)$$

in the reduced dimensional space obtained by LDA. This implies that the MSE procedure is equivalent to applying centroid-based classification with an inner product similarity measure in the reduced dimensional space obtained by LDA. If $\beta_i = 1$ ($1 \leq i \leq r$), (53) becomes

$$\arg \max_{1 \leq i \leq r} \{n_i/n + n_i(\tilde{c}_i - \tilde{c})^T(\tilde{z} - \tilde{c})\}. \quad (54)$$

On the other hand, with $\beta_i = n/n_i$, i.e.,

$$g_i(z) = w_{0i} + w_i^T z = \begin{cases} n/n_i, & \text{if } z \in \text{class } i \\ 0, & \text{otherwise.} \end{cases}, \quad (55)$$

(53) becomes

$$\arg \max_{1 \leq i \leq r} \{(\tilde{c}_i - \tilde{c})^T(\tilde{z} - \tilde{c})\}. \quad (56)$$

The difference between (54) and (56) is whether weighting by the number of elements in each class is considered or not.

The problem formulation (55) also gives a natural generalization of the relationship between the generalized MSE solution for the two-class case and FDA. Let \tilde{z} be the 1-dimensional representation obtained by FDA. Then the equivalence of (52) and (56) gives

$$\begin{aligned} g_1(z) > g_2(z) &\leftrightarrow (\tilde{c}_1 - \tilde{c})(\tilde{z} - \tilde{c}) > (\tilde{c}_2 - \tilde{c})(\tilde{z} - \tilde{c}) \\ &\leftrightarrow (\tilde{c}_1 - \tilde{c}_2)(\tilde{z} - \tilde{c}) > 0 \end{aligned}$$

indicating the decision rule (26) in FDA.

Let us consider undersampled problems where all the scatter matrices are singular and therefore we have the term U_2 in the EVD of S_i . For a given data item $z = a_i$, by Eq. (20)

$$U_2^T z = U_2^T c$$

and Theorem 1, we have

$$\begin{aligned} \begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} &= \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} (z - c) \\ &= \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} (U_1U_1^T + U_2U_2^T)(z - c) \\ &= \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} n_1\beta_1(\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r\beta_r(\tilde{c}_r - \tilde{c})^T \end{bmatrix} (\tilde{z} - \tilde{c}). \end{aligned} \quad (57)$$

Eq. (57) is exactly the same as (51) which was obtained for the case when S_w is nonsingular, implying that the above discussion regarding the nonsingular case still holds for undersampled problems. However, when a new unseen data item z is presented, the third equality in (57) becomes an approximation since Eq. (20) is based on the given training data set. If new data items come from the same distribution as the training data, Eq. (57) should hold almost exactly as the experiments in Section 6 show.

5 Performing LDA through the generalized MSE Procedure

Now we show how to obtain the reduced dimensional space of LDA through the MSE procedure without computing the transformation matrix G of the LDA procedure. From the relation (51) (and also (57)) of LDA and MSE, we have

$$\begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} [c_1 - c, \dots, c_r - c] = \begin{bmatrix} n_1\beta_1(\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r\beta_r(\tilde{c}_r - \tilde{c})^T \end{bmatrix} [\tilde{c}_1 - \tilde{c}, \dots, \tilde{c}_r - \tilde{c}], \quad (58)$$

where $\tilde{c}_i, i = 1, \dots, r$, are the class centroids in the reduced dimensional space obtained by LDA, i.e.,

$$\tilde{c}_i = G^T c_i.$$

Denoting

$$L \equiv \begin{bmatrix} \sqrt{n_1}(\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ \sqrt{n_r}(\tilde{c}_r - \tilde{c})^T \end{bmatrix} = H_b^T G, \quad (59)$$

(58) becomes

$$\begin{bmatrix} \frac{1}{\sqrt{n_1}\beta_1} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{n_r}\beta_r} \end{bmatrix} \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} [c_1 - c, \dots, c_r - c] \begin{bmatrix} \sqrt{n_1} & & \\ & \ddots & \\ & & \sqrt{n_r} \end{bmatrix} = LL^T. \quad (60)$$

Let the EVD of the left side in (60) be $Q\Lambda Q^T$ where Q is orthogonal and Λ has nonincreasing diagonal components. Then

$$LL^T = Q\Lambda Q^T \equiv \underbrace{[Q_1]}_{r-1} \underbrace{[Q_2]}_1 \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} = Q_1\Lambda_1Q_1^T. \quad (61)$$

On the other hand, from Eqs. (59) and (46-47)

$$L^T L = G^T H_b H_b^T G = (V_1^T F^T H_b)(H_b^T F V_1) = (\Gamma_{b1} S_1^T)(S_1 \Gamma_{b1}^T) = \Gamma_{b1} \Gamma_{b1}^T. \quad (62)$$

Hence from Eqs. (61) and (62),

$$\Lambda_1 = \Gamma_{b1} \Gamma_{b1}^T$$

and we can obtain the SVD of L as

$$L = Q_1 \Lambda_1^{1/2}, \quad \text{i.e.,} \quad L^+ = (\Lambda_1^{1/2})^+ Q_1^T.$$

When

$$\text{rank}(H_b) = \text{rank}([c_1 - c, \dots, c_r - c]) = r - 1,$$

L has full column rank and $L^+ L = I$. When S_w is nonsingular, from (51) we have

$$\begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} = \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} n_1 \beta_1 (\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r \beta_r (\tilde{c}_r - \tilde{c})^T \end{bmatrix} \tilde{z} = \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} \sqrt{n_1} \beta_1 & & \\ & \ddots & \\ & & \sqrt{n_r} \beta_r \end{bmatrix} L \tilde{z}, \quad (63)$$

and therefore

$$\begin{aligned} \tilde{z} &= L^+ \begin{bmatrix} \frac{1}{\sqrt{n_1}\beta_1} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{n_r}\beta_r} \end{bmatrix} \left(\begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} - \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} \right) \\ &= (\Lambda_1^{1/2})^{-1} Q_1^T \begin{bmatrix} \frac{1}{\sqrt{n_1}\beta_1} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r}\beta_r} w_r^T \end{bmatrix} z. \end{aligned} \quad (64)$$

Algorithm 1 An efficient algorithm for LDA

Given a data matrix $A \in \mathbb{R}^{m \times n}$ with r classes, this computes a $r - 1$ -dimensional representation of any data point $z \in \mathbb{R}^{m \times 1}$.

1. Compute $\begin{bmatrix} w_{01} & \cdots & w_{0r} \\ w_1 & \cdots & w_r \end{bmatrix} = \mathcal{P}^+ \mathcal{Y}$ where \mathcal{P} and \mathcal{Y} are defined in (23) and (29) respectively.
2. Compute the EVD of the left side of Eq. (60):

$$\begin{bmatrix} \frac{1}{\sqrt{n_1 \beta_1}} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r \beta_r}} w_r^T \end{bmatrix} \begin{bmatrix} \sqrt{n_1}(c_1 - c), & \cdots, & \sqrt{n_r}(c_r - c) \end{bmatrix} = \underbrace{\begin{bmatrix} Q_1 & Q_2 \end{bmatrix}}_{\substack{r-1 \\ 1}} \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}.$$

3. For any data item z , the $r - 1$ -dimensional representation is given by

$$\Lambda_{11}^{-1/2} Q_1^T \begin{bmatrix} \frac{1}{\sqrt{n_1 \beta_1}} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r \beta_r}} w_r^T \end{bmatrix} z.$$

Eq. (64) shows that the reduced dimensional representation by LDA can be obtained from the discriminant functions of the MSE solution

$$\{ g_i(z) = w_{0i} + w_i^T z \}_{(1 \leq i \leq r)}$$

and the EVD of the $r \times r$ matrix, instead of solving the generalized eigenvalue problem for LDA. For undersampled problems, Eq. (63) must change accordingly

$$\begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} = \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_2 U_2^T z + \begin{bmatrix} n_1 \beta_1 (\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r \beta_r (\tilde{c}_r - \tilde{c})^T \end{bmatrix} \tilde{z}. \quad (65)$$

The second term of the right side in (65) is invariant for any given training data item since they are transformed to the constant point by U_2^T . Hence we can obtain the reduced dimensional representation except for a translation factor as

$$\tilde{z} \approx (\Lambda_1^{1/2})^{-1} Q_1^T \begin{bmatrix} \frac{1}{\sqrt{n_1 \beta_1}} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r \beta_r}} w_r^T \end{bmatrix} z. \quad (66)$$

The new algorithm to compute the LDA solution is summarized in Algorithm 1.

This approach to LDA utilizing the relation with the MSE solution has the following properties. First, the scatter matrices S_b and S_w need not be computed explicitly. It reduces computational

Table 1: The comparison of computational complexities. m : data dimension, n : no. of data items, r : no. of classes, s : rank of Z .

	When S_w is nonsingular	Undersampled problems
Classical LDA $S_w^{-1}S_b x = \lambda x$	$S_w^{-1} : O(m^3)$ EVD of $S_w^{-1}S_b : O(m^3)$	Not applicable
LDA/GSVD in [18]	SVD of $Z = \begin{bmatrix} H_b^T \\ H_w^T \end{bmatrix} : O(\min\{m(n+r)^2, m^2(n+r)\})$ SVD of $P(1:r, 1:s) : O(sr^2)$	
Algo. 1 via relation with MSE	Pseudo-inverse of $\mathcal{P} : O(\min\{mn^2, m^2n\})$ Step 2 : $O(r^3)$	

Table 2: The description of datasets

	dataset	no. of classes	dim	no. of data
UCI Machine Learning Repository	Musk	2	166	6598
	Isolet	26	617	7797
	M-feature	10	649	2000
	B-scale	3	4	625
	B-cancer	2	9	699
	Wdbc	2	30	569
	Car	4	6	1728
	Glass	2	9	214
Text Documents	Cacmcisi	2	14409	4663
	Cranmed	2	9038	2431
	Hitech	6	13170	2301
	La1	6	17273	3204
	La2	6	15211	3075
	Tr23	6	5832	204
	Tr41	10	7454	878
	Tr45	10	8261	690

complexities and saves memory requirements. Second, in addition to the SVD of the matrix \mathcal{P} defined in (23), the eigenvalue decomposition (EVD) is only needed for an $r \times r$ matrix where the number of classes r is usually much smaller than the data dimension m or the total number of data n . Table 1 compares computational complexities among the classical LDA, LDA/GSVD [18, 19] and the new proposed algorithm 1. The cost for computing the SVD of an $m \times n$ ($m \geq n$) matrix is estimated as $O(mn^2)$ [21, pp.254]. When \mathcal{P} has full rank, QR decomposition can be used to compute the pseudo-inverse of \mathcal{P} which is cheaper than the SVD [21]. It is due to the fact that the reduced QR decomposition of $B = Q_1 R$ gives the pseudoinverse of B as $R^{-1}Q_1^T$.

6 Experimental Results

In order to verify the theoretical results for the relationship between LDA and the MSE procedure, we conducted extensive experiments. The experiments use two types of datasets: the first has

Table 3: The comparison of classification performances for the verification of the relation (67). The mean prediction accuracies (%) from 10 times random splittings of training and test sets are shown.

	$\beta_i = 1$		$\beta_i = n/n_i$	
	MSE	LDA	MSE	LDA
Musk	93.7	93.7	79.7	79.7
M-feature	98.0	98.0	98.0	98.0
B-scale	87.2	87.2	83.1	83.1
B-cancer	95.8	95.8	96.9	96.9
Wdbc	95.1	95.1	96.1	96.1
Car	76.8	76.8	45.9	45.9
Glass	91.5	91.5	91.4	91.4
Isolet	91.3	91.3	91.3	91.3
Undersampled Problems				
Cacmcisi	95.3	95.3	96.3	96.3
Cranmed	99.8	99.8	99.7	99.7
Hitech	70.5	70.5	62.7	62.7
La1	87.8	87.8	82.2	82.2
La2	89.2	89.2	84.4	84.4
Tr23	89.5	89.5	80.9	80.7
Tr41	95.7	95.7	84.7	84.7
Tr45	92.8	92.8	87.7	87.6

a nonsingular within-class scatter matrix S_w , therefore the classical LDA can be performed for these datasets; the other is from undersampled problems which have singular scatter matrices. Datasets were collected from the UCI Machine Learning Repository ² and text documents ³. A collection of text documents is represented as a term-document matrix where each document is expressed as a column vector. The term-document matrix is obtained after preprocessing with common words and rare term removal, stemming, term frequency and inverse term frequency weighting and normalization [22]. The term-document matrix representation often makes the high dimensionality inevitable. Each dataset is split randomly to training data and test data in equal size and it is repeated 10 times in order to prevent the possible bias from random splitting. The detailed description of the datasets are shown in Table 2.

For all datasets in Table 2, the relationship between the MSE procedure and LDA

$$\begin{aligned} & \arg \max_{1 \leq i \leq r} \{g_i(z) = w_{0i} + w^T z\} \\ & = \arg \max_{1 \leq i \leq r} \left\{ \frac{n_i \beta_i}{n} + n_i \beta_i (G^T c_i - G^T c)^T (G^T z - G^T c) \right\} \end{aligned} \quad (67)$$

was demonstrated by comparing the prediction accuracies. Table 3 reports the mean prediction accuracies (%) from 10 random splittings of training and test sets. The relation (67) was verified for all the datasets, subject to minor differences in Tr23 and Tr45.

²<http://www.ics.uci.edu/~mllearn/MLRepository.html>

³<http://www-users.cs.umn.edu/~karypis/cluto/download.html>

Table 4: Verification of the new efficient algorithm for LDA. The mean prediction accuracies (%) from 10 times random splittings are shown.

	LDA			Algorithm 1		
	1-NN	15-NN	29-NN	1-NN	15-NN	29-NN
Musk	91.4	93.8	93.9	91.4	93.8	93.9
M-feature	98.1	98.1	98.1	98.1	98.1	98.1
B-scale	87.3	88.1	88.5	87.2	88.1	88.5
B-cancer	95.5	96.8	96.4	95.5	96.8	96.4
Wdbc	95.2	96.2	95.9	95.2	96.2	95.9
Car	88.0	87.1	86.6	88.0	87.1	86.6
Glass	90.8	91.3	90.9	90.8	91.3	90.9
Isolet	92.0	92.5	92.2	92.0	92.5	92.2
Undersampled Problems						
Cacmcisi	95.3	95.3	95.3	95.3	95.3	95.3
Cranmed	99.8	99.8	99.8	99.8	99.8	99.8
Hitech	69.9	69.9	69.9	69.9	69.9	69.9
La1	86.4	86.4	86.4	86.4	86.4	86.4
La2	87.7	87.7	87.7	87.7	87.7	87.7
Tr23	84.6	75.9	75.9	84.6	75.9	75.9
Tr41	93.9	93.4	90.0	93.9	93.4	90.0
Tr45	88.6	88.4	86.3	88.6	88.4	86.3

Algorithm 1 was tested for all the datasets in order to verify our derivation by comparing the prediction accuracies by Algorithm 1 with those by LDA using k -NN classifier. Table 4 shows the mean prediction accuracies (%) from 10 runs. Exactly the same results were obtained by both algorithms in all the datasets used except in the B-scale dataset with 1-NN classifier which resulted in a 0.1% difference.

7 Conclusion

In this paper, we have shown a relationship between LDA and the generalized MSE solution for multi-class problems. It generalizes the relation between the MSE solution and FDA to multi-class cases and on undersampled problems. We also proposed an efficient algorithm for LDA which utilizes the relationship with the generalized MSE solution. In Algorithm 1, the generalized eigenvalue problem is solved by the SVDs of the matrix \mathcal{P} and the small $r \times r$ matrix. In addition, the proposed algorithm does not need to explicitly compute the scatter matrices thus saving computational costs as well as memory requirements.

References

- [1] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annual eugenics*, 7, Part II:179–188, 1936.

- [2] F. Rosenblatt. *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Spartan Books, 1962.
- [3] B.E. Boser, I.M. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Fifth annual workshop on computational learning theory, Pittsburgh, ACM*, 1992.
- [4] R.O. Duda and P.E. Hart. *Pattern classification and scene analysis*. Wiley Interscience, 1973.
- [5] Y-C. Ho and R.L. Kashyap. An algorithm for linear inequalities and its applications. *IEEE transactions on Elec. Comp.*, 14:683–688, 1965.
- [6] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley-interscience, New York, 2001.
- [7] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1990.
- [8] J.S. Koford and G.F.Groner. The use of an adaptive threshold element to design a linear optimal pattern classifier. *IEEE transactions on Information Theory*, 12:42–50, 1966.
- [9] T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89:1255–1270, 1994.
- [10] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixture. *J. Royal Statist. Soc. (series B)*, 58:155–176, 1996.
- [11] N. Campbell. Canonical variate analysis - a general formulation. *Australian Journal of Statistics*, 26:86–96, 1984.
- [12] J.H. Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.
- [13] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23(1):73–102, 1995.
- [14] L. Chen, H.M. Liao, M. Ko, J. Lin, and G. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33:1713–1726, 2000.
- [15] H. Yu and J. Yang. A direct lda algorithm for high-dimensional data- with application to face recognition. *Pattern Recognition*, 34:2067–2070, 2001.
- [16] X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.
- [17] J. Yang, J.-Y. Yang, and D. Zhang. What’s wrong with fisher criterion? *Pattern Recognition*, 35(11):2665–2668, 2002.
- [18] P. Howland, M. Jeon, and H. Park. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25(1):165–179, 2003.

- [19] P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE transaction on pattern analysis and machine intelligence*, 26(8):995-1006, 2004.
- [20] C.C. Paige and M.A. Saunders. Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, 18:398–405, 1981.
- [21] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, third edition, 1996.
- [22] T.G. Kolda and D.P. O’Leary. A semidiscrete matrix decomposition for latent semantic indexing in information retrieval. *ACM transactions on Information Systems*, 16(4):322–346, 1998.