Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 04-013

An Efficient Algorithm for LDA Utilizing the Relationship between LDA and the generalized Minimum Squared Error Solution

Cheonghee Park and Haesun Park

March 03, 2004

An Efficient Algorithm for LDA Utilizing the Relationship between LDA and the generalized Minimum Squared Error Solution*

Cheong Hee Park[1] and Haesun Park[2] †

Dept. of Computer Science and Engineering[1,2]
University of Minnesota
Minneapolis, MN 55455, U.S.A.
{chpark,hpark}@cs.umn.edu

The National Science Foundation[2]
4201 Wilson Boulevard, Arlington, Virginia 22230, USA

Abstract

In this paper, we study the relationship between Linear Discriminant Analysis (LDA) and the generalized Minimum Squared Error (MSE) solution. We show that the generalized MSE solution is equivalent to applying a certain classification rule in the space transformed by LDA. The relationship of the MSE solution with Fisher Discriminant Analysis (FDA) is extended to multi-class problems and also undersampled problems where the classical LDA is not applicable due to the singularity of scatter matrices. We propose an efficient algorithm for LDA that can be performed through the relationship with the MSE procedure without solving the eigenvalue problem. Extensive experiments verify the theoretical results and also demonstrate that the classification rule induced by MSE procedure can be effectively applied in the dimension reduced space by LDA.

Keywords. Dimension reduction method, Linear Discriminant Analysis, Minimum Squared Error Solution, Undersampled problems.

1 Introduction

Using linear discriminant functions for pattern classification is based on the existence of a hyperplane which can separate two classes optimally. This presumption may be naive, however the

*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 and was partly supported by IR/D from the National Science Foundation (NSF).

continuous utilization of a linear discriminant function in practice is attributed to its simple concept and easiness of computation. The optimality of the separating hyperplane can be measured by various criteria and numerous methods and algorithms have been proposed [1], originating from the paper [2] by R.A. Fisher. In Perceptron method, a linear discriminant function is obtained by iterative procedures to reduce the number of misclassified training data [3]. Support Vector Machines (SVM) searches a linear function which maximizes the margin between classes either in the original data space or the nonlinearly transformed feature space [4]. Minimum Squared-Error (MSE) solution seeks a linear discriminant function that minimizes the squared error [5] and the closed form of its solution is known [6].

While being similar in the sense of using linear functions, Linear Discriminant Analysis (LDA) has different concept from the above mentioned methods in that it 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 maximize the between-class scatter and minimize the within-class scatter. It is known that the original data space can be reduced to the $r-1$ dimensional space which satisfies the criterion in LDA, where r is the number of classes [7]. In two-class case, the relationship of the MSE solution and Fisher Discriminant Analysis (FDA) has been known [8, 5], where FDA is a special case of LDA for two-class problem. Since both FDA and the MSE solution in two-class problems deal with one linear function, the relationship between them can come naturally. On the other hand, the generalized MSE procedure for multiple-class problems set r two-class problems each of which produces a hyperplane separating each class from the rest of data points.

In this paper, we develop the relationship between LDA and the generalized MSE procedure in a multi-class problem. We show that the MSE solution can be obtained by applying a certain classification rule in the reduced dimensional space by LDA, and LDA can be performed through the MSE procedure without solving the eigenvalue problem explicitly. The classification rule induced by the MSE procedure can be effectively applied in LDA, resulting in LDA as a classifier.

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

2 Linear Discriminant Analysis

Linear Discriminant Analysis is a linear dimension reduction method which can be used as a preprocessing step for data analysis. The goal of LDA is to find a linear transformation that maximizes the between-class scatter and minimizes the within-class scatter so that the class separability can be optimized in the transformed space. We begin with the brief review of LDA and an algorithm which gives a solution for LDA.

Throughout the paper, we assume the vector space representation of a data set A

$$A = [a_1, \cdots, a_n] = [A_1, A_2, \cdots, A_r] \in \mathbb{R}^{m \times n} \tag{1}$$

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. Each class A_i $(1 \leq i \leq r)$ has n_i elements and the total number of data is $n = \sum_{i=1}^{r} n_i$.

The between-class scatter matrix S_B, within-class scatter matrix S_W, and mixture scatter matrix S_M are defined as

$$S_B = \sum_{i=1}^{r} n_i (c_i - c)(c_i - c)^T, \tag{2}$$

$$S_W = \sum_{i=1}^{r} \sum_{j \in N_i} (a_j - c_i)(a_j - c_i)^T, \tag{3}$$

$$S_M = \sum_{j=1}^{n} (a_j - c)(a_j - c)^T, \tag{4}$$

where N_i $(1 \leq i \leq r)$ denotes the index set of data items in the i-th class and

$$c_i = \frac{1}{n_i} \sum_{j \in N_i} a_j \quad \text{and} \quad c = \frac{1}{n} \sum_{j=1}^{n} a_j \tag{5}$$

are class centroids and the global centroid, respectively.

The *traces* of scatter matrices can be used to measure the quality of clustered 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 trace(S_B) quantifies the distance between classes and trace(S_W) measures the scatter within classes. The optimal dimension reducing transformation G^T for LDA is the one that maximizes

$$J(G) = \text{trace}((G^T S_W G)^{-1} (G^T S_B G)) \tag{6}$$

where $G^T S_B G$ and $G^T S_W G$ are scatter matrices in the transformed space. It is well known [1] that the criterion in (6) 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

$$S_B x = \lambda S_W x. \tag{7}$$

When S_W is nonsingular, one can solve the eigenvalue problem

$$S_W^{-1} S_B x = \lambda x \tag{8}$$

instead of the generalized eigenvalue problem (7). However, the computation of the inverse matrix can be demanding and is not numerically stable for the ill-conditioned matrix. Next we give our approach for the generalized eigenvalue problem (7), which will be used in deriving the relationship between LDA and the generalized MSE solution.

2.1 Generalized Eigenvalue Problem

Let the symmetric eigenvalue decomposition [1] (EVD) of S_M be

$$S_M = U\Sigma U^T = [\underbrace{U_1}_{s} \ \underbrace{U_2}_{m-s}] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \tag{9}$$

where $s = rank(S_M)$, U is orthogonal and Σ is a diagonal matrix with nonincreasing diagonal elements. Denoting the eigenvectors scaled by nonzero eigenvalues as

$$F = U_1\Sigma_1^{-1/2} \in \mathbb{R}^{m\times s},$$

from (9) we obtain

$$F^T S_M F = I_s. \tag{10}$$

Note that the between-class scatter matrix S_B can be expressed as a product of the smaller matrices

$$S_B = H_B H_B^T \quad \text{where} \quad H_B = [\sqrt{n_1}(c_1 - c), \cdots, \sqrt{n_r}(c_r - c)] \in \mathbb{R}^{m\times r}.$$

Let the SVD of $F^T H_B$ be

$$F^T H_B = V\Gamma Z^T = [\underbrace{V_1}_{r-1} \ \underbrace{V_2}_{s-r+1}] \begin{bmatrix} \Gamma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Z_1^T \\ Z_2^T \end{bmatrix}, \tag{11}$$

where $V \in \mathbb{R}^{s\times s}$, $Z \in \mathbb{R}^{r\times r}$ are orthogonal and $\Gamma_1 \in \mathbb{R}^{(r-1)\times(r-1)}$ is a diagonal matrix with nonincreasing diagonal elements. Then the EVD of $F^T S_B F$ is obtained by

$$F^T S_B F = (F^T H_B)(H_B^T F) = V\Sigma_B V^T \tag{12}$$

where $\Sigma_B = \Gamma\Gamma^T$. Eqs. (10) and (12) and the relation

$$S_M = S_B + S_W$$

give the EVD of $F^T S_W F$ as

$$F^T S_W F = V(I_s - \Sigma_B)V^T = V\Sigma_W V^T \tag{13}$$

where $\Sigma_W = I_s - \Sigma_B$. Now from Eqs. (12) and (13), we obtain the simultaneous diagonalizations of S_B and S_W as

$$\begin{aligned} \begin{bmatrix} V^T F^T \\ U_2^T \end{bmatrix} S_B \begin{bmatrix} FV & U_2 \end{bmatrix} &= \begin{bmatrix} \Sigma_B & 0 \\ 0 & 0 \end{bmatrix} \\ \begin{bmatrix} V^T F^T \\ U_2^T \end{bmatrix} S_W \begin{bmatrix} FV & U_2 \end{bmatrix} &= \begin{bmatrix} \Sigma_W & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned} \tag{14}$$

[1] The symmetric eigenvalue decomposition requires the equality of the left and right eigenvectors, i.e., $U = V$ in the singular value decomposition (SVD) of $X = U\Sigma V^T$.

Eqs. in (14) hold, since

$$\text{span}\{U_2\} = \text{null}(S_M) \text{ and } \quad \text{null}(S_M) \subset \text{null}(S_B) \cap \text{null}(S_W). \tag{15}$$

Let us denote the diagonal elements of Σ_B and Σ_W as η_i and ζ_i,

$$\Sigma_B = \text{diag}(\eta_1, \cdots, \eta_s), \quad \Sigma_W = \text{diag}(\zeta_1, \cdots, \zeta_s).$$

Then the diagonalizations in (14) imply that the column vectors x_i of $[FV \ U_2]$ satisfy the generalized eigenvalue problem

$$\zeta_i S_B x_i = \eta_i S_W x_i, \quad i = 1, \cdots, m. \tag{16}$$

Note that x_i, $i = s+1, \cdots, m$, belong to $\text{null}(S_B) \cap \text{null}(S_W)$. Hence η_i and ζ_i for $i = s+1, \cdots, m$ can be any arbitrary numbers in (16).

2.2 When S_W is nonsingular

Now suppose that the within-class scatter matrix S_W is nonsingular, therefore the mixture scatter matrix S_M is also nonsingular. Then

$$s = \text{rank}(S_M) = m,$$

and $U = U_1$ and U_2 does not appear in (9). Hence

$$FV = U_1 \Sigma_1^{-1/2} V = [x_1, \cdots, x_m]$$

and

$$S_B x_i = \frac{\eta_i}{\zeta_i} S_W x_i.$$

Since the diagonal elements of Σ_B are nonincreasing and those of Σ_W are nondecreasing,

$$\frac{\eta_1}{\zeta_1} \geq \cdots \geq \frac{\eta_m}{\zeta_m} \geq 0.$$

Hence the linear transformation for LDA is obtained by taking the leftmost $r-1$ columns of FV as

$$G = [x_1, \cdots, x_{r-1}] = FV_1 = U_1 \Sigma_1^{-1/2} V_1. \tag{17}$$

One can normalize x_i's in order to use the unit eigenvectors, although this scaling does not make any change in the optimization criterion in (6).

2.3 LDA on Undersampled Problems

When the dimension of data is greater than the number of data, referred to as undersampled problems, all the scatter matrices become singular and the classical LDA is difficult to apply. This situation is common in the areas of text classification or face recognition. Since the scatter matrices are all singular

$$s = \text{rank}(S_M) < m,$$

and therefore we have the term U_2 in the EVD of the mixture scatter matrix S_M in (9). Now we have the generalized eigenvectors $\{x_1, \cdots, x_s\}$ corresponding to the eigenvalues

$$\eta_1/\zeta_1 \geq \cdots \geq \eta_s/\zeta_s \geq 0.$$

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

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

and these imply that all data items are transformed to one constant point by x^T. Hence the vectors x_i, $i = s+1, \cdots, m$, 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 zeros. It can justify that we can form the linear transformation G^T for LDA by taking the first $r-1$ columns from $[FV \;\; U_2] = [FV_1 \;\; FV_2 \;\; U_2]$ even when S_M is singular. Hence on the undersampled problems, the transformation matrix G is also given as

$$G = FV_1 = U_1 \Sigma_1^{-1/2} V_1.$$

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

$$g(z) = w_0 + w^T z = \begin{cases} b_1, & \text{if } z \in A_1 \\ b_2, & \text{if } z \in A_2 \end{cases}, \tag{18}$$

where b_i is the prespecified number for each class. For the data set A given in (1), the problem (18) can be reformulated to minimize the squared error

$$\left\| \begin{bmatrix} 1 & a_1^T \\ \vdots & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \right\|_2^2 \tag{19}$$

where $y_i = b_1$ if a_i belongs to the class A_1, otherwise $y_i = b_2$. Denoting

$$P = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{bmatrix}^T, \tag{20}$$

the solution which minimizes the squared error (19) can be computed using the pseudoinverse as

$$\begin{bmatrix} w_0 \\ w \end{bmatrix} = P^+ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{21}$$

When the least squares problem in (19) is overdetermined (i.e., the number of rows in P is greater then the number of columns) and P has the full column rank, (21) gives a unique solution for the least squares problem (19) and it is also the minimal norm solution [9].

Different choices of $b = [b_1, b_2]^T$ would give different discriminant function. In particular, when $b_1 = n/n_1$ and $b_2 = -n/n_2$, the MSE solution is related with Fisher Discriminant Analysis (FDA) [5]. The vector w in (21) is same as the solution of FDA except for a scaling factor as

$$w = \alpha S_W^{-1}(c_1 - c_2) \equiv \alpha x, \tag{22}$$

and w_0 is equal to $-w^T c$ where c and c_i are the centroids defined in (5). For a new data item, it is assigned to the class A_1 if

$$w^T z + w_0 = w^T(z - c) = \alpha(x^T z - x^T c) > 0, \tag{23}$$

otherwise it is assigned to the class A_2.

The MSE procedure is generalized to multi-class cases as a set of r two-class problems [5]. For each class A_i $(1 \le i \le r)$, the MSE solution to the problem

$$g_i(z) = w_{0i} + w_i^T z = \begin{cases} b_i, \text{ if } z \in A_i \\ 0, \text{ otherwise.} \end{cases} \tag{24}$$

is sought. Compared with the problem (18), the solution of the multi-class problem (24) will be referred to as the generalized MSE solution whenever the distinction is needed. As in [5], one choice for b_i would be assigning $b_i = 1$ for $i = 1, \cdots, r$. The squared error function in the multi-class problem is expressed as

$$\left\| \begin{bmatrix} 1 & a_1^T \\ & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_{01} & \cdots & w_{0r} \\ w_1 & \cdots & w_r \end{bmatrix} - \begin{bmatrix} y_{11} & \cdots & y_{1r} \\ \vdots & & \\ y_{n1} & \cdots & y_{nr} \end{bmatrix} \right\|_2^2, \tag{25}$$

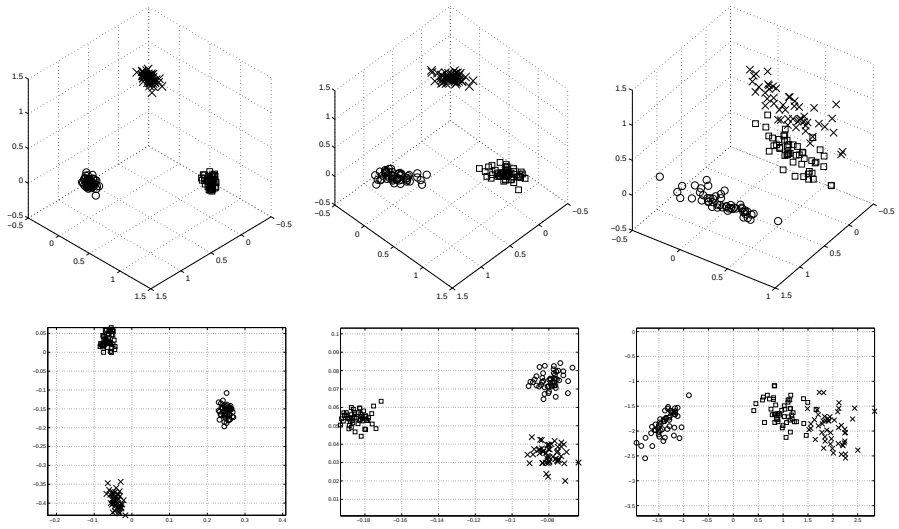

Figure 1: The visualization of the space mapped by discriminant functions of the MSE solution and LDA.

where $y_{ji} = b_i$ if a_j belongs to the class A_i, otherwise 0. Denoting

$$\mathcal{W} = \begin{bmatrix} w_{01} & \cdots & w_{0r} \\ w_1 & \cdots & w_r \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} y_{11} & \cdots & y_{1r} \\ \vdots & & \\ y_{n1} & \cdots & y_{nr} \end{bmatrix}, \tag{26}$$

the MSE solution of the problem (25) is obtained by

$$\mathcal{W} = P^{+}Y, \tag{27}$$

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

$$g_i(z) > g_j(z). \tag{28}$$

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

$$x \longrightarrow \begin{bmatrix} g_1(x), & \cdots, & g_r(x) \end{bmatrix}^T \in \mathbb{R}^{r \times 1}. \tag{29}$$

The squared error in (25) can be represented as

$$\left\| \begin{bmatrix} 1 & a_1^T \\ & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_{01} & \cdots & w_{0r} \\ w_1 & \cdots & w_r \end{bmatrix} - Y \right\|_2^2 = \sum_{1 \leq i \leq r} \sum_{j \in N_i} \left\| \begin{bmatrix} g_1(a_j) \\ \vdots \\ g_r(a_j) \end{bmatrix} - b_i e_i \right\|_2$$

where e_i ($1 \leq i \leq r$) is the column vector with 1 in the i-th position and 0 elsewhere. Hence in the space transformed by the mapping (29), the i-th class centroid will be mapped near to the point $b_i e_i$. Figure 1 visualizes the transformed spaces by LDA and the mapping (29) where b_i was set as 1 for all classes. The first two figures on the top were obtained by taking three subclasses randomly

in the Isolet dataset from UCI Machine Learning Repository and they show well-separated three classes. 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 second row figures show the transformed space by LDA which corresponds to the figures on the top. The corresponding figures look quite similar. Then what is the mathematical relationship between two methods? If we know the relationship, is it possible to take advantage of the merits from each method and combine them? In the next section, we study the relationship between FDA and the MSE solution on the undersampled problems and also between LDA and the generalized MSE solution in multi-class problems.

4 Relationships between LDA and the MSE solution

The relationship of the MSE solution and FDA given in (22) holds when the within-scatter matrix S_W is nonsingular. We generalize (22) on the undersampled problem by using the algorithm discussed in Section 2.

4.1 FDA and the MSE Solution on Undersampled Problems

Let $g(z) = w_0 + w^T z$ be the MSE solution to the problem

$$g(z) = w_0 + w^T z = \begin{cases} n/n_1, & \text{if } z \in A_1 \\ -n/n_2, & \text{if } z \in A_2 \end{cases} . \tag{30}$$

The least squares problem (19) can be solved by the normal equation

$$\begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} 1 & a_1^T \\ & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} \frac{n}{n_1} I_{n_1} \\ -\frac{n}{n_2} I_{n_2} \end{bmatrix} \tag{31}$$

where I_{n_i} is the $n_i \times 1$ column vector with elements 1. From (31), we obtain

$$\begin{cases} nw_0 + nc^T w = 0 \\ ncw_0 + (\sum_{1 \le j \le n} a_j a_j^T) w = \frac{n}{n_1} \sum_{j \in N_1} a_j - \frac{n}{n_2} \sum_{j \in N_2} a_j. \end{cases} \tag{32}$$

From the first equation in (32) we have

$$w_0 = -c^T w.$$

By substituting it in the second equation and using the alternative expressions of S_B and S_W in two-class problems

$$S_B = \sum_{1 \le i \le 2} n_i c_i c_i^T - ncc^T \quad \text{and} \quad S_W = \sum_{1 \le j \le n} a_j a_j^T - \sum_{1 \le i \le 2} n_i c_i c_i^T$$

we obtain

$$(S_B + S_W) w = n(c_1 - c_2). \tag{33}$$

Let x_1 be the first column vector of $[FV \;\; U_2]$ in (14). From the discussions given in Section 2 and the fact that 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 $\Sigma_B + \Sigma_W = I$,

$$\eta_1(S_B + S_W)x_1 = (\eta_1 + \zeta_1)S_B x_1 = S_B x_1 = \frac{n_1 n_2}{n}(c_1 - c_2)(c_1 - c_2)^T x_1. \tag{34}$$

Denoting

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

Eq. (34) becomes

$$(S_B + S_W)\beta x_1 = n(c_1 - c_2). \tag{35}$$

Then by (33) and (35), we have

$$S_M w = (S_B + S_W)w = (S_B + S_W)\beta x_1 = S_M \beta x_1. \tag{36}$$

From (36) and the EVD of S_M in (9),

$$U_1 \Sigma_1 U_1^T w = U_1 \Sigma_1 U_1^T \beta x_1 \quad \text{and} \quad U_1^T w = U_1^T \beta x_1.$$

Since any data items are transformed to the constant point by $x \in U_2$

$$U_2^T (z - c) = 0$$

and we obtain

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

It gives the generalized relation between the MSE solution and the solution of FDA, which holds regardless of the singularity of the scatter matrices.

While FDA gives 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 $r - 1$ 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

As in 4.1, the generalized MSE solution to the problem

$$g_i(z) = w_{0i} + w_i^T z = \begin{cases} b_i, \text{ if } z \in A_i \\ 0, \text{ otherwise.} \end{cases} \qquad \text{for } i = 1, \cdots, r$$

can be solved by the normal equation

$$P^T P\mathcal{W} = P^T Y. \tag{37}$$

Now from Eq. (37), we obtain

$$\begin{bmatrix} n & \sum_{j=1}^n a_j^T \\ \sum_{j=1}^n a_j & \sum_{j=1}^n a_j a_j^T \end{bmatrix} \begin{bmatrix} w_{01} & \cdots & w_{0r} \\ w_1 & \cdots & w_r \end{bmatrix} = \begin{bmatrix} n_1 b_1 & \cdots & n_r b_r \\ (\sum_{j\in N_1} a_j) b_1 & \cdots & (\sum_{j\in N_r} a_j) b_r \end{bmatrix},$$

resulting in a linear system

$$\begin{cases} n\, w_{0i} + n\, c^T w_i = n_i b_i \\ n\, c\, w_{0i} + (\sum_{j=1}^n a_j a_j^T) w_i = n_i\, b_i c_i \end{cases} \text{ for } i = 1, \cdots, r. \tag{38}$$

By substituting w_{0i} of the second equation with w_{0i} of the first equation in (38), (38) becomes

$$(n_i b_i - n c^T w_i)\, c + \left(\sum_{j=1}^n a_j a_j^T \right) w_i = n_i b_i c_i,\ \ i = 1, \cdots, r. \tag{39}$$

Now from (39) and

$$S_B = \sum_{1\le i\le r} n_i c_i c_i^T - n\, c\, c^T \quad \text{and} \quad S_W = \sum_{1\le j\le n} a_j a_j^T - \sum_{1\le i\le r} n_i c_i c_i^T,$$

we have

$$S_M w_i = (S_B + S_W) w_i = n_i b_i (c_i - c), \quad i = 1, \cdots, r. \tag{40}$$

Recall that the linear transformation matrix G for LDA was obtained by

$$G = F V_1 = U_1 \Sigma_1^{-1/2} V_1 \tag{41}$$

where

$$S_M = U_1 \Sigma_1 U_1^T \quad \text{and} \quad F^T H_B = \Sigma_1^{-1/2} U_1^T H_B = V_1 \Gamma_1 Z_1^T. \tag{42}$$

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

THEOREM 1 *Let G be the transformation matrix for LDA given in (41) and*

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

be the discriminant functions for the MSE problem (24). Then

$$\begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_1 U_1^T = \begin{bmatrix} n_1 b_1 (c_1 - c)^T \\ \vdots \\ n_r b_r (c_r - c)^T \end{bmatrix} G G^T. \tag{43}$$

PROOF. *From (40), we have*

$$\begin{aligned} S_M w_i = n_i b_i (c_i - c) \quad &\rightarrow \quad U_1 \Sigma_1 U_1^T w_i = n_i b_i (c_i - c) \\ &\rightarrow \quad w_i^T U_1 = n_i b_i (c_i - c)^T U_1 \Sigma_1^{-1}. \end{aligned} \tag{44}$$

Then by (44),

$$\begin{aligned} \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_1 U_1^T &= \begin{bmatrix} n_1 b_1 (c_1 - c)^T \\ \vdots \\ n_r b_r (c_r - c)^T \end{bmatrix} U_1 \Sigma_1^{-1} U_1^T \\ &= \mathit{diag}(\sqrt{n_1} b_1, \cdots, \sqrt{n_r} b_r)\, H_B^T U_1 \Sigma_1^{-1/2} (V_1 V_1^T + V_2 V_2^T) \Sigma_1^{-1/2} U_1^T \\ &= \mathit{diag}(\sqrt{n_1} b_1, \cdots, \sqrt{n_r} b_r)\, H_B^T F V_1 V_1^T F^T \\ &= \begin{bmatrix} n_1 b_1 (c_1 - c)^T \\ \vdots \\ n_r b_r (c_r - c)^T \end{bmatrix} G G^T. \end{aligned} \tag{45}$$

The third equality holds, since

$$V_2 = \mathit{null}(F^T S_B F) = \mathit{null}(H_B^T F) \quad \textit{and} \quad H_B^T F V_2 = 0. \quad \square$$

Let us denote the reduced dimensional representation by the linear transformation G^T for 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_M is nonsingular. It means that $U = U_1$ is orthogonal and U_2 does not appear in the EVD of S_M in (9). Then by Theorem 1 and Eqs. in (38), for any data item z

$$\begin{aligned} \begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} &= \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} z = \begin{bmatrix} n_1 b_1 / n \\ \vdots \\ n_r b_r / n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} (z - c) \\ &= \begin{bmatrix} n_1 b_1 / n \\ \vdots \\ n_r b_r / n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_1 U_1^T (z - c) \\ &= \begin{bmatrix} n_1 b_1 / n \\ \vdots \\ n_r b_r / n \end{bmatrix} + \begin{bmatrix} n_1 b_1 (\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r b_r (\tilde{c}_r - \tilde{c})^T \end{bmatrix} (\tilde{z} - \tilde{c}). \end{aligned} \tag{46}$$

On the other hand, on the undersampled problems all the scatter matrix are singular and we have the term U_2 in the EVD of S_M. Since all data items are transformed to the same point by x^T where $x \in U_2$,

$$\begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_2 U_2^T (z - c) = 0. \tag{47}$$

By Theorem 1, for any data item z

$$\begin{aligned}
\begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} &= \begin{bmatrix} n_1 b_1/n \\ \vdots \\ n_r b_r/n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} (z - c) \\
&= \begin{bmatrix} n_1 b_1/n \\ \vdots \\ n_r b_r/n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} (U_1 U_1^T + U_2 U_2^T)(z - c) \\
&= \begin{bmatrix} n_1 b_1/n \\ \vdots \\ n_r b_r/n \end{bmatrix} + \begin{bmatrix} n_1 b_1 (\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r b_r (\tilde{c}_r - \tilde{c})^T \end{bmatrix} (\tilde{z} - \tilde{c}). \qquad (48)
\end{aligned}$$

The relations in (46) and (48) show that the decision rule by the generalized MSE solution

$$\arg \max_{1 \le i \le r} \{g_i(z)\} \qquad (49)$$

is equivalent to

$$\arg \max_{1 \le i \le r} \left\{ n_i b_i/n + n_i b_i (\tilde{c}_i - \tilde{c})^T (\tilde{z} - \tilde{c}) \right\} \qquad (50)$$

in the reduced dimensional space by LDA. If $b_i = 1$ $(1 \le i \le r)$ are used, (50) becomes

$$\arg \max_{1 \le i \le r} \{ n_i/n + n_i (\tilde{c}_i - \tilde{c})^T (\tilde{z} - \tilde{c}) \}. \qquad (51)$$

On the other hand, we can set $b_i = n/n_i$, i.e.,

$$g_i(z) = w_{0i} + w_i^T z = \begin{cases} n/n_i, & \text{if } z \in A_i \\ 0, & \text{otherwise.} \end{cases} \qquad (52)$$

Then (50) becomes

$$\arg \max_{1 \le i \le r} \{ (\tilde{c}_i - \tilde{c})^T (\tilde{z} - \tilde{c}) \}. \qquad (53)$$

It implies that the MSE procedure is equivalent to applying centroid-based classification with inner product similarity measure in the reduced dimensional space by LDA. The difference between (51) and (53) is whether weighting by the number of elements in each class is considered or not. The problem formulation (52) also gives a natural generalization of the relationship between the generalized MSE solution for two-class case and FDA. Let the transformation matrix by FDA be

$$G = x \in \mathbb{R}^{m \times 1}.$$

Then the equivalence of (49) and (53) gives

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

indicating the decision rule (23) in FDA.

Eqs. (46) and (48) also give the relationship between the spaces transformed by the mapping (29) and LDA. For any z_1 and z_2,

$$\begin{bmatrix} g_1(z_1) \\ \vdots \\ g_r(z_1) \end{bmatrix} - \begin{bmatrix} g_1(z_2) \\ \vdots \\ g_r(z_2) \end{bmatrix} = \begin{bmatrix} n_1 b_1 (\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r b_r (\tilde{c}_r - \tilde{c})^T \end{bmatrix} (\tilde{z}_1 - \tilde{z}_2). \tag{54}$$

In two-class case the L_2-norm distance between data points is invariant expect for a scaling factor, since

$$\begin{aligned} \left\| \begin{bmatrix} g_1(z_1) \\ g_2(z_1) \end{bmatrix} - \begin{bmatrix} g_1(z_2) \\ g_2(z_2) \end{bmatrix} \right\| &= \left\| \begin{bmatrix} n_1 b_1 (\tilde{c}_1 - \tilde{c})^T \\ n_2 b_2 (\tilde{c}_2 - \tilde{c})^T \end{bmatrix} (\tilde{z_1} - \tilde{z_2}) \right\| \\ &= \left\| \begin{bmatrix} n_1 b_1 (\tilde{c}_1 - \tilde{c})^T \\ n_2 b_2 (\tilde{c}_2 - \tilde{c})^T \end{bmatrix} \right\| |\tilde{z_1} - \tilde{z_2}|. \end{aligned}$$

In the next section we show that

$$\begin{bmatrix} n_1 b_1 (\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r b_r (\tilde{c}_r - \tilde{c})^T \end{bmatrix}$$

can be computed through the MSE solution. It means that once we have the MSE solution any classifiers utilizing distances between data points in the space transformed by LDA can be built without performing LDA explicitly. Since the MSE solution requires only the pseudoinverse of the matrix P, it can save computational complexities compared with solving the eigenvalue problem.

5 Performing LDA Through the MSE Procedure

Now we show how to obtain the reduced dimensional space by LDA through MSE procedure without computing the transformation matrix G for LDA. From the relation (46) (and also (48)) of LDA and MSE, we have

$$\begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} \begin{bmatrix} c_1 - c, & \cdots, & c_r - c \end{bmatrix} = \begin{bmatrix} n_1 b_1 (\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r b_r (\tilde{c}_r - \tilde{c})^T \end{bmatrix} \begin{bmatrix} \tilde{c}_1 - \tilde{c}, & \cdots, & \tilde{c}_r - \tilde{c} \end{bmatrix}. \tag{55}$$

Denoting

$$L \equiv \begin{bmatrix} \sqrt{n_1} (\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ \sqrt{n_r} (\tilde{c}_r - \tilde{c})^T \end{bmatrix} = H_B^T G, \tag{56}$$

(55) becomes

$$\begin{bmatrix} \frac{1}{\sqrt{n_1} b_1} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{n_r} b_r} \end{bmatrix} \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} \begin{bmatrix} c_1 - c, & \cdots, & c_r - c \end{bmatrix} \begin{bmatrix} \sqrt{n_1} & & \\ & \ddots & \\ & & \sqrt{n_r} \end{bmatrix} = L L^T. \tag{57}$$

Let the EVD of the left side in (57) be QRQ^T where Q is orthogonal and R has nonincreasing diagonal components. Then we have the EVD of LL^T

$$LL^T = QRQ^T \equiv [\underbrace{Q_1}_{r-1} \underbrace{Q_2}_{1}] \begin{bmatrix} R_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} = Q_1 R_{11} Q_1^T. \tag{58}$$

On the other hand, (56) and (41-42) gives the EVD of $L^T L$ as

$$L^T L = G^T H_B H_B^T G = (V_1^T F^T H_B)(H_B^T F V_1) = (\Gamma_1 Z_1^T)(Z_1 \Gamma_1^T) = \Gamma_1 \Gamma_1^T. \tag{59}$$

Hence from (58) and (59),

$$R_{11} = \Gamma_1 \Gamma_1^T$$

and we can obtain the SVD of L as

$$L = Q_1 R_{11}^{1/2}, \quad \text{i.e.,} \quad L^+ = (R_{11}^{1/2})^+ Q_1^T.$$

When

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

L has full column rank and $L^+ L = I$. Hence from (46),

$$\begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} = \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} n_1 b_1 (\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r b_r (\tilde{c}_r - \tilde{c})^T \end{bmatrix} \tilde{z} = \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} \sqrt{n_1} b_1 & & \\ & \ddots & \\ & & \sqrt{n_r} b_r \end{bmatrix} L\tilde{z},$$

and we have

$$\begin{aligned} \tilde{z} &= L^+ \begin{bmatrix} \frac{1}{\sqrt{n_1} b_1} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{n_r} b_r} \end{bmatrix} \left(\begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} - \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} \right) \\ &= (R_{11}^{1/2})^{-1} Q_1^T \begin{bmatrix} \frac{1}{\sqrt{n_1} b_1} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r} b_r} w_r^T \end{bmatrix} z. \end{aligned} \tag{60}$$

Eq. (60) shows that the reduced dimensional representation by LDA can be obtained from the discriminant functions of the MSE solution

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

and the EVD of the $r \times r$ matrix, instead of solving the eigenvalue problem for LDA.

On undersampled problems, a difference takes place in the equation

$$\begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} = \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_2 U_2^T z + \begin{bmatrix} n_1 b_1 (\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r b_r (\tilde{c}_r - \tilde{c})^T \end{bmatrix} \tilde{z}. \tag{61}$$

Algorithm 1 An efficient algorithm for LDA

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

1. Compute $\begin{bmatrix} w_{01} & \cdots & w_{0r} \\ w_1 & \cdots & w_r \end{bmatrix} = P^+ Y$ where P and Y are defined in (20) and (26) respectively.

2. Compute the EVD of the left side of Eq. (57):

$$\begin{bmatrix} \frac{1}{\sqrt{n_1}b_1} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r}b_r} w_r^T \end{bmatrix} \begin{bmatrix} \sqrt{n_1}(c_1 - c), & \cdots, & \sqrt{n_r}(c_r - c) \end{bmatrix} = \begin{bmatrix} \underbrace{Q_1}_{r-1} & \underbrace{Q_2}_{1} \end{bmatrix} \begin{bmatrix} R_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}.$$

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

$$R_{11}^{-1/2} Q_1^T \begin{bmatrix} \frac{1}{\sqrt{n_1}b_1} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r}b_r} w_r^T \end{bmatrix} z.$$

Since the second term of the right side in (61) is inveriant for any data item, we can obtain the reduced dimensional representation except for a translation factor as

$$\tilde{z} \approx (R_{11}^{1/2})^{-1} Q_1^T \begin{bmatrix} \frac{1}{\sqrt{n_1}b_1} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r}b_r} w_r^T \end{bmatrix} z. \tag{62}$$

The new algorithm to compute LDA 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. Secondly, in addition to the SVD of the matrix P, the eigenvalue decomposition (EVD) is only needed for the $r \times r$ matrix where the number of classes r is usually much smaller than the data dimension m or the total number of data n. Especially, on undersampled problems where the data dimension m is higher than the total number of data n as shown in Table 1, the pseudoinverse of the matrix P in step 1 of Algorithm 1 can be computed efficiently. Instead of the SVD of $P \in \mathbb{R}^{n\times(m+1)}$, let us compute the EVD of $PP^T \in \mathbb{R}^{n\times n}$ as

$$PP^T = \begin{bmatrix} \underbrace{Z_1}_{\text{rank}(PP^T)} & Z_2 \end{bmatrix} \begin{bmatrix} \Lambda^2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Z_1^T \\ Z_2^T \end{bmatrix}. \tag{63}$$

Then we can obtain the SVD of P

$$P = Z_1 \Lambda (\Lambda^{-1} Z_1^T P),$$

Table 1: 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	7	5832	204
	Tr41	10	7454	878
	Tr45	10	8261	690

that is, the pseudoinverse of P as

$$P^+ = (P^T Z_1 \Lambda^{-1})\Lambda^{-1} Z_1^T = P^T Z_1 \Lambda^{-2} Z_1^T.$$

6 Experimental Results

In order to verify the theoretical results for the relationship between LDA and the MSE procedure, we conducted extensive experiments using real datasets. The experiments use two types of datasets: The first one has the nonsingular mixture scatter matrix S_M, 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 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 [10]. The term-document matrix representation often makes the high dimensionality inevitable. For each dataset, we split it 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 1.

[2]http://www.ics.uci.edu/∼mlearn/MLRepository.html

[3]http://www-users.cs.umn.edu/∼karypis/cluto/ download.html

Table 2: The comparison of classification performances.

	$b_i = 1$		$b_i = n/n_i$	
	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**

For all datasets in Table 1, the relationship between the MSE procedure and LDA

$$\begin{aligned} &\arg \max_{1\le i\le r} \left\{ g_i(z) = w_{0i} + w^T z \right\} \\ &= \arg \max_{1\le i\le r} \left\{ \frac{n_i b_i}{n} + n_i b_i (G^T c_i - G^T c)^T (G^T z - G^T c) \right\} \end{aligned} \tag{64}$$

was demonstrated by comparing the prediction accuracies. Table 2 reports the mean prediction accuracies (%) from 10 times random splitting to training and test sets. The relation (64) was verified for all the datasets except Tr23 and Tr45. (Compare the columns 1 and 2, 3 and 4 in Table 2.) Since the differences in those two incidents are so small, it might be caused by rounding errors in the numerical computation.

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 3 shows the mean prediction accuracies from 10 times runnings. In the B-scale dataset, 1-NN classifier produced 0.1% difference. Except that, the same results were obtained by both algorithms in all the datasets used. One interesting comment can come from the comparison of Table 2 and Table 3. They show that the classification rule induced by the MSE solution for $b_i = 1$ obtained high prediction accuracies overall. (See the 1-th column in Table 2.) It indicates that since LDA utilizes the class centroids in optimization criterion, the classification rule using class centroids might be effective in spite of its simplicity.

Table 3: The verification of new efficient algorithm for LDA.

	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

7 Conclusion

In this paper, we have showed the relationship between LDA and the generalized MSE solution for multi-class problems. It generalizes the relation of the MSE solution with FDA and explains the meaning of the MSE solution on the undersampled problems. We also proposed an efficient algorithm for LDA which utilizes the relationship with the generalized MSE solution. In the Algorithm, the eigenvalue problem is replaced by the SVDs of the data size matrix and the small $r \times r$ matrix. In addition, the proposed algorithm does not need to compute the scatter matrices so it can save computational expense and memory requirements.

References

[1] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. Wiley-interscience, New York, 2001.

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

[3] F. Rosenblatt. principles of neurodynamics: perceptrons and the theory of brain mechanisms. Spartan Books, 1962.

[4] 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.

[5] R.O. Duda and P.E. Hart. Pattern classifiaction and scene analysis. Wiley Interscience, 1973.

[6] Y-C. Ho and R.L. Kashyap. An algorithm for linear inequalities and its applications. IEEE transactions on Elec. Comp., 14:683–688, 1965.

[7] K. Fukunaga. Introduction to Statistical Pattern Recognition. Acadamic 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] G.H. Golub and C.F. Van Loan. Matrix Computations. Johns Hopkins University Press, third edition, 1996.

[10] 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.