SIAM J. MATRIX ANAL. APPL.
Vol. 27, No. 2, pp. 474–492


# A RELATIONSHIP BETWEEN LINEAR DISCRIMINANT ANALYSIS AND THE GENERALIZED MINIMUM SQUARED ERROR SOLUTION*

CHEONG HEE PARK† AND HAESUN PARK‡

**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 is extended to multiclass 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.

**Key words.** dimension reduction method, generalized singular value decomposition, linear discriminant analysis, minimum squared error solution, pseudoinverse, undersampled problems

**AMS subject classifications.** 15A09, 68T10, 62H30, 65F15, 15A18

**DOI.** 10.1137/040607599

## 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 been proposed originating from the work by 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]. Baysian 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, Baysian 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 transforma-

*Received by the editors April 29, 2004; accepted for publication (in revised form) by A. H. Sayed March 7, 2005; published electronically November 22, 2005. This material is based upon work supported in part by 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.

http://www.siam.org/journals/simax/27-2/60759.html

†Department of Computer Science and Engineering, Chungnam National University, 220 Gung-dong, Yuseong-gu, Daejeon 305-764, South Korea (cheonghee@cnu.ac.kr).

‡Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 (hpark@cs.umn.edu), and The National Science Foundation, 4201 Wilson Boulevard, Arlington, VA 22230. Research was performed while this author was at the National Science Foundation (NSF) and was partly supported by IR/D from the NSF.

tion 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 data set 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 deal with *one linear function*, the relationship between them follows naturally. The MSE procedure is generalized for multiclass 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 multiclass 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 *LDA* has also been used to denote Baysian linear classifiers resulting from the assumption of normal density functions with a common covariance. We note that in the rest of the paper LDA refers to a dimension reduction method. Many problems including generalization of Baysian linear classifiers have been studied [9, 10], and under certain restrictions the relationships between Baysian 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 multiclass 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.** 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 $A$,

$$A = [a_1, \ldots, a_n] = [A_1, A_2, \ldots, A_r] \in \mathbb{R}^{m \times n}, \tag{2.1}$$

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 $i$ $(1 \le i \le r)$ has $n_i$ elements and the total number of data items is $n = \sum_{i=1}^{r} n_i$. Let $N_i$ $(1 \le i \le r)$ be the index set of data items in the class $i$. The data set $A$ can be considered a training set on which the modeling 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

$$S_b = \sum_{i=1}^{r} n_i (c_i - c)(c_i - c)^T, \qquad S_w = \sum_{i=1}^{r} \sum_{j \in N_i} (a_j - c_i)(a_j - c_i)^T, \tag{2.2}$$

$$S_t = S_b + S_w = \sum_{j=1}^{n} (a_j - c)(a_j - c)^T,$$

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

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

The distance between classes is quantified by $\operatorname{trace}(S_b)$, and $\operatorname{trace}(S_w)$ 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

$$J(G) = \operatorname{trace}\big((G^T S_w G)^{-1} (G^T S_b G)\big), \tag{2.3}$$

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 (2.3) 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 $\lambda$ of

$$S_b x = \lambda S_w x. \tag{2.4}$$

When $S_w$ is nonsingular, one can solve the eigenvalue problem

$$S_w^{-1} S_b x = \lambda x, \tag{2.5}$$

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 used 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

$$U_b^T Z_b X = [\underbrace{\Gamma_b}_{s} \underbrace{0}_{p-s}] \quad \text{and} \quad U_w^T Z_w X = [\underbrace{\Gamma_w}_{s} \underbrace{0}_{p-s}] \quad \text{for } s = \operatorname{rank}\left(\begin{bmatrix} Z_b \\ Z_w \end{bmatrix}\right), \tag{2.6}$$

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

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

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

$$U_b^T H_b^T X = [\Gamma_b \quad 0] \qquad \text{and} \qquad U_w^T H_w^T X = [\Gamma_w \quad 0]. \tag{2.7}$$

From (2.7) and $\Gamma_b^T\Gamma_b + \Gamma_w^T\Gamma_w = I_s$,

$$X^T S_b X = \begin{bmatrix} \Gamma_b^T\Gamma_b & \\ & 0_{m-s} \end{bmatrix} \equiv \begin{bmatrix} I_\mu & & & \\ & D_\tau & & \\ & & 0_{s-\mu-\tau} & \\ & & & 0_{m-s} \end{bmatrix} \tag{2.8}$$

and

$$X^T S_w X = \begin{bmatrix} \Gamma_w^T\Gamma_w & \\ & 0_{m-s} \end{bmatrix} \equiv \begin{bmatrix} 0_\mu & & & \\ & E_\tau & & \\ & & I_{s-\mu-\tau} & \\ & & & 0_{m-s} \end{bmatrix}, \tag{2.9}$$

where the subscripts on $I$ and 0 denote the order of square identity and zero matrices. Denoting the diagonal elements in (2.8) as $\eta_i$ and the diagonal elements in (2.9) as $\zeta_i$, we have

$$\zeta_i S_b x_i = \eta_i S_w x_i, \qquad i = 1, \ldots, m, \tag{2.10}$$

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

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

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 $Z = \left[\begin{smallmatrix} H_b^T \\ H_w^T \end{smallmatrix}\right] \in \mathbb{R}^{(r+n)\times m}$: $Z = P\left[\begin{smallmatrix} \Sigma & 0 \\ 0 & 0 \end{smallmatrix}\right] U^T$, 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 $X = U\left[\begin{smallmatrix} \Sigma^{-1}V & 0 \\ 0 & I \end{smallmatrix}\right]$, and assign them to the transformation matrix $G$.

Since $\Gamma_b^T\Gamma_b + \Gamma_w^T\Gamma_w = I_s$, from (2.8) and (2.9), we have

$$X^T S_t X = X^T S_b X + X^T S_w X = \begin{bmatrix} I_s & 0 \\ 0 & 0 \end{bmatrix}, \tag{2.11}$$

where $s = \text{rank}(Z)$. Equation (2.11) implies $s = \text{rank}(S_t)$ and from step 3 in the LDA/GSVD algorithm

$$S_t = X^{-T} \begin{bmatrix} I_s & 0 \\ 0 & 0 \end{bmatrix} X^{-1} = U \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} U^T, \tag{2.12}$$

which results in the eigenvalue decomposition (EVD) of $S_t$. Partitioning $U$ as

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

we have

$$X = U \begin{bmatrix} \Sigma^{-1}V & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} U_1\Sigma^{-1}V & U_2 \end{bmatrix}. \tag{2.13}$$

By substituting $X$ in (2.8) with (2.13),

$$\Sigma^{-1}U_1^T S_b U_1 \Sigma^{-1} = V\Gamma_b^T \Gamma_b V^T. \tag{2.14}$$

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 (2.12) and (2.14) show that the solution to LDA/GSVD can be obtained as follows:

1. Compute the EVD of $S_t$:

$$S_t = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix}. \tag{2.15}$$

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 step 2 of the new approach, denoting

$$F = U_1\Sigma^{-1} \in \mathbb{R}^{m\times s}, \tag{2.16}$$

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

$$F^T H_b = V\Gamma_b^T S^T = [\underbrace{V_1}_{r-1} \underbrace{V_2}_{s-r+1}] \begin{bmatrix} \Gamma_{b1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} S_1^T \\ S_2^T \end{bmatrix}, \tag{2.17}$$

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

$$X = [U_1\Sigma^{-1}V \;\; U_2] = [FV_1 \;\; FV_2 \;\; U_2] \tag{2.18}$$

and the transformation matrix $G$ is given as

$$G = FV_1 = U_1\Sigma^{-1}V_1. \tag{2.19}$$

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

$$0 = x^T S_b x = (x^T H_b)(H_b^T x) = \|x^T H_b\|^2 = \sum_{i=1}^{r} n_i \left| x^T c_i - x^T c \right|^2$$

and

$$0 = x^T S_w x = \sum_{j=1}^{n} \left| x^T a_j - x^T c_i \right|^2, \qquad \text{where } a_j \text{ belongs to the } i\text{th class.}$$

Hence

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

therefore,

$$U_2^T z = U_2^T c \tag{2.20}$$

for any given data item $z$. This implies that the vectors $x_i$, $i = s+1, \ldots, m$, 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 (2.8) and (2.9)

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

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 $[FV \ U_2] = [FV_1 \ FV_2 \ U_2]$.

**3. Minimum squared error solution.** The 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

$$g(z) = w_0 + w^T z = \begin{cases} \beta_1 & \text{if } z \in \text{ class 1}, \\ \beta_2 & \text{if } z \in \text{ class 2}, \end{cases} \tag{3.1}$$

where $\beta_i$ is the prespecified number for each class. For the data set $A$ given in (2.1), the problem (3.1) can be reformulated to minimize the squared error

$$\left\| \begin{bmatrix} 1 & a_1^T \\ \vdots & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \right\|_2^2, \tag{3.2}$$

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

$$\mathcal{P} = \begin{bmatrix} 1 & a_1^T \\ \vdots & \vdots \\ 1 & a_n^T \end{bmatrix}, \tag{3.3}$$

a solution which minimizes the squared error (3.2) can be computed using the pseudo-inverse $\mathcal{P}^+$ of $\mathcal{P}$ as

$$\begin{bmatrix} w_0 \\ w \end{bmatrix} = \mathcal{P}^+ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{3.4}$$

When the number of columns of $P$ is greater than the number of rows, i.e., $m+1 > n$, the least squares problem of (3.2) is underdetermined and there may exist infinitely many solutions. The one given in (3.4) 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 FDA [4]. The vector $w$ in (3.4) is the same as the solution $x$ of FDA except for some scaling factor $\alpha$ as

$$w = \alpha S_w^{-1}(c_1 - c_2) \equiv \alpha x \quad \text{and} \quad w_0 = -w^T c, \tag{3.5}$$

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

$$w^T z + w_0 = w^T(z - c) = \alpha x^T(z - c) > 0; \tag{3.6}$$

otherwise it is assigned to class 2.

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

$$g_i(z) = w_{0i} + w_i^T z = \begin{cases} \beta_i & \text{if } z \in \text{ class } i, \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

is to be found. The solution of the multiclass problem (3.7) in contrast to the problem (3.1) will be referred to as the *generalized* MSE solution whenever the distinction is needed. As in [4], one choice for $\beta_i$ would be assigning $\beta_i = 1$ for $i = 1, \ldots, r$. The squared error function in the multiclass problem is expressed using the Frobenius norm as

$$\left\| \begin{bmatrix} 1 & a_1^T \\ \vdots & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_{01} & \cdots & w_{0r} \\ w_1 & \cdots & w_r \end{bmatrix} - \begin{bmatrix} y_{11} & \cdots & y_{1r} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nr} \end{bmatrix} \right\|_F^2, \tag{3.8}$$

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

$$\mathcal{W} = \begin{bmatrix} w_{01} & \cdots & w_{0r} \\ w_1 & \cdots & w_r \end{bmatrix} \quad \text{and} \quad \mathcal{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1r} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nr} \end{bmatrix} \tag{3.9}$$

and with $\mathcal{P}$ defined as in (3.3), the MSE solution of the problem (3.8) can be obtained by

$$\mathcal{W} = \mathcal{P}^+ \mathcal{Y}, \tag{3.10}$$

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

$$g_i(z) > g_j(z). \tag{3.11}$$

Let us consider the mapping defined by the discriminant functions of the MSE solution (3.10) as

$$x \longrightarrow \big[g_1(x), \ldots, g_r(x)\big]^T \in \mathbb{R}^{r \times 1}. \tag{3.12}$$

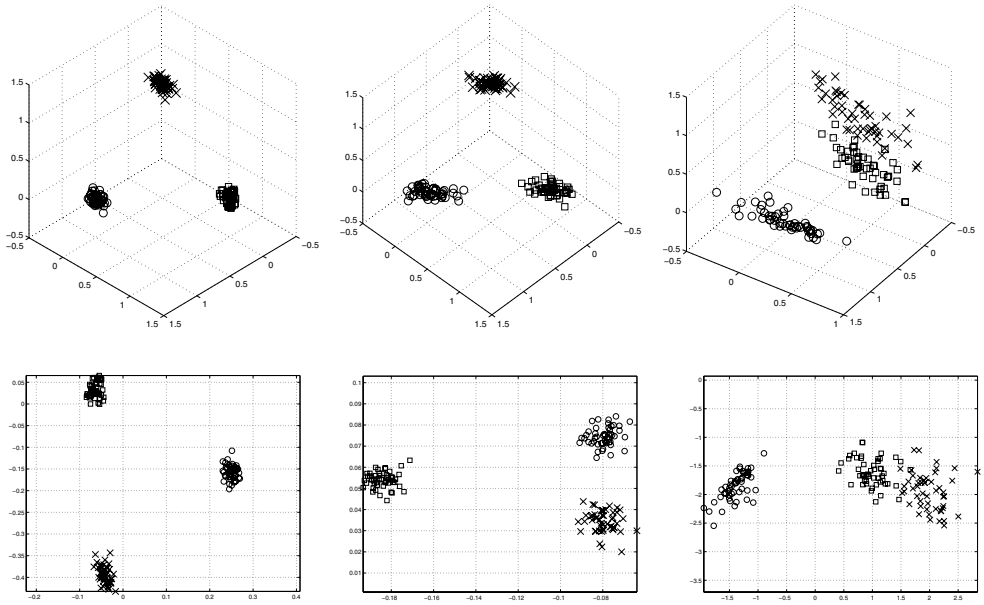

FIG. 1. *The spaces transformed by the mapping* (3.12) *induced by the generalized MSE solution (first row) and LDA (second row).*

Then (3.8) can be represented as

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

where $b_i \in \mathbb{R}^{r\times 1}$ $(1 \le i \le r)$ is the column vector with 1 in the $i$th position and 0 elsewhere. Hence in the space transformed by the mapping (3.12), the $i$th class centroid will be mapped close to the point $\beta_i b_i$. Figure 1 illustrates the transformed spaces by LDA and the mapping (3.12), where $\beta_i$ was set to 1, $1 \le i \le 3$, for a problem with three classes. The figures in the first row were obtained by the mapping (3.12) resulting in the dimension which is the same as the number of classes, while the figures in the second row 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 taking three subclasses in the Isolet data set from the UCI Machine Learning Repository.[1] The Isolet data set has 26 classes, and a detailed explanation of the data set will be given in section 6. The third figure on the top, which was obtained by the Iris data set, 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 multiclass problems.

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

**4. Relationships between LDA and the generalized MSE solution.** The relationship of the MSE solution and FDA given in (3.5) holds when the within-class scatter matrix $S_w$ is nonsingular. Now we show that the relationship (3.5) can be generalized for multiclass and undersampled problems by using the algorithm discussed in section 2.

**4.1. FDA and the MSE solution on undersampled problems.** Let

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

be the MSE solution to the problem

$$g(z) = w_0 + w^T z = \begin{cases} n/n_1 & \text{if } z \in \text{ class 1,} \\ -n/n_2 & \text{if } z \in \text{ class 2.} \end{cases} \tag{4.1}$$

The normal equations for the problem in (3.2) are

$$\begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} 1 & a_1^T \\ & \vdots \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} \frac{n}{n_1} e_{n_1} \\ -\frac{n}{n_2} e_{n_2} \end{bmatrix}, \tag{4.2}$$

where $e_{n_i}$ is the $n_i \times 1$ column vector with elements 1. From (4.2), we obtain

$$\begin{cases} nw_0 + nc^T w = 0, \\ ncw_0 + \left( \displaystyle\sum_{1 \le j \le n} a_j a_j^T \right) w = \dfrac{n}{n_1} \displaystyle\sum_{j \in N_1} a_j - \dfrac{n}{n_2} \sum_{j \in N_2} a_j. \end{cases} \tag{4.3}$$

From the first equation in (4.3) we have

$$w_0 = -c^T w. \tag{4.4}$$

By substituting (4.4) in the second equation of (4.3) and using the expressions of $S_b$ and $S_w$ for 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

$$(S_b + S_w) w = n(c_1 - c_2). \tag{4.5}$$

Let $x_1$ be the first column vector of $[FV_1 \ FV_2 \ U_2]$ in (2.18). From the discussion given in section 2 and the fact that $\operatorname{rank}(S_b) = 1$, we have

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

Since $\eta_1 + \zeta_1 = 1$,

$$\eta_1 (S_b + S_w) x_1 = (\eta_1 + \zeta_1) S_b x_1 = S_b x_1 = \frac{n_1 n_2}{n} (c_1 - c_2)(c_1 - c_2)^T x_1. \tag{4.6}$$

Denoting

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

(4.6) becomes

$$(S_b + S_w)\mu x_1 = n(c_1 - c_2). \tag{4.7}$$

Then by (4.5) and (4.7), we have

$$S_t w = (S_b + S_w)w = (S_b + S_w)\mu x_1 = S_t \mu x_1. \tag{4.8}$$

From (4.8) and the EVD of $S_t$ in (2.15),

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

and from (2.20) and (4.4) we obtain

$$\begin{aligned} w^T z + w_0 = w^T(z - c) &= w^T\big(U_1 U_1^T + U_2 U_2^T\big)(z - c) \\ &= w^T U_1 U_1^T (z - c) = \mu x_1^T U_1 U_1^T (z - c) \\ &= \mu x_1^T \big(U_1 U_1^T + U_2 U_2^T\big)(z - c) = \mu x_1^T (z - c). \end{aligned} \tag{4.9}$$

Equation (4.9) 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 one-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

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

can be solved by the normal equation

$$\mathcal{P}^T \mathcal{P} \mathcal{W} = \mathcal{P}^T \mathcal{Y}, \tag{4.10}$$

where $\mathcal{P}$, $\mathcal{Y}$, and $\mathcal{W}$ are defined in (3.3) and (3.9). From (4.10), we obtain

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

resulting in a linear system

$$\begin{cases} n\, w_{0i} + n\, c^T w_i = n_i \beta_i \\ n\, c\, w_{0i} + \left( \displaystyle\sum_{j=1}^n a_j a_j^T \right) w_i = n_i\, \beta_i c_i \end{cases} \qquad \text{for } i = 1, \ldots, r. \tag{4.11}$$

By substituting $w_{0i}$ of the second equation with $w_{0i}$ of the first equation in (4.11), (4.11) becomes

$$\big(n_i \beta_i - n c^T w_i\big)\, c + \left( \sum_{j=1}^n a_j a_j^T \right) w_i = n_i \beta_i c_i, \qquad i = 1, \ldots, r. \tag{4.12}$$

From (4.12) 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

$$S_t w_i = (S_b + S_w) w_i = n_i \beta_i (c_i - c), \qquad i = 1, \ldots, r. \tag{4.13}$$

Recall that according to (2.16), (2.17), and (2.19), the transformation matrix $G$ for LDA was obtained by

$$G = F V_1 = U_1 \Sigma^{-1} V_1, \tag{4.14}$$

where

$$S_t = U_1 \Sigma^2 U_1^T \quad \text{and} \quad F^T H_b = \Sigma^{-1} U_1^T H_b = V_1 \Gamma_{b1} S_1^T. \tag{4.15}$$

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

THEOREM 4.1. *Let $G$ be the dimension reducing transformation matrix from LDA given in* (4.14) *and let*

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

*be the discriminant functions for the MSE problem* (3.7)*. Then*

$$\begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_1 U_1^T = \begin{bmatrix} n_1\beta_1(c_1 - c)^T \\ \vdots \\ n_r\beta_r(c_r - c)^T \end{bmatrix} G G^T. \tag{4.16}$$

*Proof.* From (4.13), we have

$$\begin{aligned} S_t w_i = n_i\beta_i(c_i - c) &\to U_1\Sigma^2 U_1^T w_i = n_i\beta_i(c_i - c) \\ &\to w_i^T U_1 = n_i\beta_i (c_i - c)^T U_1 \Sigma^{-2}. \end{aligned} \tag{4.17}$$

Then by (4.17),

$$\begin{aligned} \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_1 U_1^T &= \begin{bmatrix} n_1\beta_1(c_1 - c)^T \\ \vdots \\ n_r\beta_r(c_r - c)^T \end{bmatrix} U_1 \Sigma^{-2} U_1^T \\ &= \operatorname{diag}(\sqrt{n_1}\beta_1, \ldots, \sqrt{n_r}\beta_r)\, H_b^T U_1 \Sigma^{-1} \left( V_1 V_1^T + V_2 V_2^T \right) \Sigma^{-1} U_1^T \\ &= \operatorname{diag}(\sqrt{n_1}\beta_1, \ldots, \sqrt{n_r}\beta_r)\, H_b^T F V_1 V_1^T F^T \\ &= \begin{bmatrix} n_1\beta_1(c_1 - c)^T \\ \vdots \\ n_r\beta_r(c_r - c)^T \end{bmatrix} G G^T. \end{aligned} \tag{4.18}$$

The third equality in (4.18) holds, since

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

from (2.14) and (2.17). □

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

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

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 (2.15). Then by Theorem 4.1 and in (4.11) for any data item $z$,

$$\begin{aligned}
\begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} &= \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} z = \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} (z - c) \\
&= \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_1 U_1^T (z - c) \\
&= \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} n_1\beta_1(\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r\beta_r(\tilde{c}_r - \tilde{c})^T \end{bmatrix} (\tilde{z} - \tilde{c}).
\end{aligned} \tag{4.19}$$

Equation (4.19) shows that the decision rule in the generalized MSE solution

$$\arg \max_{1 \le i \le r} \{g_i(z)\} \tag{4.20}$$

is equivalent to

$$\arg \max_{1 \le i \le r} \left\{ n_i\beta_i/n + n_i\beta_i(\tilde{c}_i - \tilde{c})^T(\tilde{z} - \tilde{c}) \right\} \tag{4.21}$$

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 $\beta_i = 1$ $(1 \le i \le r)$, then (4.21) becomes

$$\arg \max_{1 \le i \le r} \left\{ n_i/n + n_i(\tilde{c}_i - \tilde{c})^T(\tilde{z} - \tilde{c}) \right\}. \tag{4.22}$$

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

$$g_i(z) = w_{0i} + w_i^T z = \begin{cases} n/n_i & \text{if } z \in \text{ class } i, \\ 0 & \text{otherwise,} \end{cases} \tag{4.23}$$

(4.21) becomes

$$\arg \max_{1 \le i \le r} \left\{ (\tilde{c}_i - \tilde{c})^T(\tilde{z} - \tilde{c}) \right\}. \tag{4.24}$$

The difference between (4.22) and (4.24) is whether weighting by the number of elements in each class is considered or not.

The problem formulation (4.23) 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 one-dimensional representation obtained by FDA. Then the equivalence of (4.20) and (4.24) 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 (3.6) 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 $S_t$. For a given data item $z = a_i$, by (2.20)

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

and by Theorem 4.1, we have

$$\begin{aligned} \begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} &= \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} (z - c) \\ &= \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} \left(U_1U_1^T + U_2U_2^T\right)(z - c) \\ &= \begin{bmatrix} n_1\beta_1/n \\ \vdots \\ n_r\beta_r/n \end{bmatrix} + \begin{bmatrix} n_1\beta_1(\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r\beta_r(\tilde{c}_r - \tilde{c})^T \end{bmatrix} (\tilde{z} - \tilde{c}). \end{aligned} \tag{4.25}$$

Equation (4.25) is exactly the same as (4.19), 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 (4.25) becomes an approximation since (2.20) is based on the given training data set. If new data items come from the same distribution as the training data, (4.25) 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 (4.19) (and also (4.25)) of LDA and MSE, we have

$$\begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} \left[c_1 - c, \ldots, c_r - c\right] = \begin{bmatrix} n_1\beta_1(\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r\beta_r(\tilde{c}_r - \tilde{c})^T \end{bmatrix} \left[\tilde{c}_1 - \tilde{c}, \ldots, \tilde{c}_r - \tilde{c}\right], \tag{5.1}$$

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

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

Denoting

$$L \equiv \begin{bmatrix} \sqrt{n_1}(\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ \sqrt{n_r}(\tilde{c}_r - \tilde{c})^T \end{bmatrix} = H_b^T G, \tag{5.2}$$

(5.1) becomes

$$\begin{bmatrix} \frac{1}{\sqrt{n_1}\beta_1} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{n_r}\beta_r} \end{bmatrix} \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} [c_1 - c, \ldots, c_r - c] \begin{bmatrix} \sqrt{n_1} & & \\ & \ddots & \\ & & \sqrt{n_r} \end{bmatrix} = LL^T. \tag{5.3}$$

Let the EVD of the left side in (5.3) be $Q\Lambda Q^T$, where $Q$ is orthogonal and $\Lambda$ has nonincreasing diagonal components. Then

$$LL^T = Q\Lambda Q^T \equiv [\underbrace{Q_1}_{r-1} \quad \underbrace{Q_2}_{1}] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} = Q_1 \Lambda_1 Q_1^T. \tag{5.4}$$

On the other hand, from (5.2), (4.14), and (4.15)

$$L^T L = G^T H_b H_b^T G = \left(V_1^T F^T H_b\right)\left(H_b^T F V_1\right) = \left(\Gamma_{b1} S_1^T\right)\left(S_1 \Gamma_{b1}^T\right) = \Gamma_{b1}\Gamma_{b1}^T. \tag{5.5}$$

Hence from (5.4) and (5.5),

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

and we can obtain the SVD of $L$ as

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

When

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

$L$ has full column rank and $L^+ L = I$. When $S_w$ is nonsingular, from (4.19) we have

$$\begin{aligned} \begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} &= \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} n_1\beta_1(\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r\beta_r(\tilde{c}_r - \tilde{c})^T \end{bmatrix} \tilde{z} \\ &= \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} \sqrt{n_1}\beta_1 & & \\ & \ddots & \\ & & \sqrt{n_r}\beta_r \end{bmatrix} L\tilde{z}, \end{aligned} \tag{5.6}$$

and therefore

$$\begin{aligned} \tilde{z} &= L^+ \begin{bmatrix} \frac{1}{\sqrt{n_1}\beta_1} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{n_r}\beta_r} \end{bmatrix} \left( \begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} - \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} \right) \\ &= (\Lambda_1^{1/2})^{-1} Q_1^T \begin{bmatrix} \frac{1}{\sqrt{n_1}\beta_1} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r}\beta_r} w_r^T \end{bmatrix} z. \end{aligned} \tag{5.7}$$

Equation (5.7) shows that the reduced dimensional representation by LDA can be obtained from the discriminant functions of the MSE solution

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

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 (3.3) and (3.9), respectively.
2. Compute the EVD of the left-hand side of (5.3):
$$\begin{bmatrix} \frac{1}{\sqrt{n_1}\beta_1} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r}\beta_r} w_r^T \end{bmatrix} \left[\sqrt{n_1}(c_1 - c), \ldots, \sqrt{n_r}(c_r - c)\right] = [\underbrace{Q_1}_{r-1} \ \underbrace{Q_2}_{1}] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}.$$
3. For any data item $z$, the $(r-1)$-dimensional representation is given by
$$\Lambda_{11}^{-1/2} Q_1^T \begin{bmatrix} \frac{1}{\sqrt{n_1}\beta_1} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r}\beta_r} w_r^T \end{bmatrix} z.$$

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

For undersampled problems, (5.7) must change accordingly:

$$\begin{bmatrix} g_1(z) \\ \vdots \\ g_r(z) \end{bmatrix} = \begin{bmatrix} w_{01} \\ \vdots \\ w_{0r} \end{bmatrix} + \begin{bmatrix} w_1^T \\ \vdots \\ w_r^T \end{bmatrix} U_2 U_2^T z + \begin{bmatrix} n_1\beta_1(\tilde{c}_1 - \tilde{c})^T \\ \vdots \\ n_r\beta_r(\tilde{c}_r - \tilde{c})^T \end{bmatrix} \tilde{z}. \tag{5.8}$$

The second term of the right-hand side in (5.8) 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

$$\tilde{z} \approx \left(\Lambda_1^{1/2}\right)^{-1} Q_1^T \begin{bmatrix} \frac{1}{\sqrt{n_1}\beta_1} w_1^T \\ \vdots \\ \frac{1}{\sqrt{n_r}\beta_r} w_r^T \end{bmatrix} z. \tag{5.9}$$

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 complexities and saves memory requirements. Second, in addition to the SVD of the matrix $\mathcal{P}$ defined in (3.3), the EVD is needed only 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 newly 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, p. 254]. When $\mathcal{P}$ has full rank, QR decomposition can be used

TABLE 1
*The comparison of computational complexities. $m$: data dimension; $n$: number of data items; $r$: number of classes; $s$: rank of $Z$.*

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

TABLE 2
*The description of data sets.*

| | Data set | No. of classes | Dimension | 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 |

to compute the pseudoinverse of $\mathcal{P}$, which is cheaper than the SVD [21]. This is due to the fact that when $m + 1 \geq n$, the reduced QR decomposition of $\mathcal{P}^T = Q_1 R$ gives the pseudoinverse of $\mathcal{P}^T$ as $R^{-1}Q_1^T$; therefore, $\mathcal{P}^+ = Q_1(R^T)^{-1}$.

**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 data sets: the first has a nonsingular within-class scatter matrix $S_w$, and therefore the classical LDA can be performed for these data sets; the other is from undersampled problems which have singular scatter matrices. Data sets 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, and term frequency and inverse term frequency weighting and normalization [22]. The term-document matrix representation often makes the high dimensionality inevitable. Each data set is split randomly into training data and test data of equal size, and this is repeated 10 times in order to prevent any possible bias from random splitting. The detailed description of the data sets are shown in Table 2.

[2]http://www.ics.uci.edu/∼mlearn/MLRepository.html
[3]http://www-users.cs.umn.edu/∼karypis/cluto/ download.html

TABLE 3

*The comparison of classification performances for the verification of the relation* (6.1). *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 |

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 |

For all data sets in Table 2 the relationship between the MSE procedure and LDA

$$\begin{aligned} &\arg \max_{1\le i\le r} \{g_i(z) = w_{0i} + w^T z\} \\ &\quad = \arg \max_{1\le i\le r} \left\{ \frac{n_i\beta_i}{n} + n_i\beta_i \big(G^T c_i - G^T c\big)^T \big(G^T z - G^T c\big) \right\} \end{aligned} \tag{6.1}$$

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 (6.1) was verified for all the data sets, subject to minor differences in Tr23 and Tr45.

Algorithm 1 was tested for all the data sets 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 data sets used except in the B-scale data set with a 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 multiclass problems. It generalizes the relation between the MSE solution and FDA to multiclass 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*, Annu. Eugenics, 7 (Part II) (1936), pp. 179–188.

[2] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, Washington, D.C., 1962.

[3] B. E. Boser, I. M. Guyon, and V. Vapnik, *A training algorithm for optimal margin classifiers,* in Proceedings of the Fifth Annual Workshop on Computational Learning Theory (Pittsburgh, PA), ACM, New York, 1992.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley-Interscience, New York, 1973.

[5] Y-C. Ho and R. L. Kashyap, *An algorithm for linear inequalities and its applications*, IEEE Trans. Elec. Comp., 14 (1965), pp. 683–688.

[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*, 2nd ed., Academic Press, New York, 1990.

[8] J. S. Koford and G. F.Groner, *The use of an adaptive threshold element to design a linear optimal pattern classifier*, IEEE Trans. Inf. Theory, 12 (1966), pp. 42–50.

[9] T. Hastie, R. Tibshirani, and A. Buja, *Flexible discriminant analysis by optimal scoring*, J. Amer. Statist. Assoc., 89 (1994), pp. 1255–1270.

[10] T. Hastie and R. Tibshirani, *Discriminant analysis by Gaussian mixture*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 155–176.

[11] N. Campbell, *Canonical variate analysis—A general formulation*, Austral. J. Statist., 26 (1984), pp. 86–96.

[12] J. H. Friedman, *Regularized discriminant analysis*, J. Amer. Statist. Assoc., 84 (1989), pp. 165–175.

[13] T. Hastie, A. Buja, and R. Tibshirani, *Penalized discriminant analysis*, Ann. Statist., 23 (1995), pp. 73–102.

[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 Recogn., 33 (2000), pp. 1713–1726.

[15] H. Yu and J. Yang, *A direct LDA algorithm for high-dimensional data with application to face recognition*, Pattern Recogn., 34 (2001), pp. 2067–2070.

[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, Washington, D.C., 2004.

[17] J. Yang, J.-Y. Yang, and D. Zhang, *What's wrong with Fisher criterion?*, Pattern Recogn., 35 (2002), pp. 2665–2668.

[18] P. Howland, M. Jeon, and H. Park, *Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 165–179.

[19] P. Howland and H. Park, *Generalizing discriminant analysis using the generalized singular value decomposition*, IEEE Trans. Pattern Anal. Machine Intel., 26 (2004), pp. 995–1006.
[20] C. C. Paige and M. A. Saunders, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.
[21] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[22] T. G. Kolda and D. P. O'Leary, *A semidiscrete matrix decomposition for latent semantic indexing in information retrieval*, ACM Trans. Inf. Syst., 16 (1998), pp. 322–346.