<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Trace ratio criterion for feature extraction in classification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Li, Guoqi; Wen, Changyun; Wei, Wei; Xu, Yi; Ding, Jie; Zhao, Guangshe; Shi, Luping</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Li, G., Wen, C., Wei, W., Xu, Y., Ding, J., Zhao, G., et al. (2014). Trace Ratio Criterion for Feature Extraction in Classification. Mathematical Problems in Engineering, 2014, 725204-.</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>2014</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10220/20016">http://hdl.handle.net/10220/20016</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>© 2014 Guoqi Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.</td>
</tr>
</tbody>
</table>
A generalized linear discriminant analysis based on trace ratio criterion algorithm (GLDA-TRA) is derived to extract features for classification. With the proposed GLDA-TRA, a set of orthogonal features can be extracted in succession. Each newly extracted feature is the optimal feature that maximizes the trace ratio criterion function in the subspace orthogonal to the space spanned by the previous extracted features.

### 1. Introduction

Linear discriminant analysis (LDA) [1–3] has been proposed as a class separatory measure, which has been intensively used to reduce dimensionality of a classification problem as well as improve the generalization capability of a pattern classifier. Generally speaking, LDA method is to optimize the ratio criterion of the between-class distance and within-class distance constructed based on the available learning data. Such optimization can be realized by solving a generalized eigenvalue problem of the between-class and within-class scatter matrices [4].

In our opinion, there are three main problems for LDA methods. The first problem is its difficulty of dealing with high-dimensional data, where the number of observed samples is much lower than the samples’ feature dimension [5]. Many methods have been studied and proposed to address this problem; see, for example, the linear discriminant feature selection (LDFS) [6], Sparse Discriminant Analysis [5,7], and Sparse Tensor Discriminant Analysis [8].

The second problem is the well-known undersampled problem [9] in LDA method, in which scatter matrices may become singular due to insufficient samples. The solutions to this problem have also been well investigated in the following works such as the Regularized LDA [10, 11] using regularization techniques [12,13] and the Penalized LDA [14], the Pseudo Fisher Linear Discriminant [15], the Generalized Singular Value Decomposition [16], and the Uncorrelated LDA [17] and the Orthogonal LDA [17].

Basically the above two problems are quite similar and they can be unified as the same problem, which has also been extensively investigated in the above schemes. However, the third problem due to the LDA method can only extract quite limited features for classification problems [4]. For example, in two-class classification, one can only find one nonzero eigenvalue (extracted feature), as the between-class scatter matrix is a rank-one matrix. To the best of our knowledge, currently there is no good way to deal with this problem yet.

We focus on the third problem in this paper. A generalized LDA based on trace ratio criterion algorithm (GLDA-TRA) is derived to extract features from the input feature space. The algorithm first extracts a feature which maximizes the trace ratio criterion by solving a generalized eigenvalue problem. It is shown that such a generalized eigenvalue problem is the same as the generalized eigenvalue problem of LDA. Then, the learning data are projected to a subspace orthogonal to the space spanned by the extracted features. In that orthogonal subspace, the algorithm continues to extract a feature which maximizes the proposed trace ratio criterion. This process continues.
and, in this way, a set of orthogonal features is obtained iteratively. It is proven that each newly extracted feature is the optimal feature that maximizes the trace ratio criterion in the subspace orthogonal to the space spanned by the previous extracted features. Finally the extracted features are shown to give a sequence of trace ratios with magnitudes monotonically decreasing.

2. Problem Formulation

Let \((x, y) \in \mathbb{R}^d \times \mathcal{Y}\) be a sample, where \(\mathbb{R}^d\) denotes a \(d\)-dimensional feature space and \(\mathcal{Y} = \{1, 2, \ldots, C\}\) is a label set. Let \(x_{ij}\) denote the \(i\)th sample in the \(j\)th class. The within-class scatter matrix \(S_W\) and the between-class scatter matrix \(S_B\) are, respectively, defined as

\[
S_W = \frac{1}{n} \sum_{j=1}^{C} \sum_{i=1}^{n} (x_{ij} - \mu_j)(x_{ij} - \mu_j)',
\]

\[
S_B = \frac{1}{n} \sum_{j=1}^{C} \sum_{i=1}^{n} (m_j - \mu)(m_j - \mu)',
\]

where \(n_j\) is the number of samples in the \(j\)th class and \(n\) is the total number of samples and \(m_j\) is the sample mean of the \(j\)th class and \(m\) is the sample mean of all classes and the notation \(\cdot\)' means matrix transpose. Without loss of generality, we assume that \(n \gg d \gg C\). Then \(S_W\) can be constructed as a full rank matrix of rank \(d\) while \(S_B\) is at most with rank \(C - 1\). So, in this paper, \(S_W\) is considered as a symmetrical and positive definite matrix and \(S_B\) is a symmetrical and nonnegative definite matrix. The LDA method extracts features, that is, the column vectors in matrix \(W_{opt}\), in such a way that the ratio of the between-class scatter and the within-class scatter is maximized [25]. Consider

\[
W_{opt} = \arg \max_{W} \frac{\text{tr}(W'S_BW)}{\text{tr}(W'S_WW)}
\]

\[
= [\omega_1 \omega_2 \cdots \omega_m],
\]

where \(|:|\) is the determinant of a matrix and \(\{\omega_i\} \; i = 1, \ldots, m\) is the set of generalized eigenvectors of \(S_B\) and \(S_W\) corresponding to the \(m\) largest generalized eigenvalues \(\lambda_i \; i = 1, \ldots, m\) such that

\[
S_B\omega_i = \lambda_i S_W\omega_i, \quad i = 1, \ldots, m.
\]

Unfortunately, there are at most \(C - 1\) nonzero generalized eigenvalues as the rank of \(S_B\) is at most \(C - 1\). As such, for a two-class classification problem, LDA can only extract one feature.

To overcome this problem, we hope that one can continue to extract \(\omega_2\) after \(\omega_1\) is extracted. Generally speaking, we hope to extract \(\omega_{i+1}\) after the features \(\omega_1, \ldots, \omega_i\) are extracted with \(i\) starting from 0. We now use the trace ratio criterion function in [24, 26] to formulate the above-mentioned feature extraction problem. The optimization model we proposed is shown as follows:

\[
w_{i+1} = \arg \max_{\omega_{i+1}} F(W) = \frac{1}{2} \text{tr}(W'S_BW) \quad \text{subject to} \quad W = W_{i+1} = [W_i \; \omega_{i+1}]
\]

\[
W_{i+1} \text{ is orthonormal to span } \{W_i\},
\]

where \(W_i = [\omega_1 \omega_2 \cdots \omega_i]\) is a matrix denoting all the features extracted with \(W_0\) being empty, \(\omega_{i+1}\) is the feature to be determined in the space orthogonal to span\([W_i]\), and

\[
F(W) = \frac{1}{2} \omega_1 S_B\omega_1 + \cdots + \omega_i S_B\omega_i + \omega_{i+1} S_B\omega_{i+1},
\]

with \(\text{tr}(\cdot)\) denoting the trace of a matrix.

Remark 1. Later it can be shown that the first extracted vector \(\omega_1\) which maximizes \(F(W)\) in the space \(\mathbb{R}^d\) is the same \(\omega_1\) found by LDA in (2). Here it is also necessary to point out that the orthogonal constraint condition in \(W\) is needed in our formulated problem (4). As in (2), the numerator is the determinant of matrix \(W'S_BW\). Implicitly there is a constraint that \(\omega_1, \ldots, \omega_i\) cannot be the same. Because when \(\omega_1 = \cdots = \omega_i\), the numerator \(\text{tr}(W'S_BW)\) would be zero. But in (4), the numerator is the trace of \(W'S_BW\), which does not exclude the possibility that \(\omega_1, \ldots, \omega_i\) are the same. With the constraint in (4), such a possibility can be avoided.

3. Proposed Algorithm and Analysis

In this section, we present and analyze the proposed feature extraction algorithm. Our idea is summarized as follows. We first extract a feature \(\omega_1\) by maximizing the trace ratio criterion function involving \(S_W\) and \(S_B\) in (4). When the current extracted features become \(W_i = [\omega_1, \ldots, \omega_i]\), let span\([W_i]\) denote a space spanned by the linear combination of all the columns of \(W_i\) and let \(\text{span}(W_i^+)\) denote the space orthogonal to span\([W_i]\). Then, \(S_W\) and \(S_B\) are projected onto the subspace span\([W_i]\) by using projection operators \((I - W_i^+ W_i^+)\) and \(W_i^+ W_i^+\), respectively, where \(W_i^+ = (W_i^+ W_i^+) W_i^+\) is the generalized matrix inverse of a column full rank matrix \(W_i\). We continue the process to find \(\omega_{i+1}\) by optimizing (4) until all \(m \leq d\) features are extracted.

We first present the algorithm in Section 3.1. In Section 3.2., we will show how this algorithm is derived and then analyze its properties.

3.1. Proposed Algorithm. For convenience, we present the definition of a generalized eigenvalue as follows.

Definition 2. A number \(\lambda\) is called a generalized eigenvalue of matrix \(B\) with respect to \(A\) if \(\lambda\) satisfies that \(Bx = \lambda Ax\) for a nonzero vector \(x\), where \(A\) is a positive definite symmetrical matrix. When \(A = I\), \(\lambda\) is a normal eigenvalue of \(B\).
Mathematical Problems in Engineering

For an arbitrary full column rank matrix

Later in Lemma 5, it will be shown that $R_u$ with respect to $S_W$. This can be achieved by the following process:

do the Cholesky decomposition $S_W = G_i G_i^T$. Let $S_i = G_i^{-1} B_{i-1} (G_i^{-1})^T$ and obtain its maximum eigenvalue $\lambda_i$ together with its corresponding eigenvector $x_i$. Choose $w_i = (G_i^{-1}) x_i / \| (G_i^{-1}) x_i \|_2$ and $W_i = [W_{i-1}^{\top} w_i]$.\(^{(b)}\)

(b) Update stage as follows:

$$
S_{W_{i+1}} = (I - W_i W_i^+) S_W (I - W_i W_i^+) + \mu W_i W_i^+ S_W W_i W_i^+,
$$

$$
S_{B_{i+1}} = (I - W_i W_i^+) S_B (I - W_i W_i^+),
$$

where $u$ is a sufficiently small positive number.

Remark 3. (1) As $W$ is orthonormal, (6) can be rewritten as the following recursive from:

$$
S_{W_{i+1}} = (I - w_i w_i^+) S_W (I - w_i w_i^+) + \mu \sum_{j=1}^{d} w_j w_j^+ S_W w_j w_j^+,\quad (7)
$$

$$
S_{B_{i+1}} = (I - w_i w_i^+) S_B (I - w_i w_i^+).
$$

Later in Lemma 5, it will be shown that $S_{W_i}$ for $i = 1, \ldots, m$ can still be positive definite for a sufficiently small positive $u$. At each step $i$, $S_{W_i}$ is positive definite and the generalized eigenvalue of $S_B$ with respect to $S_{W_i}$ exists.

(2) The reason why we need to find the maximum eigenvalue of $S_i$ in the Calculation stage is shown in Lemma 8.

(3) Suppose that $i$ features $w_1, \ldots, w_i (i < d)$ have been extracted. Then Theorem 12 shows that $w_{i+1}$ is found to ensure the trace ratio criterion function in (4) to attain its maximum value.

(4) When $m = 1$, the LDA and GLDA-TRA extract the same feature.

(5) GLDA-TRA extracts $m \leq d$ features one by one. When $i = m = d$, $S_{B_{i+1}}$ becomes a zero matrix, and the algorithm will not extract any more features.

3.2. Derivation and Analysis of GLDA-TRA

Lemma 4. For an arbitrary full column rank matrix $W_i \in R^{d\times i}$, $W_i W_i^+ = W_i (W_i^\top W_i)^{-1} W_i^\top$ and $I - W_i W_i^+$ are projection operators which project a vector onto span$[W_i]$ and span$[W_i]^\perp$, respectively.

Proof. Suppose that $v \in R^{m\times 1}$ belongs to span$[W_i]$; there exists a vector $x \in R^{d\times 1}$ such that $v = W_i x$. It can be obtained that $W_i W_i^+ v = W_i (W_i^\top W_i)^{-1} W_i^\top x = v$ and $I - W_i W_i^+$ is positive definite since $I - W_i W_i^+$ is positive definite and the generalized eigenvalue $\lambda_i$ of $S_i$ is positive.

Lemma 5. $S_{W_{i+1}}$ in (6) for $i = 1, \ldots, m$ are positive definite for a sufficiently small positive $\mu$.

Proof. Note that $x S_{W_i} x > 0$ for any nonzero $x \in R^{m\times 1}$ as $S_W = S_W$. Let $x_1 = (I - W_j W_j^+) x$ and $x_2 = W_j W_j^+ x$. Then $x = (I - W_j W_j^+) x + W_j W_j^+ x = x_1 + x_2$. As $x \neq 0, x_1$ and $x_2$ cannot be zero at the same time. Thus $x^\top S_{W_{i+1}} x = x^\top (I - W_j W_j^+) S_W (I - W_j W_j^+) x + \mu x^\top (W_j W_j^+) S_W (W_j W_j^+) x = (I - W_j W_j^+) x + \mu (W_j W_j^+) x^\top S_W W_j W_j^+ x = x_1^\top S_W x_1 + \mu x_2^\top S_W x_2 > 0,$

for a sufficiently small positive number $\mu$.\(\square\)

Lemma 6. For matrices $W \in R^{d\times m}$, $X \in R^{d\times d}$, define $g(W) = \text{tr}(W^\top X W)$. Then $dg/dW = (X + X^\top) W$.

Proof. Let $W = [w_1 \cdots w_i \cdots w_m]$ and $X = [x_1 \cdots x_d]$, where $w_i \in R^{d\times 1}$ for $i = 1, \ldots, m$, and $x_i \in R^{d\times 1}$ for $i = 1, \ldots, d$. We have

$$
g(W) = \text{tr}(W^\top X W) = \text{tr}\left( d \sum_{i=1}^{d} w_i^\top x_1 w_1 + \cdots + d \sum_{i=1}^{d} w_i^\top x_d w_d \right)
$$

$$
= \text{tr}\left( \sum_{i=1}^{d} w_i^\top x_1 w_1 + \cdots + \sum_{i=1}^{d} w_i^\top x_d w_d \right)
$$

$$
= \sum_{i=1}^{d} \sum_{j=1}^{d} w_i^\top x_j w_j + \cdots + \sum_{i=1}^{d} \sum_{j=1}^{d} w_i^\top x_d w_d.
$$

Then, for $k_0 = 1, \ldots, m$, $j_0 = 1, \ldots, d$, we get

$$
dg(W) \left( w_{j_0}^\top w_{k_0} + \sum_{j=1}^{d} w_{j_0}^\top x_j w_{j_0} \right)
$$

$$
= d x_{j_0}^\top w_{k_0} + \sum_{i=1}^{d} x_{j_0}^\top w_{i k_0}.
$$

So (10) gives $dg/dW = (X^\top + X) W$.\(\square\)
Lemma 7. Let $w \in \mathbb{R}^{d \times 1}$ and the trace ratio function $f(w) = (1/2)(tr(w'Ww)/tr(w'Sw))$. The gradient $\nabla f(w) = df(w)/dw = (w'Ww - w'Sw)/tr(w'Sw)$. The gradient $\nabla f(w) = df(w)/dw = (w'Ww - w'Sw)/tr(w'Sw)$.

Proof. Let $g_1(w) = tr(w'Sw)$ and $g_2(w) = tr(w'Sw)$. From Lemma 6, $\nabla f(w) = df(w)/dw = (\partial g_1(w)/\partial w)g_1(w) - (\partial g_2(w)/\partial w)g_2(w) = (w'Ww - w'Sw)/tr(w'Sw)$. Defined $R(w) = tr(w'Sw)/tr(w'Sw)$, and we have the following results mentioned in the Calculation stage of step i.

Lemma 8. Assume that the maximum eigenvalue of $S_i$ is $\lambda^*$ with an eigenvector $x^*$. Then the unit vector $w^* = (G_i^{-1})'x^*/\|G_i^{-1})'x^*\|$ is an extracted feature ensuring that $R(w)$ attains its maximum value which is equal to $\lambda^*$.

Proof. Note that $f(w) = (1/2)(1/R(w)) = (1/2)(tr(w'Sw)/tr(w'Sw))$. Thus maximizing $R(w)$ is equivalent to minimizing $f(w)$. From Lemma 7, we have

$$\nabla f(w) = \frac{df(w)}{dw} = \frac{w'Sw - w'B}{(w'Sw)^2}.$$

Then one can minimize $f(w)$ by using an iterative method. Let $w(k+1) - w(k) = -\eta \nabla f(w(k))$, where $\eta$ is a small positive constant. Then $f(w(k+1)) = f(w(k)) - \eta \nabla f(w(k)) + o(\|w(k+1) - w(k)\|^2) \leq f(w(k))$. Thus $f(w(k))$ is a nonincreasing positive sequence and its limit, denoted as $f^*(w)$, exists. We have $\lim_{k \to \infty} f(w(k+1)) = f^*(w) = f^*(w)$.

As $\eta \to 0$, then $w(\infty)$ converges to an accumulation point $\bar{w}$ that satisfies $f^*(\bar{w}) = 0$. This gives

$$\frac{df(w)}{dw} \bigg|_{w=\bar{w}} = \frac{w'Sw - w'B}{(w'Sw)^2} = 0.$$

Let $\lambda = \text{tr}(w'Sw)/\text{tr}(w'Sw)$. From (12), we have $w'Sw = \lambda w'Sw$. That is, $\lambda$ is the generalized eigenvalue obtained from $S_iw = \lambda S_iw$.

Note that each eigenvector $w$ in (13) is an accumulation point of $f(w)$, since it satisfies (12). Now we hope to convert the generalized eigenvalue problem of (13) to a normal eigenvalue problem. This is achieved by Cholesky decomposition of $S_iw$, which is given as $S_iw = G_iG_i^T$, where $G_i$ is a full column rank lower triangular matrix. By doing this, (13) becomes

$$S_iw = \lambda G_iG_i^T w.$$

Defining $x = G_i'w$ and substituting $x$ into (14), it can be obtained that $S_iw = \lambda x$, where $S_i = G_i'G_iG_i'^{-1}$ is a symmetrical positive definite matrix. Let the maximum eigenvalue $\lambda^*$ and its corresponding eigenvector of $S_i$ be $\lambda^*$ and $x^*$. Finally the optimal $w^* = (G_i^{-1})'x^*/\|G_i^{-1})'x^*\|$ is a unit vector and $\lambda^* = \max \{R(w)\} = ||w'Sw||^*/||w'Sw||$.

Theorem 9. Suppose that $w_i$ is a feature extracted at step i; the trace ratio $(1/2)(tr(w_i'S_iw_i)/tr(w_i'S_iw_i)) = 0$ at step $i + 1$.

Proof. Note that $tr(w_i'S_iw_i) = w_i(I - W_iW_i')S_i(I - W_iW_i')w_i$. We have $tr(w_i'S_iw_i) = 0$ since $S_iw_i$ is positive definite by Lemma 5. Also, $(I - W_iW_i')w_i = 0$ as $(I - W_iW_i')w_i$ is a projection operator which projects to span$[W_i]$ based on Lemma 4. Thus, $(1/2)(tr(w_i'S_iw_i)/tr(w_i'S_iw_i)) = 0$.

Remark 10. As seen in Theorem 9, if the feature $w_i$ has been extracted at step i, then $w_{i+1}$ is supposed to make the trace ratio $(1/2)(tr(w_{i+1}'S_iw_{i+1})/tr(w_{i+1}'S_iw_{i+1}))$ attain its maximum in this step. While at step $i + 1$, after $S_{i+1}$ and $S_{i+1}$ are updated to $S_{i+1}$ and $S_{i+1}$ respectively, we have $(1/2)(tr(w_{i+1}'S_iw_{i+1})/tr(w_{i+1}'S_iw_{i+1})) = 0$ at $w = w_i$. This means that the algorithm needs to find $w_{i+1}$ which can maximize $(1/2)(tr(w_{i+1}'S_iw_{i+1})/tr(w_{i+1}'S_iw_{i+1}))$, and obviously, $w_{i+1}$ must be different from $w_i$.

Theorem 11. Sequence $\{F(w_i)\}_{i=1}^m$ produced by GLDA-TRA is a decreasing sequence.

Proof. Assume that $V_i$ is a space which the ith feature $w_i$ belongs to. That is, $w_i$ is the feature extracted from $V_i$ such that $F(w_i)$ attains its maximum. Now we consider $w_i \in V_i$ at step $i$ and $w_{i+1} \in V_{i+1}$ at step $i + 1$ with the respective corresponding maximum eigenvalues $\lambda_i$ and $\lambda_{i+1}$. As $S_{i+1}$ is positive definite; from Lemma 4 it can be obtained that

$$\|S_{i+1}w_i\|_2^2 = \|((I - W_iW_i')S_iw_i - \mu W_iW_i')w_i\|_2^2
\|I - W_iW_i')S_iw_i - \mu W_iW_i')w_i\|_2^2
\rightarrow \|((I - W_iW_i')S_iw_i - \mu W_iW_i')w_i\|_2^2,$$

when $\mu \to 0$. If $w_i \in \text{span}[W_i]$, $(I - W_iW_i')w_i = 0$. Then, $\|S_{i+1}w_i\|_2^2 = \|((I - W_iW_i')S_iw_i - \mu W_iW_i')w_i\|_2^2 = 0$. If $w_i \in \text{span}[W_i]$ and $(I - W_iW_i')w_i = w_i$ then $\|S_{i+1}w_i\|_2^2 > 0$. Additionally, $\lambda_i$ and $\lambda_{i+1}$ correspond to the maximum eigenvalues at steps $i$ and $i + 1$, it is obvious that

$$\|\lambda_iS_iw_i\|_2^2 \geq \mu \|w_i\|_2^2 > 0,$$

$$\|\lambda_{i+1}S_{i+1}w_{i+1}\|_2^2 \geq \mu \|w_{i+1}\|_2^2 > 0.$$

Thus, it can be concluded that $w_i \in \text{span}[W_i]$ while $w_{i+1} \in V_{i+1} \subseteq \text{span}[W_i]$. Similarly, $w_i \in V_i \subseteq \text{span}[W_{i-1}]$. As
Theorem 12. The feature $w_{i+1}$ extracted by GLDA-TRA is the optimal feature maximizing $F(W_{i+1})$ in (4) when $W_i = [w_1 \cdots w_i]$ is extracted.

Proof. Note that $w_{i+1} \in \text{span}[W_i]$ as $w_{i+1}$ is orthonormal to $\text{span}[W_i]$. Consider an arbitrary $\tilde{w}_{i+1} \notin w_{i+1} \in \text{span}[W_i]$. Construct $W_{i+1} = [W_i \ w_{i+1}]$ and $\tilde{W}_{i+1} = [W_i \tilde{w}_{i+1}]$. Note that $W_i$ is an orthonormal matrix. Then $w_i^T w_i = 0$ as long as $i_0 \neq j_0$ and we have

$$F(W_{i+1}) - F(\tilde{W}_{i+1}) = w_i^T S_B w_i + \cdots + w_i^T S_B w_i^T + w_i^T S_B w_i^T + w_i^T S_B w_i^T + w_i^T S_B w_i^T + w_i^T S_B w_i^T + w_i^T S_B w_i^T + w_i^T S_B w_i^T + w_i^T S_B w_i^T + w_i^T S_B w_i^T + w_i^T S_B w_i^T + w_i^T S_B w_i^T \cdots \cdots$$

Denote that $S_B = ((I - W_i W_i^T)S_B + W_i W_i^T S_B)$ and $S_W = (I - W_i W_i^T)S_W + W_i W_i^T S_W$. Then, $W_i W_i^T w_{i+1} = 0$ and $(I - W_i W_i^T)w_{i+1} = w_{i+1}$. Thus, we have

$$w_i^T S_B w_i^T = w_i^T ((I - W_i W_i^T)S_B + W_i W_i^T S_B) w_i^T = w_i^T ((I - W_i W_i^T)S_B + W_i W_i^T S_B) w_i^T = w_i^T S_B w_i^T.$$  \hfill (17)

Remark 13. Our main theoretical conclusions of this paper are shown in Theorems 9–12. The implication of Theorem 9 is summarized in Remark 10. In Theorem 11, the decreasing of the trace ratio sequence $\{F(w_i)\}$ means that the separability of the data set on the each newly extracted features decreases compared with the previous extracted features. This is the main result and contribution of this paper. Note that the trace ratio cost function is to evaluate whether a feature is good or not. By employing our proposed method, one can obtain $m (m \leq d$ and $m$ can be greater than $C - 1$) new extracted orthogonal features in which the $w_j$ is better than or at least equal to $w_{i+1}$, Theorem 12 gives us more concrete conclusions: if we already have $i$ features denoted as $w_1, \ldots, w_i$, the $i + 1$th feature $w_{i+1}$ extracted by GLDA-TRA is the optimal feature in the space orthogonal to the space spanned by $w_1, \ldots, w_i$. We present and prove Lemmas 4–8 as they are needed in proving Theorems 9-12. The theorems will be illustrated and verified in Section 4.

4. Simulation and Discussion

To simply illustrate our idea, two experiments are done. The first one is on Iris data set and the second one is on an artificial data set.

Example 14. Iris data set is a standard data set to verify the performance of classification algorithms. There are 150 data points belonging to three classes (C = 3): Setosa, Versicolor, and Virginica, respectively. Each class has 50 samples with four features (d = 4): Sepal length, Sepal width, Petal length, and Petal width.

One can construct $S_W$ and $S_B$ based on the Iris data set. It can be checked that rank($S_W$) = 4 and rank($S_B$) = $C - 1 = 2$, so LDA can only extract at most two features, while by employing GLDA-TRA to the Iris data set, we can extract $m (m \leq 4)$ features one by one, which are denoted as $W = [w_1, w_2, \ldots, w_m]$. When $m = 4$, the obtained $W = [w_1, w_2, w_3, w_4]$ is given as

$$W = \begin{bmatrix}
-0.0828 & 0.1717 & 0.6691 & -0.7183 \\
-0.4612 & -0.0012 & 0.6222 & 0.6325 \\
0.4446 & -0.8434 & 0.3004 & 0.0270 \\
0.7634 & 0.5090 & 0.2735 & 0.2886
\end{bmatrix}. \hfill (21)
$$

Obviously, $W'W = I$. Figure 1 shows the distribution of the three classes of the data points after projecting them onto each extracted feature $w_1, w_2, w_3,$ and $w_4$. As seen in Figure 1, the separability of the data decreases in the directions of $w_1, w_2, w_3,$ and $w_4$. One can also notice that the sequence $\{F(w_i)\}_{i=1}^{m}$ is strictly decreasing as shown in Figure 2. When using the extracted features to do classification via support vector machine (SVM), both LDA and GLDA-TRA can obtain good results in this example as the data points are well separated even if they are projected to one-dimensional space. When we extract one feature,
Table 1: Comparison of LDA and GLDA-TRA in the Iris dataset.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number of extracted features</th>
<th>Accuracy rate</th>
<th>Name</th>
<th>Accuracy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>1</td>
<td>98%</td>
<td>GLDA-TRA</td>
<td>98%</td>
</tr>
<tr>
<td>LDA</td>
<td>2</td>
<td>98%</td>
<td>GLDA-TRA</td>
<td>98%</td>
</tr>
<tr>
<td>LDA</td>
<td>3</td>
<td>Not available</td>
<td>GLDA-TRA</td>
<td>98.6667%</td>
</tr>
<tr>
<td>LDA</td>
<td>4</td>
<td>Not available</td>
<td>GLDA-TRA</td>
<td>98%</td>
</tr>
</tbody>
</table>

the accuracy is 98% (147/150) for both LDA and GLDA-TRA, since they extract the same feature when \( m = 1 \) (see Remarks 1 and 3). When we extract \( m \) features with our methods for \( m = 2, 3, \) and 4, the accuracies are, respectively, 98% (147/150) when \( m = 2 \), 98.6667% (148/150) when \( m = 3 \), and 98% (147/150) when \( m = 4 \). More details are shown in Table 1 and even when the data points are well separable, it is still observed that GLDA-TRA can be better than LDA in this case.

Example 15. This radar data was a 34-dimensional data set \( (x \in \mathbb{R}^{34}, y \in \mathcal{Y} = \{1,2\}, C = 2, d = 34) \)
The trace ratio for each extracted feature in order is decreasing; this can also be verified in Figure 4, where both trace ratio and its logarithm are plotted. And from Theorem 12, we know that each newly extracted feature is the optimal feature that maximizes the trace ratio function in the subspace orthogonal to the space spanned by the previous extracted features, which is exactly what we claimed in the abstract.

**Example 16.** In the previous examples, both LDA and GLDA-TRA perform quite well even when only one feature is extracted. In this example, we will show that the data points are inseparable if they are projected to one-dimensional space. The classification problem is given by

$$(x, y) \in \mathbb{R}^7 \times \mathcal{Y} = \begin{cases} \|x\|_2 = |1 + v_i| & \text{if } y = 1 \\ \|x\|_2 = |4 + v_i| & \text{if } y = 2, \end{cases} \quad (22)$$

where $y \in \mathcal{Y} = \{1, 2\}$ and $v_i$ is a variable following normal distribution $N(0, 1)$. It can be known that most data points with label 1 locate around the surface of a sphere with radius $r_1 = 1$ while data points with label 2 mostly locate around the surface of a sphere with radius 2.

To do the experiment, two hundred data points that are equally distributed in above two classes are generated. It is observed that LDA does not perform well if it is used to extract features in this problem. This is because LDA can only extract one feature and the data points are inseparable or well separable in an arbitrary one-dimensional feature space. By using GLDA-TRA, when we extract two features ($m = 2$), we obtain a projection matrix $W \in \mathbb{R}^{34 \times 2}$. By employing $\tilde{x} = W^T x$, one can then visualize the data points by projecting the date onto the two extracted features $\tilde{x}$. Figure 5 shows the classification using SVM on this two-dimensional extracted feature space. With GLDA-TRA, the data points can be better separated if we extract more ($m > 1$) features. The comparisons of LDA and GLDA-TRA regarding the classification accuracy rate are shown in Table 2 and Figure 6.

### 5. Conclusion

In this paper, a generalized linear discriminant analysis based on trace ratio criterion (GLDA-TRA) algorithm has been proposed. This is to overcome the problem that linear discriminant analysis (LDA) can only extract limited features in classification. It is shown that, in GLDA-TRA, a set of orthogonal features can be extracted one by one. Each newly extracted feature is the optimal feature that maximizes the trace ratio criterion function in the subspace orthogonal to the space spanned by the previous extracted features. Finally the extracted features are such that the trace ratio sequence of these features is decreasing in order. Experimental results also show the effectiveness of our proposed algorithm.

### Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.
The trace ratio (a) and logarithm of the trace ratio (b) on each extracted feature in order.

The data points projected onto the first two extracted features in Example 16.

Table 2: Comparison of LDA and GLDA-TRA in Example 16.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number of extracted features</th>
<th>Accuracy rate</th>
<th>Name</th>
<th>Accuracy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>1</td>
<td>69%</td>
<td>GLDA-TRA</td>
<td>69%</td>
</tr>
<tr>
<td>LDA</td>
<td>2</td>
<td>Not available</td>
<td>GLDA-TRA</td>
<td>77.5%</td>
</tr>
<tr>
<td>LDA</td>
<td>3</td>
<td>Not available</td>
<td>GLDA-TRA</td>
<td>88%</td>
</tr>
<tr>
<td>LDA</td>
<td>4</td>
<td>Not available</td>
<td>GLDA-TRA</td>
<td>92.5%</td>
</tr>
<tr>
<td>LDA</td>
<td>5</td>
<td>Not available</td>
<td>GLDA-TRA</td>
<td>96%</td>
</tr>
<tr>
<td>LDA</td>
<td>6</td>
<td>Not available</td>
<td>GLDA-TRA</td>
<td>97%</td>
</tr>
</tbody>
</table>
Figure 6: Classification accuracy rate for LDA and GLDA-TRA in Example 16.

References


