Uncorrelated trace ratio linear discriminant analysis for undersampled problems

Lei-Hong Zhang*

Department of Applied Mathematics, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, PR China

ARTICLE INFO

Article history:
Received 12 May 2010
Available online 13 November 2010
Communicated by F. Roli

Keywords:
Dimension reduction
Fisher criterion
Linear discriminant analysis
The ratio trace criterion
The trace ratio criterion
The uncorrelated LDA

ABSTRACT

For linear discriminant analysis (LDA), the ratio trace and trace ratio are two basic criteria generalized from the classical Fisher criterion function, while the orthogonal and uncorrelated constraints are two common conditions imposed on the optimal linear transformation. The ratio trace criterion with both the orthogonal and uncorrelated constraints have been extensively studied in the literature, whereas the trace ratio criterion receives less interest mainly due to the lack of a closed-form solution and efficient algorithms. In this paper, we make an extensive study on the uncorrelated trace ratio linear discriminant analysis, with particular emphasis on the application on the undersampled problem. Two regularization uncorrelated trace ratio LDA models are discussed for which the global solutions are characterized and efficient algorithms are established. Experimental comparison on several LDA approaches are conducted on several real world datasets, and the results show that the uncorrelated trace ratio LDA is competitive with the orthogonal trace ratio LDA, but is better than the results based on ratio trace criteria in terms of the classification performance.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

High dimensional data arises frequently in many modern applications of data mining, machine learning, bioinformatics and others. Analysis of such data is challenging since an enormous number of samples are required to perform accurate predictions on problems with a high dimensionality (Bellman, 1961; Fukunaga, 1990). Feature extraction and dimensionality reduction are common and efficient ways to handle the so-called curse of dimensionality by extracting a small number of features by removing irrelevant, redundant and noisy information. Many methods in this area have been proposed and linear discriminant analysis (LDA) is one of the most efficient statistical approaches for supervised dimensionality reduction and classification (see e.g., Bishop, 2006; Duchene and Leclercq, 1988; Fisher, 1936; Duda et al., 2001; Fukunaga, 1990; Hastie et al., 2001; Martinez and Kak, 2001; Martinez and Zhu, 2005; McLachlan, 2004; Theodoridis and Koutroambas, 1999).

Given a data matrix $A \in \mathbb{R}^{n \times d}$ in which each column $a_i \in \mathbb{R}^d$ $(i = 1, 2, \ldots, n)$ corresponds to a training sample, while each row corresponds to a particular feature, the classical LDA computes a linear transformation $G \in \mathbb{R}^{d \times d}$ (generally $d < n$), so that each original sample $a_i \in \mathbb{R}^d$ is mapped to a new reduced sample $y_i = G^\top a_i \in \mathbb{R}^d$.

Suppose $A = [A_1, \ldots, A_p] \in \mathbb{R}^{n \times d}$, where each $A_j \in \mathbb{R}^{n \times c}$ for $j = 1, \ldots, c$, each column of $A_j$ represents an independent class data set and $n_j$ denotes the number of the samples of the $j$th class in $A$ and $\sum_{j=1}^{c} n_j = n$. The centroid of cluster $A_j$ and the global centroid of all objects are defined as

$$m_j = \frac{1}{n_j} \sum_{x \in A_j} x = \frac{1}{n_j} A_j e^{(j)}, \quad \text{where } e^{(j)} = (1, \ldots, 1) \in \mathbb{R}^n,$$

$$m = \frac{1}{n} \sum_{x \in A} x = \frac{1}{n} A e,$$

respectively. Then the within-class scatter matrix $S_w$, the between-class scatter matrix $S_b$, and the total scatter matrix $S_t$ (Fukunaga, 1990) are defined respectively by

$$S_w = \frac{1}{n} \sum_{j=1}^{c} \sum_{x \in A_j} (x - m_j)(x - m_j)^\top = H_w H_w^\top \in \mathbb{R}^{n \times n},$$

$$S_b = \frac{1}{n} \sum_{j=1}^{c} n_j (m_j - m)(m_j - m)^\top = H_b H_b^\top \in \mathbb{R}^{n \times n},$$

$$S_t = \frac{1}{n} \sum_{j=1}^{c} (a_j - m)(a_j - m)^\top = H_t H_t^\top \in \mathbb{R}^{n \times n},$$

where

$$H_w = \begin{bmatrix} A_1 - m_1 & \ldots & A_c - m_c \end{bmatrix} \in \mathbb{R}^{n \times c},$$

$$H_b = \begin{bmatrix} \sqrt{n_1}(m_1 - m) & \ldots & \sqrt{n_c}(m_c - m) \end{bmatrix} \in \mathbb{R}^{n \times c},$$

* Tel.: +86 21 65904076.
E-mail address: longzhb@gmail.com

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.11.008
\[
H_t = \frac{1}{\sqrt{n}}(A - \mu^t) \in \mathbb{R}^{N \times n}.
\]

The relation
\[
S_t = S_b + S_w
\]

can be easily verified (Fukunaga, 1990) but is very important in designing efficient LDA algorithms. For a given linear transformation \(G\), the within-class cohesion as well as the between-class separation in the projected low-dimensional space are usually measured by \(tr(G^t S_b G)\) and \(tr(G^t S_w G)\), respectively. The principal in defining an optimal linear transformation is based on simultaneously maximizing the between-class separation and minimizing the within-class cohesion, in the optimization framework:

\[
\max_{G \in M} F(G),
\]

Though various optimization criteria for \(F(G)\) have been proposed in the literature, the ratio trace and trace ratio are two basic formula generated from the classical Fisher criterion (Fisher, 1936). The common ratio trace formula for \(F(G)\):

\[
F_1(G) = tr((G^t S_b G)^{-1} G^t S_b G)
\]

are very popular and widely used in LDA community (e.g., Duda et al., 2001; Fukunaga, 1990; Howland et al., 2003; Ye, 2005; Howland and Park, 2004a,b; Park et al., 2007; Ye and Xiong, 2006; Ye et al., 2004, 2006a). On the other hand, even though the following ratio trace criteria:

\[
F_2(G) = \frac{tr(G^t S_b G)}{tr(S_b G)}
\]

are natural and obvious, they receive less interest mainly due to the lack of a closed-form solution and efficient algorithms. However, it is claimed (Wang et al., 2007; Yan et al., 2005) that many popular dimensionality reduction algorithms can be formulated in the graph embedding framework, where the ratio trace criterion turns out to be the essential and original model, while the ratio trace criteria act only as its alternatives and may sacrifice the potential classification capability of the derived low-dimensional feature space. It is further stated (Wang et al., 2007) that the solutions based on the ratio trace may deviate from the original objective and suffer from the fact that it is invariant under any nonsingular transformation which may lead to uncertainty in subsequent processing such as classification and clustering. Refer to Wang et al. (2007) for more discussions on the comparison between these two criteria.

On the other hand, different constraints \(M\) for the optimal transformation \(G\) have also been imposed and lead to different LDA models. The common two constraints are the orthogonal constraint \(G^t G = I_l\) and the \(S_t\)-orthogonal constraint \(G^t S_t G = I_l\) where \(I_l\) stands for the identity matrix of size \(l\). The latter leads to the so-called uncorrelated features in the reduced space (Ye, 2005; Ye et al., 2004). For the ratio trace criteria \(F_1(G)\) and \(F_2(G)\), the corresponding orthogonal LDA models have been proposed and discussed (see Duchene and Leclercq, 1988; Foley and Sammon, 1975; Ye, 2005; Ye and Xiong, 2006), while some associated uncorrelated LDA models have also been discussed (see e.g., Howland et al., 2003; Howland and Park, 2004a; Park et al., 2007; Ye, 2005; Ye et al., 2004, 2006b; Jin et al., 2001a,b). For both constraints, efficient algorithms have been proposed and the resulted transformations are effective in classification.

The trace ratio criterion \(F_2(G)\) with the orthogonal constraint has been considered and tested (Wang et al., 2007; Guo et al., 2003; Nie et al., 2009; Zhang et al., 2010). Two iterative algorithms are available thus far for the resulted optimization problem. The first iterative method (Guo et al., 2003) is based on the bisection technique, and a much more efficient algorithm is recently suggested (Wang et al., 2007). Interestingly, the most recent work (Zhang et al., 2010) further shows that the second algorithm converges quadratically even in the two kinds of inexact computations. Due to the developments of the efficient algorithm, the trace ratio criterion begins to receive interests. For example, the uncorrelated constraint has been mentioned in (Nie et al., 2009), and if the \(S_t\)-orthogonal constraint is imposed to \(F_2(G)\), it results to the following optimization problem

\[
\max_{G \in M} \frac{tr(G^t S_b G)}{tr(S_b G)}.
\]

which is referred as the uncorrelated trace ratio discriminant analysis (UTRDA). It is interesting to note in Section 4 that the UTRDA (8) is equivalent to the classical undersampled LDA (ULDA) (Jin et al., 2001a; Ye et al., 2006b):

\[
\max_{G \in R^d} tr((G^t S_b G)^{-1} G^t S_b G).
\]

However, both the mentioned ratio trace and trace ratio criteria suffer from the so-called undersampled problem or the singularity problem (see e.g., Krzanowski et al., 1995). Undersampled problems frequently arise from the modern data such as the image databases of facial recognition, gene expression data, as well as the text documents, in which the dimension of features \(N\) could easily be up to several thousands and is larger than the number of samples \(n\). In recent years, many approaches (e.g., Fukunaga, 1990; Howland et al., 2003; Howland and Park, 2004a; Park et al., 2007; Ye, 2005; Ye et al., 2004; Friedman, 1989; Liu et al., 2008; Swets and Weng, 1996) have been proposed to overcome the undersampled problem. Among the successful and effective approaches, the regularization technique by Friedman (1989) is simply to add a scaled multiple of the identity matrix \(\mu I\) to the corresponding singular scatter matrix, where \(\mu > 0\) is known as the regularization parameter.

With the aid of the regularization technique, the ratio trace criterion subject to the orthogonal constraint is called the ROLDA (Ye and Xiong, 2006), while the following optimization problem with \(S_t\)-orthogonal constraint

\[
\max_{G \in M} \frac{tr(G^t S_b G + \mu I_l)}{tr(S_b G + \mu I_l)}
\]

leads to the so-called ULDA/NSVD (Ye, 2005; Ye et al., 2006b). Similarly, for the trace ratio criterion \(F_2(G)\), a regularization and orthogonal version, namely RGST (Zhang et al., 2010), is obviously

\[
\max_{G \in M} \frac{tr(G^t S_b G)}{tr(S_b G + \mu I_l)}.
\]

For the UTRDA model (8), on the other hand, we propose in this paper two regularization versions, namely UTRDA(R1):

\[
\max_{G \in M} \frac{tr(G^t S_b G)}{tr(S_b G + \mu I)}.
\]

and UTRDA(R2):

\[
\max_{G \in M} \frac{tr(G^t S_b G)}{tr(S_b G + \mu I_l)}.
\]

The fashion in regularizing the singular matrix for the UTRDA(R1) and the UTRDA(R2) is different in that the UTRDA(R1) adds a uniform regularization term \(\mu I\) to the reduced within-class scatter matrix for any linear transformation \(G\), while the UTRDA(R2) adds \(\mu I_l\) to the original within-class scatter matrix which has effects on the resulted optimal transformation.
Our main contributions in the present paper include:

- the equivalence relation between the UTRDA(R1) and the ULDA/ GSVD (10),
- the characterization of the global solution of the UTRDA(R2),
- an efficient algorithm for solving the UTRDA(R2) based on a reduced but equivalent problem of (13),
- an experimental comparison among several LDA approaches include the ULDA/GSVD, the OLDA (Ye, 2005), the RGFST and the UTRDA(R2).

We last mention that the regularization value $\mu$ will significantly affect the classification performance and choosing an appropriate $\mu$ turns out to be a critical issue. Cross-validation is commonly employed to estimate the optimal $\mu$ from a given finite set of candidate (Hastie et al., 2001). The computational cost of cross-validation for optimal $\mu$ can be expensive since for each $\mu$ in the candidate set, expensive matrix computations are required in order to obtain the corresponding optimal transformation. However, similar to the computation fashion for $\mu$ in ROLDA (Ye and Xiong, 2006), we will show in Section 4 that the computational burden in choosing the optimal $\mu$ in UTRDA(R2) can be reduced by a two-stage scheme and hence improves the efficiency of the cross-validation in the UTRDA(R2).

The rest of the paper is organized in the following way: we review, in Section 2, the uncorrelated and orthogonal ratio trace LDA, and the orthogonal trace ratio LDA in Section 3, by presenting efficient algorithms and some related theoretical properties. These two sections serve as not only a brief overview but the preliminaries for our proposed regularized uncorrelated trace ratio LDA discussed in Section 4. Two regularization forms together with the characterization of the global solutions and efficient algorithm are discussed. In Section 5, we report experimental results by comparing the classification performances of the ULDA/GSVD, the OLDA (Ye, 2005), the RGFST and the UTRDA(R2). Final remarks are drawn in Section 6.

2. The uncorrelated and orthogonal ratio trace LDA

Recall that the orthogonal and $S_l$-orthogonal constraints are two common conditions imposed on the transformation $G$. The $S_l$-orthogonal constraint in LDA is able to produce preferable statistical property, namely the uncorrelated features in the projected low-dimensional space as shown in the following lemma (Ye, 2005).

**Lemma 2.1.** Let the transformation matrix be $G = [g_1, \ldots, g_l]$ for some $l > 0$. The original feature vector $A$ is transformed into $Z = G^T A$, where the $i$th feature of $Z$ is $z_i = g_i^T A$. Assume that $g_i$ and $g_j$ are $S_l$-orthogonal to each other, i.e., $g_i^T S_l g_j = 0$ for $i \neq j$. Then the correlation between $z_i$ and $z_j$ is 0 for $i \neq j$. That is, $z_i$ and $z_j$ are uncorrelated to each other.

The classical uncorrelated LDA (ULDA) (Jin et al., 2001a) is proposed for the case when $S_o$ is nonsingular, where the discriminant vectors $g_1, \ldots, g_l$ are generated in a successive way. Suppose $g_1, \ldots, g_l$ have been obtained, the $(i + 1)$th column $g_{i+1}$ of $G$ maximizes the Fisher criterion function under the constraint $g_i^T S_l g_0 = 0$, for $i = 1, \ldots, l$; that is, $g_{i+1}$ is the optimal solution to the following optimization problem:

$$\max_{g^T S_l g_0} g^T S_l g_{i+1}.$$  

The algorithm in (Jin et al., 2001a) solves $g_{i+1}$ as the eigenvector corresponding to the maximum eigenvalue of a special generalized eigenvalue problem. Interestingly, instead of the successive way in generating $G$, it is shown (Ye et al., 2006b) that the variant of classical LDA (9) is equivalent to the classical uncorrelated LDA and the optimal solution of (9) is an $S_l$-orthogonal eigenbasis corresponding to the $l$-largest eigenvalues of the matrix $S_l^{-1} S_0$.

For the undersampled problem, Ye et al. (2006b) further introduces a regularized ULDA, namely ULDA/GSVD, based on the optimization problem (10). It is shown that the optimal solution of (10) is independent of the regularization parameter $\mu > 0$, and Algorithm 1 is proposed as an efficient method to obtain an optimal transformation $G$.

**Algorithm 1.** Uncorrelated LDA (Ye, 2005)

<table>
<thead>
<tr>
<th>Input: data matrix $A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output: transformation matrix $G$</td>
</tr>
<tr>
<td>1. Form matrices $H_0$ and $H_1$ as in (15) and (16), respectively;</td>
</tr>
<tr>
<td>2. Compute reduced SVD of $H_1$ as</td>
</tr>
<tr>
<td>$H_1 = U \cdot \text{diag}(\Sigma_1, 0) \cdot V$</td>
</tr>
<tr>
<td>$\Sigma_1 = [u_1, u_2] \cdot \text{diag}(\Sigma_2, 0) \cdot [v_1, v_2]^T = u_1 \Sigma_2 v_1^T$;</td>
</tr>
<tr>
<td>3. Form $B = \Sigma_1^{-1} H_1$;</td>
</tr>
<tr>
<td>4. Compute SVD of $B$ as $B = P \Sigma Q^T$;</td>
</tr>
<tr>
<td>$q = \text{rank}(B)$;</td>
</tr>
<tr>
<td>5. Form $X = U, \Sigma_1^{-1} p$;</td>
</tr>
<tr>
<td>6. Set $G = [x_1, \ldots, x_l]$.</td>
</tr>
</tbody>
</table>

Let the matrices $U$, $\Sigma_1$, $\Sigma_2$ and $P$ be given by Algorithm 1. Then it can be shown (Ye et al., 2006b) that the following matrix

$$X = U_{\Sigma_2^{-1} P 0 \ \ 0} \infty \text{rank}(\Sigma_1),$$

simultaneously diagonalizes $S_w$, $S_b$ and $S_b$, and satisfies

$$X^T S_b X = \left[ \begin{array}{cc} \Sigma_b & 0 \\ 0 & 0 \end{array} \right], \quad X^T S_w X = \left[ \begin{array}{cc} \Sigma_w & 0 \\ 0 & 0 \end{array} \right], \quad X^T S_f X = \left[ \begin{array}{cc} I & 0 \\ 0 & 0 \end{array} \right].$$

where $\Sigma_b = \Sigma_2^2 = \text{diag}(\lambda_1 \geq \cdots \geq \lambda_q > \geq 0 = \lambda_{q+1} = \cdots = \lambda_l)$, and $\Sigma_w = I_l - \Sigma_b$.

The orthogonal LDA (OLDA) (Ye, 2005) is a natural alternative to ULDA/GSVD whose optimal transformation can be computed by simply taking the orthogonal part of the QR factorization of $X$ given in Algorithm 1 and the pseudo-code is stated in Algorithm 2.

**Algorithm 2.** Orthogonal LDA (Ye, 2005)

<table>
<thead>
<tr>
<th>Input: data matrix $A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output: transformation matrix $G$</td>
</tr>
<tr>
<td>1. Compute the matrix $X$ as in ULDA (Step 1–5 of Algorithm 1);</td>
</tr>
<tr>
<td>2. Compute the QR decomposition of $X$ as $X = QR$;</td>
</tr>
<tr>
<td>3. $G = Q$.</td>
</tr>
</tbody>
</table>

3. The orthogonal trace ratio LDA

In the literature of LDA, Foley and Sammon (1975) first introduced the idea to find the orthogonal transformation for the two-class problem. The idea is extended to the multi-class problem by Duchene and Leclercq (1988) which is similar to the classical ULDA (Jin et al., 2001a) in finding the discriminant vectors successively. Though both Foley–Sammon discriminant analysis and the classical ULDA use the same Fisher criterion function to find the $i$th discriminant vector, the main difference lies in that the optimal discriminant vectors generated by the classical ULDA are $S_l$-orthogonal, while the optimal discriminant vectors by Foley–Sammon discriminant analysis are orthogonal to each other.
The generalized Foley–Sammon transformation (GFST) is later defined as the global solution of the optimization problem when \(S_w\) is nonsingular (see e.g., Guo et al., 2003; Wang et al., 2007; Zhang et al., 2010):

\[
\max_{C \in \mathbb{C}^{t \times n}} \frac{\text{tr}(C^\top S_t G)}{\text{tr}(C^\top S_w G)}.
\]

(16)

Rather than using a successive way for finding the optimal transformation \(G^*\), (16) is intended to define \(G^*\) in a global manner. For the undersampled problem, furthermore, the regularization GFST (RGFST) (11) and its efficient algorithm is recently established (Zhang et al., 2010) according to its reduced but equivalent problem

\[
\max_{C \in \mathbb{C}^{t \times n}, G \in \mathbb{R}^{m \times n}} \frac{\text{tr}(C^\top S_t G)}{\text{tr}(C^\top (S_w + \mu I) G)},
\]

where \(S_t = Q^\top S_t Q\), \(S_w = Q^\top S_w Q\), and \(Q\) is orthogonal generated from the QR decomposition of \(A\); i.e., \(A = Q R\). As the reduced problem (17) involves matrices with much smaller size, it effectively improves the computational overhead for undersampled problems.

Notice that both (16) and (11) fall into the unified optimization problem

\[
\max_{V \in \mathbb{R}^{m \times n}, \psi \in \mathbb{R}^m} \frac{\text{tr}(V^\top B V)}{\text{tr}(V^\top W V)}.
\]

(18)

where \(B \in \mathbb{R}^{m \times m}\) is positive semi-definite and \(W \in \mathbb{R}^{m \times m}\) is positive definite. It is known that the global solution can be characterized by Theorem 3.1 (see Zhang et al. (2010) and Guo et al. (2003), for more details).

**Theorem 3.1.** Let \(\psi^*\) be the global optimal objective value of (18). Then any \(V^* \in \mathbb{R}^{m \times l}\) solves (18) globally if and only if \(V^*\) is an orthonormal eigenbasis corresponding to the \(l\)-largest eigenvalues of the matrix

\[
E_{V^*} = B - \psi^* W.
\]

(19)

Moreover, the sum of the \(l\)-largest eigenvalues of the matrix \(E^*\) is zero.

Even though the global solution, as stated in Theorem 3.1, does not possess a closed-form, an efficient iterative algorithm (Wang et al., 2007) is proposed with the pseudo-code simply stated as Algorithm 3.

**Algorithm 3.** A fast iterative scheme for solving (18)

**Input:** data matrices \(B\) and \(W\), and \(l\)

**Output:** the optimal solution \(V\)

1. **Initial step:** Select any \(V_0 \in \{V \in \mathbb{R}^{m \times l} | V^\top V = I_l\}\), and the tolerance \(\varepsilon > 0\). Set \(k = 0\).

2. Compute an orthonormal eigenbasis \(V_{k+1}\) corresponding to the \(l\)-largest eigenvalues of

\[
E_{V_k} := B - \psi_k W, \quad \psi_k := \psi(V_k).
\]

(20)

3. If \(\psi_{k+1} - \psi_k < \varepsilon\), then stop; (if \(\psi_{k+1} - \psi_k = 0\), then \(V_{k+1}\) solves (18) globally.) otherwise, set \(k = k + 1\) and go to 2.

The algorithm is related with the Newton iteration (Ngo et al., 2009) and it is further shown to be of global and quadratic convergence (Zhang et al., 2010) under the mild and generic assumption that the \(l\)-largest eigenvalue \(\lambda_i(E_{V_k})\) of \(E_{V_k}\) is strictly larger than the \((l + 1)\)-largest eigenvalue \(\lambda_{l+1}(E_{V_k})\); that is, \(\lambda_i(E_{V_k}) > \lambda_{l+1}(E_{V_k})\). Under this generic assumption, the global maxima set concerns essentially the dominant eigenspace of \(E^*\) and the convergence rate of \(\{V_k\}\) is then measured by the distance between subspaces (Golub and Van Loan, 1996).

**Definition 3.1.** Let \(\mathcal{L}_1\) and \(\mathcal{L}_2\) be two subspaces of \(\mathbb{R}^m\) with the same dimension, the distance between \(\mathcal{L}_1\) and \(\mathcal{L}_2\) is defined by

\[
\text{dist}(\mathcal{L}_1, \mathcal{L}_2) = \|\pi_{\mathcal{L}_1} - \pi_{\mathcal{L}_2}\|_2,
\]

(21)

where \(\pi_{\mathcal{L}_1}\) and \(\pi_{\mathcal{L}_2}\) are the orthogonal projections onto \(\mathcal{L}_1\) and \(\mathcal{L}_2\), respectively.

**Theorem 3.2.** Let \(V^*\) be any global maximizer of (18) and suppose \(\lambda_i(E_{V^*}) > \lambda_{i+1}(E_{V^*})\) holds. Let \(V_k\) and \(\psi_k\) be generated from Algorithm 3, then

\[
\text{dist}((\text{span}(V_k), \text{span}(V^*)) \rightarrow 0
\]

(22)

quadratically, and moreover, \(\psi_k \rightarrow \psi^*\) quadratically as \(k \rightarrow +\infty\).

4. The uncorrelated trace ratio LDA

In this section, we will discuss the uncorrelated trace ratio LDA methods, putting our emphasis on the application on the undersampled problem. Two regularization uncorrelated trace ratio LDA approaches will be proposed and their corresponding optimal transformations and efficient algorithms will be presented.

4.1. The uncorrelated trace ratio discriminant analysis (UTRDA)

When the scatter matrix \(S_w\) is nonsingular, the uncorrelated version, the UTRDA (8), of the GFST (16) has been mentioned (Nie et al., 2009). By rewriting (8) into its equivalent problem

\[
\max_{C \in \mathbb{C}^{l \times m}} \frac{\text{tr}(G^\top S_t G)}{\text{tr}(G^\top S_w G)}.
\]

(23)

and by employing the Lagrange multiplier, the global solution is shown to be the \(S_t\)-orthogonal eigenbasis corresponding to the \(l\)-largest eigenvalues of \(S_t^{-1}S_w^\top\). In fact, this fact can also be derived from Theorem 3.1. Let \(G^*\) be an arbitrary global maximizer of (23). Defining

\[
V^* = S_t G^*, \quad B = S_t^{-1}S_p S_t^{-1}, \quad \text{and} \quad W = I_l,
\]

(23) then is in the form of (18). Since \(V^*\) is a global maximizer, it is an orthonormal eigenbasis of \(S_t^{-1}S_p S_t^{-1} - \psi^* V^\top V^*\) corresponding to the \(l\)-largest eigenvalues. Obviously, \(V^*\) is also an orthogonal eigenbasis of \(S_t^{-1}S_p S_t^{-1}\) corresponding to its \(l\)-largest eigenvalues, and hence there is a matrix \(M \in \mathbb{R}^{l \times l}\) (eigenblock) such that

\[
S_t^{-1}S_p S_t^{-1}V^* = V^* M, \quad \text{or} \quad S_t^{-1}S_p G^* = G^* M.
\]

Since \(S_t^{-1}S_p S_t^{-1}\) and \(S_t^{-1}S_p\) have the same eigenvalues, the previous relation implies that \(G^*\) is an \(S_t\)-orthogonal eigenbasis corresponding to the \(l\)-largest eigenvalues of \(S_t^{-1}S_p\).

**Remark 4.1.** When \(S_w\) is nonsingular, the UTRDA (8) is equivalent both to the classical ULDA (Jin et al., 2001a) and to the variant LDA (9) (Ye et al., 2006b).

As has been already pointed out, the ratio trace criterion can be viewed as an approximation of the trace ratio criterion (Wang et al., 2007; Nie et al., 2009). However, Remark 4.1 implies that under the \(S_t\)-orthogonal constraint, this approximation is exact when \(S_w\) is nonsingular. Interestingly, for the undersampled problem, we shall further show that the equivalence of these criteria still holds for our first regularization UTRDA under the \(S_t\)-orthogonal constraint.
4.2. The first type regularization uncorrelated trace ratio LDA–UTRDA(R1)

For the undersampled problem, the UTRDA(R1) defined in (12) is our first regularization LDA based on the trace ratio criterion \( E(G) \). It is worth pointing out that the regularization manner in the UTRDA(R1) (12) is the same as that in the ULDA/GSVD (10) (Ye et al., 2006b), where a uniform regularization term \( \mu_l \) is added to the reduced within-class scatter matrix \( G' = \Sigma_b G \) in the projected low-dimensional space. Because \( \mu_l \) is independent of the linear transformation \( G \), the global solution of (12) then could be expected to be independent of \( \mu \). In fact, we shall show that this is the case; moreover, we shall also show that the equivalence relation between the UTRDA(R1) and the ULDA/GSVD can be established in the sense that they share the same optimal transformation.

**Theorem 4.2.** Let \( S_m, S_b \) and \( S_l \) be defined by (1)–(3), respectively, and let the matrix \( X \) be given in (14). Then for any \( \mu > 0 \),

\[
G' := X^T G = \begin{bmatrix} Y_1 & 0 \\ Y_2 \end{bmatrix}, \quad Y_1 \in \mathbb{R}^{t \times t}, \quad Y_2 \in \mathbb{R}^{(N-t) \times t},
\]

(25)

solves the UTRDA(R1) (12).

**Proof.** Define \( Y := X^T G := \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \) where the matrix \( X \) is generated from Algorithm 1 and is defined by (14) and \( t = \text{rank}(S_b) \). From the discussion in Section 2, it follows that \( X \) simultaneously diagonalizes \( S_m, S_b \) and \( S_l \) and therefore,

\[
G' = Y_1^T Y_1, \quad G' S_b G = Y_1^T \Sigma_b Y_1, \quad \text{and} \quad G' S_m G = Y_1^T \Sigma_m Y_1,
\]

(26)

where \( \Sigma_b \) and \( \Sigma_m \) are given in (15). Thus, the UTRDA(R1) is equivalent to

\[
\max_{Y_2^T Y_2 = 1} \frac{\text{tr}(Y_1^T \Sigma_b Y_1)}{\text{tr}(Y_1^T (\Sigma_b + \mu I) Y_1)}.
\]

(27)

It is clear then that the global solution of (27) is independent of \( Y_2 \). Moreover, Theorem 3.1 claims that an arbitrary global maximizer \( Y_1 \) to (27) is an orthonormal eigenbasis corresponding to the largest eigenvalues of the diagonal matrix

\[
\Sigma_b - \phi(Y_1) \Sigma_m + \mu I,
\]

whose diagonal elements are in the non-increasing order. Therefore, \( Y_1 = [I_{t \times t}, 0] \) is obviously a global maximizer of (27). The numerator \( \text{tr}(Y_1^T \Sigma_b Y_1) \) further implies that the optimal \( l \) is \( \text{rank}(\Sigma_b) = \text{rank}(S_b) \), and consequently, by choosing a particular \( Y_2 = 0 \in \mathbb{R}^{(N-t) \times t} \), we complete the proof. \( \square \)

Based on Theorem 4.2, we claim again that in such a regularization manner and under the \( S_b \)-orthogonal constraint, the ratio trace approximation model ULDA/GSVD has no difference with the UTRDA(R1).

4.3. The second type regularization uncorrelated trace ratio LDA–UTRDA(R2)

As has been pointed out, the optimal transformation of the UTRDA(R1) is independent of the regularization parameter \( \mu > 0 \). This is resulted from the fact that the regularization term \( \mu_l \) is added to the reduced within-class scatter matrix for any transformation \( G \). As a natural generalization, our second regularized UTRDA defined by (13) focuses directly on the original data by adding a regularization term \( \mu_l \) to the within-class scatter matrix \( S_m \). Different from the UTRDA(R1), we will show that the optimal transformation from the UTRDA(R2) is closely related with the regularization parameter, and hence the classification performance can be improved by choosing an appropriate \( \mu > 0 \). Experimental results show that the UTRDA(R2) performs better than the UTRDA(R1).

However, solving directly the optimization problem (13) is impractical since large scaled matrix computations involve. For an efficient algorithm, we first give the characterization of the global maximizer of (13).

**Theorem 4.3.** Let \( S_m, S_b \) and \( S_l \) be defined by (1)–(3), respectively, and let the matrix \( X \) be given in (14). Then for any \( \mu > 0 \), any global maximizer of the optimization problem (13) is of the form

\[
G'_w := X [Y_1 \quad 0],
\]

(28)

where \( Y_1 \in \mathbb{R}^{t \times t} \) is a global solution of

\[
\max_{Y_2^T Y_2 = 1} \frac{\text{tr}(Y_1^T \Sigma_b Y_1)}{\text{tr}(Y_1^T (\Sigma_w + \mu P \Sigma_l^2 P^T) Y_1)}.
\]

(29)

where \( l = \text{rank}(S_b) \), and the matrices \( \Sigma_b, \Sigma_m, \Sigma_l \) and \( P \) are generated from Algorithm 1.

**Proof.** Using the same notations in Section 4.2, we let \( Y = X^T G \) be defined by (25) and hence (26) follows. Furthermore, one has

\[
\text{tr}(G' (\Sigma_w + \mu I) G) = \text{tr}(Y_1^T \Sigma_m Y_1) + \mu \text{tr}(G' G),
\]

and

\[
\text{tr}(G' G) = \text{tr}(Y_1^T \Sigma_l^2 P^T PY_1 + \mu \text{tr}(\Sigma_l^2 P^T PY_1)) \geq \text{tr}(Y_1^T \Sigma_l^2 P^T PY_1).
\]

where the equality holds if and only if \( Y_2 = 0 \). From

\[
\frac{\text{tr}(G' \Sigma_b)}{\text{tr}(G' (\Sigma_w + \mu I) G)} = \frac{\text{tr}(Y_1^T \Sigma_b Y_1) + \mu \text{tr}(\Sigma_l^2 P^T PY_1)}{\text{tr}(Y_1^T \Sigma_m Y_1) + \mu \text{tr}(\Sigma_l^2 P^T PY_1)} \leq \frac{\text{tr}(Y_1^T \Sigma_b Y_1) + \mu \text{tr}(\Sigma_l^2 P^T PY_1)}{\text{tr}(Y_1^T \Sigma_b Y_1)}
\]

we conclude that if \( G' \) is a global maximizer, it must follow that

\[
Y' := X^T G' = [Y_1 \quad 0],
\]

(28)

moreover, from the numerator \( \text{tr}(Y_1^T \Sigma_b Y_1) \) again, we know that the optimal \( l \) is \( \text{rank}(S_b) = \text{rank}(S_w) \), and \( Y_1 \) is a global maximizer of (29). \( \square \)

Comparing Theorem 4.3 with Theorem 4.2, we know that (a) the global solution of the UTRDA(R2) \( G'_w (28) \) approaches the global solution \( G' \) in (24) as \( \mu \to 0 \), and (b) for a given \( \mu > 0 \), the second part \( Y_2 \) of \( Y' \) in the UTRDA(R1) can be arbitrary, which may lead to uncertainty in subsequent classification and cluster processing, while \( Y_2 \) in \( Y' \) of the UTRDA(R2) is enforced to be exactly zero and hence removes uncertainty for the subsequent processing.

The characterization of the global solution of (13) naturally leads to an efficient two-stage algorithm for the UTRDA(R2), in which the first stage employs Algorithm 1 to generate the related...
matrices in (29), while the second stage uses Algorithm 3 to solve the reduced optimization problem (29) in the low-dimensional space. The pseudo-code of the algorithm can then be stated in Algorithm 4. It is clear, from Algorithm 4, that the first stage involves matrix computations of high dimensionality but independent of μ, while the second stage involves computations related with low dimensional matrices. For a given training set, the results from the first stage can be repeatedly used for different μ’s in the candidate set, and hence only the second stage requires repeated computations. This reduces the computational cost of the cross-validation in choosing the optimal μ.

Algorithm 4. The algorithm for the UTRDA(R2)

Input: data matrix A and the regularization parameter μ > 0
Output: transformation matrix G
1. Form matrices H0 and H1 as in (5) and (6), respectively;
2. Compute the reduced SVD of H0 as H0 = U1Σ1V0T: t = rank(Σ1);
3. Compute the SVD of B := Σ1U1H0 as B = PΣQ; q = rank(B);
4. Set Σq = Σq and Σw = Iq − Σq;
5. Solve the optimization problem (29) for Y1 with l = q by Algorithm 3;
6. Gq = U1Σq−1PYqT.

We conclude this section by briefly discussing the computational complexity of Algorithm 4. Comparing with Algorithm 1, we notice that Algorithm 4 has only additional computational complexity in Step 5, which from Algorithm 3 requires at most O(t3), and I is the iteration number for convergence. Fortunately, we have known that Algorithm 3 enjoys the quadratic convergence and in the practical computation, it only needs less than 10 iterations (Wang et al., 2007; Zhang et al., 2010); moreover, as the optimal solution Y1 of (29) is an orthonormal eigenbasis of the matrix

\[ \Sigma_0 - \varphi^2 \left( \Sigma_w + \mu P^T \Sigma_1^{-2}P \right) = (\Sigma_0 - \varphi^2 \Sigma_w) - \varphi^2 \mu P^T \Sigma_1^{-2}P. \]

where the diagonal matrix \( \Sigma_0 - \varphi^2 \Sigma_w \) has non-increasing diagonal elements, numerical testing in Section 5 indicates that Algorithm 3 needs about 5 iterations in meeting the stopping criterion \( \varepsilon = 10^{-7} \). The additional computation in the Algorithm 4 therefore does not increase the essential computational cost for the undersampled problem. Table 1 summarizes the complexity of each step in Algorithm 4.

5. Experiments

In this section, the ratio trace and the trace ratio criteria under the orthogonal and S-orthogonal constraints are evaluated on 13 public data sets from the categories of face image, microarray, and the text document data. The classification performance of the four discussed LDA methods: the ULDA/GSVD, the OLDA, the RGFST and the UTRDA(R2), are compared.

---

1 http://www-users.cs.umn.edu/~jeeping/ULDA/
2 http://www-users.cs.umn.edu/~jeeping/OLDA/

---

### Table 1

<table>
<thead>
<tr>
<th>Step no.</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Step 4</th>
<th>Step 5</th>
<th>Step 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order of flops</td>
<td>( O(N) )</td>
<td>( O(N^2) )</td>
<td>( O(Nc) )</td>
<td>( O(t) )</td>
<td>( I )</td>
<td>( O(Nc) )</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Data set</th>
<th>Dimension (N)</th>
<th>Number of samples</th>
<th>Number of classes (c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORL</td>
<td>10,304</td>
<td>400</td>
<td>40</td>
</tr>
<tr>
<td>Yale</td>
<td>16,384</td>
<td>165</td>
<td>15</td>
</tr>
<tr>
<td>Yale B</td>
<td>10,304</td>
<td>450</td>
<td>10</td>
</tr>
<tr>
<td>Gene</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leukemia</td>
<td>7129</td>
<td>72</td>
<td>2</td>
</tr>
<tr>
<td>Colon</td>
<td>2000</td>
<td>62</td>
<td>2</td>
</tr>
<tr>
<td>Srct</td>
<td>2308</td>
<td>63</td>
<td>4</td>
</tr>
<tr>
<td>Brain</td>
<td>5597</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>Text</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text A2</td>
<td>1145</td>
<td>200</td>
<td>2</td>
</tr>
<tr>
<td>Text B2</td>
<td>1067</td>
<td>200</td>
<td>2</td>
</tr>
<tr>
<td>Text A4</td>
<td>1795</td>
<td>400</td>
<td>4</td>
</tr>
<tr>
<td>Text B4</td>
<td>1323</td>
<td>398</td>
<td>4</td>
</tr>
<tr>
<td>Text A4-U</td>
<td>708</td>
<td>300</td>
<td>4</td>
</tr>
<tr>
<td>Text B4-U</td>
<td>1324</td>
<td>298</td>
<td>4</td>
</tr>
</tbody>
</table>

5.1. Data sets

The statistics of the 13 tested data sets are summarized in Table 2.

The ORL data set consists of 400 face images taken from 40 distinct subjects. Each person has 10 images taken at different times, varying the lighting, facial expressions and facial details (glasses/no glasses). The Yale data set contains totally 165 images from 15 individuals. Each person has 11 images with different facial expression or configuration. The third image data set in Table 2, Yale B, is from the Yale Face Database B (Georgiades et al., 2001), containing 405 viewing conditions (9 poses × 45 illumination conditions) for 10 individuals. Yale B in Table 2 contains 450 images which are 10 persons’ frontal pose images with 45 illumination conditions per person.

Four microarray data sets are colon cancer data (Alon et al., 1999), Leukemia MIT AML/ALL data (Golub et al., 1999), Srct data (Khan et al., 2001) and Brain data (Pomeroy et al., 2002). The colon cancer data set consists of 62 subject samples each of which has 2000 gene expression values. Among the data, 40 are cancer samples while the rest of them are normal samples. MIT AML/ALL data set contains 7129 genes expression of 72 samples, including 47 AML samples and 25 ALL samples. The original Srct and the Brain data appear first in (Khan et al., 2001; Pomeroy et al., 2002), respectively, and the data sets in Table 2 are the pre-processing versions based on Dettling (2004). The former contains 4 classes while the latter consists of 5 classes, representing different tumor types.

The six text document data sets are the publicly available 20-Newsroups data. The original text data was first preprocessed by using the BOW toolkit (McCallum, 1998) to strip the news messages from the e-mail headers and special tags, and eliminate the stopwords and stem words to their root forms. The words then were sorted on the inverse document frequency (IDF) and some words were removed if the IDF values were too small or too large. Each data set in Table 2 contains 2 or 4 categories. Different data sets have different class structures. For example, data sets A2 and A4 contain categories which are semantically different, while data set B2 and B4 contain semantically close categories. Each category represents an individual class and is described by a subset of words (features). The data sets A4-U and B4-U contain unbalanced documents in each category.

---

3 http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
4 http://stat.ethz.ch/~dettling/hagboost.html
5 http://people.csail.mit.edu/jrennie/20Newsgroups/.
5.2. Experimental results

For each of data set described in Table 2, we firstly randomly split the samples into the training and testing parts of ratio 2:1. The K-Nearest-Neighbor (KNN) procedure (refer to e.g., Howland et al., 2003; Theodoridis and Koutroumbas, 1999) is employed for computing the classification accuracy. To compare the performance of each LDA approach, we average the resulted classification accuracy over 10 random partitions with different $K = 1, 3, 5, 7$.

Note that both the RGFST and the UTRDA(R2) contain the regularization parameter which will be chosen by the 5-fold cross-validation (see e.g., Friedman, 1989; Ye et al., 2006a) in the following way: First, we choose $C = \{10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}\}$ as the candidate set for $l$. Then for each random partition of the data set in Table 2, the part for training is randomly divided into 5 subsets of (approximately) equal size. All subsets are mutually exclusive, and in the $i$th fold, the $i$th subset is held out for testing while the rest subsets are used for training. For every $l \in C$, we compute the cross validation accuracy, namely $Accu(l)$, which is...
Table 6
Comparison of classification accuracies and the standard deviation (in parenthesis) using 7-NN.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Trace ratio criterion</th>
<th>Ratio trace criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UTRDA(R2)</td>
<td>RGFS</td>
</tr>
<tr>
<td>ORL</td>
<td>93.9167(1.6960)</td>
<td>93.0833(8.9326)</td>
</tr>
<tr>
<td>Yale</td>
<td>87.1111(3.5985)</td>
<td>88.6667(3.6963)</td>
</tr>
<tr>
<td>Yale B</td>
<td>99.0000(0.4019)</td>
<td>99.1333(0.4049)</td>
</tr>
<tr>
<td>Leukemia</td>
<td>95.4166(3.6483)</td>
<td>95.4166(3.6483)</td>
</tr>
<tr>
<td>Colon</td>
<td>83.3333(7.9206)</td>
<td>83.3333(7.9206)</td>
</tr>
<tr>
<td>Srbct</td>
<td>97.3884(4.4728)</td>
<td>98.9474(2.2192)</td>
</tr>
<tr>
<td>Brain</td>
<td>92.6667(4.9191)</td>
<td>90.6667(4.6613)</td>
</tr>
<tr>
<td>Text A2</td>
<td>94.2647(2.4214)</td>
<td>94.2647(2.4214)</td>
</tr>
<tr>
<td>Text B2</td>
<td>90.0000(3.5136)</td>
<td>90.0000(3.5136)</td>
</tr>
<tr>
<td>Text A4</td>
<td>90.5147(1.7139)</td>
<td>90.8824(1.9972)</td>
</tr>
<tr>
<td>Text B4</td>
<td>90.9999(3.1269)</td>
<td>90.0000(3.1269)</td>
</tr>
<tr>
<td>Text B4-U</td>
<td>95.9184(1.6663)</td>
<td>98.5714(1.1977)</td>
</tr>
</tbody>
</table>

Table 7
Average iteration numbers of Algorithm 3 and average values of the regularization parameter \( \mu \) for the RGFS and UTRDA(R2) over 10 random testings.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Average iter. #</th>
<th>Average value of ( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UTRDA(R2)</td>
<td>RGFS</td>
</tr>
<tr>
<td>ORL</td>
<td>3.0</td>
<td>6.8</td>
</tr>
<tr>
<td>Yale</td>
<td>2.8</td>
<td>8.8</td>
</tr>
<tr>
<td>Yale B</td>
<td>2.4</td>
<td>9.9</td>
</tr>
<tr>
<td>Leukemia</td>
<td>2.0</td>
<td>8.9</td>
</tr>
<tr>
<td>Colon</td>
<td>2.0</td>
<td>10.7</td>
</tr>
<tr>
<td>Srbct</td>
<td>2.6</td>
<td>6.6</td>
</tr>
<tr>
<td>Brain</td>
<td>2.6</td>
<td>5.0</td>
</tr>
<tr>
<td>Text A2</td>
<td>5.1</td>
<td>4.0</td>
</tr>
<tr>
<td>Text B2</td>
<td>4.3</td>
<td>5.5</td>
</tr>
<tr>
<td>Text A4</td>
<td>5.2</td>
<td>4.1</td>
</tr>
<tr>
<td>Text B4</td>
<td>5.9</td>
<td>4.7</td>
</tr>
<tr>
<td>Text A4-U</td>
<td>3.9</td>
<td>3.3</td>
</tr>
<tr>
<td>Text B4-U</td>
<td>5.7</td>
<td>4.1</td>
</tr>
</tbody>
</table>

The experimental results in these tables lead us to the following main observations:

- For the image and gene expression data sets, the classification performance (both the classification accuracy and the standard deviation) of the LDA approaches (the ULDA/GSVD and the OLDA) is better than the results based on ratio trace criteria in terms of the classification performance.

6. Conclusions
In this paper, we discussed in details the trace ratio criterion under the \( S \)-orthogonal constraint, which is able to generate the uncorrelated features in the reduced space. To deal with the undersampled problem, by employing the regularization technique, two regularized uncorrelated trace ratio discriminant analysis are proposed, one of which is shown to be equivalent to the regularized ratio trace discriminant analysis, the ULDA/GSVD. The global solution of the other, the UTRDA(R2), is characterized and an efficient algorithm is proposed based on its equivalent but reduced form. Experimental testings of four LDA approaches (the ULDA/GSVD, the UTRDA(R2), the RGFST and the UTRDA(R2)) are conducted on 13 public data sets and show that the uncorrelated trace ratio LDA is competitive with the orthogonal trace ratio LDA, but is better than the results based on ratio trace criteria in terms of the classification performance.

References


