THE TRACE RATIO OPTIMIZATION PROBLEM

T. T. NGO*, M. BELLALIJ†, AND Y. SAAD*

Abstract. This paper considers the problem of optimizing the ratio $\mathrm{Tr}\,[W^TAW]/\mathrm{Tr}\,[W^TBW]$ over all unitary matrices W with p columns, where A, B are two positive definite matrices. This problem is common in supervised learning techniques. However, because its numerical solution is typically expensive it is often replaced by the simpler optimization problem which consists of optimizing $\mathrm{Tr}\,[W^TAW]$ under the constraint that $W^TBW = I$, the identity matrix. The goal of this paper is to examine this trace ratio optimization problem in detail, to consider different algorithms for solving it, and to illustrate the use of these algorithms for face recognition problems.

Key words. Trace optimization, Classification, Linear Dimension Reduction, Face recognition.

1. Introduction.

A number of techniques in machine learning are based on optimizing a trace of the form $\mathrm{Tr}\,[V^TAV]$ under certain constraints on V. This defines a projector with the basis V, which is then used for various dimension reduction tasks.

A particular case of this scenario is the well-known Fisher Linear Discriminant Analysis (LDA). The method which is a prototypical approach of supervised learning, defines a linear hyperplane which best separates two or more data-sets. This is achieved by trying to maximize the ratio of two traces. The first of these (numerator) represents the in-between scatter which measures how well the classes are separated in the projected space. The second (denominator) represents the within scatter which measures how well clustered each class is in the projected space.

When the desired dimension of the projected space is two or more, the problem is then to maximize a ratio of the form,

$$\frac{\mathrm{Tr}\,\left[V^TAV\right]}{\mathrm{Tr}\,\left[V^TBV\right]}, \tag{1.1}$$

where $V \in \mathbb{R}^{n\times p}$ is subjected to having orthonormal columns. This problem is seldom solved in practice. It has been considered too difficult to solve and is commonly replaced by the simpler, but not equivalent problem :

$$\max_{\left\{\begin{array}{l} V \in \mathbb{R}^{n\times p} \\ V^TBV = I \end{array}\right.} \mathrm{Tr}\,\left[V^TAV\right] \tag{1.2}$$

Yet, recent publications do indicate that methods based on optimizing the trace ratio (1.1) yield better results in general than those based on their simplified analogues. As a result, several papers have recently addressed the problem of how to optimize this ratio. Though the problem has been judged difficult, it was observed that the results may warrant the extra cost. In fact as will be shown in this paper, the cost of solving this problem need not be high.

The goal of this paper is to explore this problem from a few different avenues. From a practical point of view we will show that maximizing the ratio (1.1) can be done quite efficiently. The computational cost can be drastically reduced by exploiting

[*] University of Minnesota, Dept. of Computer Science and Engineering `E-mail: {saad, thango}@cs.umn.edu`. Work supported in part by NSF under grants DMS-0810938, NSF –DMR 0940218, and by the Minnesota Supercomputer Institute.

[†] Université Lille Nord de France, F-59000 Lille, France ; UVHC, LAMAV, F-59313 Valenciennes, France. `E-mail: Mohammed.Bellalij@univ-valenciennes.fr`.

a combination of techniques: Newton iteration, Lanczos procedure, etc. In fact, we will argue that contrary to widespread belief, *this problem is in fact less costly to solve than that based on the standard constrained trace optimization of equation (1.2).* In a nuthshell this is because optimizing (1.1) will require solving a few *standard* eigenvalue problems, while (1.2) leads to a *generalized* eigenvalue problem. We will argue that with the help of the Lanczos procedure, the former will in fact be typically less expensive than the second.

Throughout the paper, $\mathcal{U}_p$ denotes the set of unitary matrices (matrices with orthonormal columns) with p columns, (i.e., of size $n \times p$). The identity matrix will be denoted by I.

2. Preliminaries. Given a symmetric matrix A, of dimension $n \times n$ and an arbitrary unitary matrix V of dimension $n \times p$ it is known that the trace of V^TAV reaches its maximum (resp. minimum) when V is an orthogonal basis of the eigenspace of A associated with the p algebraically largest (resp. smallest) eigenvalues. In particular, it is achieved for the eigenbasis itself: if eigenvalues are labeled decreasingly and $u_1, \cdots, u_p$ are eigenvectors associated with the first p eigenvalues $\lambda_1, \cdots, \lambda_p$, and $U = [u_1, \cdots, u_p]$, with $U^TU = I$, then,

$$\max_{\begin{cases} V \in \mathbb{R}^{n\times p} \\ V^TV = I \end{cases}} \text{Tr}\left[V^TAV\right] = \text{Tr}\left[U^TAU\right] = \lambda_1 + \cdots + \lambda_p. \tag{2.1}$$

This result is seldom explicitly stated on its own in standard textbooks, but it is an immediate consequence of the Courant-Fisher characterization, see, e.g., [10, 12]. The optimal V is not unique since any system V that is an orthonormal basis of the eigenspace associated with the first p eigenvalues will be optimal. In other words, it is the subspace that matters rather than any specific particular orthonormal basis for the subspace.

Maximizing the trace in (2.1), requires the solution of a standard eigenvalue problem. Sometimes it is necessary to maximize $\text{Tr}\,[V^TAV]$ subject to a new normalization constraint for V, one that requires that V be B-orthogonal, i.e., $V^TBV = I$. Assuming that A is symmetric and B positive definite, we know that there are n real eigenvalues for the generalized problem $Au = \lambda Bu$, with B-orthogonal eigenvectors. If these eigenvalues are labeled decreasingly, and if $U = [u_1, \cdots, u_p]$ is the set of eigenvectors associated with the first p eigenvalues, with $U^TBU = I$, then we have

$$\max_{\begin{cases} V \in \mathbb{R}^{n\times p} \\ V^TBV = I \end{cases}} \text{Tr}\left[V^TAV\right] = \text{Tr}\left[U^TAU\right] = \lambda_1 + \cdots + \lambda_p. \tag{2.2}$$

In reality, Problem (2.2) often arises as a simplification of an objective function that is more difficult to maximize, namely:

$$\max_{\begin{cases} V \in \mathbb{R}^{n\times p} \\ V^TCV = I \end{cases}} \frac{\text{Tr}\left[V^TAV\right]}{\text{Tr}\left[V^TBV\right]} \,. \tag{2.3}$$

Here B and C are assumed to be symmetric and positive definite for simplicity. The matrix C defines the desired orthogonality and in the simplest case it is just the identity matrix. The original version shown above has resurfaced in recent years, see, e.g., [5, 15, 9, 14, 16, 17, 14, 13] among others. One of the main reasons for the regained interest in this problem, is that it seems to yield markedly improved results for supervised learning tasks than its simplified counterpart (2.2).

3. Existence and uniqueness of a solution. There is no loss of generality in assuming that C is the identity matrix. Problem (2.3) may not have a solution when B is not positive definite. This is because in this situation it will be possible to find subspaces for which $\mathrm{Tr}\,[V^TBV]$ is zero while $\mathrm{Tr}\,[V^TAV]$ is nonzero, making the maximum ratio (2.3) infinite. A simple example is

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad V = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

It is helpful to examine the trace $\mathrm{Tr}\,[V^TBV]$ in detail. Let $B = Q\Lambda_BQ^T$ the diagonalization of B, where Q is unitary and $\Lambda_B = diag(\mu_1, \mu_2, \cdots, \mu_n)$. Let $v_1, \ldots, v_n$ be the columns of V, and define $\tilde{v}_j = Qv_j$. Then clearly

$$\mathrm{Tr}\,[V^TBV] = \sum_{j=1}^{p}\sum_{i=1}^{n} \mu_i \tilde{v}_{ij}^2 = \sum_{i=1}^{n} \mu_i \sum_{j=1}^{p} \tilde{v}_{ij}^2. \tag{3.1}$$

The following lemma examines under which conditions $\mathrm{Tr}\,[V^TBV]$ is nonzero in the situation when B is positive semi-definite.

LEMMA 3.1. *Assume that B is positive semi-definite and let p be the number of columns of V. If B has at most $p-1$ zero eigenvalues then $Tr[V^TBV]$ is nonzero for any unitary V.*

Proof. Using the previous notation $\tilde{V} = [\tilde{v}_1, \cdots, \tilde{v}_p]$. has at least one $p \times p$ submatrix which is nonsingular, so it has at least p rows that have a nonzero norm. Then in the sum (3.1) at least one of the $n-p+1$ nonzero eigenvalues μ_i will coincide with one of these row norms and this sum will be nonzero. □

Therefore, the problem is well-posed under the condition that the null space of B is of dimension less than p, i.e., that its rank be at least $n - p + 1$. In this case the maximum is finite.

Another situation that leads to difficulties is when the two traces have a zero value for the same V. This situation should be excluded from consideration as it leads to an indefinite ratio of 0/0. For this we must assume that $Null(A) \cap Null(B) = \{0\}$.

PROPOSITION 3.2. *Let A, B be two symmetric matrices and assume that B is semi-positive definite with rank $> n - p$ and that $Null(A) \cap Null(B) = \{0\}$. Then the ratio (2.3) admits a finite maximum (resp. minimum) value ρ_*. The maximum is reached for a certain V that is unique up to unitary transforms of the columns.*

Proof. The set of matrices V such that $V^TV = I$ is closed and, under the assumptions, the ratio trace function in the right-hand side of (2.3) is continuous function of its argument. Therefore, using Lemma 3.1 the maximum of the trace ratio (2.3) is reached. □

4. Conversion to a scalar problem. In the remainder of the paper we will assume that C is the identity and that B satisfies the conditions of Proposition 3.2. From Proposition 3.2 we know that there is a maximum ρ_* that is reached for a certain (non-unique) orthogonal matrix, which we will denote by U_*. Then, for any orthogonal V we have $\mathrm{Tr}\,[V^TAV]/\mathrm{Tr}\,[V^TBV] \leq \rho_*$ and hence,

$$\mathrm{Tr}\,[V^TAV] - \rho_*\ \mathrm{Tr}\,[V^TBV] \leq 0.$$

This means that for this ρ_* we have $\text{Tr}\,[V^T(A-\rho_* B)V] \leq 0$ for any orthogonal V, and also $\text{Tr}\,[U_*^T(A-\rho_* B)U_*] = 0$. Therefore, we have the following necessary condition for the pair ρ_*, U_* to be optimal:

$$\max_{V^TV=I} \text{Tr}\,[V^T(A-\rho_* B)V] = \text{Tr}\,[U_*^T(A-\rho_* B)U_*] = 0. \tag{4.1}$$

According to (2.1), the maximum trace of $V^T(A-\rho_* B)V$ over all unitary matrices V of size $n \times p$, is simply the sum of the largest p eigenvalues of $A-\rho_* B$ and U_* is the set of corresponding eigenvectors. If ρ_* maximizes the trace ratio (2.3) (with $C = I$), then the sum of the largest p eigenvalues of the pencil $A-\rho_* B$ must be equal to zero, and the corresponding eigenvectors form the desired optimal solution of (2.3).

Consider now the function

$$f(\rho) = \max_{V^TV=I} \text{Tr}\,[V^T(A-\rho B)V] \ . \tag{4.2}$$

Note that the matrices V that reach the above maximum are not unique: any orthogonal transformation of the columns of V will not change the trace. We can select the optimal V to be a set of eigenvectors of the matrix $A-\rho B$. We will denote by $V(\rho)$ a set of the p eigenvectors which reach the above maximum and by $G(\rho)$ the matrix:

$$G(\rho) \equiv A - \rho B \ , \tag{4.3}$$

whose n eigenvalues labeled decreasingly are:

$$\mu_1(\rho) \geq \mu_2(\rho) \geq \cdots \geq \mu_n(\rho) \ . \tag{4.4}$$

With this notation, it is clear that

$$f(\rho) = \mu_1(\rho) + \mu_2(\rho) + \cdots + \mu_p(\rho) \ . \tag{4.5}$$

Another useful expression for $f(\rho)$ which will be exploited later is one that is based on the eigenprojector. Indeed if we set $P(\rho) = V(\rho)V(\rho)^T$ then clearly

$$f(\rho) = \text{Tr}\,[V(\rho)^T G(\rho) V(\rho)] = \text{Tr}\,[G(\rho)V(\rho)V(\rho)^T] = \text{Tr}\,[G(\rho)P(\rho)]. \tag{4.6}$$

It is also possible to exploit the Dunford integral for expressing $P(\rho)$:

$$P(\rho) = \frac{-1}{2\pi i}\int_\Gamma (G(\rho) - zI)^{-1}\, dz$$

where Γ is a Jordan curve containing the p eivenvalues of interest. We will denote by $R_\rho(z)$ the resolvant

$$R_\rho(z) = (G(\rho) - zI)^{-1} = (A - \rho B - zI)^{-1}. \tag{4.7}$$

From this we obtain the following expression for $f(\rho)$:

$$\begin{aligned} f(\rho) &= \frac{-1}{2\pi i}\text{Tr} \int_\Gamma G(\rho)(G(\rho)-zI)^{-1}\, dz \qquad (4.8) \\ &= \frac{-1}{2\pi i}\text{Tr} \int_\Gamma (G(\rho) - zI + zI)\ (G(\rho)-zI)^{-1}\, dz \\ &= \frac{-1}{2\pi i}\text{Tr} \int_\Gamma z(G(\rho)-zI)^{-1}\, dz \qquad (4.9) \end{aligned}$$

The following properties of f can now be proved.

LEMMA 4.1.

1. f is a non-increasing function of ρ;

2. $f(\rho) = 0$ iff $\rho = \rho_*$.

Proof. To prove (1) we need to compare the sums of the p largest eigenvalues of $A - \rho_2 B$ and $A - \rho_1 B$ for $\rho_2 \geq \rho_1$. We have

$$G(\rho_2) - G(\rho_1) = -(\rho_2 - \rho_1)B.$$

Since B is positive semi-definite, classical monotonicity results show that the p largest eigenvalues of $G(\rho_2)$ will not exceed those of $G(\rho_1)$.

To prove (2), we start by observing that the sufficient condition is trivial, i.e., according to (4.1), $\rho = \rho_*$ implies $f(\rho) = 0$. Next, since $\mathrm{Tr}\,[V^T BV] > 0$ for any $V \in \mathcal{U}_p$ we can write

$$\mathrm{Tr}\,[V^T AV - \rho V^T BV] \leq 0,\ \forall\ V \in\ \mathcal{U}_p \quad \text{iff} \quad \frac{\mathrm{Tr}\,[V^T AV]}{\mathrm{Tr}\,[V^T BV]} \leq \rho,\ \forall\ V \in\ \mathcal{U}_p\ .$$

This can be restated as

$$f(\rho) \leq 0 \quad \text{iff} \quad \rho_* \leq \rho\ . \tag{4.10}$$

Suppose now that $f(\rho) > 0$ for a certain ρ. Then, there is a V_0 such that

$$\mathrm{Tr}\,[V_0^T AV_0 - \rho V_0^T BV_0] > 0\ \rightarrow\ \frac{\mathrm{Tr}\,[V_0^T AV_0]}{\mathrm{Tr}\,[V_0^T BV_0]} > \rho.$$

This means that

$$\max_{V \in \mathcal{U}_p} \frac{\mathrm{Tr}\,[V^T AV]}{\mathrm{Tr}\,[V^T BV]} > \rho,$$

and therefore $\rho_* > \rho$. This can be restated as

$$f(\rho) > 0 \quad \rightarrow \quad \rho_* > \rho\ . \tag{4.11}$$

Equations (4.10) and (4.11) together, along with the continuity of f, show that $f(\rho) = 0$ implies $\rho = \rho_*$. This completes the proof. □

It is to be noted that the function f is actually strictly decreasing as will be shown later. This will provide another way to prove the second part of the proposition.

4.1. Localization of the optimum. We now know that the optimal trace ratio can be found as the root of a decreasing function $f(\rho)$. One may ask if it is possible to find an interval where the root lies. When A is positive definite, then $f(\rho) \geq 0$ for $\rho = 0$, since $G(0) = A$. For $\rho > \lambda_1(A, B)$ we have $f(\rho) < 0$, where $\lambda_1(A, B)$ is the largest generalized eigenvalue of the pencil (A, B). Therefore, the root belongs to the interval $[0,\ \lambda_1(A, B)]$.

A more refined location interval for the root may be found by exploiting Sylvester's inertia theorem. For simplicity we assume that B is positive definite. Let Z be the matrix which diagonalizes the pencil A, B:

$$Z^T AZ = \Lambda, \quad Z^T BZ = I\ .$$

Here the diagonal entries $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ of Λ are the generalized eigenvalues of the pencil A, B. Then,

$$Z^T[A - \rho B]Z = \Lambda - \rho I.$$

According to the Sylvester inertia theorem, the number of negative and positive eigenvalues for the matrices $G(\rho)$ and those of $\Lambda - \rho I$ are the same. Thus for $\rho = \lambda_p$, the first (largest) p eigenvalues of $G(\rho)$ will be nonnegative and so their sum $f(\rho)$ is nonnegative. On the other side of the spectrum, for $\rho = \lambda_1$ all eigenvalues of $G(\rho)$ will be negative and so $f(\rho) \leq 0$. We have just proved the following proposition.

Proposition 4.2. *The root ρ_* of $f(\rho)$ is located in the interval $[\lambda_p,\ \lambda_1]$, where λ_i is the i-th largest eigenvalue of the pair (A, B).*

An alternative to the above bound uses eigenvalues of A and B instead of those of the generalized eigenvalue problem.

Proposition 4.3. *Assume that B is positive definite. Then the root ρ_* of $f(\rho)$ is such that*

$$\frac{\sum_{i=1}^{p} \lambda_i(A)}{\sum_{i=1}^{p} \lambda_i(B)} \leq \rho_* \leq \frac{\sum_{i=1}^{p} \lambda_i(A)}{\sum_{i=1}^{p} \lambda_{n-i+1}(B)}, \tag{4.12}$$

where $\lambda_i(A)$, and $\lambda_i(B)$ are the i-th largest eigenvalues of the matrices A and B respectively.

Proof. Let U be the unitary matrix whose column are the eigenvectors of A associated with $\lambda_1(A), \cdots, \lambda_p(A)$. Then clearly $\mathrm{Tr}\,[U^T AU] = \lambda_1(A) + \cdots + \lambda_p(A)$, $\mathrm{Tr}\,[U^T BU] \leq \lambda_1(B) + \cdots + \lambda_p(B)$, so

$$\frac{\lambda_1(A) + \cdots + \lambda_p(A)}{\lambda_1(B) + \cdots + \lambda_p(B)} \leq \frac{\mathrm{Tr}\,[U^T AU]}{\mathrm{Tr}\,[U^T BU]} \leq \max_{V^T V = I} \frac{\mathrm{Tr}\,[V^T AV]}{\mathrm{Tr}\,[V^T BV]} = \rho_*.$$

For the right-hand side inequality, we exploit the fact that for any unitary matrix U we have $\mathrm{Tr}\,[U^T AU] \leq \lambda_1(A) + \cdots + \lambda_p(A)$, $\mathrm{Tr}\,[U^T BU] \geq \lambda_n(B) + \lambda_{n-1}(B) + \cdots + \lambda_{n-p+1}(B)$. Hence, for any unitary U,

$$\frac{\mathrm{Tr}\,[U^T AU]}{\mathrm{Tr}\,[U^T BU]} \leq \frac{\lambda_1(A) + \cdots + \lambda_p(A)}{\lambda_n(B) + \lambda_{n-1}(B) + \cdots + \lambda_{n-p+1}(B)},$$

which is therefore an upper bound for ρ_*. □

Finding the optimal solution will involve a search for the (unique) root of $f(\rho)$. In [15] and [5] algorithms were proposed to solve (2.3) by computing this root and by exploiting the above relations. No matter what method is used it appears at the outset that it will be more complicated to solve (2.3) than (2.2), because the search for the root ρ_* may involve solving several eigenvalue problems instead of just one. However, this does not necessarily mean that it will be more costly. The use of Newton's method combined with the Lanczos procedure will alleviate this search. In this regard, an interesting connection to known methods can be established and this is taken up in the next section.

4.2. The derivative of f. To obtain the derivative of the function f, we first assume that the eigenvalues of $G(\rho)$ are all simple. Then the derivative of each individual eigenvalue $\mu_i(\rho)$ with respect to ρ is explicitly known in terms of the associated eigenvector. When ρ is perturbed to $\rho + \delta$, the matrix $G(\rho)$ is perturbed by $-\delta B$.

The corresponding infinitesimal perturbation to the individual eigenvalue $\mu_i(\rho)$ is then given by (see, e.g., [12]):

$$\mu_i(\rho + \delta) - \mu_i(\rho) = -\delta(Bv_i(\rho), v_i(\rho))$$

where $v_i(\rho)$ is a unit eigenvector of $G(\rho)$ associated with $\mu_i(\rho)$, and (x, y) denotes the inner product of the two vectors x and y. As a result the derivative of $\mu_i(\rho)$ is simply

$$\mu_i'(\rho) = -(Bv_i(\rho), v_i(\rho))$$

and this is translated for $f(\rho)$ by

$$f'(\rho) = -\mathrm{Tr}\,[V(\rho)^T BV(\rho)].$$

We now consider the extension of this expression to the general case where there may be multiple eigenvalues. For this, we will consider the differential of $V(\rho)^T(A - \rho B)V(\rho)$ which is the diagonal matrix of eigenvalues. This is doable provided some care is exercised in defining $V(\rho)$. Indeed, we need to define the eigenvectors so the mapping $V(\rho)$ is differentiable.

In what follows the notation is simplified: $V(\rho)$ which is assumed to be a differentiable function of ρ, is denoted simply by V. In addition, we assume that V diagonalizes $A - \rho B$ and that we have $(A - \rho B)V = VD$ where D is a diagonal matrix of size $p \times p$. (note that D is a function of ρ).

First observe that from the equality $V^TV = I$ it follows that

$$0 = \frac{d}{d\rho}[V^TV] = \frac{dV^T}{d\rho}V + V^T\frac{dV}{d\rho} = 0 \quad \rightarrow \quad \mathrm{Diag}\left[V^T\frac{dV}{d\rho}\right] = 0. \tag{4.13}$$

This means that the matrix $V^T\ dV/d\rho$ has a zero diagonal, a property which will be expoited shortly. Next we proceed with the differentiation of $f(\rho)$. First, consider

$$\begin{aligned}
\frac{d}{d\rho}[V^T(A - \rho B)V] &= \frac{d}{d\rho}[V^TAV] - \frac{d}{d\rho}[V^T\rho B)V] \\
&= \frac{dV^T}{d\rho}AV + V^TA\frac{dV}{d\rho} - \frac{dV^T}{d\rho}\rho BV - V^T[BV + \rho B\frac{dV}{d\rho}] \\
&= \frac{dV^T}{d\rho}[A - \rho B]V + V^T[A - \rho B]\frac{dV}{d\rho} - V^TBV \\
&= \frac{dV^T}{d\rho}VD + DV^T\frac{dV}{d\rho} - V^TBV\ .
\end{aligned}$$

Now, taking the trace in the above final expression yields:

$$\begin{aligned}
\frac{df(\rho)}{d\rho} &= \mathrm{Tr}\left[\frac{dV^T}{d\rho}VD + DV^T\frac{dV}{d\rho} - V^TBV\right] \\
&= 2\ \mathrm{Tr}\left[DV^T\frac{dV}{d\rho}\right] - \mathrm{Tr}\,[V^TBV] \\
&= -\mathrm{Tr}\,[V^TBV].
\end{aligned}$$

The last equality comes from the fact that the matrix $V^TdV/d\rho$ has a zero diagonal as was established above, see Eq. (4.13). Therefore, we can state the following result.

Proposition 4.4. *The function $f(\rho)$ admits the derivative $-Tr[V(\rho)^T BV(\rho)]$. In particular, under the assumption that B is positive semi-definite with fewer than p zero eigenvalues, f is a strictly decreasing function.*

Proof. Only the second part remains to be shown which is a consequence of Lemma 3.1. □

There is an alternative for deriving the above result - which is based on the Dunford integral formula. The advantage of this viewpoint is that it bypasses the need to restrict the mapping $V(\rho)$ to being differentiable. This is based on the alternative expression (4.9). Taking the derivative of $f(\rho)$ from the expression yields:

$$\begin{aligned}
f'(\rho) &= \frac{-1}{2\pi i}\mathrm{Tr}\int_\Gamma z\frac{d}{d\rho}R_\rho(z)dz \\
&= \frac{-1}{2\pi i}\mathrm{Tr}\int_\Gamma zR_\rho(z)BR_\rho(z)dz \\
&= \frac{-1}{2\pi i}\int_\Gamma z\mathrm{Tr}\,[R_\rho(z)BR_\rho(z)]dz \\
&= \frac{-1}{2\pi i}\int_\Gamma z\mathrm{Tr}\,[R_\rho(z)^2B]dz \\
&= \frac{-1}{2\pi i}\int_\Gamma \mathrm{Tr}\,\big([(A-\rho B)-(A-\rho B-zI)]\,R_\rho(z)^2B\big)\,dz \\
&= \frac{-1}{2\pi i}\int_\Gamma \mathrm{Tr}\,\big[(A-\rho B)R_\rho(z)^2B - R_\rho(z)B\big]\,dz \\
&= \frac{-1}{2\pi i}\mathrm{Tr}\int_\Gamma (A-\rho B)R_\rho(z)^2Bdz - \frac{-1}{2\pi i}\mathrm{Tr}\int_\Gamma R_\rho(z)Bdz \\
&= 0 - \mathrm{Tr}\,[P(\rho)B]
\end{aligned}$$

The integral in the first term of the above expression is zero because the term $(R_\rho(z))^2$ in the integrand is the exact derivative (with respect to z) of $R_\rho(z)$. The integral in the second bracketed term is just $P(\rho)$. This gives the expression:

$$f'(\rho) = -\mathrm{Tr}\,[P(\rho)B] = -\mathrm{Tr}\,[V(\rho)V(\rho)^TB] = -\mathrm{Tr}\,[V(\rho)^TBV(\rho)]\ .$$

4.3. Practical implementation via Newton's method. From the expression of the differential of f, Newton's method takes the form

$$\rho_{new} = \rho - \frac{\mathrm{Tr}\,[V(\rho)^T(A-\rho B)V(\rho)]}{-\mathrm{Tr}\,[V(\rho)^TBV(\rho)]} = \frac{\mathrm{Tr}\,[V(\rho)^TAV(\rho)]}{\mathrm{Tr}\,[V(\rho)^TBV(\rho)]}$$

Remarkably, Newton's method for finding the zero of f amounts to a form of fixed point iteration. The function on the right side of the above equality is

$$g(\rho) = \frac{\mathrm{Tr}\,[V^T(\rho)AV(\rho)]}{\mathrm{Tr}\,[V^T(\rho)BV(\rho)]},$$

in which $V(\rho)$ was defined above. An approach of this type was proposed in the literature and it was observed that convergence is fast. The reason for this is that it is in essence a Newton method.

It is possible to exploit the Lanczos algorithm to provide a highly effective procedure.

Algorithm 4.1. *Newton-Lanczos algorithm for Trace Ratio maximization*

```
1. Input: $A, B$ and a dimension $p$.
2. Select initial $n \times p$ unitary matrix $V$; compute $\rho = \mathrm{Tr}\,[V^T AV]/\mathrm{Tr}\,[V^T BV]$ .
3. Until convergence Do:
4.     Call the Lanczos algorithm to compute the $p$ largest eigenvalues
5.     of $G(\rho) = A - \rho B$ and associated eigenvectors $[v_1, v_2, \cdots, v_p] \equiv V$
6.     Set $\rho := \frac{\mathrm{Tr}\,[V^T AV]}{\mathrm{Tr}\,[V^T BV]}$
7. EndDo
```

A number of practical refinements can make the above procedure highly effective. The most important of these is based on the observation that variable accuracy techniques can exploited to reduce cost. Initially, when we are away from the solution, there is no need to compute the eigenspace accurately at all. As we get closer to the solution ρ_*, it becomes essential to tighten the accuracy of the eigenvectors in order for the procedure to enjoy a superlinear convergence. The well-known paper [2] discusses the theory and the practical application of these inexact Newton methods.

4.4. Relation to repulsion Laplaceans. Graph-based methods for supervised learning employ a Laplacean graph based on classes: edges (i, j) of the graph are associated with the binary relation "i and j belong to the same class". The Laplacean weights are often defined simply as $W_{ij} = 1$ if i and j are adjacent and $W_{ij} = 0$ otherwise. The graph Laplacean is the matrix defined as $L = D - W$, where D is the diagonal of the row-sums of W. As a result of these definitions $D - W$ is singular and admits the vector of all ones as a null vector.

A dimension reduction technique based on these graphs and called Locality Preserving Projections (LPP) produces a set Y of data from the original set X by minimizing the objective function, see [6]:

$$\Psi(Y) \equiv \mathrm{Tr}\,[YLY^\top] = \frac{1}{2} \sum_{i,j=1}^{n} W_{ij} \|y_i - y_j\|_2^2. \tag{4.14}$$

Intuitively, when two points x_i and x_j are similar, the corresponding weight W_{ij} will be large. Then, minimizing (4.14) will tend to force the distances $\|y_i - y_j\|_2^2$ to be small, i.e., it encourages points y_i and y_j to be placed close by in the low dimensional space. A similar principle was advocated in the Orthogonal Neighborhood Preserving Projections (ONPP) approach [7] by using a similarity graph borrowed from the Locally Linear Embedding approach [11]. In ONPP (as in LLE), a weighted graph is built by writing each data point x_i as best as possible as a convex combination of its k nearest neighbors,

$$x_i \approx \sum_{j \,\in\, N(i)} w_{ij} x_j. \tag{4.15}$$

Once this is done, projected points $y_i = V^T x_i$ are sought in a low dimensional space, so that the relationships (4.15) between the original points are 'optimally' preserved. This means that the following objective function:

$$\Phi_\rho(Y) = \sum_i \|y_i - \sum_j w_{ij} y_j\|_2^2 \quad \text{with} \quad Y = V^T X, \tag{4.16}$$

is minimized with respect to all possible unitary matrices V of size $n \times p$. The minimization of the objective functions (4.14) and (4.16) will yield points that are close by in the low-dimensional space, when they are close-by in the original space.

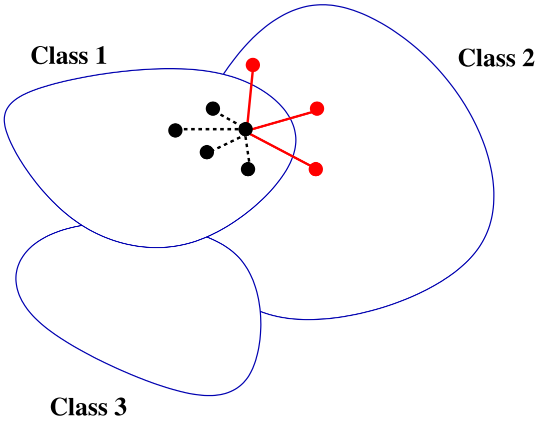


Fig. 4.1. *Illustration of the repulsion graph. The repulsion graph is obtained by only retaining among the k nearest neighbors of a node i, those nodes that are not in the same class as i (illustrated in red).*

In [8], it was observed that this mechanism was often inadequate because it did not take into account nearness of points that are from different classes. Two points x_i and x_j may exist that are close by, but which do not belong to the same class. When a projection is performed, there is a risk that these two points which are close-by, will get incorrectly projected to the same class. To remedy this the paper introduced a method based on the concept of *repulsion graphs*.

A *repulsion graph* is one that is extracted from the kNN graph, based on class label information, and whose goal is to create repulsion forces between nearby points which are not from the same class. For example, when a k-nearest neighbor graph is used, a repulsion graph can be created by only *retaining among the k-nearest neighbors of a node i, those nodes that are not in the same class as i* (see Fig. 4.1). For simplicity, *we assume that the kNN graph is symmetrized by including the edge (j, i) whenever the edge (i, j) exists.* The weight matrix can be defined in the same way as for Laplacean graphs.

In the following, the original graph is only referenced as the class graph and requires no further notation. Its associated Laplacean is denoted by L. The repulsion graph is derived from a certain kNN graph, which we denote by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The repulsion graph itself is denoted by $\mathcal{G}^{(r)} = (\mathcal{V}^{(r)}, \mathcal{E}^{(r)})$, and its associated Laplacean by $L^{(r)}$. Accordingly, the adjacency list for a given node i is now $\mathcal{N}^{(r)}(i)$. Assume for a moment that the weights are of the simple 'uniform' type, i.e., $l_{ij}^{(r)} = -1$ for $(i, j) \in \mathcal{E}^{(r)}$ and $i \neq j$ and $l_{ii}^{(r)} = -\sum_j l_{ij}^{(r)}$. In other words, if we denote by $\ell(k)$ the class label of item k, then the Laplacean matrix $L^{(r)}$ is derived from the weight matrix

$$W_{ij}^{(r)} = \begin{cases} 1 & \text{for} \quad (i,j) \in \mathcal{E},\ i \neq j, \quad \text{and } \ell(i) \neq \ell(j) \\ 0 & \text{otherwise} \end{cases} \tag{4.17}$$

by defining $L^{(r)} = D^{(r)} - W^{(r)}$, in which $D^{(r)}$ is the matrix of row-sums of $W^{(r)}$. This is a valid graph Laplacean as the row sums of the matrix are all zero, and the off-diagonal entries are non-positive. By the assumption of the indirection of the kNN graph (see above), $L^{(r)}$ is symmetric.

The key idea is that any objective function which will utilize the repulsion graph will tend to *maximize, rather than minimize* (4.14), where the Laplacean matrix now is associated with the repulsion graph. This 'repulsion Laplacean' will model a force - or rather an energy - which will tend to repel near-by points in different classes away from each other. This is achieved by adding a negative term to the original objective function (4.16):

$$\Phi_\rho(Y) = \sum_i \|y_i - \sum_j w_{ij} y_j\|_2^2 - \rho \sum_i \sum_{j \in N^{(r)}(i)} \|y_i - y_j\|_2^2. \tag{4.18}$$

Here, $N^{(r)}(i)$ represents the set of neighbors of node i, with respect to the repulsion graph. A similar device was used for the LPP approach [8]. The second term in the above expression was referred to as the *penalty term* and to the parameter ρ as the *penalty parameter*. If two projected entries x_i and x_j are not in the same class but they are close, then the edge (i, j) is part of the graph $\mathcal{G}^{(r)}$ and there is a penalty for having the two nodes close in Y. Due to the negative sign, the penalty term will tend to be larger in absolute value in order to minimize the objective function.

It can be shown from relation (4.18) that the above objective function can be expressed as (see [8]):

$$\Phi_\rho(V) = \mathrm{Tr}\left[V^T X (M - \rho L^{(r)}) X^T V\right], \tag{4.19}$$

where $M = (I - W^T)(I - W)$. If we impose orthogonality constraints i.e., $V^T V = I$, then V is obtained from the d bottom eigenvectors of matrix $X(M - \rho L^{(r)})X^T$.

In the end the optimization problem solved when repulsion Laplaceans are used is of the form:

$$\min_{V^T V = I} \mathrm{Tr}\left[V^T (A - \rho B) V\right], \tag{4.20}$$

with $A = XLX^T$, and $B = XL^{(r)}X^T$ and $V \in \mathcal{U}_p$. This problem is clearly of the form (4.2), except that the max is replaced by the min. In other words, *the method of repulsion Laplaceans amounts to just selecting ρ arbitrarily (i.e., not optimally) and computing the optimal subspace of $G(\rho)$*. For the problems and applications seen in [8], the performance of the method varied very smoothly in terms of ρ. The numerical experiments section will illustrate how this technique compares with one based on the techniques presented in earlier sections.

5. Necessary conditions for optimality. In this section we consider the optimization problem (1.1) under the common framework of necessary conditions of optimality. The Lagrangian function of the problem (2.3) (where C is the identity and B satisfies the conditions of Lemma 3.1) is

$$L(W, \Gamma) = \frac{\mathrm{Tr}\,[W^T A W]}{\mathrm{Tr}\,[W^T B W]} - \mathrm{Tr}\,[\Gamma(W^T W - I)].$$

According to the Karush-Kuhn-Tucker (KKT) optimality conditions, since (2.3) has a global maximizer W_* then there exist a Lagrangian multiplier matrix Γ_* such that,

$$\frac{\partial L(W_*, \Gamma_*)}{\partial W} = 0 \quad \text{with} \quad W_*^T W_* = I.$$

Given a matrix M we need to derive an expression for the partial derivative of $\varphi_M(W) = \mathrm{Tr}\,[W^T M W]$ with respect to W.

The function $\varphi_M(W)$ is a scalar function which depends on W. When M is symmetric, the gradient of $\varphi_M(W)$ with respect to W satisfies

$$\nabla\varphi_M(W).E = 2\text{Tr}\ \left[W^T M E\right]$$

so that $\partial\varphi_M(W)/\partial W = 2MW$. Therefore, we obtain ($A$ and B are symmetric)

$$\frac{\partial L(W,\Gamma)}{\partial W} = \frac{2\varphi_B(W)AW - 2\varphi_A(W)BW}{(\varphi_B(W))^2} - W(\Gamma^T + \Gamma).$$

Hence, the optimal solutions W_* and Γ_* verify

$$(A - \rho_* B)\ W_* = \frac{\varphi_B(W_*)}{2} W_*(\Gamma_*^T + \Gamma_*), \tag{5.1}$$

where $\rho_* = \varphi_A(W_*)/\varphi_B(W_*)$. Let Q be the matrix which diagonalizes $\Gamma_*^T + \Gamma_*$:

$$\Gamma_*^T + \Gamma_* = Q\ \Sigma_*\ Q^T, \quad Q^T Q = I,$$

where Σ_* is a diagonal matrix. Observe that

$$\text{Tr}\,[\Gamma_*^T + \Gamma_*] = 2\frac{\text{Tr}\,[W_*^T(A - \rho_* B)W_*]}{\varphi_B(W_*)} = 0 \quad \text{ and } \quad \text{Tr}\,[\Sigma_*] = 0.$$

Define $U_* = W_* Q$. We have $U_*^T U_* = 1$ and we can rewrite Equation 5.1 as

$$(A - \rho_* B)\ U_* = U_* \Lambda_*, \quad \text{where} \quad \Lambda_* = \frac{\varphi_B(W_*)}{2}\ \Sigma_*. \tag{5.2}$$

Equation (5.2) above is the necessary condition for the pair ρ_*, U_* to be optimal with $\text{Tr}\,[U_*^T(A - \rho_* B)\ U_*] = \text{Tr}\,[\Lambda_*] = 0$. This provides another viewpoint to the analysis seen in earlier sections.

6. Experiments. This section illustrates the methods discussed in this paper with applications in dimensionality reduction for face recognition and handwritten digit recognition.

The classical dimensionality reduction technique leading to optimizing the trace ratio is Fisher Linear Discriminant Analysis (LDA) [3], where A corresponds to the between-class covariance matrix and B corresponds to the within-class covariance matrix. LDA can be seen as a global approach to supervised dimensionality reduction since the computation of the covariance matrices involves all data points. From a graph-based point of view, the method employs two globally-binary-relationship graphs: within-class graph (or class graph), G_W, where edge (i, j) exists if i and j belong to the same class; and between-class graph (or repulsion graph), G_B, where edge (i, j) exists if i and j belong to different classes. In this view, A and B correspond to the graph Laplacean of G_B and G_W, respectively. Local Discriminant Embedding (LDE) [1], ONPP-R and OLPP-R [8] are local versions of LDA in which local variances are exploited by using only k nearest neighbors to form between-class graphs and within-class graphs. Moreover, the weights can be generalized to any similarity measures other than binary relationships and different methods construct the weights in different ways.

LDE optimizes the ratio by relying on the eigenvectors of $B^{-1}A$. In contrast, ONPP-R and OLPP-R rely on the eigenvectors of $A - \rho B$, where ρ is penalty term set a priori. In our experiments, we will compare the results of the iterative method based on maximizing the trace ratio, and represented by Algorithm 4.1, against these two methods. We will consider both the global (or non-local) way and a local way of forming A and B.

Here, by *local way* we simply mean a method based on some graph (e.g., kNN) to capture local structures. By *global way* we mean a method, such as LDA, which uses dense matrices to capture similarities between data samples in the data set.

The notation used for the various methods tested is as follows:

- LDA and LDE refer to methods that rely on the eigenvectors of $B^{-1}A$. LDA uses non-local matrices and LDE uses a local matrices.
- LDA-ITR and LDE-ITR refer to methods which optimize the trace ratio iteratively, using the Newton approach described in Section 4.3. The matrices A and B are formed in a non-local way for LDA-ITR and in a local way for LDE-ITR.
- LDA-R and LDE-R refer to methods which exploit repulsion Laplaceans. They utilize the eigenvectors of $A - \rho B$. Again, A and B are nonlocal for LDA-R and local for LDE-R.

6.1. Experimental setup. We experimented on six different datasets: ORL, AR, UMIST, PIE, Essex (can be found at http://face-rec.org/databases/) and USPS hanwritten dataset. Each image in ORL, AR and UMIST is downsampled to 38×31 and is represented as a 1178 dimensional vector. Similarly, each image in PIE is represented as a 1024 dimensional vector. The ORL dataset contains 40 individuals with 10 images for each individual under variation in facial expression and pose. From the AR dataset, we use 126 subjects, each of which has 8 images taken under different facial expressions and lightning conditions. The UMIST database contains 20 people with different poses and the number of images per person varies from 19 to 48. The PIE dataset contains 68 subjects, each of them has about 170 images. The Essex database contains 4 different sets, from easy to hard: face94, face95, face96, and grimaces. We use the collection face95 which has 72 individuals, 20 images per individual. Each individual was photographed using a fixed camera, while the subject took one step towards the camera. Each image in Essex dataset is cropped by removing the top 20 rows, 10 bottom rows and 20 columns each side and yields a 36×43 image or equivalently, a 1548-dimensional vector. We use a portion of the USPS dataset which consists of 1100 grayscale images of handwritten digits with 110 images for each digit. Each image is represented as a 256-dimensional vector.

Images of the same subject are divided randomly into training sets and test sets. We perform 5 different random realizations of the training/testing sets and average the errors. The numbers of training samples for ORL, AR, UMIST, PIE, Essex and USPS are 5, 4, 10, 10, 10 and 20, respectively.

For local methods, within-class graphs and repulsion graphs are formed separately using k-nearest neighbors. We use $k = 3$ for within graphs and $k = 10$ for repulsion graphs. The heat kernel is used to compute distances between nodes in graphs.

In all experiments, both matrices A and B are scaled to have unit trace before optimization. In addition, a preliminary PCA using Lanczos algorithm (see e.g., [4]) is employed before all of these methods to reduce the dimensionality of data to $n - c$, where n is the number of training samples and c is the number of classes. To make sure that A and B are positive definite, we regularize them by adding small numbers to their diagonals.

Nearest neighbor classifier is used on the reduced spaces to classify images into

Dims	10	20	30	40	50	60	70
LDE-ITR	32.4648	19.3766	13.6758	11.7107	28.2914	16.9605	13.7062
LDE	23.5373	13.5477	9.4640	8.0027	20.0822	12.7405	9.5377

FIG. 6.1. *Values of* $tr[W^T AW]/tr[W^T BW]$.

subjects. Classification rates are shown to illustrate the performance of dimensionality reduction.

6.2. Results and discussions. Because they optimize the ratio of the traces, the Newton-Lanczos iterative procedure yielded significantly better trace ratios than did non-iterative ones. This is depicted in table 6.1 for the PIE dataset. However, in a few cases, results generated by the Newton-Lanczos iterative method were slightly worse than those generated by the other techniques. This usually happens when the dimension of the reduced space d is small and does not happen when $d > 50$. One possible explanation for this is that in those cases, largest eigenvalues of intermediate matrix $A - \rho B$ are very close to each other and Matlab (the algorithm under Matlab) fails to capture the eigenvectors corresponding to true largest eigenvalues.

Figure 6.2 compares LDA-ITR and LDE-ITR against LDA and LDE for face recognition on different datasets with the dimensions of reduced spaces ranging from 10 to 100. We can see that the LDA-ITR/LDE-ITR outperforms LDA/LDE for ORL, AR, and UMIST. The improvement gets more significant as the dimension gets bigger.

However, for the PIE dataset, the iterative method based on Algorithm 4.1 performed worse than LDA and LDE for low dimensional spaces and started to perform better when $d \geq 50$. This is in spite of the fact that in all experiments with the PIE dataset, the optimal ratios obtained by the trace-ratio based variants LDA-ITR and LDE-ITR are always better than those of their non-ratio-based sibblings LDA and LDE.

Figure 6.3 shows the results of 2-D projection for PIE dataset using LDA and LDA-ITR. On the left hand side are projected training data and on the right handside are projected testing data. Only 3 random subjects are displayed, but the whole 68 subjects give similar pattern. We can see that trace-ratio based iterative methods tends to minimize inner-class variance. They also tend to overfit the training data . In this case, the projected data on 2-D space almost lie on a 1-D subspace and this may provide a clue for the poor results seen in the low dimensional case.

We can also see that for some small datasets such as ORL, the results of LDA and LDE are very similar due to the fact that training sets are small and therefore local information used in LDE is roughly the same with global information used in LDA.

LDA-ITR and LDE-ITR usually take 6-11 iterations to converge to the optima, which is quite fast. One implementation issue we may mention when using Matlab is that often Matlab's `eigs` function, drops eigenvalues down to 0 which causes convergence difficulties for the Newton-based iterative methods. We do not expect this to be an issue in a production-type procedure implemented in C, C++, or Fortran.

Finally, we compare LDA, LDE, LDA-ITR and LDE-ITR against LDA-R and LDE-R. Surprisingly, LDA-R and LDE-R with suitable penalty terms (0.2 in all of our experiments) give better results in all datasets. The results tend to be the same as LDA-ITR and LDE-ITR for high dimensional spaces. Especially, LDA-R and LDE-R quickly reach high recognition rates at very low dimensions. Figure 6.4 shows recognition rates for ORL, PIE, UMIST and USPS with dimensions ranging from 10

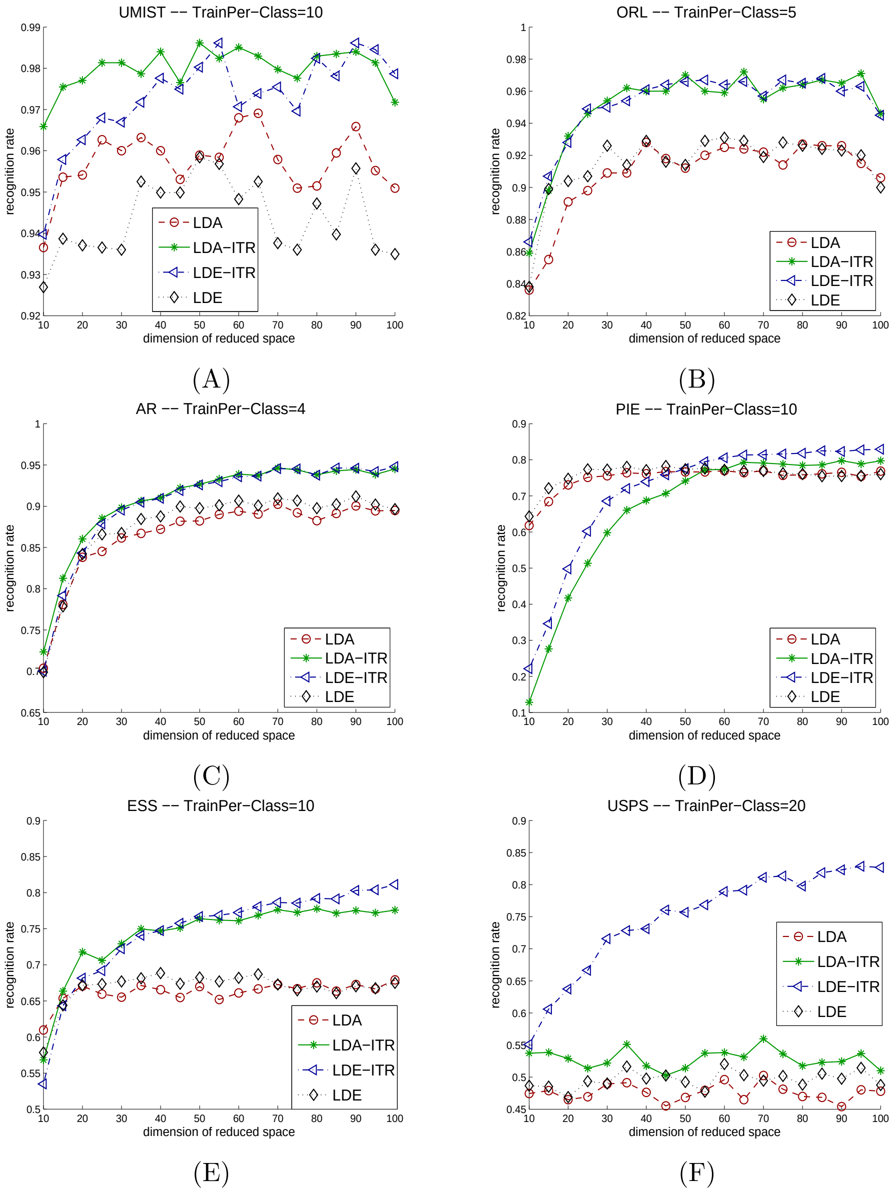


Fig. 6.2. *Recognition rates for the (A) UMIST, (B) ORL, (C) AR, (D) PIE, (E) Essex and (F) USPS.*

to 100 and from 2 to 30. 2-D projections for ORL in figure 6.5 show that LDE-R gives similar point clouds for both training and testing data (LDA-R gives similar results). Meanwhile, other methods demonstrate overfitting with very sticky clouds for training data but scatter clouds for testing data.

7. Conclusion.

We conclude with three observations. First, maximizing the trace ratio in (1.1) need not be expensive. In fact our experiments show that with a judicious use of the Lanczos procedure, a good initialization, and inexact eigenvector calculations in the early stages of the Newton procedure, the overall procedure may

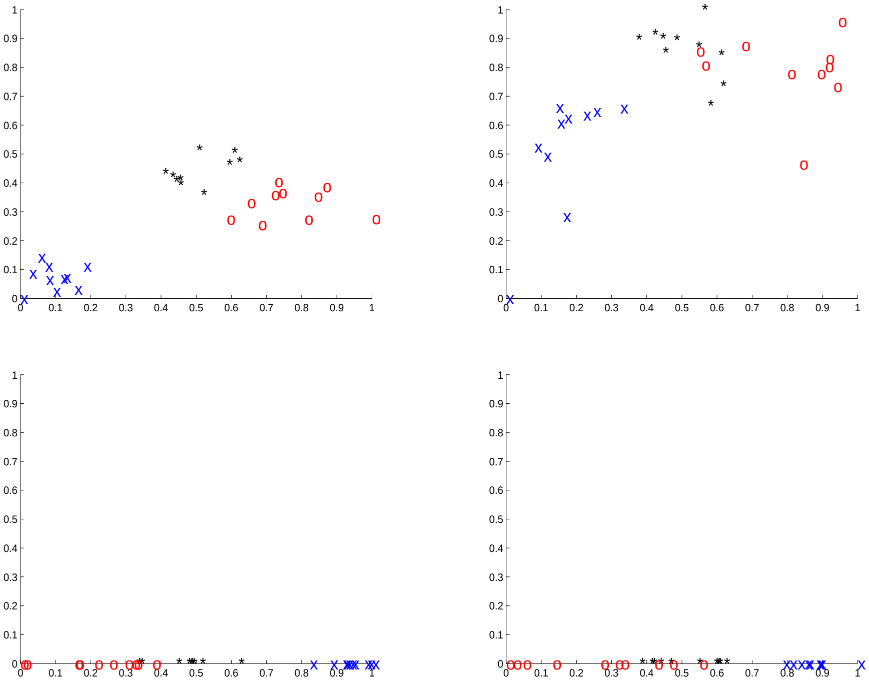


FIG. 6.3. *2-D projections of the PIE dataset. Top: LDA; Bottom: LDA-ITR. Left side: Training samples; Right side: test samples.*

be much less expensive than one which relies on solving the generalized eigenvalue problem associated with the common approach based on (1.2). Secondly, if one compares two methods based on the same principle, one of which maximizes the trace ratio (1.1) and the other the constrained trace (1.2), then generally the former will do better. This confirms observations made by other researchers. Our third observation is that when a good penatly parameter is used, the technique of repulsion Laplaceans appears to perform generally better than one based on optimizing the trace ratio. In other words, in most cases, there are values of $\rho \neq \rho_*$ which will yield better observed performance than when using the optimal ρ_*. This is rather suprising, and merits further investigation, because the method of repulsion Laplaceans can be viewed as a simplification of the trace ratio optimization approach.

REFERENCES

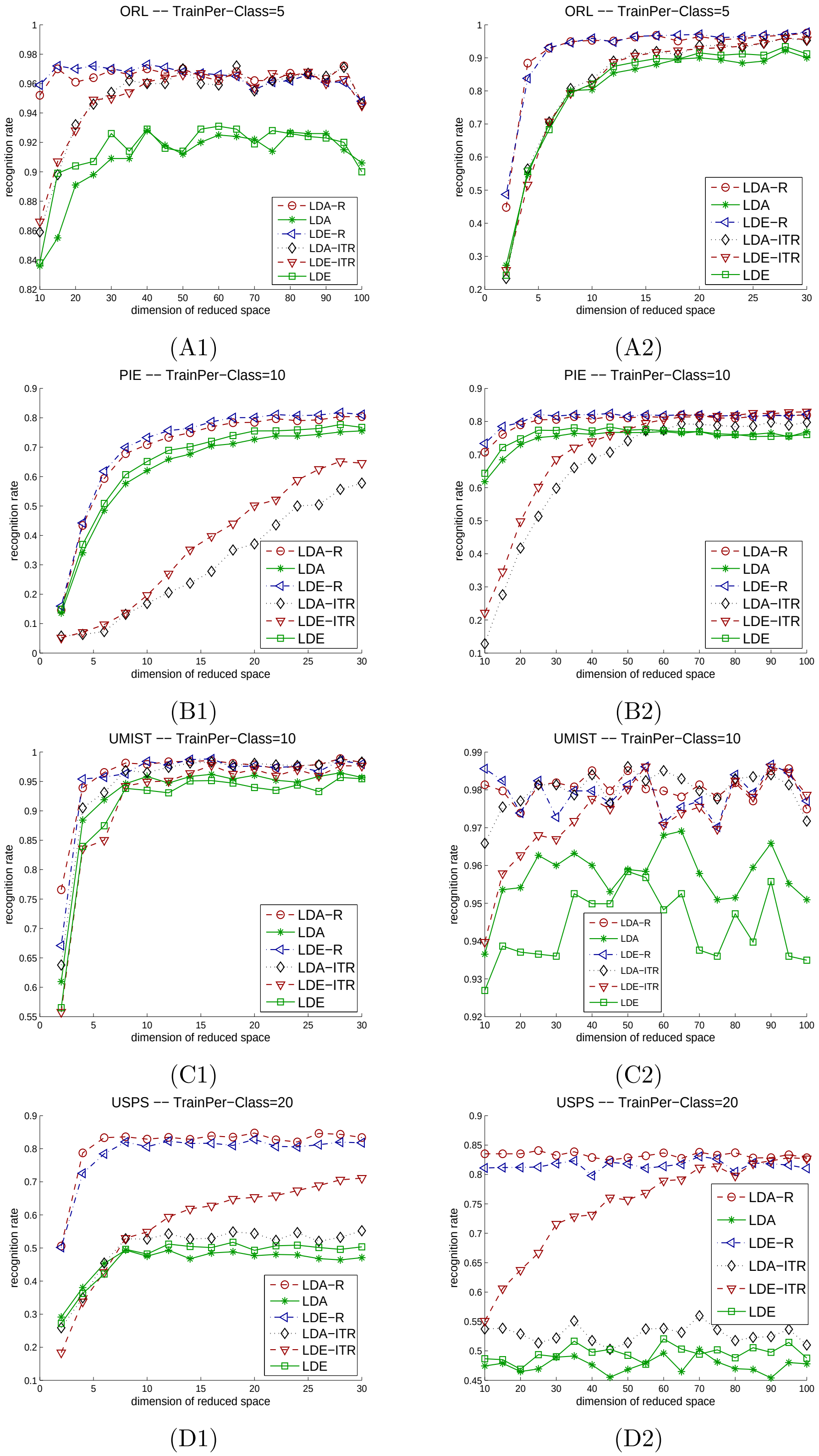
[1] Hwann-Tzong Chen, Huang-Wei Chang, and Tyng-Luh Liu. Local discriminant embedding and its variants. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 846–853, Washington, DC, USA, 2005. IEEE Computer Society.

[2] R. S. Dembo, S. C. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM J. Numer. Anal.*, 18(2):400–408, 1982.

[3] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1991.

[4] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 3rd edition, 1996.

[5] Yue-Fei Guo, Shi-Jin Li, Jing-Yu Yang, Ting-Ting Shu, and Li-De Wu. A generalized Foley-



(A1) (A2)

(B1) (B2)

(C1) (C2)

(D1) (D2)

Fig. 6.4. *Recognition rates for the (A) ORL (B) PIE, (C) UMIST and (D) USPS.*

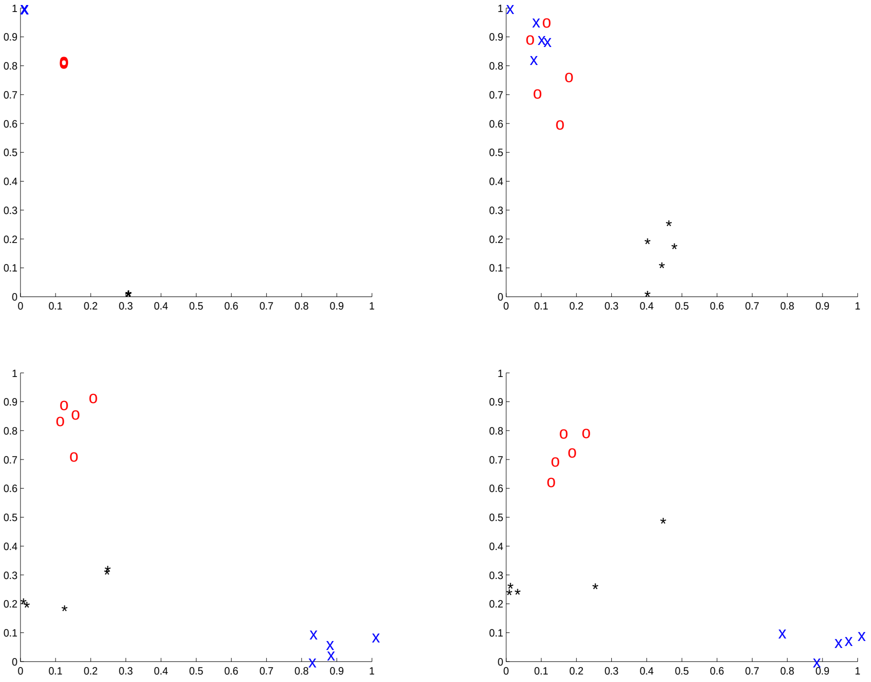


Fig. 6.5. *2-D projection of ORL dataset. LDA at the top and LDA-R at the bottom. Training samples on the left and testing samples on the right.*

Sammon transform based on generalized Fisher discriminant criterion and its application to face recognition. *Pattern Recogn. Lett.*, 24(1-3):147–158, 2003.

[6] X. He and P. Niyogi. Locality preserving projections. *In Proc. Conf. Advances in Neural Information Processing Systems*, 2003.

[7] E. Kokiopoulou and Y. Saad. Orthogonal neighborhood preserving projections. In J. Han et al., editor, *IEEE 5th Int. Conf. on Data Mining (ICDM05), Houston, TX, Nov. 27-30th*, pages 234–241. IEEE, 2005.

[8] E. Kokiopoulou and Y. Saad. Enhanced graph-based dimensionality reduction with repulsion Laplaceans. Technical Report umsi-2008-278, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN, 2008. To Appear- Pattern Recognition [accp. Apr. 8th, 2009].

[9] Feiping Niei, Shiming Xiang, Yangqing Jia, Changshui Zhang, and Shuicheng Yan. Trace ratio criterion for feature selection. In *AAAI*, pages 671–676, 2008.

[10] B. N. Parlett. *The Symmetric Eigenvalue Problem*. Prentice Hall, Englewood Cliffs, 1980.

[11] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[12] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Halstead Press, New York, 1992.

[13] Chunhua Shen, Hongdong Li, and Michael J. Brooks. A convex programming approach to the trace quotient problem. In *ACCV (2)*, pages 227–235, 2007.

[14] Chunhua Shen, Hongdong Li, and Michael J. Brooks. Supervised dimensionality reduction via sequential semidefinite programming. *Pattern Recognition*, 41(12):3644–3652, 2008.

[15] Huan Wang, Shuicheng Yan, Dong Xu, and Xiaoou Tang Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *CVPR '07. IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[16] Shiming Xiang, Feiping Nie, and Changshui Zhang. Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, 41(12):3600 – 3612, 2008.

[17] Shuicheng Yan and Xiaoou Tang. Trace quotient problems revisited. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Proceedings of the European Conference on Computer*

Vision, volume 2 of *Lecture Notes in Computer Science, Number 3952*, pages 232–244, Berlin-Heidelberg, 2006. Springer Verlag.