Trace Quotient Problems Revisited

Shuicheng Yan[1] and Xiaoou Tang[1,2]

[1] Department of Information Engineering,
The Chinese University of Hong Kong, Hong Kong
[2] Microsoft Research Asia, Beijing, China

Abstract. The formulation of *trace quotient* is shared by many computer vision problems; however, it was conventionally approximated by an essentially different formulation of *quotient trace,* which can be solved with the generalized eigenvalue decomposition approach. In this paper, we present a direct solution to the former formulation. First, considering that the feasible solutions are constrained on a Grassmann manifold, we present a necessary condition for the optimal solution of the trace quotient problem, which then naturally elicits an iterative procedure for pursuing the optimal solution. The proposed algorithm, referred to as Optimal Projection Pursuing (OPP), has the following characteristics: 1) OPP directly optimizes the trace quotient, and is theoretically optimal; 2) OPP does not suffer from the solution uncertainty issue existing in the quotient trace formulation that the objective function value is invariant under any non-singular linear transformation, and OPP is invariant only under orthogonal transformations, which does not affect final distance measurement; and 3) OPP reveals the underlying equivalence between the trace quotient problem and the corresponding trace difference problem. Extensive experiments on face recognition validate the superiority of OPP over the solution of the corresponding quotient trace problem in both objective function value and classification capability.

1 Introduction

In recent decades, a large family of algorithms [19]—supervised or unsupervised; stemming from statistical or geometry theory — has been proposed to provide different solutions to the problem of dimensionality reduction [2][4][12][15][16][19]. Many of them, such as Linear Discriminant Analysis (LDA) [1] and Locality Preserving Projection (LPP) [6], eventually come down to the trace quotient problem [17][20] as follows

$$W^* = arg \max_{W^{\mathrm{T}}CW=I} \frac{Tr(W^{\mathrm{T}}AW)}{Tr(W^{\mathrm{T}}BW)}. \tag{1}$$

Here A, B, and C are all symmetric positive semidefinite; $Tr(\cdot)$ denotes the trace of a matrix; I is an identity matrix and W is the pursued transformation matrix for dimensionality reduction. Commonly, the null space of matrix C lies within the null space of both A and B, that is, $null(C) \in null(A) \bigcap null(B)$. Due to the lack of a direct efficient

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part II, LNCS 3952, pp. 232 – 244, 2006.

solution for Eq. (1), the quotient trace problem $Tr((W^{\mathrm{T}}BW)^{-1}(W^{\mathrm{T}}AW))$ is often discussed instead and the generalized eigenvalue decomposition (GEVD) [20] approach is applied for a direct closed-form solution.

If W is a vector, it is theoretically guaranteed that the optimal solution of (1) is the eigenvector corresponding to the largest eigenvalue of GEVD by using the Lagrange Multiplier method. GEVD can provide an optimal solution to the quotient trace problem, yet it is not necessarily optimal for the trace quotient problem when W is in the form of a matrix. Moreover, the solution from GEVD is unstable when matrix B is singular; and Principal Component Analysis (PCA) [14] is often used beforehand to avoid the singularity issue. However, it is often observed that the algorithmic performance is extremely sensitive to the retained dimension of PCA. All these motivate us to pursue an efficient and theoretically sound procedure to solve the trace quotient problem.

More specifically, our contributions are as follows. First, we prove that GEVD cannot provide an optimal solution to the trace quotient problem. Then, we present a necessary condition for the optimal solution of the trace quotient problem by taking into account the fact that the feasible solutions are constrained to lie on a Grassmann manifold. Finally, by following the necessary condition, an efficient procedure is proposed to pursue the optimal solution of the trace quotient problem. As a product, the necessary condition indicates the underlying equivalence between the trace quotient problem and the corresponding trace difference problem.

The rest of the paper is organized as follows. In section 2, we introduce the trace quotient problem and the corresponding quotient trace problem, and then discuss the infeasibility of the GEVD method in solving the trace quotient problem. In Section 3, a necessary condition for the optimal solution of the trace quotient problem is presented, which naturally elicits an iterative procedure to pursue the optimal solution. Extensive experiments on face recognition are demonstrated in Section 4 to show the superiority of our proposed algorithm over GEVD. Finally, in Section 5, we conclude the paper and provide discussions of future work.

2 Trace Quotient Problem

Denote the sample set as matrix $X = [x_1, x_2, \cdots, x_N], x_i \in \mathbb{R}^m$ is an m-dimensional vector. For supervised learning tasks, the class label of the sample x_i is assumed to be $c_i \in \{1, 2, \cdots, N_c\}$ and n_c denotes the sample number of the c-th class.

2.1 Trace Quotient Problem vs. Quotient Trace Problem

A large family of algorithms for subspace learning [6] ends with solving a trace quotient problem as in (1). Among them, the most popular ones are the Linear Discriminant Analysis (LDA) [17] algorithm and its kernel extension. LDA searches for the most discriminative directions that maximize the quotient of the inter-class scatter and the intra-class scatter

$$W^* = arg \max_{W^TW=I} \frac{Tr(W^T S_b W)}{Tr(W^T S_w W)}$$
$$S_b = \sum_{c=1}^{N_c} n_c (m_c - m)(m_c - m)^T, S_w = \sum_{i=1}^{N} (x_i - m_{c_i})(x_i - m_{c_i})^T \quad (2)$$

Here m_c is the mean of the samples belonging to the c-th class and m is the mean of all samples; $W \in \mathbb{R}^{m \times k}$ is the pursued transformation matrix for dimensionality reduction. The objective function of (2) has explicit semantics for both numerator and denominator and they characterize the scatters measured by the Euclidean distances in the low dimensional feature space

$$Tr(W^T S_b W) = \sum_{c=1}^{N_c} n_c \| W^T m_c - W^T m \|^2, \quad Tr(W^T S_w W) = \sum_{i=1}^{N} \| W^T x_i - W^T m_{c_i} \|^2 . \quad (3)$$

A direct way to extend a linear algorithm to a nonlinear case is to utilize the kernel trick [5][9][18]. The intuition of the kernel trick is to map the data from the original input space to a higher dimensional Hilbert space as $\phi : x \to \mathcal{F}$ and then the linear algorithm is performed in this new feature space. It can be well applied to the algorithms that only need to compute the inner products of the data pairs $k(x, y) = \phi(x) \cdot \phi(y)$. For LDA, provided that $W = [\phi(x_1), \phi(x_2), ..., \phi(x_N)]M$, where $M \in \mathbb{R}^{N \times k}$ and $K \in \mathbb{R}^{N \times N}$ is the kernel Gram matrix with $K_{ij} = k(x_i, x_j)$, we have

$$M^* = arg \max_{M^TKM=I} \frac{Tr(M^T K S_b K M)}{Tr(M^T K S_w K M)} . \quad (4)$$

Obviously, LDA and its kernel extension both follow the formulation of trace quotient as in (1); generally, there is no closed-form solution for (2) and (4) when k >1.

Instead of directly solving the trace quotient problem, many researchers study another formulation, called the quotient trace problem here, to pursue the most discriminative features as follows

$$W^* = arg \max_W Tr((W^T B W)^{-1} (W^T A W)) . \quad (5)$$

Notice that commonly there is no constraint on matrix W in the quotient trace problem and it is solved by the generalized eigenvalue decomposition (GEVD) method

$$A w_i = \lambda_i B w_i, \quad i = 1, ..., k . \quad (6)$$

Here w_i is the eigenvector corresponding to the i-th largest eigenvalue λ_i . Despite extensive study of the quotient trace problem, it suffers the following disadvantages: 1) it is invariant under any nonsingular linear transformation, which results in the uncertainty of the Euclidean metric on the derived low dimensional feature space; and 2) unlike the trace quotient problem, there does not exist explicit semantics for the objective function of quotient trace problem. Therefore, compared with the quotient trace formulation, the trace quotient formulation is more reasonable; and in the following, we study the problem of how to directly solve the trace quotient problem.

2.2 Is Generalized Eigenvalue Decomposition Approach Feasible?

Based on the constraint on the transformation matrix and the *Lagrange Multiplier* method, the trace quotient problem (1) is equivalent to maximizing

$$F(W,\lambda) = Tr(W^{\mathrm{T}}AW) - \lambda(Tr(W^{\mathrm{T}}BW) - c). \tag{7}$$

Here, c is a constant and λ is the Lagrange Multiplier. When W is a vector, *i.e.* $k = 1$, the problem (1) is simplified to maximizing $F(W,\lambda) = W^{\mathrm{T}}AW - \lambda(W^{\mathrm{T}}BW - c)$. It is easy to prove that the optimal solution is the eigenvector corresponding to the largest eigenvalue calculated from the generalized eigenvalue decomposition method as in (6). Yet, when W is a matrix, *i.e.* $k > 1$, the problem is much more complex, and intuitively it was believed that the leading eigenvectors from GEVD were more valuable in discriminating power than the later ones, since the individual trace quotient, namely eigenvalue, from the leading eigenvector is larger than those from later ones. However, no theoretical proof was ever presented to justify using GEVD for solving the trace quotient problem. Here, we show that GEVD is infeasible for the following reasons. For simplicity, we discuss the LDA formulation with the constraint $W^{\mathrm{T}}W = I$.

Orthogonality: The derived eigenvectors from GEVD are not necessarily orthogonal. Let the Singular Value Decomposition of the final projection matrix W be

$$W = U\Lambda V^{\mathrm{T}}. \tag{8}$$

The right orthogonal matrix V is free for the trace quotient, thus the derived solution is equal to $U\Lambda$ in the sense of rotation invariance. In this point, GEVD does not find a set of unit projection directions, but weighted ones. The left column vector of U maybe is more biased when the original feature is transformed to the low dimensional space, which conflicts with the unitary constraint.

Theoretical Guarantee: There is no theoretical proof to guarantee that the derived projection matrix can optimally maximize the trace quotient. Actually, the projection vector from GEVD is evaluated in an individual manner and the collaborative trace quotient will be easily biased by the projection direction with larger values of $(w^{\mathrm{T}}Bw, w^{\mathrm{T}}Aw)$. For example, for projection directions w_1, w_2, w_3, if their trace values are as follows (*e.g.* A=diag{10.0, 100.0, 2.0} and B=diag{1.0, 20.0, 1.0})

	w_1	w_2	w_3
$w^{\mathrm{T}}Aw$	10.0	100.0	2.0
$w^{\mathrm{T}}Bw$	1.0	20.0	1.0

then the combination of w_1 and w_3 (with trace quotient 6) is better than that of w_1 and w_2 (with trace quotient 5.24) although the single trace quotient from w_2 is larger than that from w_3. Thus, *it is not true that the eigenvector corresponding to the larger eigenvalue of GEVD is always superior to that from a smaller one in the trace quotient problem.*

Necessary Condition: It was commonly believed that the optimal solution of (1) should satisfy

$$\partial F(W,\lambda)/\partial W = 0\,. \tag{9}$$

Yet, the solution may not exist at all if directly setting the gradient as zero,

$$\partial F(W,\lambda)/\partial W = 2(A-\lambda B)W = 0\,. \tag{10}$$

It means W is the null subspace of the weighted difference of matrix B and A, *i.e.* $A-\lambda B$. In the LDA formulation, when $m < N - N_c$, matrix B is of full rank, and for most λ, $A-\lambda B$ is also of full rank; consequently there does not exist matrix $W \in \mathbb{R}^{m\times k}$ with independent columns that satisfies (7). As we will analyze later, the fundamental reason that GEVD fails to find the optimal solution is that it does not consider that the feasible solution of (1) is constrained to lie on a lower dimensional Grassmann manifold (or a transformed one when matrix C is not equal to I), not the whole matrix space, and the derivative should also be constrained to lie on the Grassmann manifold, instead of the matrix space.

All the above analyses show that the GEVD cannot provide an optimal solution for the trace quotient problem. In the following, we will present our solution to this problem.

3 Optimal Solution to Trace Quotient Problem

For the trace quotient problem (1), let the Singular Value Decomposition of matrix C be

$$C = U_c \Lambda_c U_c^{\mathrm{T}},\ U_c \in \mathbb{R}^{m\times n},\quad n \geq k\,. \tag{11}$$

Here Λ_c only contains positive diagonal elements, and denote $Q = \Lambda_c^{1/2} U_c^{\mathrm{T}} W$. As we have the assumption that $null(C) \in null(A) \bigcap null(B)$, we can constrain the matrix W in the space spanned by the column vectors of U_c and we have $W = U_c \Lambda_c^{-1/2} Q$, then

$$Q^* = arg \max_{Q^{\mathrm{T}}Q=I} \frac{Tr(Q^{\mathrm{T}}\Lambda_c^{-1/2} U_c^{\mathrm{T}} A U_c \Lambda_c^{-1/2} Q)}{Tr(Q^{\mathrm{T}}\Lambda_c^{-1/2} U_c^{\mathrm{T}} B U_c \Lambda_c^{-1/2} Q)}\,. \tag{12}$$

It is still a trace quotient problem, yet with the unitary and orthogonal constraints; hence in the following, we only discuss the trace quotient problem with the unitary and orthogonal constraints.

3.1 Necessary Condition for Optimal Solution

When the solution of the trace quotient problem is constrained to be columnly orthogonal and unitary, the solution space is not the whole matrix space any more, instead, mathematically, all the feasible solutions constitute a Grassmann manifold [3]. Before describing the procedure to solve the trace quotient problem, we introduce the concepts of the Grassmann manifold and its tangent space.

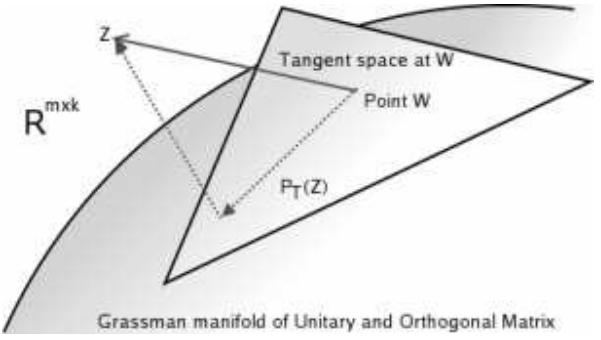

Fig. 1. The illustration of the relationship between the original matrix space, Grassmann manifold, and the projection to the tangent space. Note that it is unnecessary for gradient vector Z to be zero, instead, only its projection to the horizontal space of the tangent space is required to be zero for the trace quotient problem.

Grassmann Manifold [3]**:** All feasible matrices $W \in \mathbb{R}^{m\times k}$ with unit and orthogonal column vectors, *i.e.* $W^{\mathrm{T}}W = I$, constitute a continuous curved hyper-surface in the original matrix space, namely a Grassmann manifold, as shown in Figure 1. Commonly, a Grassmann manifold is associated with an objective function $F(W)$, such as the objective function in (7), yielding $F(WR) = F(W)$ for any orthogonal matrix $R \in \mathbb{R}^{k\times k}$.

If for two columnly orthogonal matrices W_1 and W_2, there exists an orthogonal matrix R so that $W_1 = W_2 R$, then we call W_1 and W_2 homogeneous, denoted as $W_1 \sim W_2$. Thus, on the Grassmann manifold, the objective function $F(W)$ is invariant to all matrices that are homogeneous.

Projection on the Tangent Space [3]**:** As a curved hyper-surface, the movement of any point on the manifold always follows a direction in the tangent space as shown in Figure 1. All matrices M in the tangent space at point W satisfy

$$W^{\mathrm{T}}M + M^{\mathrm{T}}W = 0\,. \tag{13}$$

And for any matrix Z, its projection on the tangent space is defined as

$$P_T(Z) = \tfrac{1}{2}W(W^{\mathrm{T}}Z - Z^{\mathrm{T}}W) + (I - WW^{\mathrm{T}})Z\,. \tag{14}$$

Considering the homogeneity condition, not all variations in tangent space will result in a change of the objective function. The tangent space is decomposed into the direct sum of a vertical space and a horizontal space, where only the directions in the horizontal space actually contribute to the change of the objective function. It is proved [3] that the projection of any vector Z on the horizontal space at W is

$$P_H(Z) = (I - WW^{\mathrm{T}})Z\,. \tag{15}$$

For the equivalent objective function (7) of the trace quotient problem, the gradient vector $Z = \partial F(W,\lambda)/\partial W = 2(A-\lambda B)W$. Its projection on the tangent space directly lies within the horizontal space, since *A* and *B* are both symmetric and

$$P_T(Z) = W(W^{\mathrm{T}}(A-\lambda B)W - W^{\mathrm{T}}(A^{\mathrm{T}}-\lambda B^{\mathrm{T}})W) + P_H(Z) = 0 + P_H(Z) = P_H(Z)\,. \quad (16)$$

Also, it is easy to prove that the function (7) satisfies the homogeneity condition

$$\begin{aligned} F(WR,\lambda) &= Tr(R^{\mathrm{T}}W^{\mathrm{T}}AWR) - \lambda Tr(R^{\mathrm{T}}W^{\mathrm{T}}BWR) = Tr(WRR^{\mathrm{T}}W^{\mathrm{T}}A) - \lambda Tr(WRR^{\mathrm{T}}W^{\mathrm{T}}B) \\ &= Tr(WW^{\mathrm{T}}A) - \lambda Tr(WW^{\mathrm{T}}B) = Tr(W^{\mathrm{T}}AW) - \lambda Tr(W^{\mathrm{T}}BW) = F(W,\lambda) \end{aligned} \quad (17)$$

The second and fourth steps are derived from the fact that, for any two matrices $M_1 \in \mathbb{R}^{m\times k}$, $M_2 \in \mathbb{R}^{k\times m}$, we have $Tr(M_1M_2) = Tr(M_2M_1)$.

As the solution space is constrained on a Grassmann manifold, the necessary condition for the optimality of the projection matrix is that the projection on the horizontal space at point *W* of the gradient vector $\partial F(W,\lambda)/\partial W = 2(A-\lambda B)W$ is zero, *i.e.*

$$(I - WW^{\mathrm{T}})(AW - \lambda BW) = 0\,. \quad (18)$$

Then, the column vectors of the matrix $AW - \lambda BW$ all lie in the space spanned by the column vectors of *W*, and there exists a matrix $P \in \mathbb{R}^{k\times k}$ satisfying

$$AW - \lambda BW = WP\,. \quad (19)$$

By multiplying W^{T} on the left side of (19), we have

$$W^{\mathrm{T}}AW - \lambda W^{\mathrm{T}}BW = W^{\mathrm{T}}WP = P\,. \quad (20)$$

Therefore, P is a symmetric matrix. Let its singular value decomposition be

$$P = U_p\Lambda_p U_p{}^{\mathrm{T}}\,. \quad (21)$$

Then, there exists a homogeneous solution $W_p = WU_p$ satisfying

$$(A-\lambda B)W_p = W_p\Lambda_p\,. \quad (22)$$

It means that the projection vectors are the eigenvectors of a weighted difference matrix; consequently, we have the following claim.

Theorem. (Necessary condition for the optimal solution) For the trace quotient problem, there exists an optimal solution whose column vectors are the eigenvectors of the corresponding weighted trace difference problem, *i.e.* $(A-\lambda B)W_p = W_p\Lambda_p$.

The above theorem reveals a very interesting point that the trace quotient problem is equal to a properly weighted trace difference problem in objective function. However, these two problems are still different in some aspects. First, for the weighted trace difference problem, such as the work in MMC [8] for discriminant analysis, the solution is directly the leading eigenvectors, while in the trace quotient problem the optimal projection does not always consist of the leading eigenvectors. Secondly, there is no

criterion to guide selection of the weight in the trace difference problem; while in the trace quotient problem the weight can be determined by maximizing the trace quotient, which directly motivates our following procedure to pursue the optimal solution of the trace quotient problem.

3.2 Procedure to Optimal Projection Pursuing

From (22), the optimal solution can be directly determined by Lagrange Multiplier λ; thus we can rewrite the optimal transformation matrix corresponding to λ as $W(\lambda)$. Then, the objective function in (1) is changed to a function only related to λ,

$$G(\lambda) = \frac{Tr(W(\lambda)^{\mathrm{T}} AW(\lambda))}{Tr(W(\lambda)^{\mathrm{T}} BW(\lambda))}. \tag{23}$$

The objective function is nonlinear and it is intractable to directly compute the gradient. However, the experiments show that the objective function is of a single peak, and some plots of the trace quotient distribution with respect to the Lagrange Multiplier λ are plotted in Figure 2. The observations encourage us apply multi-scale search to pursue the optimal weight. The details are listed in procedure-1.

Note that in procedure-1, for each Lagrange Multiplier λ, the column vectors of the optimal projection matrix $W(\lambda)$ are not exactly the leading eigenvectors corresponding to the largest eigenvalues of (22). Thus we utilize a backward elimination method to search for the optimal solution for a given weight parameter, *i.e.* the eigenvector is omitted, if the remaining ones lead to the largest trace quotient, in each step until reduced to the desired feature dimension.

Procedure to pursue optimal solution of trace quotient problem

---

1. Set parameter range: Set a proper parameter range $[a^0, b^0]$ for parameter search. In this work, a^0 is set as *0*, and b^0 is experientially set as the quotient of the largest eigenvalue of *A* and the smallest positive eigenvalue of *B,* which makes most eigenvalues of (22) negative.

2. Multi-scale search: For t = 1, 2, … , $T_{\max}$, Do

a) Segment $[a^{t-1}, b^{t-1}]$ into L parts by $\lambda_i^t = a^{t-1} + (i-1)(b^{t-1} - a^{t-1})/(L-1)$, i=1,.., L.

b) Compute the optimal $W(\lambda_i^t)$ and the corresponding trace quotient Tr_i^t.

c) From the left side, if $\lambda_{i_a}^t$ is the first point satisfying $Tr_{i_a}^t < Tr_{i_{a+1}}^t$ and $Tr_{i_{a+1}}^t \geq Tr_{i_{a+2}}^t$, then set $a^t = \lambda_{i_a}^t$; from the right side, if $\lambda_{i_b}^t$ is the first point satisfying $Tr_{i_b}^t < Tr_{i_{b-1}}^t$ and $Tr_{i_{b-1}}^t \geq Tr_{i_{b-2}}^t$, then set $b^t = \lambda_{i_b}^t$.

d) If $b^t - a^t < \varepsilon$ (= 0.1 in this work), then exit.

3. Output the final optimal solution from (22) by setting $\lambda = (Tr_{i_a}^t + Tr_{i_b}^t)/2$.

Procedure-1. Optimal Solution Pursuing of Trace Quotient Problem

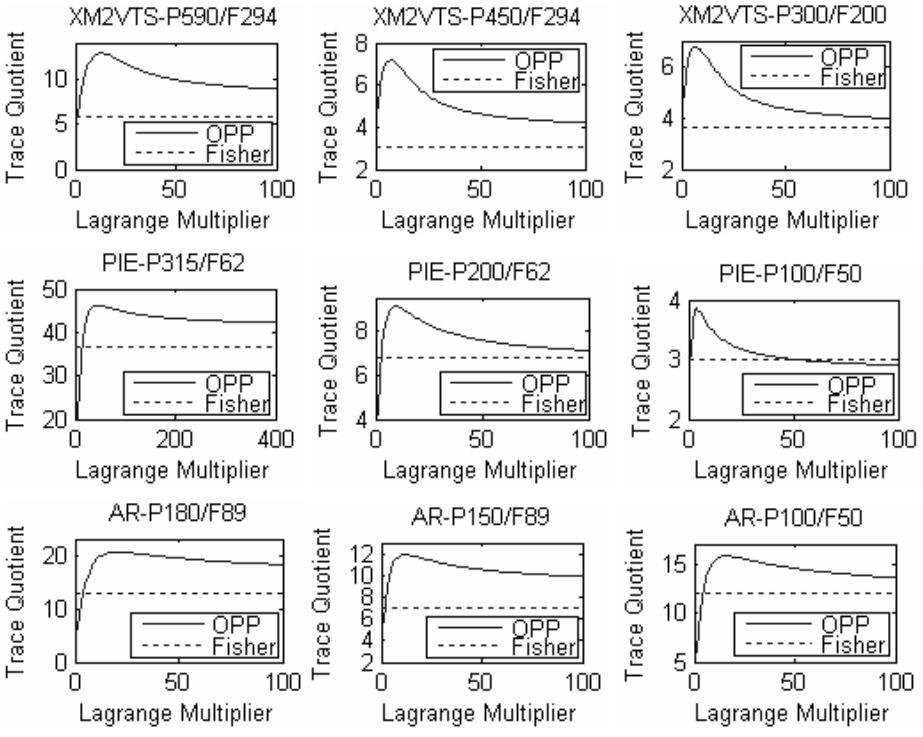

Fig. 2. The trace quotient vs. the Lagrange Multiplier (λ) in the databases XM2VTS, PIE and AR. Note that we plot the trace quotient of PCA+LDA (plotted as *Fisher* in the legends) as a line as its trace quotient value is fixed and free to the Lagrange Multiplier. We can find that the maximum trace quotient of OPP is consistently larger than that from PCA+LDA; while the trace quotient comparison between PCA+LDA and MMC (namely $\lambda = 1$) is not so clear, neither one is consistently better than the other one.

4 Experiments

In this section, three benchmark face databases, XM2VTS [10], CMU PIE [13] and AR [11] are used to evaluate the effectiveness of the proposed procedure to solve the trace quotient problem. The objective function of Linear Discriminant Analysis (2) is applied owing to its popularity; and the new procedure, referred to as OPP (**O**ptimal **P**rojection **P**ursuing), is compared with the popular PCA+LDA [1] and MMC [8], *i.e.* OPP with $\lambda = 1$. In all the experiments, the nearest neighbor method is used as a classifier for final classification based on the Euclidian distance. The trace quotient distributions with respect to the Lagrange Multiplier on the three databases are plotted in Figure 2. The results show that the derived optimal trace quotient from OPP is consistently larger than that from GEVD. In all the experiments, the parameter L in the procedure to pursue the optimal solution of the trace quotient problem is set to 8 and generally we need about three iterations to converge.

XM2VTS Database [10]**:** The XM2VTS database contains 295 persons and each person has four frontal images taken in four sessions. All the images are aligned by fixing the locations of two eyes and normalized in size of 64*64 pixels. In our experiments, we use 295*3 images from the first three sessions for model training; the first session is used as a gallery set and the probe set is composed of the 295 images from the fourth session. Three sets of experiments are conducted to compare the performances of OPP, PCA+LDA and MMC. In each experiment, we use different combinations of Principal Component Analysis (PCA) [7][14] dimension ($N - N_c$, moderate and small number) and final dimension, denoted as P*m*/F*n* in all experiments. Note that actually OPP and MMC need no PCA step, so for a fair comparison with PCA+LDA, PCA is conducted before both OPP and MMC. Table 1 shows the recognition accuracies of the three algorithms. The comparison results show that OPP outperforms MMC and PCA+LDA in all cases.

Table 1. Recognition rates (%) of PCA+LDA, MMC and OPP on XM2VTS database

	P590/F294	P450/F294	P300/F200
PCA+LDA	79.0	75.3	84.4
MMC	83.7	83.4	82.0
OPP	94.2	88.8	88.8

CMU PIE Database [13]**:** The CMU PIE (Pose, Illumination and Expression) database contains more than 40,000 facial images of 68 persons. In our experiment, five near frontal poses (C27, C05, C29, C09 and C07) and illuminations indexed as 08, 10, 11 and 13 are used. 63 persons are used for data incompleteness. Thus, each person has twenty images and all the images are aligned by fixing the locations of two eyes and normalizing to size 64*64 pixels. The data set is randomly partitioned into the gallery and probe sets. Six images of each person are randomly selected for training and also used for the gallery set, and the remaining fourteen images are used for testing. We also conduct three experiments on the PIE database. Table 2 lists the comparison results and it again shows that OPP is consistently superior to the other two algorithms.

Table 2. Recognition rates (%) of PCA+LDA, MMC and OPP on PIE database

	P315/F62	P200/F62	P100/F50
PCA+LDA	88.9	88.0	87.6
MMC	88.1	87.8	85.0
OPP	92.1	94.1	91.6

AR Database [11]**:** The AR face database contains over 4,000 frontal face images of 126 people. We use 90 persons with three images from the first session and another three images from the second session. All the images are aligned by fixing the locations of two eyes and normalizing in size to 72*64 pixels. The data set is randomly partitioned into gallery and probe sets. Three images of each person are randomly selected

Table 3. Recognition rates (%) of PCA+LDA, MMC and OPP on AR database

	P180/F89	P150/F89	P100/F50
PCA+LDA	94.1	90.7	95.2
MMC	78.9	78.5	76.3
OPP	98.2	95.9	96.7

Table 4. Recognition rates (%) of KDA, kernel MMC and OPP on three databases

	XM2VTXS	CMU PIE	AR
KDA	92.2	87.8	94.4
KMMC	86.4	88.3	81.5
KOPP	97.0	93.1	98.5

for training and as the gallery set; and the remaining three images are used for testing. The experimental details are listed in Table 3. The results show that MMC does not obtain satisfactory performance and OPP is the best.

We also apply the OPP algorithm to optimize the objective function of Kernel Discriminant Analysis, compared with the traditional method as reported in [18]. The Gaussian kernel is applied and the final feature dimension is set to $N_c - 1$ in all the experiments. Table 4 lists all the experimental results on the three databases. From the results, we can see that the solution from OPP is much better than the other two algorithms in classification capability.

From the above experimental results, we can have some interesting observations:

1. The quotient value derived from OPP is much larger than that from PCA+LDA and MMC; meanwhile, the comparison between PCA+LDA and MMC is unclear, neither one is consistently superior to the other one.
2. In all the experiments, the recognition rate of OPP is consistently superior to that of PCA+LDA and MMC in all the cases. Similar to the trace quotient value, the performances of PCA+LDA and MMC are comparable.
3. All the results show that the trace quotient criterion is more suitable than the quotient trace criterion for feature extraction owing to its explicit semantics of the numerator and denominator.
4. Recently, many other formulations of matrices A and B in the trace quotient problem were proposed [19]; the advantage of the OPP solution can be easily generalized to these new algorithms.

5 Conclusions

In this paper, we studied the problem of directly solving the trace quotient problem. First, we derived a necessary condition for the optimal solution based on the fact that the feasible solution is constrained to lie on a Grassmann manifold and the final solution is rotation invariant. Then, we presented a procedure to pursue the optimal solution based on the necessary condition. An interesting point is that the necessary condition reveals the underlying equivalence between the trace quotient problem and the

corresponding trace difference problem, which provides theoretical guidance on how to select the optimal weight for the trace difference problem. Moreover, the study of how to pursue a solution on the Grassmann manifold is general, and can be easily extended to optimize general objective functions with solutions constrained on the Grassmann manifold.

Acknowledgement

Here, we would like to thank Chunjing Xu for the valuable discussion and the help to draw the figure 1; also we would like to thank Dong Xu for the help of the experiment section. This work is supported by a joint grant from HKSAR Research Grant Council and Natural Sciences Foundation of China (RGC N_CUHK409/03).

References

[1] P. Belhumeur, J. Hespanha and D. Kriegman. “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, No. 7, 1997, pp. 711-720.

[2] M. Belkin and P. Niyogi. “Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering”, Advances in Neural Information Processing System 15, Vancouver, British Columbia, Canada, 2001.

[3] A. Edelman, T. A. Arias and S. T. Simth. “The Geometry of Algorithms with Orthogonality Constraints,” SIAM J. Matrix Anal. Appl. Vol. 20, No. 2, pp.303-353.

[4] K. Fukunaga, Statistical Pattern Recognition, Academic Press, 1990.

[5] D. Hand. “Kernel Discriminant Analysis”. Research Studies Press, Chichester, 1982

[6] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. "Face Recognition using Laplacianfaces”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 3, Mar. 2005.

[7] I. Joliffe. “Principal Component Analysis”. Springer-Verlag, New York, 1986.

[8] H. Li, T. Jiang, K. Zhang. “Efficient and Robust Feature Extraction by Maximum Margin Criterion”, Advances in Neural Information Processing Systems 16, 2004.

[9] J. Lu, K. N. Plataniotis and N. Venetsanopoulos, “Face Recognition Using Kernel Direct Discriminant Analysis Algorithms”, IEEE Trans. On Neural Networks, Aug. 2002.

[10] J. Luettin and G. Maitre. “Evaluation Protocol for the Extended M2VTS Database (XM2VTS),” DMI for Perceptual Artificial Intelligence, 1998.

[11] A. Martinez and R. Benavente. “The AR Face Database”, http://rvl1.ecn.purdue.edu/~aleix/aleix_faceDB.html, 2003.

[12] A. Shashua and A. Levin. “Linear Image Coding for Regression and Classification using the Tensor-rank Principle”, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Dec. 2001, Hawaii.

[13] T. Sim, S. Baker, and M. Bsat. “The CMU Pose, Illumination, and Expression (PIE) Database”, Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, May, 2002.

[14] M. Turk and A. Pentland. “Face Recognition Using Eigenfaces”, IEEE Conference on Computer Vision and Pattern Recognition, Maui, Hawaii, 1991.

[15] M. Vasilescu and D. Terzopoulos, "Multilinear Subspace Analysis for Image Ensembles", Proc. Computer Vision and Pattern Recognition Conf. (CVPR '03), Vol.2, June, 2003, 93-99.
[16] X. Wang and X. Tang. “A unified framework for subspace face recognition,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, No. 9, 2004, pp. 1222 – 1228.
[17] W. Wilks. “Mathematical Statistics”, Wiley, New York, 1963.
[18] M. Yang, “Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods”, in Proc. of the 5th Int. Conf. on Automatic Face and Gesture Recognition, Washington D. C., May 2002.
[19] S. Yan, D. Xu, B. Zhang and H. Zhang. "Graph Embedding: A General Framework for Dimensionality Reduction", Proc. Computer Vision and Pattern Recognition Conf. 2005.
[20] J. Ye, R. Janardan, C. Park, and H. Park. “An optimization criterion for generalized discriminant analysis on undersampled problems”, IEEE Transactions on Pattern Analysis and Machine Intelligence, V. 26, pp. 982-994, 2004.