Pattern Recognition 44 (2011) 307–319

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Unified formulation of linear discriminant analysis methods and optimal parameter selection

Senjian An [a,]*, Wanquan Liu [a], Svetha Venkatesh [a], Hong Yan [b,c]

[a] Department of Computing, Curtin University of Technology, WA 6102, Australia
[b] Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong
[c] The School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 24 November 2009
Received in revised form
25 July 2010
Accepted 23 August 2010

Keywords:
Linear discriminant analysis
Model selection
Under-sampled problem
Face recognition
Principal component analysis
Constrained ridge regression

ABSTRACT

In the last decade, many variants of classical linear discriminant analysis (LDA) have been developed to tackle the under-sampled problem in face recognition. However, choosing the variants is not easy since these methods involve eigenvalue decomposition that makes cross-validation computationally expensive. In this paper, we propose to solve this problem by unifying these LDA variants in one framework: principal component analysis (PCA) plus constrained ridge regression (CRR). In CRR, one selects the target (also called class indicator) for each class, and finds a projection to locate the class centers at their class targets and the transform minimizes the within-class distances with a penalty on the transform norm as in ridge regression. Under this framework, many existing LDA methods can be viewed as PCA+CRR with particular regularization numbers and class indicators and we propose to choose the best LDA method as choosing the best member from the CRR family. The latter can be done by comparing their leave-one-out (LOO) errors and we present an efficient algorithm, which requires similar computations to the training process of CRR, to evaluate the LOO errors. Experiments on Yale Face B, Extended Yale B and CMU-PIE databases are conducted to demonstrate the effectiveness of the proposed methods.

1. Introduction

Linear discriminant analysis (LDA) [1] is a well-known feature extraction method for classification and has been widely used in many applications such as face recognition and text categorization. LDA aims to maximize the between-class variance and minimize the within-class variance and thus to achieve maximum separability of the training patterns. The classical LDA maximizes the following well-known *Fisher criterion* [1]:

$$\max_G J(G) = \text{trace}\{(G^T \mathbf{S}_w G)^{-1}(G^T \mathbf{S}_b G)\} \tag{1}$$

where G is a linear transformation matrix with a lower dimension, $\mathbf{S}_b$ and $\mathbf{S}_w$ are between-class and within-class scatter matrices, respectively (defined in (2)). The classical LDA requires that the within-class cluster matrix is nonsingular and this limits its application to many classification problems with limited training patterns. For example, in face recognition, due to limited training images, the within-class scatter matrix is usually not of full rank. This is the well-known *small sample size problem* (which is also called *under-sampled problem*). To tackle the small sample size problem, many variants of LDA have been proposed in recent years, such as Fisherface [2], null space LDA [3], LDA/QR [4], LDA/GSVD [5], LDA/FKT [6], Direct LDA and its variants [7–10]. To exploit the nonlinear structure of data such as face images, the kernel versions of these LDA algorithms are also developed in literature [11–18]. In [19], a unified framework is formulated to connect principle component analysis (PCA) [20], LDA and Bayes analysis [21,22]. The recent progress in the investigation of LDA methods includes an eigenfeature regularization method [23] for face recognition and a method based on the Bayes optimality under the Gaussian assumption [24].

In practice, one needs to find the best LDA method and its parameters based on training patterns only. A popular way to estimate the generalization performance of a model is cross-validation [25]. In *l*-fold cross-validation, one divides the data into *l* subsets of (approximately) equal size and trains the classifier *l* times, each time leaving out one of the subsets from training, and using the omitted subset to compute the classification errors. If *l* equals the sample size, this is called leave-one-out cross-validation (LOO-CV). The implementation of LOO-CV of LDA requires $O(n^4)$ computations and is too expensive computationally to be applied in practice, where n is the training size. Recently, [26] developed a unified representation of the LDA variants and addressed the efficient cross-validation problem.

* Corresponding author.
E-mail addresses: S.An@curtin.edu.au (S. An), W.Liu@curtin.edu.au (W. Liu), S.Venkatesh@curtin.edu.au (S. Venkatesh), h.yan@cityu.edu.hk (H. Yan).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.08.026

However, the efficiency is in terms of the number of parameters in the trial instead of the training size. The computational complexity for LOO-CV is still $O(n^4)$.

It is well-known that the LOO errors of ridge regression (RR) and Kernel ridge regression can be computed efficiently [27]. If one can connect the various LDA methods to RR, it may be possible to use the efficient RR cross-validation to estimate the generalization performance of the LDA methods. It is shown in [28] that the classical LDA is equivalent to linear regression with certain class indicator under a strict rank condition $Rank(\mathbf{S}_w)+Rank(\mathbf{S}_b)-Rank(\mathbf{S}_t)=0$ (see definition in (2)). Similarly, the equivalence of canonical correlation analysis (CCA) and least regression, with certain class indicator under a similar rank condition, is established in [29]. However, the connection between ridge regression (linear regression with regularization) and regularized LDA has not been established. In this paper, we propose constrained RR (CRR) and prove that regularized LDA, LDA/GSVD and some other new LDA variants can be connected to CRR. Also we develop an efficient algorithm to compute the LOO errors of CRR and propose to select the LDA variants in practice by comparing the LOO errors (an estimate of generalization performance) of their associated CRR formulations. In CRR, one selects the target (also called class indicator) for each class, and finds a transform to locate the class centers at their class targets, and the transform minimizes the within-class distances with a penalty on the transform norm as in ridge regression (RR) [1]. Under this framework, Fisherface, LDA/QR, direct LDA, null space LDA, LDA/GSVD, LDA/FKT and regularized LDA (RLDA), can be viewed as PCA+CRR with particular regularization numbers and class indicators. Note that the class indicators and regularization number are hyper-parameters which govern the generalization performance of CRR. One can use cross validation to estimate generalization performances and choose the best hyper-parameters and thus choose the best LDA methods. We will present an efficient algorithm to evaluate the leave-one-out (LOO) errors of CRR. The proposed algorithm requires $O(n^3)$ computations, similar to the training process of CRR. Our experimental results demonstrate that the LOO errors can be used effectively to select the LDA variants.

There are two major contributions in this paper. First, a general framework is developed based on constrained ridge regression for multi-category linear discriminant analysis. Under this framework, the classic LDA and many of its recent variants are special cases with specific class indicators. Second, an efficient cross-validation algorithm is developed to evaluate the LOO errors of CRR. The significance of this contribution is that our proposed efficient cross-validation algorithm can be used to choose the optimal LDA variant that corresponds to the CRR member with minimal LOO errors.

The layout of the rest of the paper is as follows. In Section 2, we address the formulation of constrained ridge regression (CRR) and present an efficient algorithm to evaluate its LOO errors. Section 3 addresses the connections of CRR to various LDA algorithms. And Section 4 proposes to use CRR LOO errors to choose the optimal LDA algorithms. The experimental results on Yale Face B and CMU-PIE databases are provided in Section 5 to illustrate the effectiveness of the proposed algorithms.

Notations: We adopt similar notations in [30,5]. $X=[x_1,x_2,\ldots,x_n]\in\mathbb{R}^{d\times n}$ denotes the data matrix where n is the sample size and d is the dimension of the pattern x_i. Assume that we have m classes and we use $l(x)$ to denote the label of pattern x. We use X_k to denote the data matrix of the n_k training patterns for class k and rewrite the data matrix as $X=[X_1,X_2,\ldots,X_m]$. Let $c_k\in\mathbb{R}^d$ denote the centroid (mean) of class k and c be the global centroid. The between-class scatter matrix $\mathbf{S}_b$, the within-class scatter matrix $\mathbf{S}_w$ and the total scatter matrix $\mathbf{S}_t$ are defined as follows:

$$\mathbf{S}_b=\sum_{i=1}^{m} n_i(c_i-c)(c_i-c)^T=H_bH_b^T$$

$$\mathbf{S}_w=\sum_{i=1}^{m}\sum_{x\in\mathcal{X}_i}(x-c_i)(x-c_i)^T=H_wH_w^T$$

$$\mathbf{S}_t=\sum_{i=1}^{n}(x_i-c)(x_i-c)^T=H_tH_t^T \tag{2}$$

where $\mathcal{X}_i$ is the training set of class i, and

$$H_b=\hat{C}\,\mathrm{diag}\{\sqrt{n_1},\sqrt{n_2},\ldots,\sqrt{n_m}\}$$

$$H_w=[X_1-c_1\mathbf{1}_{n_1},X_2-c_2\mathbf{1}_{n_2},\ldots,X_m-c_m\mathbf{1}_{n_m}]$$

$$H_t=X-c\mathbf{1}_n. \tag{3}$$

where

$$\hat{C}=[c_1-c,c_2-c,\ldots,c_m-c] \tag{4}$$

is the class mean matrix of the centered data.

We will use $C=[c_1,c_2,\ldots,c]$ to denote the class mean matrix and use $l(x_i)(\in\{1,2,\ldots,m\})$ to denote the label of x_i. I denotes an identity matrix with dimension of m.

2. Constrained ridge regression

Linear ridge regression (RR) [1] is a classical statistical problem that aims to find a linear function that models the dependencies between covariates $\{x_i\}_{i=1}^n$ and response variables $\{y_i\}_{i=1}^n$ in $\mathbb{R}$. The classical method to solve this problem is the ordinary least squares (OLS) formulation which minimizes the squared loss:

$$\sum_i (y_i-g^Tx_i)^2. \tag{5}$$

Due to limited training examples, the variance of the estimate g by OLS may be large and thus the estimate is not reliable. An effective way to overcome this problem is to penalize the norm of g. Instead of minimizing squared errors, RR minimizes the following cost function:

$$J(g)=\sum_i (y_i-g^Tx_i)^2+\gamma\|g\|^2 \tag{6}$$

where γ is a fixed positive number. By introducing the regularization parameter γ, the ridge regression can reduce the estimate variance at the expense of increasing training errors. The regularization parameter γ controls the trade-off between the bias and variance of the estimate. In practice, one can use cross-validation [25] to find the optimal regularization parameter that minimizes the cross-validation errors.

For multivariate RR, the responses Y_i are vectors in $\mathbb{R}^r$ and the task is then to find a matrix $G\in\mathbb{R}^{d\times r}$ which minimizes

$$\begin{aligned}J(G)&=\sum_i \|Y_i-G^Tx_i\|^2+\gamma\|G\|^2\\&=\mathrm{tr}(YY^T+G^TXX^TG-2G^TXY^T)+\gamma\,\mathrm{tr}(G^TG),\end{aligned} \tag{7}$$

where $Y=[Y_1,Y_2,\ldots,Y_n]$.

Taking derivatives and equaling them to zero, we have the solution

$$G=(XX^T+\gamma I)^{-1}XY^T. \tag{8}$$

2.1. Class indicators

RR is developed to model the dependencies between covariates $\{x_i\}_{i=1}^n$ and response variables y_i $(=f(x_i))$. To apply it to classification problem where only the labels of training patterns are available, we need to design the response variables. Since we prefer to make the similar patterns (from the same class) close, it is desirable to set equal responses for similar training patterns. On the other hand, we also prefer to make dissimilar patterns separate and thus the class indicators are required to be far from each other. For each class, we need to design a desirable response for its patterns. This desirable response is called *class indicator* [1] in the literature. When nothing except the labels are known, one usually chooses $L_i = e_i \in \mathbb{R}^m$ as the class indicator for class $i(=1,2,\ldots,m)$, where e_i is a vector whose i-th entry is 1 and other entries are zeros.

In order to connect to LDA methods, this paper considers the general class indicators where L_i can be any real vector with any dimension to associate with the dimension of the LDA methods. Next, we will propose a new framework called *constrained ridge regression* (CRR). In this framework, the class means' responses are required to equal their class indicators.

2.2. Constrained ridge regression

The main idea of CRR is to make the class means equal to their class indicators in the reduced low-dimensional subspace. Assume that we have the class indicators $\{L_k\}_{k=1}^m$. The task of CRR is to seek a projection to transform the class means $(c_i - c)$ to their class indicators and then minimize the within-class distances with a penalty on the norm of the transform, that is

$$\begin{aligned} \min_G \quad & J(G;\gamma) = \mathrm{tr}\{G^T \mathbf{S}_w G\} + \gamma \mathrm{tr}\{G^T G\} \\ \text{s.t.} \quad & G^T \hat{C} = L \end{aligned} \tag{9}$$

where $L = [L_1, L_2, \ldots, L_m]$ is the class indicator matrix.

From (2) and (3), it follows

$$\mathbf{S}_b = \hat{C}\,\mathrm{diag}\{n_1, n_2, \ldots, n_m\}\hat{C}^T \tag{10}$$

and therefore

$$G^T \mathbf{S}_b G = L\,\mathrm{diag}\{n_1, n_2, \ldots, n_m\}L^T \tag{11}$$

is a constant under the condition $G^T\hat{C} = L$.

Note that the class indicator matrix L is a hyper-parameter (as the regularization number γ is) in CRR.

Eq. (9) is equivalent to the following minimization problem

$$\begin{aligned} \min_G \quad & J(G) = \mathrm{tr}\{G^T \mathbf{S}_w G\} + \mathrm{tr}\{G^T \mathbf{S}_b G\} + \gamma \mathrm{tr}\{G^T G\} \\ & \qquad = \mathrm{tr}\{G^T(\mathbf{S}_t + \gamma I)G\} \\ s.t. \quad & G^T \hat{C} = L. \end{aligned} \tag{12}$$

While (9) and (12) are equivalent, we will use (12) for numerical stability since their solutions involve the inverse of $(\mathbf{S}_w + \gamma I)$ or $(\mathbf{S}_t + \gamma I)$, respectively. For the solution of (12), we have the following result (the proof is delegated to Appendix A):

Theorem 1. *The minimization problem* (12) *has a solution only if*

$$\mathrm{span}(L^T) \subset \mathrm{span}(\hat{C}^T). \tag{13}$$

Assume that $\mathrm{span}(L^T) \subset \mathrm{span}(\hat{C}^T)$ *and* $\mathbf{S}_t$ *is nonsingular, the solution of* (12), *denoted by* G_γ, *is then given by*

$$G_\gamma = \mathbf{S}_\gamma^{-1}\hat{C}(\hat{C}^T S_\gamma^{-1}\hat{C})^\dagger L^T \tag{14}$$

where $\mathbf{S}_\gamma = \mathbf{S}_t + \gamma I$ *is positive definite and* † *denotes pseudo-inverse of matrices.*

Note that one can always make $\mathbf{S}_t$ nonsingular by removing its null space without loss of any discriminant power. From Theorem 1, we know that the selection of the class indicator matrix is limited by the class mean matrix $\hat{C}$. Next, we consider the characterization of all admissible class indicators for a given $\hat{C}$. Let $V \in \mathbb{R}^{m \times r_c}$ be an orthogonal base of $\mathrm{span}\{\hat{C}\}$ where $r_c = rank(C)$. Then $\mathrm{span}(L^T) \subset \mathrm{span}(C^T)$ holds if and only if there exists $M \in \mathbb{R}^{r_c \times r}$ such that

$$L = M^T V^T \tag{15}$$

where r is the dimension of L_i. So the admissible set of L can be characterized as

$$\mathcal{L} = \{L | L = M^T V^T, M \in \mathbb{R}^{r_c \times r}, r \geq 1\}. \tag{16}$$

The optimal selection of M (or an admissible L equivalently) will depend on and this is needed to be learnt from the data. In the next section, we will show that many existing LDA algorithms can be described as CRR with specific regularization number γ and specific class indicators L. The generalization performance of CRR is governed by the hyper-parameters γ and L and one can estimate the generalization performance by cross-validation.

Before we address the relationships between CRR and the LDA methods, we first discuss the efficient evaluation of the leave-one-out cross-validation for CRR. The efficient algorithm is useful to select L and γ for CRR and will be used to choose the LDA variants in the consequent sections.

2.3. Efficient LOO cross-validation

Since L is limited by the class mean matrix $\hat{C}$ which may change when some training patterns are removed, the cross-validation of CRR needs a slight modification in the standard cross-validation. Suppose some training patterns are removed and the new class mean matrix is $\overline{C}$. From (16), the class indicator matrix $\overline{L}$ must satisfy $\overline{L} = \overline{M}^T\overline{V}^T$ where $\overline{V}$ is an orthogonal basis matrix of $\mathrm{span}\{\overline{C}^T\}$. If $\mathrm{span}\{\overline{C}^T\} \neq \mathrm{span}\{\hat{C}^T\}$, $L = M^T V^T$ is not admissible for CRR with these training patterns being removed. In this case, we choose $\overline{M}$ to minimize $\|\overline{L} - L\|$ so that $\overline{L}$ is as closest as possible to L. This results in $\overline{M} = M^T V^T \overline{V}$ and

$$\overline{L} = M^T V^T \overline{V}\,\overline{V}^T. \tag{17}$$

When the original class mean matrix C, where data is not centered yet, is of full column rank, $\hat{C}$ is of rank $(m-1)$ and $\hat{C}[n_1, n_2, \ldots, n_m]^T = 0$. Therefore $\mathrm{span}\{\hat{C}^T\} = null\{[n_1, n_2, \ldots, n_m]\}$ where $null\{a\}$ denotes the null space of a vector a. In the leave-one-out cross-validation where only one training pattern is removed, if n_i is not very small, the spaces $\mathrm{span}\{\hat{C}\}$ $(= null\{[n_1, \ldots, n_m]\})$ and $\mathrm{span}\{\overline{C}\}$ $(null\{[n_1, \ldots, n_i - 1, \ldots, n_m]\})$ will be very close and therefore $\overline{L} \approx L$. In this paper, we only consider the LOO evaluation of CRR. For general l–fold cross-validation, one needs to group the training patterns properly so that the difference between $\mathrm{span}\{\hat{C}^T\}$ and $\mathrm{span}\{\overline{C}^T\}$ is not significant. For example, if each class has equal training size, one needs to group the training patterns so that each class has equal (or approximately equal) training size for each fold.

Let V be an orthogonal base matrix of $\mathrm{span}\{\hat{C}^T\}$. The matrix $\hat{C}V$ is then of full column rank and $V^T\hat{C}^T S_\gamma^{-1}\hat{C}V$ is nonsingular. Note that $L = M^T V^T$, the solution (14) can be described as

$$G_\gamma = \mathbf{S}_\gamma^{-1}\hat{C}V(V^T\hat{C}^T S_\gamma^{-1}\hat{C}V)^{-1}M. \tag{18}$$

Suppose x_k, belonging to class i, is left out for testing. The new global mean $\overline{c}$, the new mean $\overline{c}_i$ of class i, the new total scatter matrix $\overline{\mathbf{S}}_t$ and the new class center matrix $\overline{C}$ become

$$\overline{c}_i = c_i - \frac{x_k - c_i}{n_i - 1}$$

$$\overline{c} = c - \frac{x_k - c}{n-1}$$

$$\overline{\mathbf{S}}_t = \mathbf{S}_t - \frac{n}{n-1}(x_k - c)(x_k - c)^T$$

$$\overline{C} = \hat{C} + \frac{1}{n-1}(x_k - c)\mathbf{1}_m^T - \frac{1}{n_i - 1}(x_k - c_i)e_i^T. \tag{19}$$

Let $\overline{V}$ be an orthogonal base matrix of $\overline{C}^T$ and, from (17), $\overline{L} = MV^T\overline{VV}^T$ is the desired class indicator matrix for training with x_k being excluded. Then the solution of CRR with x_k being removed is then

$$\overline{G}_\gamma = \overline{\mathbf{S}}_\gamma^{-1}\overline{CV}(\overline{V}^T\overline{C}^T\overline{\mathbf{S}}_\gamma^{-1}\overline{CV})^{-1}MV^T\overline{V} \tag{20}$$

where $\overline{\mathbf{S}}_\gamma = \overline{\mathbf{S}}_t + \gamma I$.

Since span$\{\hat{C}^T\}$ and span$\{\overline{C}^T\}$ are very close, we will omit the difference between V and $\overline{V}$, and replace $\overline{V}$ by V in the above equation. Then we have

$$\overline{G}_\gamma = \overline{\mathbf{S}}_\gamma^{-1}\overline{C}V(V^T\overline{C}^T\overline{\mathbf{S}}_\gamma^{-1}\overline{C}V)^{-1}M. \tag{21}$$

Now we derive the formula to compute the predicted response $p(x_k) = \overline{G}_\gamma^T(x_k - \overline{c})$. The formula will be based on $\mathbf{S}_\gamma^{-1}$. Let

$$Q \triangleq \hat{C}^T\mathbf{S}_\gamma^{-1}\hat{C}, \tag{22}$$

$$\overline{x}_k \triangleq \mathbf{S}_\gamma^{-1}(x_k - c), \tag{23}$$

and

$$\hat{x}_k \triangleq \mathbf{S}_\gamma^{-1}(x_k - c_i). \tag{24}$$

Note that

$$\begin{aligned}\overline{\mathbf{S}}_\gamma^{-1} &= (\overline{\mathbf{S}}_t + \gamma I)^{-1}\\ &= \{\mathbf{S}_\gamma - d(x_k - c)(x_k - c)^T\}^{-1}\\ &= \mathbf{S}_\gamma^{-1} + \mu_k\mathbf{S}_\gamma^{-1}(x_k - c)(x_k - c)^T\mathbf{S}_\gamma^{-1}\end{aligned} \tag{25}$$

where

$$d = \frac{n}{n-1}$$

$$\mu_k = \frac{d}{1 - d\beta_k}$$

$$\begin{aligned}\beta_k &= (x_k - c)^T\mathbf{S}_\gamma^{-1}(x_k - c)\\ &= \overline{x}_k^T(x_k - c).\end{aligned} \tag{26}$$

Denote $\overline{Q} = V^T\overline{C}^T\overline{\mathbf{S}}_\gamma^{-1}\overline{C}V$. Then $\overline{G}_\gamma = \overline{\mathbf{S}}_\gamma^{-1}\overline{C}V\overline{Q}^{-1}M$ and the predicted response of $p(x_k)$ satisfies

$$\begin{aligned}p(x_k)^T &= (x_k - \overline{c})^T\overline{G}_\gamma\\ &= (x_k - \overline{c})^T\overline{\mathbf{S}}_\gamma^{-1}\overline{C}V\overline{Q}^{-1}M\\ &= \frac{n}{n-1}(x_k - c)^T\overline{\mathbf{S}}_\gamma^{-1}\overline{C}V\overline{Q}^{-1}M\\ &= \frac{n}{n-1}(x_k - c)^T\mathbf{S}_\gamma^{-1}(I + \mu_k(x_k - c)(x_k - c)^T\mathbf{S}_\gamma^{-1})\overline{C}V\overline{Q}^{-1}M\\ &= \frac{n}{n-1}(1 + \mu_k\beta_k)(x_k - c)^T\mathbf{S}_\gamma^{-1}\overline{C}V\overline{Q}^{-1}M\\ &= \frac{n}{(n-1)(1 - d\beta_k)}(x_k - c)^T\mathbf{S}_\gamma^{-1}\overline{C}V\overline{Q}^{-1}M\\ &= \frac{n}{(n-1)(1 - d\beta_k)}z_k^T\overline{Q}^{-1}M\end{aligned} \tag{27}$$

where

$$\begin{aligned}z_k^T &\triangleq (x_k - c)^T\mathbf{S}_\gamma^{-1}\overline{C}V\\ &= \overline{x}_k^T\overline{C}V\\ &= \overline{x}_k^TCV + \frac{m}{n-1}\beta_k v^T - \frac{1}{n_i - 1}\gamma_k v_i^T\end{aligned}$$

$$\begin{aligned}\gamma_k &= (x_k - c)^T\mathbf{S}_\gamma^{-1}(x_k - c_i)\\ &= \overline{x}_k^T(x_k - c_i)\end{aligned}$$

$$v_i = V^Te_i$$

$$v = V^T\mathbf{1}_m/m. \tag{28}$$

With $\mathbf{S}_\gamma^{-1}$ and $\overline{x}_k$ being ready for use, the computation of β_k,γ_k and z_k can be done in $O(n)$ time. Later, we will show that $z_k^T\overline{Q}^{-1}$ can be computed in $O(m^2)$ time when Q^{-1} is available. Hence, if we compute $\mathbf{S}_\gamma^{-1}$ and Q^{-1} in advance. They can be used for all x_k and the remaining computation of $p(x_k)$ is dominated by the $O(n^2)$ computations of $\overline{x}_k$ in (23).

Now we address the efficient computation of $z_k^T\overline{Q}$. Note that

$$\begin{aligned}\overline{Q} &= V^T\overline{C}^T\overline{\mathbf{S}}_\gamma^{-1}\overline{C}V\\ &= V^T\overline{C}^T\mathbf{S}_\gamma^{-1}\overline{C}V + \mu_kV^T\overline{C}^T\mathbf{S}_\gamma^{-1}(x_k - c)(x_k - c)^T\mathbf{S}_\gamma^{-1}\overline{C}V\\ &= V^T\overline{C}^T\mathbf{S}_\gamma^{-1}\overline{C}V + \mu_kQz_kz_k^TQ\\ &= Q + \mu_kz_kz_k^T + \frac{m}{n-1}V^T\mathbf{1}_m(x_k - c)^T\mathbf{S}_\gamma^{-1}\overline{C}V\\ &\quad + \frac{1}{n-1}V^TC\mathbf{S}_\gamma^{-1}(x_k - c)\mathbf{1}_m^TV\\ &\quad - \frac{1}{n_i - 1}V^Te_i(x_k - c_i)^T\mathbf{S}_\gamma^{-1}\overline{C}V - \frac{1}{n_i - 1}V^TC\mathbf{S}_\gamma^{-1}(x_k - c_i)e_i^TV\\ &= Q + \mu_kz_kz_k^T + \frac{m}{n-1}vz_k^T + \frac{m}{n-1}Qy_kv^T\\ &\quad - \frac{1}{n_i - 1}v_i\hat{w}_k^T - \frac{1}{n_i - 1}w_kv_i^T\end{aligned} \tag{29}$$

where

$$\begin{aligned}y_k^T &= (x_k - c)^T\mathbf{S}_\gamma^{-1}CVQ^{-1}\\ &= \overline{x}^TCVQ^{-1}\end{aligned}$$

$$\begin{aligned}w_k &= V^TC^T\mathbf{S}_\gamma^{-1}(x_k - c_i)\\ &= V^TC^T(\hat{x})_k\end{aligned}$$

$$\begin{aligned}\hat{w}_k &= V^T\overline{C}\mathbf{S}_\gamma^{-1}(x_k - c_i)\\ &= w_k + \frac{m}{n-1}\gamma_kv - \frac{1}{n_i - 1}\delta_kv_i\end{aligned}$$

$$\delta_k = (x_k - c_i)^T\hat{x}_k. \tag{30}$$

Let $a_1 = \mu_kz_k + m/(n-1)v$, $b_1 = z_k$, $a_2 = -1/(n_i - 1)v_i$, $b_2 = \hat{w}_k$, $a_3 = -1/(n_i - 1)w_k$, $b_3 = v_i$, $a_4 = m/(n-1)Qy_k$, $b_4 = v$ and define $Q_0 \triangleq Q$,

$$Q_1 \triangleq Q_0 + a_1b_1^T$$

$$Q_2 \triangleq Q_1 + a_2b_2^T$$

$$Q_3 \triangleq Q_2 + a_3b_3^T$$

$$Q_4 \triangleq Q_3 + a_4 b_4^T \tag{31}$$

Note that $\overline{Q} = Q_4$ and

$$Q_i^{-1} = Q_{i-1}^{-1} - (1 + b_i^T Q_{i-1} a_i)^{-1} Q_{i-1}^{-1} a_i b_i^T Q_{i-1}^{-1} \tag{32}$$

$z_k^T \overline{Q}^{-1}$ can be computed using (32) by the following four steps where each step uses the terms obtained from the previous step:

- Compute $z_k^T Q^{-1}$,$a_i^T Q^{-1}$ and $b_i^T Q^{-1}$ for i=1,2,3,4;
- Compute $z_k^T Q_1^{-1}$,$a_i^T Q_1^{-1}$ and $b_i^T Q_1^{-1}$ for i=2,3,4;
- Compute $z_k^T Q_2^{-1}$,$a_i^T Q_2^{-1}$ and $b_i^T Q_2^{-1}$ for i=3,4;
- Compute $z_k^T Q_3^{-1}$, $a_i^T Q_3^{-1}$ and $b_i^T Q_3^{-1}$ for i=4;
- Compute $z_k^T Q_4^{-1}$. Note that $\overline{Q} = Q_4$.

We only need to compute the inverse Q^{-1} (that is Q_0^{-1} by definition). When Q_0^{-1} and a_i,b_i, i=1,2,3,4 are ready, we can compute $z_k^T Q^{-1}$ in $O(m^2)$ time. Then based on $z_k^T Q^{-1}$, k=1,2,3,4 and using (32), we can compute $z_k^T Q_1^{-1}$, k=2,3,4 in $O(m^2)$ time. Sequentially, one can get $z^k \overline{Q}^{-1}$ in $O(m^2)$ time. The computations of a_i,b_i are based on (30) which requires $O(n)$ computations if $\hat{x}_k$ is available. Hence the computation of $z_k^T \overline{Q}$ is dominated by that of $\hat{x}_k$ which requires $O(n^2)$ computations. In summary, $p(x_k)$ can be computed in $O(n^2)$ time when $\mathbf{S}_\gamma^{-1}$ and Q^{-1} are computed in advance.

Given the total scatter matrix $\mathbf{S}_t$ and the class mean matrix $\hat{C}$, the procedure to compute the LOO errors of CRR can be summarized as follows:

1. Compute V via SVD on $\hat{C}^T \hat{C} = V^T \Sigma V$;
2. Compute $\mathbf{S}_\gamma^{-1}$, $Q = V^T C^T \mathbf{S}_\gamma^{-1} C V$ and Q^{-1};
3. For each training pattern x_k, compute the predicted response $p(x_k)$ from (27) based on (23) and (24), predict the label of pattern x_k by comparing the distances of $p(x_k)$ to all training examples other than x_k and compare the predicted label with the true label to decide if there is an error;
4. Sum up all errors and compute the error rate.

The first step requires $O(m^3)$ computations while the second and third steps require $O(n^3)$ computations. The computations are dominated by that of computing $\mathbf{S}_\gamma^{-1}$ and $\overline{x}_k$,$\hat{x}_k$ in (23) and (24). Hence the computational complexity for LOO errors of CRR is cubic and about three times of CRR training.

3. Connections to LDA methods

In this section, we summarize some popular LDA variants for under-sampled problems and show their connections to CRR. The connections are made as follows: given the solution of some LDA variant, we find the class indicator L and regularization number for CRR so that its solution (14) is equal to the solution of that LDA variant.

3.1. Regularized LDA

Regularization is a classical method to deal with ill-posed problems and it can be used to solve the singularity problem in under-sample discriminant analysis. The method is called regularized LDA (RLDA). RLDA adds γI on $\mathbf{S}_w$ and then applies the classic LDA. It is equivalent to optimize the following *regularized Fisher criterion*

$$\max_G J(G) = \text{trace}\{(G^T(\mathbf{S}_w + \gamma I)G)^{-1}(G^T \mathbf{S}_b G)\} \tag{33}$$

or equivalently

$$\max_G J(G) = \text{trace}\{(G^T(\mathbf{S}_t + \gamma I)G)^{-1}(G^T \mathbf{S}_b G)\} \tag{34}$$

where γ is usually a small positive number.

RLDA uses the large non-zero eigenvectors of $(\mathbf{S}_w + \gamma I)^{-1}\mathbf{S}_b$ or, equivalently, of $(\mathbf{S}_t + \gamma I)^{-1}\mathbf{S}_b$, as the projection matrix G_{RLDA}. In this case, one needs to find a suitable regularization parameter γ which is critical for the successful application of RLDA. Note that the solutions of either (33) or (34) are not unique. They are equivalent in the sense that their solutions are identical under same scale constraint, say the unit norm constraint on the eigenvectors. We suggest (34) instead of (33) since computing $(\mathbf{S}_t + \gamma I)^{-1}$ is more stable numerically than computing $(\mathbf{S}_w + \gamma I)^{-1}$. In particular when $\gamma = 0$, for the solution of (33), one needs to include the null eigenvectors of $\mathbf{S}_w$ and the leading eigenvectors of $\mathbf{S}_w^\dagger \mathbf{S}_b$. In this case, (33) and (34) are still equivalent under same scale constraint.

Now let G_{RLDA}=[g_1, g_2,..., g_r] be the solution of RLDA. That is, g_1, g_2,..., g_r are the r leading eigenvectors of $\mathbf{S}_\gamma^{-1}\mathbf{S}_b$. From (2) and (3), we have

$$\mathbf{S}_b = \hat{C}\,\text{diag}\{n_1, n_2, \ldots, n_m\}\hat{C}^T \tag{35}$$

and therefore

$$\begin{aligned} \text{span}\{\mathbf{S}_\gamma^{-1}\mathbf{S}_b\} &= \text{span}\{\mathbf{S}_\gamma^{-1}\hat{C}\} \\ &= \text{span}\{\mathbf{S}_\gamma^{-1} U_c \Sigma_c V_c^T\} \\ &= \text{span}\{\mathbf{S}_\gamma^{-1} U_c\} \end{aligned} \tag{36}$$

where $U_c \Sigma_c V_c^T$ is the singular value decomposition of $\hat{C}$. Let H_γ be an orthogonal basis of span$\{\mathbf{S}_\gamma^{-1} U_c\}$. Note that the eigenvectors

$$g_i \in \text{span}\{\mathbf{S}_\gamma^{-1}\mathbf{S}_b\} = \text{span}\{H_\gamma\}. \tag{37}$$

There exists M such that $G_{RLDA} = H_\gamma M$.

Substituting $\hat{C} = U_c \Sigma_c V_c^T$ into (14), we have the CRR's solution

$$G_\gamma = \mathbf{S}_\gamma^{-1} U_c (U_c^T \mathbf{S}_\gamma^{-1} U_c)^{-1} \Sigma_c^{-1} V_c^T L^T \tag{38}$$

and therefore, there exists a nonsingular matrix Q such that

$$G_\gamma = H_\gamma Q V_c^T L^T. \tag{39}$$

Hence if we choose

$$L = M^T Q^{-T} V_c^T \tag{40}$$

Then we have $QV_c^T L^T$=M and thus $G_\gamma = G_{RLDA}$. That is, CRR is equivalent to RLDA if we choose the class indicator as in (40).

3.2. Fisherface [2]

In order to apply LDA for face recognition where the within-class scatter matrix $\mathbf{S}_w$ is usually singular, Fisherface was proposed to conduct PCA first on the data matrix so that the within-class scatter matrix $\hat{\mathbf{S}}_w$ is nonsingular in the dimension-reduced subspace wherein classical LDA can be applied. Fisherface is also called PCA+LDA. The within-class variance matrix is usually nonsingular if we reduce the data dimensionality to be equal or less than $(n-m+1)$. However, applying PCA to reduce the data dimensionality may lose discriminative information. Also, our experiments demonstrate that the performance of Fisherface is sensitive to the dimensions of PCA and it is critical to find a suitable dimension of PCA for Fisherface. Fisherface is the classical LDA (or RLDA with $\gamma = 0$) with PCA processing. Hence Fisherface is equivalent to PCA preprocessing followed by CRR with $\gamma = 0$ and the class indicators being chosen as in last section (with $\gamma = 0$).

3.3. LDA/GSVD [5] and LDA/FKT [6]

While Fisherface, LDA/QR, DLDA and NLDA acquired the matrix nonsingularity by reducing the data space, which may lose some discriminant power, LDA/GSVD was proposed to gain the full discriminant power. LDA/GSVD finds a linear transform to optimize the *generalized Fisher criterion* [5]:

$$\max_{G} J(G) = \mathrm{trace}\{(G^T\mathbf{S}_tG)^{-1}(G^T\mathbf{S}_bG)\}. \tag{41}$$

which is equivalent to (1) when $\mathbf{S}_w$ is nonsingular (note that $\mathbf{S}_t{=}\mathbf{S}_b{+}\mathbf{S}_w$).

The singularity problem is solved by using generalized singular value decomposition (GSVD). Based on the GSVD theory, there exist orthogonal matrices $Y \in \mathbb{R}^{m\times m}, Z \in \mathbb{R}^{n\times n}$ and a nonsingular matrix $W \in \mathbb{R}^{d\times d}$ such that

$$Y^TH_b^TW = [\Sigma_b,\mathbf{0}]$$

$$Z^TH_w^TW = [\Sigma_w,\mathbf{0}] \tag{42}$$

where

$$\Sigma_b = \begin{bmatrix} I_b & & \\ & D_b & \\ & & 0_b \end{bmatrix}, \quad \Sigma_w = \begin{bmatrix} 0_w & & \\ & D_w & \\ & & I_w \end{bmatrix} \tag{43}$$

The matrices $I_b \in \mathbb{R}^{(r_t-r_w)\times(r_t-r_w)}, I_w \in \mathbb{R}^{(r_t-r_b)\times(r_t-r_b)}$ are identity matrices, $\mathbf{0}, 0_b, 0_w$ are matrices with all elements equal to zero. D_b, D_w are diagonal matrices with decreasing and increasing diagonals, respectively, and $\Sigma_b^2+\Sigma_w^2$ is an identity matrix.

The transformation matrix G of LDA/GSVD is composed by the leading columns of W. These leading columns are the generalized singular vectors of the matrix pair (H_b^T, H_w^T). It is easy to check that

$$G^T\mathbf{S}_tG = I. \tag{44}$$

This indicates that *LDA/GSVD* scales the eigenvector by whitening the total variance matrix $\mathbf{S}_t$.

Since the null space of $\mathbf{S}_t$ does not include any discriminative information, one can always remove its null space so that the total variance matrix is nonsingular. Now assume that $\mathbf{S}_t$ is nonsingular. Note that Y and Z are square orthogonal matrices. From (42), we have $W^T\mathbf{S}_bW$ and $W^T\mathbf{S}_wW$ are diagonal and therefore $W^T\mathbf{S}_tW$ is diagonal as well. Hence, the columns of W are the eigenvectors of $\mathbf{S}_t^{-1}\mathbf{S}_b$. That is, LDA/GSVD uses the leading eigenvectors of $\mathbf{S}_t^{\dagger}\mathbf{S}_b$ and scales the eigenvectors with (44). Therefore, LDA/GSVD is RLDA with a novel scale constraint (44) and it can be connected to CRR similarly as RLDA.

It is proven in [6] that GSVD and Fukunaga–Koontz Transform (FKT) are equivalent on LDA and proposes the less computationally expensive LDA/FKT method than LDA/GSVD. Since LDA/FKT has the same solution as LDA/GSVD and thus it is connected to CRR with the same parameters as well.

3.4. Null space LDA [3]

Based on the intuition that the null space of $\mathbf{S}_w$ is the most discriminant subspace and the null space of $\mathbf{S}_t$ does not include any discriminative information, null space LDA (NLDA) first removes the null space of $\mathbf{S}_t$ and then projects the data into the null space of $\mathbf{S}_w$. Note that $\mathbf{S}_t{=}\mathbf{S}_w{+}\mathbf{S}_b$. The null eigenvectors of $\mathbf{S}_w$ are exactly the same set of eigenvectors of $\mathbf{S}_t^{\dagger}\mathbf{S}_b$ with eigenvalue equal to 1. NLDA is actually a special case of LDA/GSVD when using only the eigenvectors of eigenvalue 1 and scaling them to be unit norm. Therefore NLDA can be connected to CRR similarly as LDA/GSVD.

3.5. LDA/QR [4] and direct LDA [8]

In [4], a two-stage LDA method called LDA/QR is proposed to overcome the singularity problems and to achieve computational efficiency. The first stage applies QR decomposition on $\mathbf{H}_b$: $\mathbf{H}_b{=}QR$, where Q has orthogonal columns that spans the space span($\mathbf{H}_b$) and R is an square upper triangular matrix. By projecting the data with transform Q, the between-class and within-class scatter matrices become $\overline{\mathbf{S}}_b = Q^T\mathbf{S}_bQ$ and $\overline{\mathbf{S}}_w = Q^T\mathbf{S}_wQ$. At the second stage, we apply LDA on $\overline{\mathbf{S}}_b$ and $\overline{\mathbf{S}}_w$, and keep the eigenvectors associated with the smallest eigenvalues of $\overline{\mathbf{S}}_b^{-1}\overline{\mathbf{S}}_w$. LDA/QR achieves computational efficiency at the cost that the linear transform is constrained to be in the class mean subspace span($\mathbf{H}_b$).

Direct LDA (DLDA) also conducts LDA within the centroid space span($\mathbf{H}_b$). The difference is that DLDA uses the eigenvectors which whitens the within-class scatter matrix, while LDA/QR uses the unit norm eigenvectors.

LDA/QR conducts classical LDA in the class mean subspace. That is, LDA/QR chooses the r leading eigenvectors, denoted by $M{=}[M_1,M_2,\ldots,M_r]$, of

$$(H_c^T\mathbf{S}_tH_c)^{-1}(H_c^T\mathbf{S}_bH_c) \tag{45}$$

and the projection matrix is H_cM where the eigenvectors are of unit norm.

Note that H_cR is the QR decomposition of $\hat{C}$. R is of full row rank and thus $R = \Delta V^T$ for some orthogonal matrix V and nonsingular $\varDelta$.

From Theorem 1, if we choose $\gamma = +\infty$, we have

$$\begin{aligned} G_\gamma &= \hat{C}(\hat{C}^T\hat{C})^{\dagger}L^T \\ &= H_cR(R^TR)^{\dagger}L^T \\ &= H_c\varDelta^{-T}V^TL^T. \end{aligned} \tag{46}$$

Hence if we choose the regularization number $\gamma = +\infty$ and the class indictor $L = M^T\Delta V^T$, then CRR is equivalent to LDA/QR. It will be equivalent to DLDA if we scale the eigenvectors such that $M_i^TQ^T\mathbf{S}_wQM_i{=}1$.

3.6. The unifying view on the connections

In this section, we present an unifying view on the connections of the LDA methods to CRR and this is useful for CRR to connect other similar methods. From the connections, we can view CRR and the various LDA methods in two stages. In the first stage, both CRR and the LDA methods find the maximal discriminant subspace $\mathcal{H}_\gamma = \mathrm{span}(\mathbf{S}_\gamma^{-1}\hat{C})$. In the second stage, the LDA methods find the transform by applying RLDA (with some regularization number) in the space $\mathcal{H}_\gamma$, and CRR finds the transform in $\mathcal{H}_\gamma$ by designing some class indicator. Since they are done in the same subspace, CRR and the LDA methods can be connected. Note that the regularization numbers may be different for some LDA methods, say LDA/QR, in the two stages. We denote them by γ_1 and γ_2, respectively. Also, the class indicators connected to the LDA methods depends on three types of scale constraints (SCs) for the eigenvectors associated with the LDA methods:

1. *Type* 1: unit norm,
2. *Type* 2: $M_i^T\mathbf{S}_tM_i{=}1$
3. *Type* 3: $M_i^T\mathbf{S}_wM_i{=}1$

To reduce the computational complexity or remove noise, one may also apply PCA before CRR. Based on the above two-stages view, we show the connections of CRR to the popular LDA variants as below.

1. Fisherface: There is no regularization ($\gamma_1 = \gamma_2 = 0$). PCA is applied to completely remove the effect of small eigenvectors of the data matrix. The dimensionality of PCA is crucial for the performance of Fisherface. Type 1 scale constraint is usually applied.
2. RLDA: The two regularization numbers are equal ($\gamma_1 = \gamma_2$). One usually applies type 1 scale constraint. The selection of the regularization numbers is crucial for the performance of RLDA.
3. LDA/GSVD: There is no regularization ($\gamma_1 = \gamma_2 = 0$) and no PCA. Type 2 scale constraint is recommended.
4. Null space LDA: there is no regularization. Type 1 scale constraint is recommended. The dimension is chosen as that of the null space of $\mathbf{S}_w$;
5. LDA/QR: The first stage regularization number is $\gamma_1 = +\infty$ and the second stage $\gamma_2 = 0$. Type 1 scale constraint is recommended.
6. D-LDA: The first stage regularization number is $\gamma_1 = +\infty$ and the second stage $\gamma_2 = 0$. Type 3 scale constraint is recommended.

Note that PCA can be used to remove noise or reduce the computational complexity in other LDA algorithms beside Fisherface and one can develop new type of LDA algorithms by choosing new parameters on two regularization numbers and the scale constraints. The importance of PCA in LDA for face recognition, is discussed recently in [31].

3.7. Comparison to the unification in [26]

The unification problem of various LDA methods is addressed recently in [26]. The key idea of [26] is to unify different LDA algorithms using a simple transfer function $\phi(\lambda_i)$. They have shown how $\phi(\lambda_i)$ can be chosen for different LDA algorithms. However, in their work, the transfer function is used more or less for a symbolic or representation purpose. That is, this function describes how the singular values are modified in existing algorithms. So [26] provides a unified representation, but not a unified cost function or optimization procedure to derive existing algorithms or develop new ones. Our unified framework is based on a novel method *constrained ridge regression* and we have a unified cost function and optimization procedure. The various LDA methods are reformulated as CRR with specific regularization numbers and specific class indicators.

In [26], the efficient model selection algorithm is addressed for PCA dimensionality selection and RLDA regularization number selection. However, the efficiency is related to the number of parameter grids instead of the training size n. The computational complexity is still $O(n^4)$ for leave-one-out cross-validation. Our efficient LOO algorithm for CRR is of complexity $O(n^3)$ and we propose to select the best LDA method using this algorithm in the next section. Furthermore, our framework can also be used to develop new algorithms by designing new class indicators.

3.8. Remarks

Although most of the LDA variants can be reformulated in the CRR framework, the discriminant analysis methods in [23,31] cannot be formulated as CRR. The LDA methods which can be formulated in CRR framework share the same property that they try to minimize the within-class variance in some way. However, the main idea of [23] is to approximately whiten the within-class variance and then maximize the between-class variance. There is no direct task to minimize the within-class variance here and thus it cannot be formulated as CRR. It is interesting to note that [23] aims to maximize the between-class distance by fixing (approximately whitening) the within-class variance while CRR aims to minimize the within-class variance by fixing the class-centers (and thus the between-class variance as well). Similarly, the idea of [31] is to regularize the eigenvalues but there is no direct effort to minimize the within-class variance and thus it is not be related to CRR.

4. Choosing LDA methods by CRR cross-validation

In the preceding section, we have shown that the various LDA methods can be unified in a general framework PCA+CRR. Each method corresponds to CRR with a specific regularization number $\lambda^{(k)}$ and a specific class indicator matrix $L^{(k)}$. Thus, the various LDA methods can be treated as members of the CRR family and the optimization of the LDA methods is then a problem to choose the best member from the CRR family. The generalization performance of CRR can be estimated by the LOO errors and we propose to choose the best member with the minimal LOO errors. The LOO errors can be computed using the efficient algorithm we proposed in Section 2.3. Also, we treat the LDA methods with different dimensions (including dimensions in the preprocessing stage PCA), different regularization numbers and different scale constraints, as different members in the CRR family and thus the parameters can be optimized similarly.

In summary, we can optimize the dimension of PCA, the regularization numbers, and type of scale constraints corresponding to various LDA variants by minimizing the LOO error rates of their corresponding member in CRR. Since the scaling of $\mathbf{S}_t$ affects the regularization number selection, we scale $\mathbf{S}_t$ so that its trace is equal to the data dimension in our experiments.

Note that the naive implementation of each LDA method repeats the training n times, requires $O(n^4)$ computations and thus is much more expensive computationally . It should also note that the cross-validation errors of CRR is an approximate estimate on the LDA cross-validation errors.

As shown in Section 2, the major task involved in the LOO errors estimation of CRR is the computation of $\mathbf{S}_\gamma^{-1}$ and $\overline{x}_k,\hat{x}_k$ in (23) and (24). Hence the computational complexity of choosing the LDA methods is cubic and is about three times of the sum of the computations corresponding to the training of the LDA methods.

5. Experimental results

In this section, we report the experimental results on Yale Face Database B (YaleB) [32,33] and CMU PIE [34,35] databases. The LDA algorithms are implemented in our PCA+CRR framework. To void numerical instability of matrix pseudo-inverse, we removed the null space of the data matrix so that $\mathbf{S}_\gamma$ is invertible when $\gamma = 0$. The associated class indicators are computed using the formula provided in Section 3 and these class indicators are data-dependent. Note that CRR is a tool to unify and optimize LDA methods and their parameters. The first experiment will test the effectiveness of using CRR LOO errors to choose the LDA methods and their parameters. The second experiment compares the optimal LDA method with other two popular linear projection methods Laplacianfaces (LPP) [36] and orthogonal Laplacianfaces (OLPP) [37]. Note that the subspace methods are used to extract the features in face recognition and nearest neighbor method is followed to identify the labels of the testing images in our experiments.

5.1. Experiment 1: The effectiveness of using CRR LOO errors to choose LDA methods and their parameters

In this experiment, in addition to test our algorithms, we also want to show LDA methods can work well for face recognition under complex viewing conditions (e.g. various poses and lighting conditions). We first conduct experiments on the original Yale Face database B [32] and the extended Yale Face Database B [33] which contain images of 10 and 28 human subjects, respectively, where each person has 576 images taken under nine poses and 64 illumination conditions (see Figs. 1–2 for an example of pose and lighting variations). We found that most of the excluded bad images in the studies [32,33,37] are identifiable after histogram equalization [38] and we include all the images from all 38 human subjects in our experiment. We manually find the positions of eyes and mouths for each person under each pose, and then align and crop all the images according to these positions. Then all the images are re-sized to 32×32 images and preprocessed by histogram equalization.

We will test the performance using all the 576 images of each person under nine poses and 64 lighting conditions. Since the variations of poses and lighting are very large, we cannot expect to achieve good performance by randomly select a small number of images from each person as training images. In order to achieve good performance with a small number of training images, we introduce a new concept *critical viewing conditions*. Consider the images of a person under a given set of viewing conditions (eg. pose and lighting). We can use the clusters of these images to describe their structures. We find these cluster centers by applying the well-known affinity propagation method [39]. Each cluster center is a real image which corresponds to certain viewing conditions. We call these viewing conditions critical since they produce the cluster centers of the image set. Furthermore, we believe the critical viewing conditions are similar for different persons. So we propose to estimate the critical viewing conditions by using the images of a small number of people and apply them on others. The original Yale Face database B and the extended Yale Face Database B are ideal to test this idea since the images are taken under same viewing condition structure (nine poses, 64 lighting conditions) for each person. In our experiments, we use the images of 10 people in the original Yale B to estimate the critical viewing conditions. Latter we will show that good performance can be achieved using 35 critical viewing conditions and this demonstrate that the images of people do share similar critical viewing conditions. Our procedure is as follows: First, we choose 10 people in the original Yale B database to understand the critical configurations for varying pose and lighting conditions. We do this by applying affinity propagation clustering [39] and find a universal configuration of 35 cluster centers of the total 576 images for each person. (The 35 cluster-central images from one subject are shown in Fig. 3). These 35 cluster centers represent 35 critical viewing conditions among nine poses and 64 lighting conditions. We consider these 35 critical viewing conditions are universal for each person in the full Yale B database. Next, for each subject, we take the 35 images associated with these 35 critical viewing conditions as training images to train the classifiers. Finally, we compare the distances of the test images to all the training images, identify them using the nearest neighbor method and then compute the test error rate. The classifier is trained on 38×35 images and the test is conducted on the remaining $38\times(576\text{-}35)$ images.

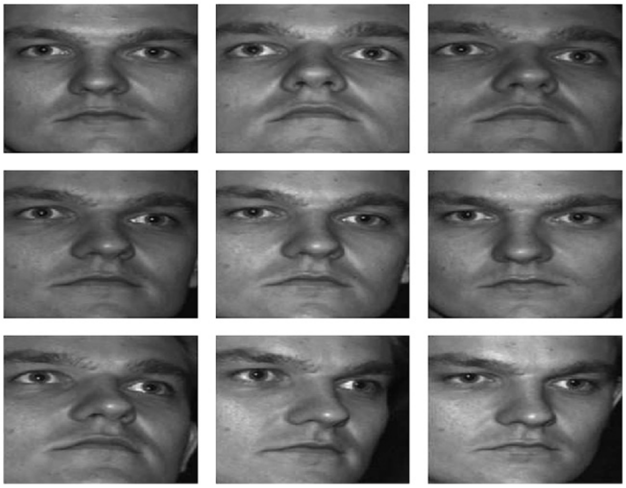

Fig. 1. The images of one person under nine poses with front lighting.

5.1.1. Dimensionality selection in PCA for Fisherface

In Fisherface, PCA is used to preprocess the data so that the within-class variance matrix is nonsingular. One can simply choose the dimension to be $(m-1)$ less than the number of training patterns and the nonsingularity requirement will be satisfied in most cases. However, our experiments demonstrate that this may not be the optimal choice.

Fig. 4 compares the LOO error rates and the error rates on the testing images. Note that the LOO errors are computed based only on the 35 training images from each person while the test error rate is the error rate of the left 541 images for each person. One can see that they fit quite well. The minimal LOO error rate and the error rate on testing images are achieved at dimension 300. It shows the effectiveness of choosing optimal dimension by comparing LOO errors.

With the selected dimension (300) by comparing LOO errors, Fisherface achieves a low error rate 4.43% which is much lower than 21%, the error rate of Fisherface without PCA (which is actually the classical LDA without any preprocessing and regularization).

It is surprising that the optimal performance of Fisherface is achieved with a significantly reduced dimension in the PCA stage.

5.1.2. Regularization number and dimensionality selection in RLDA

For regularization numbers, we applied a transform $\lambda=1/(1+\gamma)$ so that the range of λ is [0,1] is finite. Note that $\lambda=1$ when $\gamma=0$. Fig. 5 compares the LOO and test error rates under varying regularized number and varying projection dimensions. The LOO and test errors are highly related. By minimizing the LOO errors, we found $\lambda=0.96$. The performance of optimized RLDA (OptRLDA) and optimized Fisherface (OptFisher) are reported in Table 1 with a comparison to other LDA variants, which clearly demonstrate the advantages of OptRLDA and OptFisher. The LOO error rate of OptRLDA is much smaller than those of other variants and it shows LOO error rates can also be used to select the variants.

Table 2 shows the performances of the optimal CRR with class indicators derived from LDA/QR, D-LDA,LDA/GSVD and NLDA. Comparing with the performance of LDA/QR, D-LDA,LDA/GSVD and NLDA in Table 1, the performances have been improved significantly by properly optimizing the regularization numbers. Hence, from the perspective of CRR, the bad performances of LDA/GSVD and NLDA are due to no regularization ($\gamma=0$) while the bad performances of LDA/QR and D-LDA are due to over regularization ($\gamma=+\infty$).

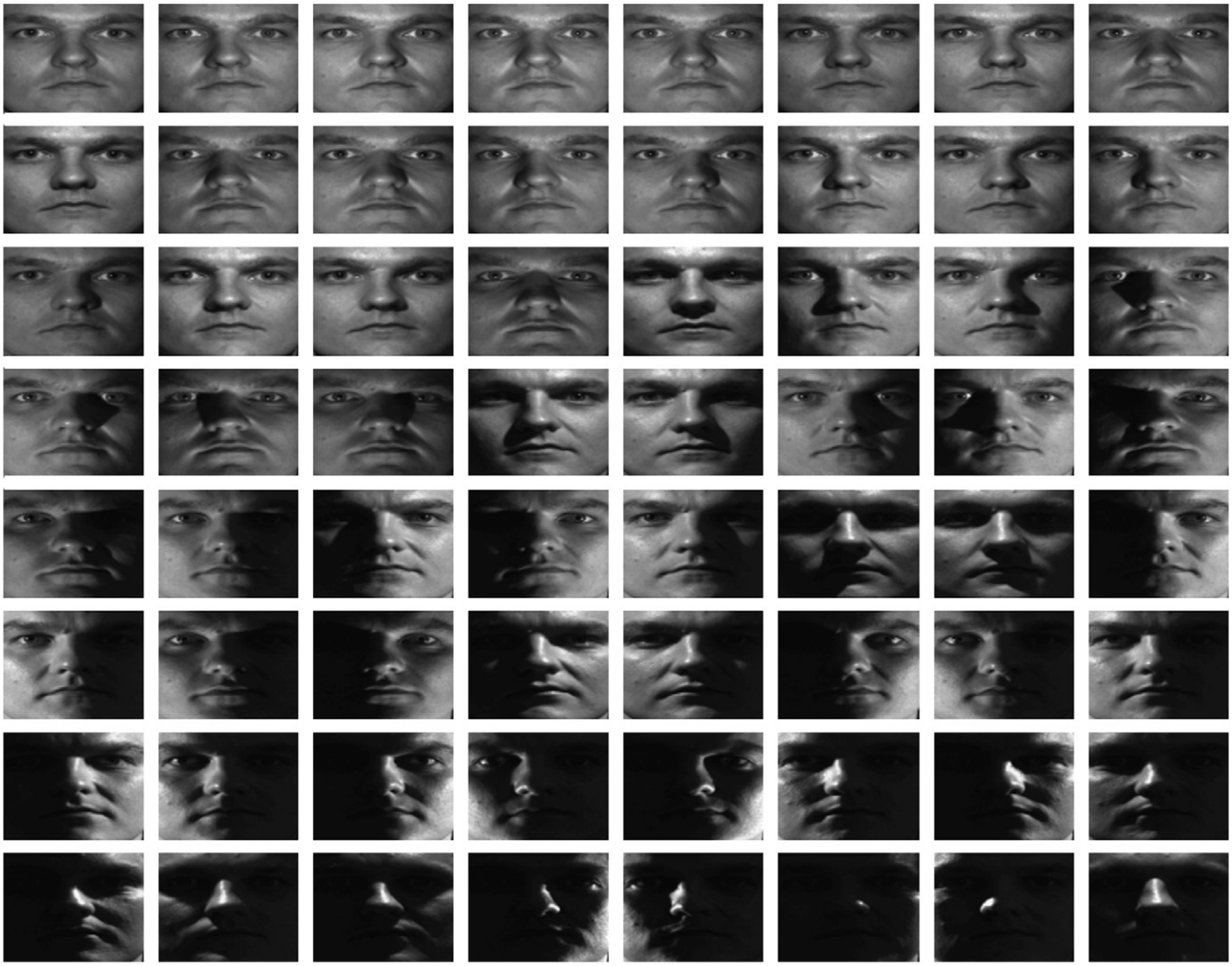

Fig. 2. The front images of one person under 64 lighting conditions.

5.1.3. *Experimental results on CMU-PIE*

The CMU PIE face database contains 68 individuals with 41 368 face images as a whole. The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination and expression. We used the cropped data used in [37][1]. This dataset includes five near frontal poses (C05,C07,C09,C27,C29) and each individual has 170 images except for one person. We exclude the person who has less number of images and use the images of all the other 67 individuals. The procedure is similar to the experiments on Yale B. We first use the images of 10 people to find 20 critical viewing conditions, and then use the 20×67 images corresponding to the critical viewing conditions as the training examples. The test performance is evaluated on the other 150×67 images.

The performance of optimized RLDA (OptRLDA) and optimized Fisherface (OptFisher) are reported in Table 3 with a comparison to other LDA variants, which demonstrate the advantages of OptRLDA and OptFisher, similar to the results on Yale B database. By using LOO errors to choose the optimal regularization number for the CRRs with class indicators derived from LDA/QR, D-LDA,LDA/GSVD and NLDA, the performances are also improved significantly as shown in Table 4.

5.2. *Experiment 2: Performance comparison to Laplacianfaces (LPP), orthogonal Laplacianfaces (OLPP) and marginal Fisher analysis (MFA)*

In this experiment, we compare the performance of the optimal RLDA (OptRLDA) and optimal Fisherface (OptFisher) to Laplacianfaces (LPP), orthogonal Laplacianfaces (OLPP) and marginal Fisher analysis (MFA) [40]. For a fair comparison, we adopt the same procedure as that in the study by [37]. A random subset with $l(=5,10,20,30)$ images per individual was taken with labels to form the training set, and the rest of the database was used as the testing set. For each given l, we average the results over 50 random splits and we used the same splits and the same matlab data files[2] which were used in [37]. The regularization parameter λ in RLDA and PCA dimension in Fisherface are selected based on leave-one-out errors of the training images in the first five splits. For the parameters selection of MFA, grid search was used to minimize the average test errors on the first five splits.

The performances are shown in Tables 5 and 6. The performances for Laplacianfaces (LPP) and orthogonal Laplacianfaces (OLPP) are taken from [37] for CMU PIE database and from http://ews.uiuc.edu/dengcai2/Data/data.html. for the Extended Yale Face Database B. The numbers in the brackets are the projection dimensions. The best performances are shown in bold

[1] Downloaded from http://ews.uiuc.edu/dengcai2/Data/data.html

[2] Which were downloaded from http://ews.uiuc.edu/dengcai2/Data/data.html

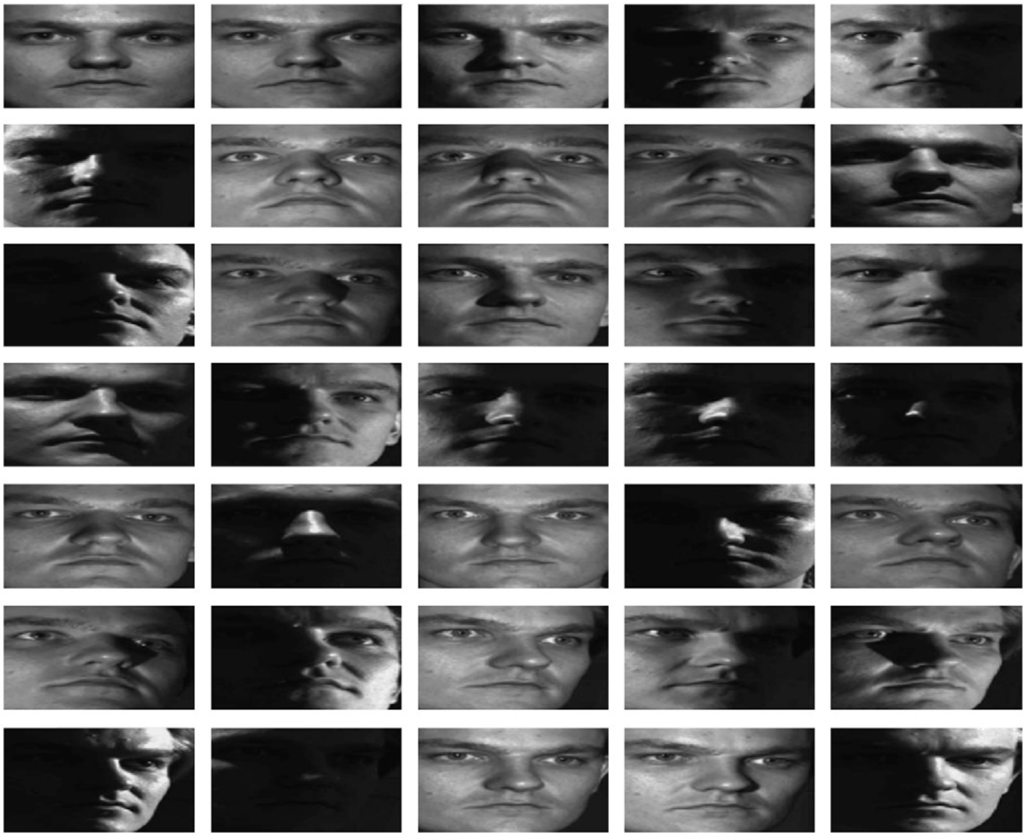

Fig. 3. The 35 cluster-central images.

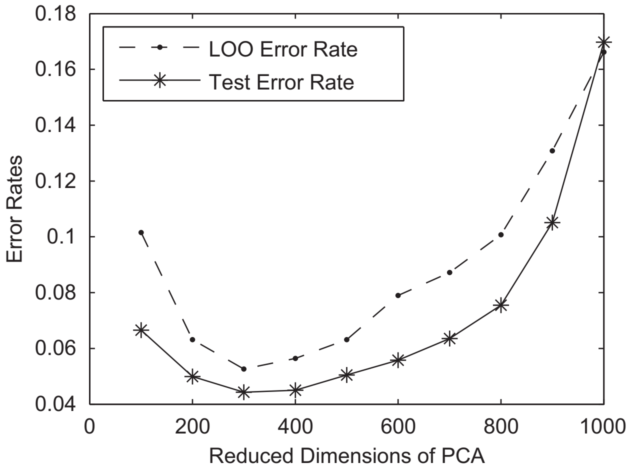

Fig. 4. Comparison of LOO errors and real test errors of Fisherface with varying dimensions in PCA.

numbers. Overall, OptRLDA performs best and it achieves the best performances in three out of four cases for CMU-PIE database and two out of four cases for Yale B database. The second best is MFA which achieves the best performance in the cases of 5 Train and 10 Train for Yale B database and its performances on other cases are also fairly good. OLPP also performs quite good in all the four cases and much better than LPP, which shows that orthogonal constraint on the project matrix improves performance. However, training OLPP is much more expensive than the other four methods.

In the cases of 5 Train and 10 Train, OptRLDA and OptFisher perform worse than MFA and OLPP. This is because RLDA and Fisherface are closely related to the class means' estimation which is not very reliable when the training size is too small.

Note also that the parameters of LPP, OLPP and MFA are selected based on the test performances while the parameters of OptFisher and OptRLDA are selected based only on the training images. The performances of OptRLDA and OptFisher can be better if we choose the parameters based on the testing performances. However, in practice, one needs to find the parameters based only on the training images and it is important to find efficient algorithms to evaluate the cross-validation errors for OLPP, LPP and MFA.

5.3. Remarks

The experimental results on Yale B and CMU PIE demonstrate that face recognition under varying lighting and pose can be achieved with very low error rate by using RLDA. However, in the training procedure, we need images under the critical viewing conditions. In case these images are not available, one may apply face synthesis techniques [41] to generate these images.

As shown in Tables 1–4, proper regularization can significantly improve performance. This may be due to the relation between regularization and robustness against model uncertainties. Recently, the relation between regularization and robustness against model uncertainty has been found for support vector machines [42]. It is interesting to investigate the relation between regularization and robustness in CRR.

The class indicators derived from the LDA variants may be applied to ridge regression [1] and discriminatively regularized least squares classification [43], and these new class indicators may provide useful alternative representations to the widely used class indicator.

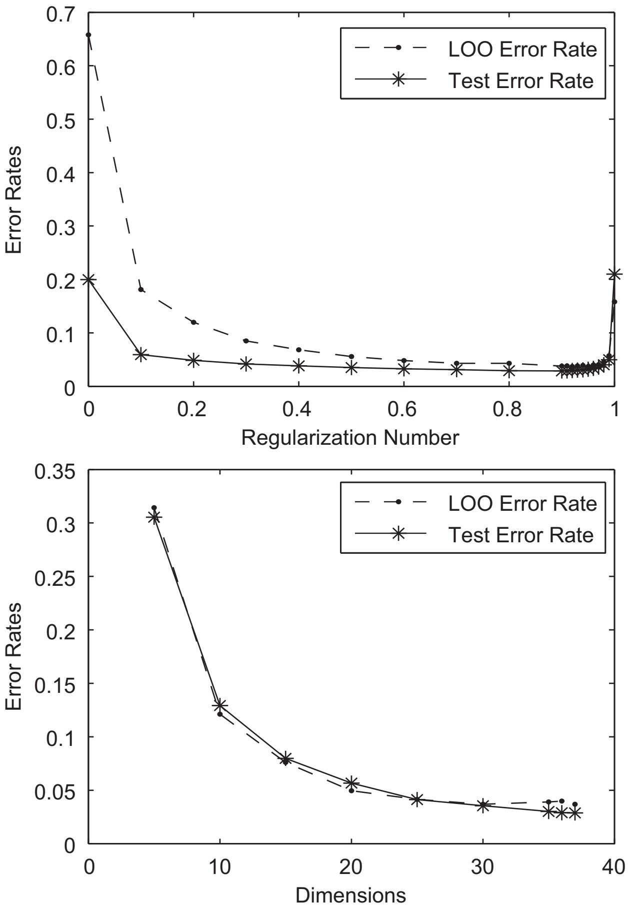

Fig. 5. Comparison of LOO errors and testing errors of RLDA against dimensions and regularization numbers $\lambda = 1/(1+\gamma)$.

Table 1
Performance of LDA's various variants on Yale B database.

Variants	Test error rates (%)	LOO error rates (%)
LDA/QR	12.47	57.89
DLDA	7.19	34.59
LDA/GSVD	22.83	23.53
NLDA	17.64	22.41
OptRLDA	4.0	4.06
OptFisher	4.43	5.26

Table 2
Performance, on Yale B, of CRR with optimal regularization and class indicators from LDA's various Variants, denoted by OptCRR (LDA Variants).

Variants	Test error rates (%)	LOO error rates (%)
OptCRR(LDA/QR)	3.77	4.36
OptCRR(DLDA)	4.38	4.21
OptCRR(LDA/GSVD)	6.54	6.92
OptCRR(NLDA)	3.82	3.91

Table 3
Performance of LDA's various Variants on CMU-PIE.

Variants	Test error rates (%)	LOO error rates (%)
LDA/QR	5.97	70.97
DLDA	2.83	40.3
LDA/GSVD	8.58	31.42
NLDA	5.83	22.01
OptRLDA	2.35	4.78
OptFisher	3.23	7.24

Table 4
Performance, on CMU-PIE, of CRR with optimal regularization and class indicators from LDA's various Variants, denoted by OptCRR (LDA Variants).

Variants	Test error rates (%)	LOO error rates (%)
OptCRR(LDA/QR)	2.77	5.15
OptCRR(DLDA)	3.12	5.67
OptCRR(LDA/GSVD)	3.89	14.25
OptCRR(NLDA)	2.73	5.30

Table 5
Performance (error rate) comparison on CMU PIE face database.

Method	5 Train	10 Train	20 Train	30 Train
LPP	30.8%(67)	21.1%(134)	14.1%(146)	7.13%(131)
OLPP	**21.4%**(108)	11.4%(265)	6.51%(493)	4.83%(423)
OptRLDA	22.6%(67)	**11.3%**(67)	**5.43%**(67)	**4.01%**(67)
OptFisher	34.07%(67)	18.97%(67)	6.8%(67)	4.68%(67)
MFA	24.17%(90)	13.81%(75)	6.29%(75)	4.19%(60)

Table 6
Performance (error rate) comparison on the extended Yale face database B.

Method	5 Train	10 Train	20 Train	30 Train
LPP	24%(37)	11.4%(76)	7.1%(193)	7.5%(251)
OLPP	22.1%(108)	9.7%(111)	3.8%(247)	1.9%(406)
OptRLDA	22.6%(37)	10.03%(37)	**3.36%**(37)	**1.74%(37)**
OptFisher	26.07%(37)	12.55%(37)	4.55%(37)	2.18%(37)
MFA	**21.17%**(136)	**8.59%**(165)	3.70%(150)	1.86%(120)

6. Conclusions

In this paper we have proposed a new ridge regression framework, called constrained ridge regression (CRR) and an efficient cross-validation algorithm to optimize the hyper-parameters in CRR. The proposed algorithms can be used to unify the various LDA methods and to choose the best amongst them. Experimental results demonstrate that the regularization of LDA and optimizing PCA in Fisherface can significantly improve performance in face recognition. The CRR is very flexible for designing the class indicators and it has further potential by designing new types of class indicators. The design of class indicators to optimize the CRR performance is a research problem which needs further investigation.

Appendix A. Proof of Theorem 1

Proof. Since we require $G^T\hat{C} = L$ in (12), the necessity of $\text{span}(L^T) \subset \text{span}(\hat{C}^T)$ is obvious. Now we prove (14). The main idea of this proof is as follows: we use the positive definite property of $\mathbf{S}_\gamma^{-1}$ to transform the constrained optimization problem (12) to be a minimum norm solution problem of linear equations.

Since $\mathbf{S}_\gamma$ is positive definite, there is a nonsingular H such that $\mathbf{S}_\gamma = HH^T$. Denote $\hat{G} = H^TG$. Then the minimization problem becomes

$$\begin{aligned} \min \quad & J(\hat{G}) = \text{tr}\{\hat{G}^T\hat{G}\} \\ \text{s.t.} \quad & \hat{G}^T H^{-1}\hat{C} = L. \end{aligned} \tag{47}$$

Now we simplify the constraint. Let $\hat{C} = U_c \Sigma_c V_c^T$ be its singular value decomposition and $r_c = rank(\hat{C})$. Then the class indicator matrix satisfies

$$L = M^T V_c^T, \quad \text{for some } M \in \mathbb{R}^{r_c \times r} \tag{48}$$

and therefore (47) becomes

$$\begin{aligned} \min \quad & J(\hat{G}) = \mathrm{tr}\{\hat{G}^T \hat{G}\} \\ \text{s.t.} \quad & \hat{G}^T H^{-1} U_c \Sigma_c = M^T. \end{aligned} \tag{49}$$

That is, $\hat{G}$ is the minimum norm solution of the equation $\hat{G}^T H^{-1} U_c \Sigma_c = M^T$.

Let $X = H^{-1} U_c \Sigma_c$ and $\overline{X}$ be a basis matrix of the null space of span$\{X\}$. Since $[X,\overline{X}]$ is nonsingular, the solution of $\hat{G}^T X = M^T$ can be described as

$$\hat{G} = XP + \overline{X}Q \tag{50}$$

for some matrices P and Q. Substituting $\hat{G} = XP + \overline{X}Q$ into $\hat{G}^T X = M^T$, we have

$$\begin{aligned} P^T &= M^T (X^T X)^{-1} \\ &= M^T (\Sigma_c U_c^T H^{-T} H^{-1} U_c \Sigma_c)^{-1} \\ &= M^T (\Sigma_c U_c^T \mathbf{S}_\gamma^{-1} U_c \Sigma_c)^{-1} \end{aligned} \tag{51}$$

and Q is arbitrary since $\overline{X}^T X = 0$. So the minimal norm solution of $\hat{G}^T X = M^T$ is

$$\begin{aligned} \hat{G} &= XP \\ &= H^{-1} U_c \Sigma_c P \\ &= H^{-1} U_c \Sigma_c (\Sigma_c U_c^T \mathbf{S}_\gamma^{-1} U_c \Sigma_c)^{-1} M \end{aligned} \tag{52}$$

and therefore

$$\begin{aligned} G &= H^{-T} \hat{G} \\ &= H^{-T} H^{-1} U_c \Sigma_c (\Sigma_c U_c^T \mathbf{S}_\gamma^{-1} U_c \Sigma_c)^{-1} M \\ &= \mathbf{S}_\gamma^{-1} U_c \Sigma_c (\Sigma_c U_c^T S_\gamma^{-1} U_c \Sigma_c)^{-1} M \\ &= \mathbf{S}_\gamma^{-1} U_c (U_c^T S_\gamma^{-1} U_c)^{-1} \Sigma_c^{-1} M \\ &= \mathbf{S}_\gamma^{-1} \hat{C} (\hat{C}^T S_\gamma^{-1} \hat{C})^\dagger L^T. \end{aligned} \tag{53}$$

Thus Theorem 1 is proved. □

References

[1] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, New York, 2001.

[2] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[3] R. Huang, Q. Liu, H. Lu, S. Ma, Solving the small sample size problem of LDA, in: Proceedings of IEEE International Conference on Pattern Recognition, ICPR, 2002.

[4] J. Ye, Q. Li, A two-stage linear discriminant analysis via QR-decomposition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (6) (2005) 929–941.

[5] P. Howland, H. Park, Generalizing discriminant analysis using generalized singular value decomposition, IEEE Trans. Pattern Anal. Mach. Intell. 26 (8) (2004) 995–1006.

[6] S. Zhang, T. Sim, Discriminant subspace analysis: a Fukunaga–Koontz approach, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1732–1745.

[7] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data with application to face recognition, Pattern Recognition 34 (11) (2001) 2067–2070.

[8] J. Yang, J.Y. Yang, Why can LDA be performed in PCA transformed space, Pattern Recognition 36 (2) (2000) 563–566.

[9] L.F. Chen, H.Y.M. Liao, M.T. Ko, J.C. Lin, G.J. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (10) (2000) 1713–1726.

[10] H. Cevikalp, M. Neamtu, M. Wilkes, A. Barkana, Discriminant common vectors for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (1) (2005) 4–13.

[11] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.-R. Müller, Fisher discriminant analysis with kernels, in: Y.-H. Hu, J. Larsen, E. Wilson, S. Douglas (Eds.), Neural Networks for Signal Processing IX, IEEE1999, pp. 41–48.

[12] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Computation 12 (10) (2000) 2385–2404.

[13] M.-H. Yang, Kernel eigenfaces vs. Kernel Fisherfaces: face recognition using kernel methods, in: FGR '02: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, IEEE Computer Society, Washington, DC, USA2002, p. 215.

[14] G. Dai, Y.T. Qian, Kernel generalized nonlinear discriminant analysis algorithm for pattern recognition, in: Proceedings of IEEE International Conference on Image Processing, 2004.

[15] Q.S. Liu, H.Q. Lu, S.D. Ma, Improving kernel Fisher discriminant analysis for face recognition, IEEE Trans. Circuits Syst. Video Tech. 14 (1) (2004) 42–49.

[16] J.W. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Face recognition using kernel direct discriminant analysis algorithms, IEEE Trans. Neural Networks 14 (1) (2003) 117–126.

[17] J.W. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, J. Wang, An efficient kernel discriminant analysis method, Pattern Recognition 38 (10) (2005) 1788–1790.

[18] J. Yang, A.F. Frangi, J.Y. Yang, D. Zhang, Z. Jin, KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2) (2005) 230–244.

[19] X. Wang, X. Tang, A unified framework for subspace face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 26 (9) (2004) 1222–1228.

[20] M. Turk, A.P. Pentland, Face recognition using eigenfaces, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 1991.

[21] B. Moghaddam, T. Jebara, A. Pentland, Bayesian face recognition, Pattern Recognition 33 (2000) 1771–1782.

[22] B. Moghaddam, Principal manifolds and probabilistic subspace for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 24 (6) (2002) 780–788.

[23] X. Jiang, B. Mandal, A. Kot, Eigenfeature regularization and extraction in face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 30 (3) (2008) 383–394.

[24] O.C. Hamsici, A.M. Martinez, Bayes optimality in linear discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 30 (4) (2008) 647–657.

[25] M. Plutowski, Survey: cross-validation in theory and in practice, Research Report. Department of Computational Science Research, David Sarnoff Research Center, Princeton, New Jersey, 1996.

[26] S. Ji, J. Ye, Generalized linear discriminant analysis: a unified framework and efficient model selection, IEEE Trans. Neural Networks 19 (10) (2008) 1768–1782.

[27] S. An, W. Liu, S. Venkatesh, Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression, Pattern Recognition 40 (8) (2007) 2154–2162.

[28] J. Ye, Least squares linear discriminant analysis, in: Proceedings of International Conference on Machine Learning, ICML, 2007.

[29] L. Sun, S. Ji, J. Ye, A least squares formulation for canonical correlation analysis, in: Proceedings of International Conference on Machine Learning, ICML, 2008.

[30] P. Howland, M. Jeon, H. Park, Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition, SIAM J. Matrix Anal. Appl. 25 (1) (2003) 165–179.

[31] X. Jiang, Asymmetric principal component and discriminant analysis for pattern classification, IEEE Trans. Pattern Anal. Mach.Intell. 31 (5) (2009) 931–937.

[32] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2005) 643–660.

[33] K. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 684–698.

[34] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression (PIE) database, in: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, IEEE Computer Society, Washington, DC, USA2002, p. 215.

[35] T. Sim, S. Baker, M. Bsat, The CMU pose illumination and expression (PIE) database, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1615–1618.

[36] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.

[37] D. Cai, X. He, J. Han, H.-J. Zhang, Orthogonal Laplacianfaces for face recognition, IEEE Trans. Image Process. 15 (11) (2006) 3608–3614.

[38] B. Jahne, Digital Image Processing, Springer, Berlin, 2005.

[39] B. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.

[40] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.

[41] Y. Sheng, K. Kucharski, A. Sadka, W. Skarbek, Automatic face synthesis and analysis: a quick survey, in: Computer Vision and Graphics, Springer, Netherlands, 2006.

[42] H. Xu, C. Caramanis, S. Mannor, Robustness and regularization of support vector machines, J. Mach. Learn. Res. 10 (2009) 1485–1510.

[43] H. Xue, S.C. Chen, Q. Yang, Discriminatively regularized least-squares classification, Pattern Recognition 42 (1) (2009) 93–104.

Senjian An received the B.S degree from Shandong University, Jinan, China, in 1989, the M.S. degree from the Chinese Academy of Sciences, Beijing, in 1992, and the Ph.D. degree from Peking University, Beijing, in 1996. He was with the Institute of Systems Science, Chinese Academy of Sciences, Beijing, where he was a postdoctoral Research Fellow from 1996 to 1998. In 1998, he joined the Beijing Institute of Technology, Beijing and he was an associate Professor from 1998 to 2001. From 2001 to 2004, he was a Research Fellow with The University of Melbourne, Parkville, Australia. Since 2004, he has been working at Curtin University of Technology, where he is currently a Research Fellow. His research interests include machine learning, face recognition and object detection.

Wanquan Liu received the B.Sc. degree in Applied Mathematics from Qufu Normal University, P.R. China, in 1985, the M.Sc. degree in Control Theory and Operation Research from Chinese Academy of Science in 1988, and the Ph.D. degree in Electrical Engineering from Shanghai Jiaotong University, in 1993. He once hold the ARC Fellowship and JSPS Fellowship and attracted research funds from different resources. He is currently an associate professor in the Department of Computing at Curtin University of Technology. His research interests include large scale pattern recognition, control systems, signal processing, machine learning, and intelligent systems.

Svetha Venkatesh is the John Curtin Distinguished Professor at Curtin University of Technology. Her research interests are large-scale pattern recognition, image understanding and applications of computer vision to surveillance and multimedia. She directs the university's Institute of Multi-Sensor Processing and Content Analysis, whose core areas are large scale pattern recognition and machine learning. She is a Senior Member of the IEEE and a Fellow of the IAPR.

Hong Yan received a B.E. degree from Nanking University of Posts and Telecommunications in 1982, an M.S.E. degree from the University of Michigan in 1984, and a Ph.D. degree from Yale University in 1989, all in Electrical Engineering. In 1982 and 1983 he worked on signal detection and estimation as a graduate student and research assistant at Tsinghua University. From 1986 to 1989 he was a research scientist at General Network Corporation, New Haven, CT, USA, where he worked on design and optimization of computer and telecommunications networks. He joined the University of Sydney in 1989 and became Professor of Imaging Science in 1997. He is currently Professor of Computer Engineering at City University of Hong Kong. His research interests include image processing, pattern recognition and bioinformatics. He is author, co-author or editor of two books and 300 journal and conference papers in these areas. Professor Yan is a fellow of the Institute of Electrical and Electronic Engineers (IEEE) and the International Association for Pattern Recognition (IAPR).