Linear Discriminant Dimensionality Reduction

Quanquan Gu, Zhenhui Li, and Jiawei Han

Department of Computer Science,
University of Illinois at Urbana-Champaign
Urbana, IL 61801, US
qgu3@illinois.edu,zli28@uiuc.edu,hanj@cs.uiuc.edu

Abstract. Fisher criterion has achieved great success in dimensionality reduction. Two representative methods based on Fisher criterion are *Fisher Score* and *Linear Discriminant Analysis* (LDA). The former is developed for feature selection while the latter is designed for subspace learning. In the past decade, these two approaches are often studied independently. In this paper, based on the observation that Fisher score and LDA are complementary, we propose to integrate Fisher score and LDA in a unified framework, namely *Linear Discriminant Dimensionality Reduction* (LDDR). We aim at finding a subset of features, based on which the learnt linear transformation via LDA maximizes the Fisher criterion. LDDR inherits the advantages of Fisher score and LDA and is able to do feature selection and subspace learning simultaneously. Both Fisher score and LDA can be seen as the special cases of the proposed method. The resultant optimization problem is a mixed integer programming, which is difficult to solve. It is relaxed into a $L_{2,1}$-norm constrained least square problem and solved by accelerated proximal gradient descent algorithm. Experiments on benchmark face recognition data sets illustrate that the proposed method outperforms the state of the art methods arguably.

1 Introduction

In many applications in machine learning and data mining, one is often confronted with very high dimensional data. High dimensionality increases the time and space requirements for processing the data. Moreover, in the presence of many irrelevant and/or redundant features, learning methods tend to over-fit and become less interpretable. A common way to resolve this problem is dimensionality reduction, which has attracted much attention in machine learning community in the past decades. Generally speaking, dimensionality reduction can be achieved by either feature selection [8] or subspace learning [12] [11] [25] (a.k.a feature transformation). The philosophy behind feature selection is that not all the features are useful for learning. Hence it aims to select a subset of most informative or discriminative features from the original feature set. And the basic idea of subspace learning is that the combination of the original features may be more helpful for learning. As a result, it aims at transforming the original features to a new feature space with lower dimensionality.

Fisher criterion [6] [22] [9] plays an important role in dimensionality reduction. It aims at finding a feature representation by which the within-class distance is minimized and the between-class distance is maximized. Based on Fisher criterion, two representative methods have been proposed. One is *Fisher Score* [22], which is a feature selection method. The other is *Linear Discriminant Analysis* (LDA) [6] [22] [9], which is a subspace learning method. Although there are many other feature selection methods [8] [10] [23], Fisher score is still among the state of the art [29]. And LDA has received great success in face recognition [2], which is known as *Fisher Face*. In the past decades, both Fisher score and LDA have been studied extensively [10] [20] [26] [24] [4] [17] [5]. However, they study Fisher score or LDA independently, ignoring the close relation between them.

In this paper, we propose to study Fisher score and LDA together. The key motivation is that, although it is based on Fisher criterion, Fisher score is not able to do feature combination such as LDA. The features selected by Fisher score are a subset of the original features. However, as we mentioned before, the transformed features may be more discriminative than the original features. On the other hand, although LDA admits feature combination, it transforms all the original features rather than only those useful ones as in Fisher score. Furthermore, since LDA uses all the features, the resulting transformation is often difficult to interpret. It can be seen that Fisher score and LDA are actually complementary to some extent. If we combine Fisher score and LDA in a systematic way, they could mutually enhance each other. One intuitive way is performing Fisher score before LDA as a two-stage approach. However, since these two stages are conducted individually, the whole process is likely to be suboptimal. This motivates us to integrate Fisher score and LDA in a principled way to complement each other.

Based on the above motivation, we propose a unified framework, namely *Linear Discriminant Dimensionality Reduction* (LDDR), integrating Fisher score and LDA. In detail, we aim at finding a subset of features, based on which the learnt linear transformation via LDA maximizes the Fisher criterion. LDDR performs feature selection and subspace learning simultaneously based on Fisher criterion. It inherits the advantages of Fisher score and LDA to overcome their individual disadvantages. Hence it is able to discard the irrelevant features and transform the relevant ones simultaneously. Both Fisher score and LDA can be seen as the special cases of LDDR. The resulting optimization problem is a mixed integer programming [3], which is difficult to solve. We relax it into a $L_{2,1}$-norm constrained least square problem and solved by accelerated proximal gradient descent algorithm [18]. It is worth noting that $L_{2,1}$-norm has already been successfully applied in Group Lasso [28], multi-task feature learning [1] [14], joint covariate selection and joint subspace selection [21]. Experiments on benchmark face recognition data sets demonstrate the effectiveness of the proposed approach.

The remainder of this paper is organized as follows. In Section 2, we briefly review Fisher score and LDA. In Section 3, we present a framework for joint feature selection and subspace learning. In Section 4, we review some related

works. Experiments on benchmark face recognition data sets are demonstrated in Section 5. Finally, we draw a conclusion in Section 6.

1.1 Notations

Given a data set that consists of n data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$, and $y_i \in \{1, 2, \ldots, c\}$ denotes the class label of the i-th data point. The data matrix is denoted by $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, and the linear transformation matrix is denoted by $\mathbf{W} \in \mathbb{R}^{d\times m}$, projecting the input data into an m-dimensional subspace. Given a matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$, we denote the i-th row of $\mathbf{W}$ by $\mathbf{w}^i$, and the j-th column of $\mathbf{W}$ by $\mathbf{w}_j$. The Frobenius norm of $\mathbf{W}$ is defined as $||\mathbf{W}||_F = \sqrt{\sum_i^d ||\mathbf{w}^i||_2^2}$, and the $L_{2,1}$-norm of $\mathbf{W}$ is defined as $||\mathbf{W}||_{2,1} = \sum_i^d ||\mathbf{w}^i||_2$. $\mathbf{1}$ is a vector of all ones with an appropriate length. $\mathbf{0}$ is a vector of all zeros. $\mathbf{I}$ is an identity matrix with an appropriate size. Without loss of generality, we assume that X has been centered with zero mean, i.e., $\sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$.

2 A Review of LDA and Fisher Score

In this section, we briefly introduce two representative dimensionality reduction methods: Linear Discriminant Analysis [6] [22] [9] and Fisher Score [22], both of which are based on Fisher criterion.

2.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA) [6] [22] [9] is a supervised subspace learning method which is based on *Fisher Criterion*. It aims to find a linear transformation $\mathbf{W} \in \mathbb{R}^{d\times m}$ that maps $\mathbf{x}_i$ in the d-dimensional space to a m-dimensional space, in which the between class scatter is maximized while the within-class scatter is minimized, i.e.,

$$\arg\max_{\mathbf{W}} \mathrm{tr}((\mathbf{W}^T\mathbf{S}_w\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})), \tag{1}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are the between-class scatter matrix and within-class scatter matrix respectively, which are defined as

$$\mathbf{S}_b = \sum_{k=1}^{c} n_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T, \mathbf{S}_w = \sum_{k=1}^{c}\sum_{i\in\mathcal{C}_k}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T, \tag{2}$$

where $\mathcal{C}_k$ is the index set of the k-th class, $\boldsymbol{\mu}_k$ and n_k are mean vector and size of k-th class respectively in the input data space, i.e., $\mathbf{X}$, $\boldsymbol{\mu} = \sum_{k=1}^c n_k\boldsymbol{\mu}_k$ is the overall mean vector of the original data. It is easy to show that Eq. (1) is equivalent to

$$\arg\max_{\mathbf{W}} \mathrm{tr}((\mathbf{W}^T\mathbf{S}_t\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})), \tag{3}$$

where $\mathbf{S}_t$ is the total scatter matrix, defined as follows,

$$\mathbf{S}_t = \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T. \tag{4}$$

Note that $\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b$.

According to [6], when the total scatter matrix $\mathbf{S}_t$ is non-singular, the solution of Eq. (3) consists of the top eigenvectors of the matrix $\mathbf{S}_t^{-1}\mathbf{S}_b$ corresponding to nonzero eigenvalues. When the total class scatter matrix $\mathbf{S}_t$ does not have full rank, the solution of Eq. (3) consists of the top eigenvectors of the matrix $\mathbf{S}_t^{\dagger}\mathbf{S}_b$ corresponding to nonzero eigenvalues, where $\mathbf{S}_t^{\dagger}$ denotes the pseudo-inverse of $\mathbf{S}_t$ [7]. Note that when $\mathbf{S}_t$ is nonsingular, $\mathbf{S}_t^{\dagger}$ equals $\mathbf{S}_t^{-1}$.

LDA has been successfully applied to face recognition [2]. Following LDA, many incremental works have been done, e.g., Uncorrelated LDA and Orthogonal LDA [26], Local LDA [24], Semi-supervised LDA [4] and Sparse LDA [17] [5]. Note that all these methods suffer from the weakness of using all the original features to learn the subspace.

2.2 Fisher Score for Feature Selection

The key idea of Fisher score [22] is to find a subset of features, such that in the data space spanned by the selected features, the distances between data points in different classes are as large as possible, while the distances between data points in the same class are as small as possible. In particular, given the selected m features, the input data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ reduces to $\mathbf{Z} \in \mathbb{R}^{m\times n}$. Then the *Fisher Score* is formulated as follows,

$$\arg\max_{\mathbf{Z}} \operatorname{tr}\left\{\tilde{\mathbf{S}}_t^{-1}\tilde{\mathbf{S}}_b\right\}, \tag{5}$$

where $\tilde{\mathbf{S}}_b$ and $\tilde{\mathbf{S}}_t$ are defined as

$$\tilde{\mathbf{S}}_b = \sum_{k=1}^{c} n_k(\tilde{\boldsymbol{\mu}}_k - \tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}_k - \tilde{\boldsymbol{\mu}})^T, \tilde{\mathbf{S}}_t = \sum_{i=1}^{n} (\mathbf{z}_i - \tilde{\boldsymbol{\mu}})(\mathbf{z}_i - \tilde{\boldsymbol{\mu}})^T, \tag{6}$$

where $\tilde{\boldsymbol{\mu}}_k$ and n_k are the mean vector and size of the k-th class respectively in the reduced data space, i.e., $\mathbf{Z}$, $\tilde{\boldsymbol{\mu}} = \sum_{k=1}^{c} n_k\tilde{\boldsymbol{\mu}}_k$ is the overall mean vector of the reduced data. Note that there are $\binom{d}{m}$ candidate $\mathbf{Z}$'s out of $\mathbf{X}$, hence Fisher score is a combinatorial optimization problem.

We introduce an indicator variable $\mathbf{p}$, where $\mathbf{p} = (p_1, \ldots, p_d)^T$ and $p_i \in \{0, 1\}, i = 1, \ldots, d$, to represent whether a feature is selected or not. In order to indicate that m features are selected, we constrain $\mathbf{p}$ by $\mathbf{p}^T\mathbf{1} = m$. Then the *Fisher Score* in Eq. (5) can be equivalently formulated as follows,

$$\begin{aligned} \arg\max_{\mathbf{p}} \ & \operatorname{tr}\{(\operatorname{diag}(\mathbf{p})\mathbf{S}_t\operatorname{diag}(\mathbf{p}))^{-1}(\operatorname{diag}(\mathbf{p})\mathbf{S}_b\operatorname{diag}(\mathbf{p}))\}, \\ \text{s.t.} \ & \mathbf{p} \in \{0, 1\}^d, \mathbf{p}^T\mathbf{1} = m, \end{aligned} \tag{7}$$

where diag($\mathbf{p}$) is a diagonal matrix whose diagonal elements are p_i's, $\mathbf{S}_b$ and $\mathbf{S}_t$ are the between-class scatter matrix and total scatter matrix, defined as in Eq. (2) and Eq. (4).

As can be seen, like other feature selection approaches [8], Fisher score only does binary feature selection. It does not admit feature combination like LDA does.

Based on the above discussion, we can see that LDA suffers from the problem which Fisher score does not have, while Fisher score has the limitation which LDA does not have. Hence, if we integrate LDA and Fisher score in a systematic way, they could complement each other and be benefited from each other. This motivates the proposed method in this paper.

3 Linear Discriminant Dimensionality Reduction

In this section, we will integrate Fisher score and Linear Discriminant Analysis in a unified framework. The key idea of our method is to find a subset of features, based on which the learnt linear transformation via LDA maximizes the Fisher criterion. It can be mathematically formulated as follows,

$$\begin{aligned} \arg\max_{\mathbf{W},\mathbf{p}} \ & \mathrm{tr}\{(\mathbf{W}^T\mathrm{diag}(\mathbf{p})\mathbf{S}_t\mathrm{diag}(\mathbf{p})\mathbf{W})^{-1}(\mathbf{W}^T\mathrm{diag}(\mathbf{p})\mathbf{S}_b\mathrm{diag}(\mathbf{p})\mathbf{W})\}, \\ \text{s.t.} \ & \mathbf{p} \in \{0,1\}^d, \mathbf{p}^T\mathbf{1} = m, \end{aligned} \tag{8}$$

which is a mixed integer programming [3]. Eq. (8) is called as *Linear Discriminant Dimensionality Reduction* (LDDR) because it is able to do feature selection and subspace learning simultaneously. It inherits the advantages of Fisher score and LDA. That is, it is able to find a subset of useful original features, based on which it generates new features by feature transformation. Given $\mathbf{p} = \mathbf{1}$, Eq. (8) reduces to LDA as in Eq. (3). Letting $\mathbf{W} = \mathbf{I}$, Eq. (8) degenerates to Fisher score as in Eq.(7). Hence, both LDA and Fisher score can be seen as the special cases of the proposed method. In addition, the objective functions corresponding to LDA and Fisher score are lower bounds of the objective function of LDDR.

Recent studies [9] [27] established the relationship between LDA and multivariate linear regression problem, which provides a regression-based solution for LDA. This motivates us to solve the problem in Eq.(8) in a similar manner. In the following, we present a theorem, which establishes the equivalence relationship between the problem in Eq.(8) and the problem in Eq.(9).

Theorem 1. *The optimal* $\mathbf{p}$ *that maximizes the problem in Eq. (8) is the same as the optimal* $\mathbf{p}$ *that minimizes the following problem*

$$\begin{aligned} \arg\min_{\mathbf{p},\mathbf{W}} \ & \frac{1}{2}||\mathbf{X}^T diag(\mathbf{p})\mathbf{W} - \mathbf{H}||_F^2 \\ s.t. \ & \mathbf{p} \in \{0,1\}^d, \mathbf{p}^T\mathbf{1} = m, \end{aligned} \tag{9}$$

where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_c] \in \mathbb{R}^{n \times c}$*, and* $\mathbf{h}_k$ *is a column vector whose* i*-th entry is given by*

$$h_{ik} = \begin{cases} \sqrt{\frac{n}{n_k}} - \sqrt{\frac{n_k}{n}}, & \text{if } \mathbf{y}_i = k \\ -\sqrt{\frac{n_k}{n}}, & \text{otherwise.} \end{cases} \tag{10}$$

In addition, the optimal $\mathbf{W}_1$ *of Eq. (8) and the optimal* $\mathbf{W}_2$ *of Eq. (9) have the following relation*

$$\mathbf{W}_2 = [\mathbf{W}_1, \mathbf{0}]\mathbf{Q}^T, \tag{11}$$

under a mild condition that

$$rank(\mathbf{S}_t) = rank(\mathbf{S}_b) + rank(\mathbf{S}_w), \tag{12}$$

and $\mathbf{Q}$ *is a orthogonal matrix.*

Proof. Due to space limit, we only give the sketch of the proof. On the one hand, given the optimal $\mathbf{W}$, the optimization problem in Eq. (8) with respect to $\mathbf{p}$ is equivalent to the optimization problem in Eq. (9) with respect to $\mathbf{p}$. On the other hand, for any feasible $\mathbf{p}$, the optimal $\mathbf{W}$ that maximizes the problem in Eq. (8) and the optimal $\mathbf{W}$ that minimizes the problem in Eq. (9) satisfy the relation in Eq. (11) according to Theorem 5.1 in [27]. The detailed proof will be included in the longer version of this paper.

Note that the above theorem holds under the condition that $\mathbf{X}$ is centered with zero mean. Since $\text{rank}(\mathbf{S}_t) = \text{rank}(\mathbf{S}_b) + \text{rank}(\mathbf{S}_w)$ holds in many applications involving high-dimensional and under-sampled data, the above theorem can be applied widely in practice.

According to theorem 1, the difference between $\mathbf{W}_1$ and $\mathbf{W}_2$ is the orthogonal matrix $\mathbf{Q}$. Since the Euclidean distance is invariant to any orthogonal transformation, if a classifier based on the Euclidean distance (e.g., K-Nearest-Neighbor and linear support vector machine [9]) is applied to the dimensionality-reduced data obtained by $\mathbf{W}_1$ and $\mathbf{W}_2$, they will achieve the same classification result. In our experiments, we use K-Nearest-Neighbor classifier.

Suppose we find the optimal solution of Eq. (9), i.e., $\mathbf{W}^*$ and $\mathbf{p}^*$, then $\mathbf{p}^*$ is a binary vector, and $\text{diag}(\mathbf{p})\mathbf{W}$ is a matrix where the elements of many rows are all zeros. This motivate us to absorb the indicator variables $\mathbf{p}$ into $\mathbf{W}$, and use $L_{2,0}$-norm on $\mathbf{W}$ to achieve feature selection, leading to the following problem

$$\begin{aligned} \arg\min_{\mathbf{W}} \ & \frac{1}{2}||\mathbf{X}^T\mathbf{W} - \mathbf{H}||_F^2, \\ \text{s.t.} \ & ||\mathbf{W}||_{2,0} \leq m. \end{aligned} \tag{13}$$

However, the feasible region defined by $||\mathbf{W}||_{2,0} \leq m$ is not convex. We relax $||\mathbf{W}||_{2,0} \leq m$ to its convex hull [3], and obtain the following relaxed problem,

$$\begin{aligned} \arg\min_{\mathbf{W}} \ & \frac{1}{2}||\mathbf{X}^T\mathbf{W} - \mathbf{H}||_F^2, \\ \text{s.t.} \ & ||\mathbf{W}||_{2,1} \leq m. \end{aligned} \tag{14}$$

Note that Eq. (14) is no longer equivalent to Eq. (8) due to the relaxation. However, the relaxation makes the optimization problem computationally much easier. In this sense, the relaxation can be seen as a tradeoff between the strict equivalence and computational tractability.

Eq. (14) is equivalent to the following regularized problem,

$$\arg\min_{\mathbf{W}} \frac{1}{2}||\mathbf{X}^T\mathbf{W} - \mathbf{H}||_F^2 + \mu||\mathbf{W}||_{2,1}, \tag{15}$$

where $\mu > 0$ is a regularization parameter. Given an m, we could find a μ, such that Eq. (14) and Eq. (15) achieve the same solution. However, it is difficult to give an analytical relationship between m and μ. Fortunately, such a relationship is not crucial for our problem. Since it is easier to tune μ than an integer m, we consider Eq. (15) in the rest of this paper.

Eq. (15) is mathematically similar to Group Lasso problem [28] and multi-task feature selection [14]. However, the motivations of our method and those methods are essentially different. Our method aims at integrating feature selection and subspace learning in a unified framework, while Group Lasso and multi-task feature selection aim at discovering common feature patterns among multiple related learning tasks. The objective function in Eq. (15) is a non-smooth but convex function. In the following, we will present an algorithm for solving Eq. (15). Similar algorithm has been used for multi-task feature selection [14].

3.1 Proximal Gradient Descent

The most natural approach for solving the problem in Eq. (15) is the sub-gradient descent method [3]. However, its convergence rate is very slow, i.e., $O(\frac{1}{\epsilon^2})$ [19].

Recently, proximal gradient descent has received increasing attention in the machine learning community [13] [14]. It achieves the optimal convergence rate, i.e., $O(\frac{1}{\epsilon})$ for the first-order method and is able to deal with large-scale non-smooth convex problems. It can be seen as an extension of gradient descent, where the objective function to minimize is the composite of a smooth part and a non-smooth part. As to our problem, let

$$\begin{aligned} f(\mathbf{W}) &= \frac{1}{2}||\mathbf{X}^T\mathbf{W} - \mathbf{H}||_F^2 \\ F(\mathbf{W}) &= f(\mathbf{W}) + \mu||\mathbf{W}||_{2,1}. \end{aligned} \tag{16}$$

It is easy to show that $f(\mathbf{W})$ is convex and differentiable, while $\mu||\mathbf{W}||_{2,1}$ is convex but non-smooth.

In each iteration of the proximal gradient descent algorithm, $F(\mathbf{W})$ is linearized around the current estimate $\mathbf{W}_t$, and the value of $\mathbf{W}$ is updated as the solution of the following optimization problem,

$$\mathbf{W}_{t+1} = \arg\min_{\mathbf{W}} G_{\eta_t}(\mathbf{W}, \mathbf{W}_t), \tag{17}$$

where $G_{\eta_t}(\mathbf{W}, \mathbf{W}_t)$ is called proximal operator, which is defined as

$$G_{\eta_t}(\mathbf{W}, \mathbf{W}_t) = f(\mathbf{W}_t) + \langle \nabla f(\mathbf{W}_t), \mathbf{W} - \mathbf{W}_t \rangle + \frac{\eta_t}{2} ||\mathbf{W} - \mathbf{W}_t||^2 + \mu ||\mathbf{W}||_{2,1}. \quad (18)$$

In our problem, $\nabla f(\mathbf{W}_t) = \mathbf{X}\mathbf{X}^T\mathbf{W}_t - \mathbf{X}\mathbf{H}$. The philosophy under this formulation is that if the optimization problem in Eq. (17) can be solved by exploiting the structure of the $L_{2,1}$ norm, then the convergence rate of the resulting algorithm is the same as that of gradient descent method, i.e., $O(\frac{1}{\epsilon})$, since no approximation on the non-smooth term is employed. It is worth noting that the proximal gradient descent can also be understood from the perspective of auxiliary function optimization [15].

By ignoring the terms in $G_{\eta_t}(\mathbf{W}, \mathbf{W}_t)$ that is independent of $\mathbf{W}$, the optimization problem in Eq. (17) boils down to

$$\mathbf{W}_{t+1} = \arg\min_{\mathbf{W}} \frac{1}{2} ||\mathbf{W} - (\mathbf{W}_t - \frac{1}{\eta_t} \nabla f(\mathbf{W}_t))||_F^2 + \frac{\mu}{\eta_t} ||\mathbf{W}||_{2,1}. \quad (19)$$

For the sake of simplicity, we denote $\mathbf{U}_t = \mathbf{W}_t - \frac{1}{\eta_t} \nabla f(\mathbf{W}_t)$, then Eq. (19) takes the following form

$$\mathbf{W}_{t+1} = \pi_{\eta_t}(\mathbf{W}_t) = \arg\min_{\mathbf{W}} \frac{1}{2} ||\mathbf{W} - \mathbf{U}_t||_F^2 + \frac{\mu}{\eta_t} ||\mathbf{W}||_{2,1}, \quad (20)$$

which can be further decomposed into c separate subproblems of dimension d

$$\mathbf{w}_{t+1}^i = \arg\min_{\mathbf{w}^i} \frac{1}{2} ||\mathbf{w}^i - \mathbf{u}_t^i||_2^2 + \frac{\mu}{\eta_t} ||\mathbf{w}^i||_2, \quad (21)$$

where $\mathbf{w}_{t+1}^i$, $\mathbf{w}^i$ and $\mathbf{u}_t^i$ are the i-th rows of $\mathbf{W}_{t+1}$, $\mathbf{W}$ and $\mathbf{U}_t$ respectively. It has a closed form solution [14] as follows

$$\mathbf{w}^{i*} = \begin{cases} (1 - \frac{\mu}{\eta_t ||\mathbf{u}_t^i||}) \mathbf{u}_t^i, & \text{if } ||\mathbf{u}_t^i|| > \frac{\mu}{\eta_t} \\ \mathbf{0}, & \text{otherwise.} \end{cases} \quad (22)$$

Thus, the proximal gradient descent in Eq. (17) has the same convergence rate of $O(\frac{1}{\epsilon})$ as gradient descent for smooth problem.

3.2 Accelerated Proximal Gradient Descent

To achieve more efficient optimization, we employ Nesterov's method [19] to accelerate the proximal gradient descent in Eq. (17), which owns the convergence rate as $O(\frac{1}{\sqrt{\epsilon}})$. More specifically, we construct a linear combination of $\mathbf{W}_t$ and $\mathbf{W}_{t+1}$ to update $\mathbf{V}_{t+1}$ as follows:

$$\mathbf{V}_{t+1} = \mathbf{W}_t + \frac{\alpha_t - 1}{\alpha_{t+1}} (\mathbf{W}_{t+1} - \mathbf{W}_t), \quad (23)$$

where the sequence $\{\alpha_t\}_{t \geq 1}$ is conventionally set to be $\alpha_{t+1} = \frac{1 + \sqrt{1 + 4\alpha_t^2}}{2}$. For more detail, please refer to [13]. Here we directly present the final algorithm for optimizing Eq. (15) in Algorithm 1.

The convergence of this algorithm is stated in the following theorem.

Algorithm 1 Linear Discriminant Dimensionality Reduction

Initialize: $\eta_0, \mathbf{W}_1 \in \mathbb{R}^{d\times m}, \alpha_1 = 1$;
repeat
  while $F(\mathbf{W}_t) > G_{\eta_{t-1}}(\pi_{\eta_{t-1}}(\mathbf{W}_t), \mathbf{W}_t)$ **do**
    Set $\eta_{t-1} = \gamma\eta_{t-1}$
  end while
  Set $\eta_t = \eta_{t-1}$
  Compute $\mathbf{W}_{t+1} = \arg\min_{\mathbf{W}} G_{\eta_t}(\mathbf{W}, \mathbf{V}_t)$
  Compute $\alpha_{t+1} = \frac{1+\sqrt{1+4\alpha_t^2}}{2}$
  Compute $\mathbf{V}_{t+1} = \mathbf{W}_t + \frac{\alpha_t - 1}{\alpha_{t+1}}(\mathbf{W}_{t+1} - \mathbf{W}_t)$
until convergence

Theorem 2. *[19] Let $\{\mathbf{W}_t\}$ be the sequence generated by Algorithm 1, then for any $t \geq 1$ we have*

$$F(\mathbf{W}_t) - F(\mathbf{W}^*) \leq \frac{2\gamma L||\mathbf{W}_1 - \mathbf{W}^*||_F^2}{(t+1)^2}, \tag{24}$$

where L is the Lipschitz constant of the gradient of $f(\mathbf{W})$ in the objective function, $\mathbf{W}^* = \arg\min_{\mathbf{W}} F(\mathbf{W})$.

Theorem 2 shows that the convergence rate of the accelerated proximal gradient descent method is $O(\frac{1}{\sqrt{\epsilon}})$.

4 Related Work

In this section, we discuss some approaches which are closely related to our method.

In order to pursue sparsity and interpretability in LDA, [17] proposed both exact and greedy algorithms for binary class sparse LDA as well as its spectral bound. For multi-class problem, [5] proposed a sparse LDA (SLDA) based on ℓ_1-norm regularized *Spectral Regression*,

$$\arg\min_{\mathbf{w}} ||\mathbf{X}^T\mathbf{w} - \mathbf{y}||_2^2 + \mu||\mathbf{w}||_1, \tag{25}$$

where $\mathbf{y}$ is the eigenvector of $\mathbf{S}_b\mathbf{y} = \lambda\mathbf{S}_t\mathbf{y}$. Due to the nature of the ℓ_1 penalty, some entries in $\mathbf{w}$ will be shrunk to exact zero if λ is large enough, which results in a sparse projection. However, SLDA does not lead to feature selection, because each column of the linear transformation matrix is optimized one by one, and their sparsity patterns are independent. In contrast, our method is able to do feature selection.

On the other hand, [16] proposed another feature selection method based on Fisher criterion, namely *Linear Discriminant Feature Selection* (LDFS), which modifies LDA to admit feature selection as follows,

$$\arg\min_{\mathbf{W}} \mathrm{tr}((\mathbf{W}^T\mathbf{S}_w\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})) + \mu\sum_{i=1}^{d}||\mathbf{w}^i||_\infty, \tag{26}$$

where $\sum_{i=1}^{d} ||\mathbf{a}^i||_\infty$ is the ℓ_1/ℓ_∞ norm of $\mathbf{W}$. The optimization problem is convex and solved by quasi-Newton method [3]. Although LDFS involves structured sparse transformation matrix $\mathbf{W}$ as in our method, it use it to select features rather than doing feature selection and transformation together. Hence it is fundamentally a feature selection method. In comparison, our method uses the structured sparse transformation matrix for both feature selection and combination.

5 Experiments

In this section, we evaluate the proposed method, i.e., LDDR, and compare it with the state of the art subspace learning methods, e.g. PCA, LDA and Locality Preserving Projection (LPP) [12], sparse LDA (SLDA) [5]. We also compare it with the feature selection methods, e.g., Fisher score (FS) and *Linear Discriminant Feature Selection* (LDFS) [16]. Moreover, we study Fisher score followed with LDA (FS+LDA), which is the most intuitive way to conduct Fisher score and LDA together. We use K-Nearest Neighbor classifier where $K = 1$ as the baseline method. All the experiments were performed in Matlab on a Intel Core2 Duo 2.8GHz Windows 7 machine with 4GB memory.

5.1 Data Sets

We use two standard face recognition databases which are used in [11] [5].

ORL face database[1] contains 10 images for each of the 40 human subjects, which were taken at different times, varying the lighting, facial expressions and facial details. The original images (with 256 gray levels) have size 92×112, which are resized to 32×32 for efficiency.

Extended Yale-B database[2] contains 16128 face images of 38 human subjects under 9 pose and 64 illumination conditions. In our experiment, we choose the frontal pose and use all the images under different illumination, thus we get 2414 image in total. All the face images are manually aligned and cropped. They are resized to 32×32 pixels, with 256 gray levels per pixel. Thus each face image is represented as a 1024-dimensional vector.

5.2 Parameter Settings

For ORL data set, $p = 2, 3, 4$ images were randomly selected as training samples for each person, and for Yale-B data set, $p = 10, 20, 30$ images were randomly selected as training samples for each person. The rest images were used for testing. The training set was used to learn a subspace, and the recognition was performed in the subspace by *K-Nearest Neighbor* classifier where $K = 1$ according to [5]. Since the training set was randomly chosen, we repeated each experiment 20

[1] http://www.cl.cam.ac.uk/Research/DTG/attarchive:pub/data
[2] http://vision.ucsd.edu/ leekc/ExtYaleDatabase/ExtYaleB.html

times and calculated the average recognition accuracy. In general, the recognition rate varies with the dimensionality of the subspace. The best average performance obtained as well as the corresponding dimensionality is reported. It is worth noticing that for LDDR, the dimensionality of the subspace is exactly the same as the number of classes, i.e., c according to Eq.(9).

For LDA, as in [2], we first use PCA to reduce the dimensionality to $n - c$ and then perform LDA to reduce the dimensionality to $c - 1$. This is also known as *Fisher Face* [2]. For FS+LDA, we first use Fisher Score to select 50% features and then perform LDA to reduce the dimensionality. For LPP, we use the cosine distance to compute the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. For SLDA, we tune μ by searching the grid $\{10, 20, \ldots, 100\}$ on the testing set according to [5]. For LDFS and LDDR, the regularization parameter μ is tuned by searching the grid $\{0.01, 0.05, 0.1, 0.2, 0.5\}$ on the testing set. As we know, tuning parameter on the testing set could be biased. However, since we did this for all the methods as long as they have parameters to tune, it is still a fair comparison.

5.3 Recognition Results

Table 1. Face recognition accuracy on the ORL data set

Data set	2 training		3 training		4 training	
	Acc	Dim	Acc	Dim	Acc	Dim
Baseline	66.81±3.41	–	77.02±2.55	–	81.73±2.27	–
PCA	66.81±3.41	79	77.02±2.55	119	81.73±2.27	159
FS	69.06±3.04	197	79.07±2.71	200	84.42±2.41	199
LDFS	62.69±3.43	198	75.45±2.28	192	81.96±2.56	188
LDA	71.27±3.58	28	83.36±1.84	39	89.63±2.01	39
LPP	72.41±3.17	39	84.20±1.73	39	90.42±1.41	39
FS+LDA	71.81±3.36	28	84.13±1.35	39	88.56±2.16	39
SLDA	74.14±2.92	39	84.86±1.82	39	91.44±1.53	39
LDDR	**76.88±3.49**	40	**86.89±1.91**	40	**92.77±1.61**	40

The experimental results are shown in Table 1 and Table 2. We can observe that (1) On some cases, Fisher score is better than LDA, while on more cases, LDA outperforms Fisher score. This implies feature transformation may be more essential than feature selection; (2) LDFS is worse than Fisher score on the ORL data set, while it is better than Fisher score on the Yale-B data set. This indicates the performance gain by doing feature selection under slightly different criterion is limited; (3) SLDA is better than LDA, which implies sparsity is able to improve the classification performance of LDA; (4) FS+LDA improves both FS and LDA at most cases. It is even better than SLDA at some cases. This implies the potential performance gain of combining Fisher score and LDA. However, at some cases, FS+LDA is not as good as FS or LDA. This is because Fisher score and LDA are conducted individually in FS+LDA. The selected features

Table 2. Face recognition accuracy on the Yale-B data set

Data set	10 training		20 training		30 training	
	Acc	Dim	Acc	Dim	Acc	Dim
Baseline	53.44±0.82	–	69.24±1.19	–	77.39±0.98	–
PCA	52.41±0.89	200	67.04±1.18	200	74.57±1.07	200
FS	64.34±1.40	200	76.53±1.19	200	82.15±1.14	200
LDFS	66.86±1.17	182	80.50±1.17	195	83.16±0.90	197
LDA	78.33±1.31	37	85.75±0.84	37	81.19±2.05	37
LPP	79.70±2.96	76	80.24±5.49	75	86.40±1.45	78
FS+LDA	77.89±1.82	37	87.89±0.88	37	93.91±0.69	37
SLDA	81.56±1.38	37	89.68±0.85	37	92.88±0.68	37
LDDR	**89.45±1.11**	38	**96.44±0.85**	38	**98.66±0.43**	38

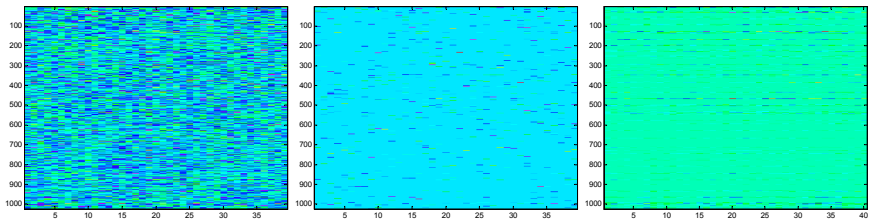
by Fisher score are not necessarily useful for LDA; (5) LDDR outperforms FS, LDA, SLDA and FS+LDA consistently and overwhelmingly, which indicates that by performing Fisher score and LDA simultaneously to maximize the Fisher criterion, Fisher score and LDA can enhance each other greatly. The selected features by LDDR should be more useful than those selected by Fisher score. We will illustrate this point latter.

5.4 Projection Matrices

To get a better understanding of our approach, we plot the linear transformation matrices of our method and related methods on the ORL and Yale-B data sets in Fig. 1 and Fig . 2 respectively. Clearly, the linear transformation matrix of LDA is very dense, which is not easy to interpret. Each column of the linear transformation matrix of SLDA is sparse. However, the sparse patterns of each column are not coherent. In other word, for different dimensions of the subspace, the selected features by SLDA are different. Therefore, it is unclear which features are useful for the whole transformation. In contrast, each row of the linear transformation matrix of LDDR tends to be zero simultaneously, which leads to joint feature selection and transformation. This is exactly what we pursue. Note that the sparsity of the linear transformation matrix of LDDR is controlled by the regularization parameter μ. That is, the number of selected features in LDDR is indirectly controlled by μ. We will show that the performance of LDDR is not sensitive to μ latter.

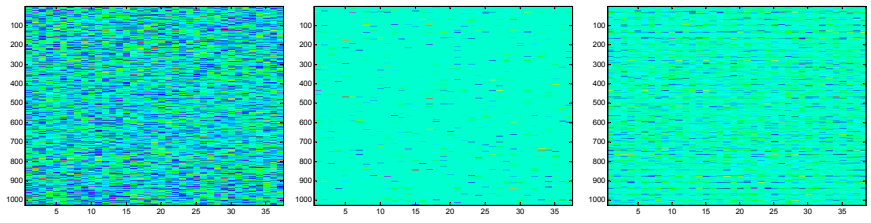
5.5 Selected Features

We are also interested in the features selected by LDDR. We plot the top 50 selected features (pixels) of our method and Fisher score on the ORL and Yale-B data sets in Fig. 3 and Fig. 4 respectively. It is shown that the distribution of selected features (pixels) by Fisher score is highly skewed. Most features distribute in only one or two regions. Many features even reside on the non-face region.



(a) LDA (b) SLDA (c) LDDR

Fig. 1. The linear transformation matrix learned by (a) LDA, (b) SLDA ($\mu = 50$) and (c) LDDR ($\mu = 0.5$) with 3 training samples per person on the ORL database. For better viewing, please see it in color pdf file.



(a) LDA (b) SLDA (c) LDDR

Fig. 2. The linear transformation matrix learned by (a) LDA, (b) SLDA ($\mu = 50$) and (c) LDDR ($\mu = 0.5$) with 20 training samples per person on the Yale-B database. For better viewing, please see it in color pdf file.

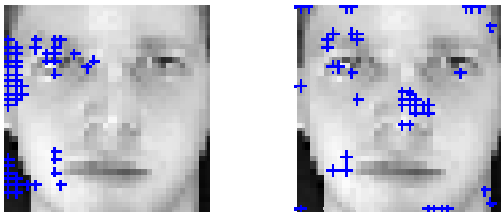
It implies that the features selected by Fisher score are not discriminative. In contrast, the features selected by LDDR distribute widely across the face region.

From another perspective, we can see that the features (pixels) selected by LDDR are asymmetric. In other word, if one pixel is selected, its axis symmetric one will not be selected. This is because the face image is roughly axis symmetry, so one in a pair of axis symmetric pixels is redundant given the other one is selected. Moreover, the selected pixels are mostly around the eyebrow, the boundary of eyes, nose and cheek, which are discriminative for distinguishing face images of different people. This is accord with our life common sense.

5.6 Sensitivity to the Regularization Parameter

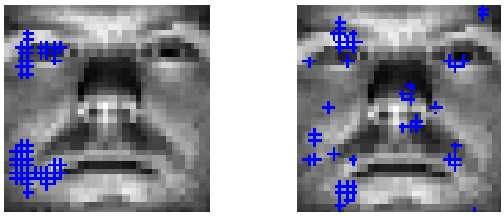
LDDR only has one parameter, which is the regularization parameter μ. It indirectly controls the number of selected features. Here we will investigate the recognition accuracy with respect to the regularization parameter μ. We vary the value of μ, and plot the recognition accuracy with respect to μ on the ORL and Yale-B data sets in Fig. 5 and Fig. 6 respectively.

As can be seen, LDDR is not sensitive to the regularization parameter μ in a wide range of μ. In detail, LDDR achieves consistently good performance



(a) Fisher Score (b) LDDR

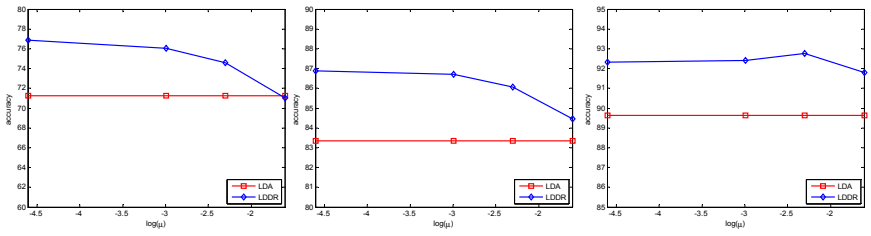
Fig. 3. Selected features (marked by blue cross) by (a) Fisher score and (b) LDDR ($\mu = 0.5$) with 3 training samples per person on the ORL database. For better viewing, please see it in color pdf file.



(a) Fisher Score (b) LDDR

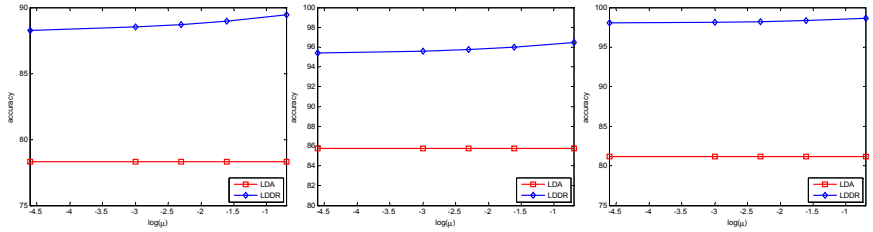
Fig. 4. Selected features (marked by blue cross) by (a) Fisher score and (b) LDDR ($\mu = 0.5$)with 20 training samples per person on the Yale-B database. For better viewing, please see it in color pdf file.

with the μ varying from 0.01 to 0.1 on the ORL data set. LDDR is even more stable on the Yale-B data set, where it gets overwhelmingly good result with the μ changing from 0.01 to 0.5. This shows that in certain range, the number of useful features does not affect the performance of the jointly learnt linear transformation very much. It is an appealing property because we do not need to tune the regularization parameter painfully in the application.



(a) 2 training (b) 3 training (c) 4 training

Fig. 5. Recognition accuracy with respect to the regularization parameter μ on the ORL database



(a) 10 training (b) 20 training (c) 30 training

Fig. 6. Recognition accuracy with respect to the regularization parameter μ on the Yale-B database

6 Conclusion

In this paper, we propose to integrate Fisher score and LDA in a unified framework, namely *Linear Discriminant Dimensionality Reduction*. We aim at finding a subset of features, based on which the learnt linear transformation via LDA maximizes the Fisher criterion. LDDR inherits the advantages of Fisher score and LDA and is able to do feature selection and subspace learning simultaneously. Both Fisher score and LDA can be seen as the special cases of the proposed method. The resultant optimization problem is relaxed into a $L_{2,1}$-norm constrained least square problem and solved by accelerated proximal gradient descent algorithm. Experiments on benchmark face recognition data sets illustrate the efficacy of the proposed framework.

Acknowledgments

The work was supported in part by NSF IIS-09-05215, U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265, and the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA). We thank the anonymous reviewers for their helpful comments.

References

1. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Machine Learning 73(3), 243–272 (2008)
2. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. 19(7), 711–720 (1997)
3. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press, Cambridge (2004)
4. Cai, D., He, X., Han, J.: Semi-supervised discriminant analysis. In: ICCV. pp. 1–7 (2007)
5. Cai, D., He, X., Han, J.: Spectral regression: A unified approach for sparse subspace learning. In: ICDM. pp. 73–82 (2007)

6. Fukunaga, K.: Introduction to statistical pattern recognition (2nd ed.). Academic Press Professional, Inc., San Diego, CA, USA (1990)
7. Golub, G.H., Loan, C.F.V.: Matrix computations (3rd ed.). Johns Hopkins University Press, Baltimore, MD, USA (1996)
8. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
9. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer (2001)
10. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: NIPS (2005)
11. He, X., Cai, D., Yan, S., Zhang, H.: Neighborhood preserving embedding. In: ICCV. pp. 1208–1213 (2005)
12. He, X., Niyogi, P.: Locality preserving projections. In: NIPS (2003)
13. Ji, S., Ye, J.: An accelerated gradient method for trace norm minimization. In: ICML. p. 58 (2009)
14. Liu, J., Ji, S., Ye, J.: Multi-task feature learning via efficient $l_{2,1}$-norm minimization. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI 2009) (2009)
15. Luo, D., Ding, C.H.Q., Huang, H.: Towards structural sparsity: An explicit l2/l0 approach. In: ICDM. pp. 344–353 (2010)
16. Masaeli, M., Fung, G., Dy, J.G.: From transformation-based dimensionality reduction to feature selection. In: ICML. pp. 751–758 (2010)
17. Moghaddam, B., Weiss, Y., Avidan, S.: Generalized spectral bounds for sparse lda. In: ICML. pp. 641–648 (2006)
18. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. 103(1), 127–152 (2005)
19. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers (2003)
20. Nie, F., Xiang, S., Jia, Y., Zhang, C., Yan, S.: Trace ratio criterion for feature selection. In: AAAI. pp. 671–676 (2008)
21. Obozinski, G., Taskar, B., Jordan, M.I.: Joint covariate selection and joint subspace selection for multiple classification problems. Statistics and Computing 20, 231–252 (April 2010)
22. R. O. Duda, P.E.H., Stork, D.G.: Pattern Classification. Wiley-Interscience Publication (2001)
23. Song, L., Smola, A.J., Gretton, A., Borgwardt, K.M., Bedo, J.: Supervised feature selection via dependence estimation. In: ICML. pp. 823–830 (2007)
24. Sugiyama, M.: Local fisher discriminant analysis for supervised dimensionality reduction. In: ICML. pp. 905–912 (2006)
25. Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q., Lin, S.: Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. 29(1), 40–51 (2007)
26. Ye, J.: Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. Journal of Machine Learning Research 6, 483–502 (2005)
27. Ye, J.: Least squares linear discriminant analysis. In: ICML. pp. 1087–1093 (2007)
28. Yuan, M., Yuan, M., Lin, Y., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68, 49–67 (2006)
29. Zhao, Z., Wang, L., Liu, H.: Efficient spectral feature selection with minimum redundancy. In: AAAI (2010)