Efficient Model Selection for Regularized Linear Discriminant Analysis

Jieping Ye
Arizona State University
Tempe, AZ 85287
jieping.ye@asu.edu

Tao Xiong
University of Minnesota
Minneapolis, MN 55455
txiong@ece.umn.edu

Qi Li
Western Kentucky University
Bowling Green, KY 42101
qi.li@wku.edu

Ravi Janardan
University of Minnesota
Minneapolis, MN 55455
janardan@cs.umn.edu

Jinbo Bi
CAD group, Siemens Medical Solutions, Inc.
Malvern, PA 19355
jinbo.bi@siemens.com

Vladimir Cherkassky
University of Minnesota
Minneapolis, MN 55455
cherkass@ece.umn.edu

Chandra Kambhamettu
University of Delaware
Newark, DE 19716
chandra@cis.udel.edu

ABSTRACT

Classical Linear Discriminant Analysis (LDA) is not applicable for small sample size problems due to the singularity of the scatter matrices involved. Regularized LDA (RLDA) provides a simple strategy to overcome the singularity problem by applying a regularization term, which is commonly estimated via cross-validation from a set of candidates. However, cross-validation may be computationally prohibitive when the candidate set is large. An efficient algorithm for RLDA is presented that computes the optimal transformation of RLDA for a large set of parameter candidates, with approximately the same cost as running RLDA a small number of times. Thus it facilitates efficient model selection for RLDA.

An intrinsic relationship between RLDA and Uncorrelated LDA (ULDA), which was recently proposed for dimension reduction and classification is presented. More specifically, RLDA is shown to approach ULDA when the regularization value tends to zero. That is, RLDA without any regularization is equivalent to ULDA. It can be further shown that ULDA maps all data points from the same class to a common point, under a mild condition which has been shown to hold for many high-dimensional datasets. This leads to the overfitting problem in ULDA, which has been observed in several applications. The theoretical analysis presented provides further justification for the use of regularization in RLDA. Extensive experiments confirm the claimed theoretical estimate of efficiency. Experiments also show that, for a properly chosen regularization parameter, RLDA performs favorably in classification, in comparison with ULDA, as well as other existing LDA-based algorithms and Support Vector Machines (SVM).

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications - Data Mining

General Terms: Algorithms

Keywords: Dimension reduction, Linear Discriminant Analysis, regularization, model selection

1. INTRODUCTION

Linear Discriminant Analysis (LDA) is a well-known classification method that projects high-dimensional data onto a low-dimensional space where the data is reshaped to maximize class separability [7, 9, 15]. The optimal projection or transformation in classical LDA is obtained by minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum discrimination. Classical LDA involves three scatter matrices, i.e., the within-class, between-class, and total scatter matrices. The total scatter matrix is a multiple of the sample covariance matrix and is required to be nonsingular. However, in many applications such as text mining, microarray data classification, and face recognition, all scatter matrices in question can be singular since the data points are in a very high-dimensional space and the sample size does not exceed this dimension. This is known as the *singularity* or *undersampled* problem [17].

Regularized LDA (RLDA) provides an effective solution for the singularity problem. The idea is to add a constant λ to the diagonal elements of the total scatter matrix, where $\lambda > 0$ is known as the *regularization parameter*. Regularization stabilizes the sample covariance matrix estimation and improves the classification performance of LDA. RLDA

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
Copyright 2006 ACM 1-59593-433-2/06/0011 ...$5.00.

has applications in many areas, including face recognition [4, 18], microarray classification [11], medical image analysis [6], etc.

Choosing an appropriate regularization value is a critical issue in RLDA, as a large λ may significantly disturb the information in the scatter matrix, while a small λ may not be effective enough to solve the singularity problem. Cross-validation is commonly used to estimate the optimal λ from a finite set, $\Lambda = \{\lambda_1, \cdots, \lambda_m\}$, of m candidates. Selecting an optimal value for a parameter such as λ is called *model selection* [15]. The computational cost of model selection for RLDA can be high, especially when m is large, since it requires expensive matrix computations for each $\lambda_i \in \Lambda$. However, a large m is often desirable in practice to obtain a good λ.

1.1 Related work

Besides regularized LDA, other methods have been brought to bear on such high-dimensional, small sample size problems, including Penalized LDA (PLDA) [12], Diagonal LDA (DLDA) [5], Uncorrelated LDA (ULDA) [16], Orthogonal LDA (OLDA) [24], and PCA+LDA [2], where PCA stands for Principal Component Analysis. Many of these LDA-based methods have the same computational cost. We show an interesting relationship between RLDA and ULDA in this paper. This relationship implies that single-model RLDA (with $m = 1$) and ULDA are of the same time complexity. Thus our experimental studies on efficiency concentrate on the comparison between single-model RLDA and multiple-model RLDA (with $m > 1$).

From the perspective of computing the discriminant score for classification, rather than feature extraction, Hastie *et al.* [14, 11] proposed an efficient algorithm for RLDA. However, these algorithms tend to be numerically unstable when the regularization parameter λ is close to 0. Friedman [8] considered a more general formulation of RLDA and proposed an efficient algorithm when leave-one-out cross-validation was applied. It did not address the high computational cost associated with estimating the best regularization parameter from a large set of candidates, which has recently been addressed in [25].

Regularization is the key to many other machine learning methods such as Support Vector Machines (SVM) [22], spline fitting [23], Quadratic Discriminant Analysis (QDA) [8], etc. The tuning of the regularization parameter also consumes time in SVM training. Hastie *et al.* [13] proposed an algorithm for SVM, which fits the entire path of SVM solutions for every value of the regularization parameter, with essentially the same computational cost as fitting one SVM model. This dramatically reduces the computational cost of model selection in SVM training.

1.2 Contributions

This paper aims to reduce the computational cost of the regularized LDA approach on high-dimensional, small sample size problems. The primary contributions of this work include the following:

- A theoretical property of RLDA is established, that is, regularizing the total scatter matrix is equivalent to regularizing its nonzero eigenvalues. This result shows the essential dimension where the regularization takes place. We call this property the *essential regularization property.*

- An efficient algorithm for solving RLDA is proposed, using the essential regularization property to speed up the model selection process for RLDA.

- The proposed algorithm also makes RLDA model more stable. Note that the optimization in traditional RLDA involves the calculation of the inverse of the regularized total scatter matrix, and is thus subject to numerical instability problems as $\lambda \to 0$.

- RLDA is shown to approach Uncorrelated LDA, as $\lambda \to 0$. Thus, the range of λ for RLDA is extended to $[0, \infty)$, overcoming the limitation of the traditional RLDA algorithms.

- Uncorrelated LDA (ULDA) is shown to map all points from the same class to a common point, under a mild condition which has been shown to hold for many high-dimensional data. This leads to the overfitting problem in ULDA, which further justifies the need for regularization applied in RLDA.

Besides the theoretical results that guarantee the efficiency of the proposed model selection algorithm for RLDA, experiments on computational efficiency confirm our theoretically-established bounds. Moreover, our experiments also show favorable performance of the algorithm in terms of classification, in comparison of several other LDA-based methods and SVM. In summary, we propose in this paper a new implementation of RLDA, which allows model selection to be optimized over a large set of candidates with low computational effort.

The rest of the paper is organized as follows. An overview of classical LDA and regularized LDA is given in Section 2. The essential regularization property of Regularized LDA is presented in Section 3. Efficient RLDA algorithms are described in Section 4. Section 5 includes the experimental results. We conclude in Section 6.

2. REVIEW OF CLASSICAL LDA AND REGULARIZED LDA

We briefly review the classical LDA formulation and the regularized LDA formulation in this section.

2.1 Classical LDA

Given a data matrix $A \in \mathbb{R}^{d\times n}$, classical LDA computes a linear transformation $G \in \mathbb{R}^{d\times \ell}$ that maps each column a_i of A, for $1 \le i \le n$, in the d-dimensional space to a vector y_i in the ℓ-dimensional space: $G : a_i \in \mathbb{R}^d \to y_i = G^T a_i \in \mathbb{R}^\ell$ $(\ell < d)$.

Let the data matrix A be partitioned into k classes as $A = [A_1, \cdots, A_k]$, where $A_i \in \mathbb{R}^{d\times n_i}$, and $\sum_{i=1}^k n_i = n$. In discriminant analysis [9], three scatter matrices, i.e., *within-class*, *between-class*, and *total* scatter matrices are defined as follows:

$$S_w = \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in A_i}(x - c^{(i)})(x - c^{(i)})^T, \quad (1)$$

$$S_b = \frac{1}{n}\sum_{i=1}^{k} n_i(c^{(i)} - c)(c^{(i)} - c)^T, \quad (2)$$

$$S_t = \frac{1}{n}\sum_{j=1}^{n}(a_j - c)(a_j - c)^T, \quad (3)$$

where the *centroid* $c^{(i)}$ of the i-th class is defined as $c^{(i)} = \frac{1}{n_i}A_i e^{(i)}$ with $e^{(i)} = (1, 1, \cdots, 1)^T \in \mathbb{R}^{n_i}$, and the *global centroid* c is defined as $c = \frac{1}{n}Ae$ with $e = (1, 1, \cdots, 1)^T \in \mathbb{R}^n$. It follows from the definition that $S_t = S_b + S_w$.

Define the matrices

$$\begin{aligned} H_w &= \frac{1}{\sqrt{n}}[A_1 - c^{(1)}(e^{(1)})^T, \cdots, A_k - c^{(k)}(e^{(k)})^T], && (4) \\ H_b &= \frac{1}{\sqrt{n}}[\sqrt{n_1}(c^{(1)} - c), \cdots, \sqrt{n_k}(c^{(k)} - c)], && (5) \\ H_t &= \frac{1}{\sqrt{n}}(A - ce^T). && (6) \end{aligned}$$

Then the three scatter matrices: S_w, S_b, and S_t in Eqs. (1)–(3) can be expressed as

$$S_w = H_w H_w^T, \quad S_b = H_b H_b^T, \quad S_t = H_t H_t^T. \qquad (7)$$

In the lower-dimensional space resulting from the linear transformation G, the scatter matrices S_w, and S_b, and S_t become $G^T S_w G$, $G^T S_b G$, and $G^T S_t G$, respectively. An optimal transformation G^* in classical discriminant analysis can be computed by solving the following optimization problem [9]:

$$G^* = \arg\max_G \left\{ \mathrm{trace}((G^T S_w G)^{-1} G^T S_b G) \right\}, \qquad (8)$$

which can be shown to be equivalent to

$$G^* = \arg\max_G \left\{ \mathrm{trace}((G^T S_t G)^{-1} G^T S_b G) \right\}, \qquad (9)$$

using the following equality: $S_t = S_b + S_w$.

The optimization problem in Eq. (9) is equivalent to finding x that satisfies $S_b x = \lambda S_t x$, for $\lambda \neq 0$ [9]. The solution can be obtained by applying an eigen-decomposition on the matrix $S_t^{-1} S_b$, if S_t is nonsingular. Note that there exist no more than $k - 1$ eigenvectors corresponding to nonzero eigenvalues, since the rank of the matrix S_b is bounded from above by $k-1$. Therefore, the reduced dimension of classical LDA is at most $k - 1$.

Classical LDA is equivalent to maximum likelihood classification assuming normal distribution for each class with the common covariance matrix. Although relying on heavy assumptions which are not true in many applications, LDA has been proved to be effective. This is mainly due to the fact that a simple, linear model is more robust against noise, and most likely will not overfit.

2.2 Regularized LDA

Classical discriminant analysis requires the total scatter matrix S_t to be nonsingular, which may not hold for small sample size data. A simple way to deal with the singularity of S_t is to apply regularization, by adding some constant value to the diagonal elements of S_t as $\tilde{S}_t = S_t + \lambda I_d$, for some $\lambda > 0$, where I_d is the identity matrix of size d. Since S_t is positive semi-definite, $S_t + \lambda I_d$ is positive definite [10], and hence nonsingular.

An optimal transformation of RLDA can be computed by solving the following optimization problem:

$$G^* = \arg\max_G \left\{ \mathrm{trace}((G^T\left(S_t + \lambda I\right)G)^{-1} G^T S_b G) \right\}. \qquad (10)$$

Similarly, the solution to Eq. (10) can be achieved by computing the eigen-decomposition of $(S_t + \lambda I)^{-1} S_b$. The computation can be divided into two stages. First the Singular Value Decomposition (SVD) [10] of H_t in Eq. (6) is computed:

$$H_t = \hat{U}\hat{\Sigma}\hat{V}^T,$$

where $\hat{U} \in R^{d\times d}$ and $\hat{V} \in R^{n\times n}$ are orthonormal square matrices, and $\hat{\Sigma} \in R^{d\times n}$ is diagonal. It follows from Eq. (7) that

$$S_t = H_t H_t^T = \hat{U}\hat{\Sigma}\hat{\Sigma}^T\hat{U}^T.$$

Therefore

$$\tilde{S}_t = S_t + \lambda I_d = \hat{U}(\hat{\Sigma}\hat{\Sigma}^T + \lambda I_d)\hat{U}^T = \hat{U}\tilde{\Sigma}_l\hat{U}^T, \qquad (11)$$

where $\tilde{\Sigma}_l = \hat{\Sigma}\hat{\Sigma}^T + \lambda I_d$. For high-dimensional data, the size of $\tilde{\Sigma}_l$ is large. However, we will show that $\tilde{\Sigma}_l$ can be replaced by a much smaller matrix whose size is the same as the rank of S_t, a number much smaller than d for high-dimensional, small sample size data.

Next, let $U_b \Sigma_b V_b^T$ be the SVD of

$$\tilde{\Sigma}_l^{-1/2}\hat{U}^T H_b,$$

where H_b is defined in Eq. (5). Let

$$X = \hat{U}\tilde{\Sigma}_l^{-1/2} U_b.$$

Then from Eq. (7), we have

$$\begin{aligned} X^T \tilde{S}_t X &= \left(\hat{U}\tilde{\Sigma}_l^{-1/2} U_b\right)^T \left(\hat{U}\tilde{\Sigma}_l\hat{U}^T\right) \hat{U}\tilde{\Sigma}_l^{-1/2} U_b = I_d, \\ X^T S_b X &= \left(\hat{U}\tilde{\Sigma}_l^{-1/2} U_b\right)^T \left(H_b H_b^T\right) \hat{U}\tilde{\Sigma}_l^{-1/2} U_b = \Sigma_b^2. \end{aligned}$$

The optimal transformation G for RLDA consists of the first q columns of X, where $q = \mathrm{rank}(S_b)$.

The time complexity of the above algorithm is $O(nd^2)$, which can be expensive for high-dimensional data. Especially, when v-fold cross-validation ($v = 5$ in our experiments) is performed for choosing the best λ, the above algorithm needs to be repeated v times. It is often computationally prohibitive, and thus restricts the application of RLDA to data of small size.

Remark 2.1. *Note that in the traditional RLDA formulation, S_w is applied instead of S_t as in Eq. (10). The regularization value λ is thus required to be positive and the algorithm is subject to numerical instability problems, when λ is close to 0. The modified formulation for RLDA used in this paper overcomes this limitation, by showing that the limit of the solution to RLDA exists, when $\lambda \to 0$, and is equal to ULDA. More details can be found in Section 4.4.*

3. ESSENTIAL REGULARIZATION PROPERTY

In this section, we present a key property of regularized LDA, establishing that regularizing the total scatter matrix S_t, in the context of LDA, is equivalent to regularizing the nonzero eigenvalues of S_t. This property has significant implications in designing an efficient RLDA algorithm for small sample size problems.

The result of this section is motivated by the following fact: if the rank of H_t is r ($r \ll d$), the orthonormal matrices in the SVD decomposition have redundant columns. Only the first r columns of $\hat{U}$ and $\hat{V}$ corresponding to the first r rows and the first r columns in $\hat{\Sigma}$ where diagonal entries are nonzero play a role in the reconstruction of H_t. Let

$U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{d\times r}$ denote the first r columns of $\hat{U}$ and $\hat{V}$, respectively. Let the square matrix Σ consist of the first r rows and the first r columns of $\hat{\Sigma}$. Then we have

$$H_t = \hat{U}\hat{\Sigma}\hat{V}^T = U\Sigma V^T.$$

It is clear that

$$S_t + \lambda I_d = \hat{U}(\hat{\Sigma}\hat{\Sigma}^T + \lambda I_d)\hat{U}^T \neq U(\Sigma^2 + \lambda I_r)U^T.$$

However, our main result of this section shows that

$$\begin{aligned}(S_t + \lambda I_d)^{-1}S_b &= \hat{U}(\hat{\Sigma}\hat{\Sigma} + \lambda I_r)^{-1}\hat{U}^T S_b \\ &= U(\Sigma^2 + \lambda I_r)^{-1}U^T S_b.\end{aligned}$$

One of the basic tools used in our proof is the Sherman-Woodbury-Morrison formula [10]: Let $P \in \mathbb{R}^{d\times d}$, and $Q, R \in \mathbb{R}^{d\times n}$. Assuming that both the matrices P and $(I+R^TP^{-1}Q)$ are nonsingular, we have

$$(P + QR^T)^{-1} = P^{-1} - P^{-1}Q(I + R^TP^{-1}Q)^{-1}R^TP^{-1}. \tag{12}$$

Proposition 3.1. *Let the scatter matrices S_t and S_b be defined as above, and let $U\Sigma V^T$ be the skinny SVD of H_t, where $U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{n\times r}$ have orthonormal columns, $\Sigma \in \mathbb{R}^{r\times r}$ is diagonal, and $r = rank(S_t)$. We have*

$$(S_t + \lambda I_d)^{-1}S_b = U(\Sigma^2 + \lambda I_r)^{-1}U^TS_b + \lambda^{-1}U_\perp U_\perp^T S_b. \tag{13}$$

where d is the dimension of data points, and $U_\perp$ is the orthogonal complement of U.

Proof. From $H_t = U\Sigma V^T$, we have

$$S_t = H_tH_t^T = U\Sigma^2U^T.$$

Substituting $P = \lambda I_d$, and $Q = R = U\Sigma$ into the Sherman-Woodbury-Morrison formula as in Eq. (12), we have

$$\tilde{S}_t^{-1} = \lambda^{-1}I_d - \lambda^{-2}U(\Sigma(I_r + \lambda^{-1}\Sigma^2)^{-1}\Sigma)U^T.$$

Note that $\Sigma(I_r + \lambda^{-1}\Sigma^2)^{-1}\Sigma$ is a diagonal matrix. Using the equality

$$\frac{\sigma^2}{\lambda(\lambda+\sigma^2)} = \frac{1}{\lambda} - \frac{1}{\lambda+\sigma^2},$$

we can show that

$$\lambda^{-2}\Sigma(I_r + \lambda^{-1}\Sigma^2)^{-1}\Sigma = \lambda^{-1}I_r - (\Sigma^2 + \lambda I_r)^{-1},$$

and thus

$$\begin{aligned}\tilde{S}_t^{-1} &= \lambda^{-1}I_d + U((\Sigma^2 + \lambda I_r)^{-1} - \lambda^{-1}I_r)U^T \\ &= U(\Sigma^2 + \lambda I_r)^{-1}U^T + \lambda^{-1}(I_d - UU^T).\end{aligned}$$

Note that $U_\perp \in \mathbb{R}^{d\times(d-r)}$ is the orthogonal complement of U, that is, $[U, U_\perp] \in \mathbb{R}^{d\times d}$ is orthogonal. Using the fact that

$$[U, U_\perp][U, U_\perp]^T = UU^T + U_\perp U_\perp^T = I_d,$$

we have

$$\tilde{S}_t^{-1} = U(\Sigma^2 + \lambda I_r)^{-1}U^T + \lambda^{-1}U_\perp U_\perp^T,$$

and the result follows by multiplying by S_b on both sides. □

The computation of the transformation G via the eigen-decomposition of the matrix in Eq. (13) may be sensitive to numerical disturbances as $\lambda \to 0$, due to the presence of λ^{-1} in the second term. Interestingly, it can be overcome using the result in the following lemma:

Lemma 3.1. *Let $U_\perp$, S_t, and S_b be defined as above. Then, the null space of S_t, denoted as Null(S_t) is a subset of the null space, Null(S_b), of S_b. That is, Null(S_t) $\subseteq$ Null(S_b). Furthermore, $U_\perp^TS_b = 0$.*

Proof. The proof directly follows from the fact that $S_t = S_b + S_w$, and both S_b and S_w are positive semi-definite. □

With Proposition 3.1 and Lemma 3.1, we have the following main result of this section:

Theorem 3.1. *Let S_t, S_b, U, V, Σ, d, and r be defined as in Proposition 3.1. Then for any $\lambda > 0$, the following equality holds:*

$$(S_t + \lambda I_d)^{-1}S_b = U(\Sigma^2 + \lambda I_r)^{-1}U^TS_b. \tag{14}$$

Theorem 3.1 implies that regularizing the total scatter S_t, in the context of LDA, is equivalent to regularizing the nonzero eigenvalues of S_t. The eigenvectors of $\tilde{S}_t^{-1}S_b$ can be computed as follows:

Theorem 3.2. *Let y be an eigenvector of $\tilde{S}_t^{-1}S_b$ corresponding to a nonzero eigenvalue μ, then $y = Ux$ for some x, where x is an eigenvector of $(\Sigma^2 + \lambda I_r)^{-1}U^TS_bU$.*

Proof. Let y be an eigenvector of $\tilde{S}_t^{-1}S_b$ corresponding to a nonzero eigenvalue μ. Then

$$y = \frac{1}{\mu}U(\Sigma^2 + \lambda I_r)^{-1}U^TS_by = Ux,$$

for some x.

Next, we show that x is an eigenvector of $(\Sigma^2+\lambda I_r)^{-1}U^TS_bU$. Multiplying both sides of the following equation by U^T:

$$U(\Sigma^2 + \lambda I_r)^{-1}U^TS_by = \mu y,$$

we have

$$(\Sigma^2 + \lambda I_r)^{-1}U^TS_b(Ux) = \lambda U^T(Ux) = \mu x.$$

This completes the proof of the theorem. □

Denote $\tilde{\Sigma}_s = \Sigma^2 + \lambda I_r$. In contrast to $\tilde{\Sigma}_l$ used in traditional RLDA, as given in Eq. (11), the size of $\tilde{\Sigma}_s$ is typically small for high-dimensional small sample size data. So is the size of $\tilde{\Sigma}_s^{-1}U^TS_bU$, an important matrix introduced in Theorem 3.2. It is also worth noting that computing U is independent of the regularization value λ. Thus, Theorem 3.2 leads to a two-step computation of the eigenvectors of $\tilde{S}_t^{-1}S_b$:

- compute U; and
- compute the eigen-decomposition of $\tilde{\Sigma}_s^{-1}U^TS_bU$.

In the following, we describe an efficient way of computing the eigenvectors of $\tilde{\Sigma}_s^{-1}U^TS_bU$, which is

$$\tilde{\Sigma}_s^{-1/2}(\tilde{\Sigma}_s^{-1/2}U^TH_b)(\tilde{\Sigma}_s^{-1/2}U^TH_b)^T\tilde{\Sigma}_s^{1/2},$$

since $S_b = H_bH_b^T$ as in Eq. (7), where H_b is defined in Eq. (5). Let $U_b\Sigma_bV_b^T$ be the SVD of $\tilde{\Sigma}_s^{-1/2}U^TH_b$, then we have

$$\begin{aligned}\tilde{\Sigma}_s^{-1}U^TS_bU &= \tilde{\Sigma}_s^{-1/2}U_b\Sigma_b^2U_b^T\tilde{\Sigma}_s^{1/2} \\ &= (\tilde{\Sigma}_s^{-1/2}U_b)\Sigma_b^2(\tilde{\Sigma}_s^{-1/2}U_b)^{-1}.\end{aligned}$$

That is, $\tilde{\Sigma}_s^{-1/2}U_b$ diagonalizes the matrix $\tilde{\Sigma}_s^{-1}U^TS_bU$. Thus, the columns of $\tilde{\Sigma}_s^{-1/2}U_b$ form the eigenvectors of $\tilde{\Sigma}_s^{-1}U^TS_bU$. The above computation is more efficient than directly applying the eigen-decomposition to $\tilde{\Sigma}_s^{-1}U^TS_bU$ since the size of $\tilde{\Sigma}_s^{-1/2}U^TH_b$ is much smaller (i.e., $r \times k$).

4. EFFICIENT RLDA ALGORITHMS

In this section, we present efficient RLDA algorithms for both the single model, where $|\Lambda| = 1$ (Λ is the candidate set for regularization) and the multiple model, where $|\Lambda| > 1$.

4.1 Single model ($|\Lambda|$ = 1)

Given a fixed training dataset and a fixed regularization value λ, our proposed single-model RLDA algorithm is summarized in **Algorithm 1** below.

The time complexity of this algorithm is dominated by Lines 2 and 3. For small sample size problems, the cost of Lines 4-6 is significantly smaller than the cost of Line 2. (More details will be given in Section 4.2.) Note that Lines 2 and 3 are independent of λ. This observation is the key for the efficient multiple-model RLDA algorithm proposed next.

```
Algorithm 1: Single-model RLDA
1.    Construct $H_b$ and $H_t$ as in Eqs. (5) and (6),
      $r \leftarrow rank(H_t)$;
2.    Compute the SVD of $H_t$: $H_t = U\Sigma V^T$;
3.    $H_{b,L} \leftarrow U^T H_b$;
4.    $\tilde{\Sigma}_s \leftarrow (\Sigma^2 + \lambda I_r)$;
5.    Compute SVD of $\tilde{\Sigma}_s^{-1/2} H_{b,L} = U_b \Sigma_b V_b^T$;
6.    $G \leftarrow U \tilde{\Sigma}_s^{-1/2} U_b$.
```

4.2 Multiple model ($|\Lambda| > 1$)

Let $\Lambda = \{\lambda_1, \cdots, \lambda_m\}$ be the candidate set for the regularization λ. In multiple-model RLDA, v-fold cross-validation is applied, where the data is divided into v subsets of (approximately) equal size. All subsets are mutually exclusive, and in the i-th fold, the i-th subset is held out for test and all other subsets are used for training. For each λ_j, $j = 1, \cdots, m$, we compute the cross-validation accuracy, Accu(j), defined as the mean of the accuracies for all folds. The best regularization value λ_{j^*} is the one with

$$j^* = \arg\max_j \text{Accu}(j).$$

The pseudo-code for multiple-model RLDA is given in **Algorithm 2**. Note that the k-nearest neighbor ($k = 1$), called 1-NN, is used for classification as in [24].

4.3 Time complexity

Line 4 takes $O(n^2 d)$ time for the SVD computation. Lines 5 and 6 take $O(drk)$ and $O(rdn)$ time, respectively, for the matrix multiplication. For each choice λ_j, Line 9 and 10 take $O(rk^2)$ time for the eigen-decomposition and matrix multiplication. Line 11 takes $O(krn)$ time for the matrix multiplication. The computation of the classification accuracy by 1-NN in Line 12 takes $O(n^2 k)$ time. Thus, the total time complexity, $T(m)$, for estimating the best parameter is

$$\begin{aligned} T(m) &= O\left(v\left(n^2 d + rnd + drk \right.\right. \\ &\quad \left.\left. +\ m(rk^2 + krn + n^2 k)\right)\right) \\ &= O\left(v(n^2 d + mn^2 k)\right) \\ &= O\left(vn^2(d + mk)\right). \end{aligned}$$

We can compare $T(m)$ with $T(1)$, where $m = 1$, and obtain

$$\frac{T(m)}{T(1)} \approx \frac{vn^2(d + mk)}{vn^2(d + k)} \approx 1 + \frac{mk}{d}.$$

For small sample size problems, where the number, k, of classes is much smaller than the dimension d, i.e., $k \ll d$, the overhead of estimating the optimal regularization value among a large set is small.

```
Algorithm 2: Multiple-model RLDA
1. For $i = 1 : v$            / / $v$-fold cross validation
2.     Construct $A^i$ and $A^{\hat{i}}$;
   / / $A^i$ = $i$-th fold, for training
   / / $A^{\hat{i}}$ = rest, for testing
3.     Construct $H_t$ and $H_b$ using $A^i$;
4.     Compute the SVD of $H_t$: $H_t = U\Sigma V^T$;
5.     $H_{b,L} \leftarrow U^T H_b$, $r = \text{rank}(H_t)$;
6.     $A_L^i \leftarrow U^T A^{\hat{i}}$; $A_L^{\hat{i}} \leftarrow U^T A^{\hat{i}}$;
7.     For $j = 1 : m$    / / $m$ choices for $\lambda$
8.         $\tilde{\Sigma} \leftarrow (\Sigma^2 + \lambda_j I_r)^{-1/2}$;
9.         Compute SVD of $\tilde{\Sigma} H_{b,L} = U_b \Sigma_b V_b^T$;
10.        $G \leftarrow \tilde{\Sigma} U_b$;
11.        $A_L^i \leftarrow G^T A_L^i$; $A_L^{\hat{i}} \leftarrow G^T A_L^{\hat{i}}$;
12.        Run 1-NN on $\left(A_L^i, A_L^{\hat{i}}\right)$ and compute the
            accuracy, denoted as Accu($i, j$);
13.     EndFor
14. EndFor
15. Accu($j$) $\leftarrow \frac{1}{v} \sum_{i=1}^{v}$ Accu($i, j$);
16. $j^* \leftarrow \arg\max_j$ Accu($j$);
17. Output $\lambda_{j^*}$ as the best parameter.
```

4.4 Relationship between RLDA and ULDA

Uncorrelated LDA (ULDA) [24] was proposed for feature extraction on small sample size problems. One key property of ULDA is that the features in the transformed space are uncorrelated, thus ensuring minimum redundancy among the features in the reduced space. It was shown [24] that the transformation by ULDA consists of the first q eigenvectors of $S_t^+ S_b$, where $q = \text{rank}(S_b)$. Interestingly, we can show that the limit of RLDA when $\lambda \rightarrow 0$ is equivalent to ULDA based on the following lemma:

THEOREM 4.1. *Let S_t and S_b be defined as above and $\lambda >$ 0. Then*

$$\lim_{\lambda \rightarrow 0} (S_t + \lambda I_d)^{-1} S_b = S_t^+ S_b.$$

PROOF. The proof follows directly from Theorem 3.1. □

REMARK 4.1. *Theorem 4.1 implies that the range of the parameter λ in RLDA is $[0, \infty)$. Note that for traditional RLDA algorithms, the range of λ is $[\eta, \infty)$ for some positive η. When λ is close to 0, these algorithms tend to have numerical instability problems, since they follow the optimization problem in Eq. (8), while the proposed RLDA algorithm follow the one in Eq. (9).*

Theorem 4.1 implies that ULDA is a special case of RLDA when $\lambda = 0$. With a properly chosen λ through multiple-model RLDA in Section 4.2, RLDA is expected to outperform ULDA, which is confirmed by the empirical results presented in the next section. It was shown [24] that ULDA has the same computational cost as many other competitive LDA methods, including Orthogonal LDA. So in our experimental study on efficiency, we will concentrate on the comparison between the single-model RLDA and multiple-model RLDA.

One interesting property of ULDA [26] is that under a mild condition that

$$\mathrm{rank}(S_b) + \mathrm{rank}(S_w) = \mathrm{rank}(S_t), \tag{15}$$

the optimal transformation G lies in the null space of the within-class scatter matrix, that is, $G^T S_w = 0$. It has been shown [26] that the condition in Eq. (15) holds for many high-dimensional data, including most datasets used in our studies in Section 5. We show in the following that if $G^T S_w = 0$, then ULDA maps all points from the same class to a common vector.

Proposition 4.1. *Let G be the transformation in ULDA, and let x be a data point from the i-th class. Assume $G^T S_w = 0$. Then $G^T x = G^T c^{(i)}$, where $c^{(i)}$ is the centroid of the i-th class. That is, all data points from the i-th class are mapped to the common vector $G^T c^{(i)}$.*

Proof. Since $G^T S_w = 0$, we have

$$0 = G^T S_w G = G^T H_w H_w^T G, \tag{16}$$

where H_w is defined as in Eq. (4):

$$H_w = [(A_1 - c^{(1)}(e^{(1)})^T), \cdots, (A_k - c^{(k)}(e^{(k)})^T)],$$

where A_i is the data matrix of the i-th class, and $e^{(i)}$ is the vector of all ones. It follows from Eq. (16) that $G^T H_w = 0$. Considering the i-th block of $G^T H_w$, we have that

$$G^T \left(A_i - c^{(i)}(e^{(i)})^T\right) = \left(G^T A_i - G^T c^{(i)}(e^{(i)})^T\right) = 0.$$

Hence, $G^T x = G^T c^{(i)}$, for each column x in A_i. This completes the proof of the proposition. □

Proposition 4.1 above shows that under a mild condition, ULDA maps all points from the same class to a common point. This leads to perfect separation between different classes, however, this may also lead to overfitting. The problem may be worse especially when the data is noisy. Regularization applied in RLDA is thus expected to alleviate this problem, provided that a good regularization parameter can be estimated.

5. EXPERIMENTS

In this section, we experimentally evaluate the performance of RLDA. All of our experiments are performed on a P4 2.80GHz Linux machine with 1GB memory. As in [5], the data is randomly partitioned into a training set consisting of two-thirds of the whole set and a test set consisting of one-third of the whole set. 1-Nearest-Neighbor (1-NN) algorithm is applied for classification. To give a better estimation of accuracy, the splitting is repeated 50 times and the resulting accuracies are averaged.

5.1 Datasets

We use three types of data in our studies: text documents (Doc1 and Doc2), gene expression data (GCM and ALL), and face images (ORL and PIX).

- Doc1 and Doc2 are from *Reuters-21578* text categorization test collection Distribution 1.0.[1] Doc1 has 4 classes, each with 80 instances; its dimension is 2887. Doc2 has 5 classes, each with 98 instances; its dimension is 3759.
- GCM has 14 classes (cancer types) and totally 198 instances (human tumor samples); its dimension is 16063. The dataset was first studied in [20, 27]. ALL [28] has 6 classes (diagnostic groups) and totally 248 instances; its dimension is 12558.
- ORL[2] has 40 classes (persons), each with 10 instances; its dimension is 10304 (same as original image size 92 × 112). PIX[3] has 30 classes (persons), each with 10 instances; its dimension is 10000 (subsampled from original image size 512 × 512).

The statistics of the test datasets are summarized in table 1.

Table 1: Statistics for our test datasets.

dataset	size (n)	dimensionality (d)	# of classes (k)
Doc1	320	2887	4
Doc2	490	3759	5
GCM	198	16063	14
ALL	248	12558	6
ORL	400	10304	40
PIX	300	10000	30

5.2 Efficiency

In this experiment, we test the efficiency of multiple-model RLDA. Table 2 shows the computational time (in seconds) of RLDA on different m, i.e., the size of Λ, ranging from 1 to 1024. It is clear that the cost of multiple-model RLDA grows slowly as m increases. For example, we can observe that $T(1024)/T(1)$ on different datasets ranges from 1.35 to 4.42. Among the six datasets, the two image datasets have relatively larger increasing rate than the others. Note that the number of classes for the image datasets is relatively larger than the others, and so is the ratio k/d. The experimental results on efficiency evaluation are consistent with the theoretical estimation given in Section 4.2.

5.3 Classification performance

In this experiment, we evaluate RLDA in classification and compare it with other three LDA-based methods: ULDA, DLDA [5], and OLDA [24], as well as SVM[4]. The results are summarized in Table 3. We set m to be 1000 for all cases. We have found that using a small value of m usually degrades the classification performance for our datasets. This confirms the effectiveness of using large candidate set to choose a good regularization value.

The results in Table 3 show that RLDA is competitive with other methods in classification. RLDA outperforms ULDA in most cases. For Doc1, GCM, and ORL, RLDA outperforms ULDA by a large margin. Interestingly, ULDA

[1] www.research.att.com/∼lewis

[2] www.uk.research.att.com/facedatabase.html

[3] peipa.essex.ac.uk/ipa/pix/faces/manchester/test-hard/

[4] Linear SVM is used as it is shown to be comparable to nonlinear SVM using kernels [3, 21] in most cases due to the high dimensionality of the data. The value of the regularization parameter in SVM is estimated through cross-validation.

Table 2: Computational time (in seconds) of RLDA for different $m = |\Lambda|$.

m	Doc1	Doc2	GCM	ALL	ORL	PIX
1	6.68	19.14	16.04	20.85	39.95	22.66
2	6.68	19.19	16.10	22.23	39.98	22.57
4	6.77	19.23	16.15	22.11	40.42	22.67
8	6.86	19.32	16.33	22.34	40.75	23.15
16	6.96	19.53	16.43	22.48	41.84	23.72
32	7.13	20.35	16.46	22.92	44.15	24.38
64	7.32	21.47	17.18	23.31	48.25	26.30
128	7.80	22.90	17.92	23.63	56.85	30.24
256	8.87	26.84	19.91	23.99	74.24	37.59
512	11.01	34.36	23.36	24.66	107.9	52.92
1024	15.36	49.59	30.15	28.14	176.7	81.74
$T(1024)/T(1)$	2.30	2.59	1.88	1.35	4.42	3.61
$k/d(\times$ 1e3)	1.39	1.33	0.87	0.48	3.88	3.00

has a better performance than RLDA for Doc2. However, the difference is not significant. Recall from Section 4.4 that ULDA is equivalent to RLDA with $\lambda = 0$. The experimental results further confirm the effectiveness of choosing the best λ from a large set of candidates.

OLDA appears to be very competitive with RLDA in many cases. However, there exist certain cases (such as the Doc1 dataset), where the difference is significant. It appears that RLDA, with a sufficiently large set of regularization candidates, is more robust to the diversity of training data than other LDA-based methods. Overall, RLDA is competitive with SVM.

6. CONCLUSIONS

An efficient algorithm for RLDA is proposed for small sample size problems. A key advantage of the proposed algorithm is that the optimal transformation of RLDA for a set of different regularization values can be computed with approximately the same cost as running the RLDA algorithm a very small number of times. Thus it dramatically reduces the computational cost for RLDA.

We analyze the intrinsic relationship between RLDA and ULDA. More specifically, we show that RLDA without any regularization is equivalent to ULDA, while ULDA maps all data points from the same class to a common point, under a mild condition which has been shown to hold for many high-dimensional data. The theoretical analysis presented provides insights on the use of regularization in RLDA. Experiments on a variety of data show that RLDA is competitive with several other LDA-based methods and SVM, in terms of classification, which shows the effectiveness of regularization applied in RLDA.

Discriminant analysis can also be studied in the non-linear fashion, so-called kernel discriminant analysis [1, 18, 19]. It is desirable if the data has weak linear separability. One of the directions for future work is to extend the current work to the nonlinear case.

7. REFERENCES

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

[2] P.N. Belhumeour, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[3] N. Cristianini and J.S. Taylor. *Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[4] D.Q. Dai and P.C. Yuen. Regularized discriminant analysis and its application to face recognition. *Pattern Recognition*, 36:845–847, 2003.

[5] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.

[6] M. Dundar, G. Fung, J. Bi, S. Sathyakama, and B. Rao. Sparse Fisher discriminant analysis for computer aided detection. In *Proceedings of the SIAM International Conference on Data Mining*, 2005.

[7] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[8] J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[9] K. Fukunaga. *Introduction to Statistical Pattern Classification*. Academic Press, San Diego, California, USA, 1990.

[10] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.

[11] Y. Guo, T. Hastie, and R. Tibshirani. Regularized discriminant analysis and its application in microarrays. *Technical report, Stanford University*, 2003.

[12] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23:73–102, 1995.

[13] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, 2004.

Table 3: Comparison of classification accuracy: The mean and standard deviation (in parenthesis) of accuracies from fifty runs are reported.

Datasets[a]	RLDA[b]	ULDA[b]	DLDA[b]	OLDA[b]	SVM[b]
Doc1	85.28 (3.21)	79.94 (4.16)	75.09 (3.71)	80.04 (4.18)	**86.14 (3.16)**
Doc2	94.67 (1.61)	**94.69 (1.67)**	83.04 (5.79)	94.11 (1.87)	94.54 (1.88)
GCM	**81.82 (3.66)**	75.05 (6.32)	66.19 (5.25)	81.14 (3.98)	70.31 (5.15)
ALL	**98.07 (1.26)**	97.20 (2.01)	97.42 (1.44)	**98.07 (1.26)**	97.23 (1.64)
ORL	**97.88 (1.45)**	92.72 (2.07)	97.77 (1.21)	97.15 (1.45)	97.55 (1.34)
PIX	98.56 (1.53)	96.29 (1.91)	98.51 (1.78)	98.42 (1.54)	**98.84 (1.48)**

[a] The partition of the data into training and test set in GCM and ALL is different from that used in [24].

[b] RLDA, ULDA, DLDA, OLDA, and SVM stand for regularized LDA, uncorrelated LDA, diagonal LDA, orthogonal LDA, and Support Vector Machines, respectively.

[14] T. Hastie and R. Tibshirani. Efficient quadratic regularization for expression arrays. *Biostatistics*, 5(3):329–340, 2004.

[15] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer, 2001.

[16] Z. Jin, J.Y. Yang, Z.S. Hu, and Z. Lou. Face recognition based on the uncorrelated discriminant transformation. *Pattern Recognition*, 34:1405–1416, 2001.

[17] W.J. Krzanowski, P. Jonathan, W.V McCarthy, and M.R. Thomas. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied Statistics*, 44:101–115, 1995.

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

[19] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, pages 41–48. IEEE, 1999.

[20] S. Ramaswamy et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Science*, 98(26):15149–15154, 2001.

[21] S. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines,Regularization, Optimization and Beyond*. MIT Press, 2002.

[22] V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[23] G. Wahba. *Spline Models for Observational Data*. Society for Industrial & Applied Mathematics, 1998.

[24] J. Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6:483–502, 2005.

[25] J. Ye and T. Wang. Regularized discriminant analysis for high-dimensional, low sample size data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 454–463, 2006.

[26] J. Ye and T. Xiong. Null space versus orthogonal linear discriminant analysis. In *Proceedings of the International Conference on Machine Learning*, pages 1073–1080, 2006.

[27] C.H. Yeang et al. Molecular classification of multiple tumor types. *Bioinformatics*, 17(1):1–7, 2001.

[28] E.J. Yeoh et al. Classification, subtype discovery, and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143, 2002.