Sparse Uncorrelated Linear Discriminant Analysis

Delin Chu MATCHUDL@NUS.EDU.SG
Xiaowei Zhang G0800887@NUS.EDU.SG
Department of Mathematics, National University of Singapore, Singapore 119076

Abstract

In this paper, we develop a novel approach for sparse uncorrelated linear discriminant analysis (ULDA). Our proposal is based on characterization of all solutions of the generalized ULDA. We incorporate sparsity into the ULDA transformation by seeking the solution with minimum ℓ_1-norm from all minimum dimension solutions of the generalized ULDA. The problem is then formulated as a ℓ_1-minimization problem and is solved by accelerated linearized Bregman method. Experiments on high-dimensional gene expression data demonstrate that our approach not only computes extremely sparse solutions but also performs well in classification. Experimental results also show that our approach can help for data visualization in low-dimensional space.

1. Introduction

Linear discriminant analysis (LDA) is a popular tool for both classification and dimension reduction that seeks an optimal linear transformation of the data into a low-dimensional space, where the data achieves maximum class separability (Fukunaga, 1990; Hastie et al., 2009). The optimal transformation is computed by solving a generalized eigenvalue problem. LDA has been widely employed in numerous applications in science and engineering, including microarray data analysis, information retrieval and face recognition. Despite the simplicity and popularity of LDA, there are two deficiencies that restrict its application in high-dimensional data analysis, where the dimension of the data space is usually thousands. One deficiency is that the classical LDA can not be applied directly to undersampled problems, that is, the dimension of the data space is larger than the number of data samples, due to singularity of the scatter matrices; the other is the lack of sparsity in the LDA solution.

To overcome the first problem, many extensions of the classical LDA have been proposed. These extensions include uncorrelated LDA (ULDA) (Jin et al., 2001; Ye, 2005; Chu et al., 2011), regularized LDA (Friedman, 1989), GSVD-based LDA (LDA/GSVD)(Howland & Park, 2004), and least squares LDA (Ye, 2007). Of these approaches, ULDA has an advantage over other approaches, that is, the feature vectors extracted by ULDA are mutually uncorrelated in the low-dimensional space. This property is highly desirable for feature extraction in many applications in order to contain minimum redundancy (Ye et al., 2004; 2006).

Sparsity in the LDA solution is generally desirable for high-dimensional data analysis as it makes the interpretation of the extracted features much easier. For LDA, each extracted feature in the transformed space is a linear combination of all the features of original data and the coefficients of such linear combination are generally nonzero, which makes the interpretation of the extracted features difficult. To overcome this problem, many attempts have been made to incorporate sparsity into the LDA transformation, for instance, the greedy algorithms ESLDA and GSLDA (Moghaddam et al., 2006), the Penalized LDA (PLDA) (Witten & Tibshirani, 2011), and the Sparse Discriminant Analysis (SLDA) (Clemmensen et al., 2011). Almost all existing sparse LDA algorithms introduce sparsity by adding ℓ_1 penalty (i.e., Lasso penalty (Tibshirani, 1996)) or its variants of the transformation matrix to objective functions, and thus, the computed sparse transformation is not a solution of LDA but an approximation. Besides interpretability, sparse LDA may be motivated by robustness to the noise, or computational efficiency in prediction. Some significant applications of sparse LDA can be found in (Dundar et al., 2005; Fung & Ng, 2007; Wu et al., 2009).

In this paper we study sparse ULDA (SULDA) which

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

extracts mutually uncorrelated features and computes sparse LDA transformation, simultaneously. We first characterize all solutions of the generalized ULDA via solving the optimization problem proposed in (Ye, 2005), then compute the sparse solution of ULDA by finding the minimum ℓ_1-norm solution from all the solutions with minimum dimension. Finding minimum ℓ_1-norm solution can be formulated as a ℓ_1-minimization problem, which is solved by Accelerated Linearized Bregman method (Yin et al., 2008; Cai et al., 2009; Yin, 2010; Huang et al., 2011). Different from existing sparse LDA algorithms, our approach seeks a sparse solution of ULDA directly from the solution set of ULDA, so the computed sparse transformation is a solution of ULDA, which further implies that the extracted features by SULDA are mutually uncorrelated.

This paper is organized as follows. We briefly review LDA and ULDA in Section 2 and derive a characterization of all solutions of generalized ULDA In Section 3. Based on the characterization we develop a novel sparse ULDA algorithm SULDA in Section 4, then test SULDA on real world data and compare its performance with existing sparse LDA algorithms in Section 5. Finally, we conclude this paper in Section 6.

2. Overview of LDA and ULDA

Given a data matrix $A \in \mathbf{R}^{m\times n}$ consisting of n samples from $\mathbf{R}^m$. We assume $A = [a_1\ a_2\ \cdots\ a_n] = \left[\ \mathcal{A}_1\quad \mathcal{A}_2\quad \cdots\quad \mathcal{A}_k\ \right]$, where $a_j \in \mathbf{R}^m$ $(1 \le j \le n)$, n is the sample size, k is the number of class and $\mathcal{A}_i \in \mathbf{R}^{m\times n_i}$ with n_i denoting the number of data in the ith class. So we have $n = \sum_{i=1}^k n_i$. Classical LDA aims to compute an optimal linear transformation $G^T \in \mathbf{R}^{l\times m}$ that maps a_i in the m-dimensional space to a vector a_i^L in the l-dimensional space

$$G^T : a_i \in \mathbf{R}^m \to a_i^L \in \mathbf{R}^l,$$

where $l \ll m$, so that the class structure in the original data is preserved in the l-dimensional space.

In discriminant analysis (Fukunaga, 1990), the between-class scatter matrix S_b, within-class scatter matrix S_w and total scatter matrix S_t are defined as:

$$S_b = \frac{1}{n}\sum_{i=1}^{k} n_i(c^{(i)} - c)(c^{(i)} - c)^T,$$
$$S_w = \frac{1}{n}\sum_{i=1}^{k}\sum_{a_j\in\mathcal{A}_i}(a_j - c^{(i)})(a_j - c^{(i)})^T,$$
$$S_t = \frac{1}{n}\sum_{j=1}^{n}(a_j - c)(a_j - c)^T,$$

where $c^{(i)} = \frac{1}{n_i}\mathcal{A}_i e_i$ with $e_i = [1\ \cdots\ 1]^T \in \mathbf{R}^{n_i}$ denotes the centroid of class i and $c = \frac{1}{n}\mathcal{A}e$ with $e = [1\ \cdots\ 1]^T \in \mathbf{R}^n$ denotes the global centroid. It follows from the definition that $S_t = S_b + S_w$. Moreover, let

$$H_b = \frac{1}{\sqrt{n}}[\sqrt{n_1}(c^{(1)} - c)\ \cdots\ \sqrt{n_k}(c^{(k)} - c)] \in \mathbf{R}^{m\times k},$$
$$H_w = \frac{1}{\sqrt{n}}[\mathcal{A}_1 - c^{(1)}e_1^T\ \cdots\ \mathcal{A}_k - c^{(k)}e_k^T] \in \mathbf{R}^{m\times n},$$
$$H_t = \frac{1}{\sqrt{n}}[a_1 - c\ \cdots\ a_n - c] = A - ce^T \in \mathbf{R}^{m\times n},$$

then the scatter matrices can be expressed as

$$S_b = H_bH_b^T,\quad S_w = H_wH_w^T,\quad S_t = H_tH_t^T. \tag{1}$$

Trace of the scatter matrices can be used to measure the quality of the class structure, where Trace(S_b) measures the distance between classes and Trace(S_w) measures the closeness of the data within the classes over all k classes.

In the low-dimensional space mapped by the linear transformation $G^T \in \mathbf{R}^{l\times m}$, the between-class, within-class and total scatter matrices are of the forms

$$S_b^L = G^TS_bG,\quad S_w^L = G^TS_wG,\quad S_t^L = G^TS_tG.$$

An optimal transformation G^T should maximize Trace(S_b^L) and minimize Trace(S_w^L) simultaneously, which results in a common criterion for classical LDA:

$$G^* = \arg\max_G\{\mathrm{Trace}((S_t^L)^{-1}S_b^L)\}. \tag{2}$$

In classical LDA (Fukunaga, 1990), the optimization problem above is solved by computing all the eigenpairs

$$S_bx = \lambda S_tx,\quad \lambda \ne 0,$$

and thus, the optimal G^* consists of eigenvectors of $S_t^{-1}S_b$ corresponding to all nonzero eigenvalues, provided that S_t is nonsingular. Since rank$(S_b) \le k-1$, the reduced dimension by classical LDA is at most $k-1$. Classical LDA can not be applied directly when S_t is singular, which is the case for undersampled problems.

To deal with the singularity of S_t, many generalizations of classical LDA have been proposed. One popular generalization is the generalized ULDA (Ye, 2005; Jin et al., 2001)

$$\begin{aligned} G^* &= \arg\max_{G^TS_tG=I}\mathrm{Trace}((S_t^L)^{(+)}S_b^L) \\ &= \arg\max_{G^TS_tG=I}\mathrm{Trace}(S_b^L), \end{aligned}\tag{3}$$

where $(S_t^L)^{(+)}$ denotes the pseudo-inverse (Golub & Loan, 1996) of S_t^G. Due to the constraint $G^T S_t G = I$, the extracted features are mutually uncorrelated in the l-dimensional space. An algorithm, based on the simultaneous diagonalization of the scatter matrices, was proposed in (Ye, 2005) for computing the optimal solution of optimization problem (3). Recently, an eigendecomposition-free and SVD-free ULDA algorithm was developed in (Chu et al., 2011) to improve the efficiency of the generalized ULDA. Some applications of ULDA can be found in (Jin et al., 2001; Ye et al., 2004; 2006).

3. Characterization of all solutions of generalized ULDA

We characterize all solutions of the optimization problem (3) explicitly in Theorem 1, which is based on singular value decomposition (SVD) (Golub & Loan, 1996) and simultaneous diagonalization of scatter matrices. The detailed proof is given in the Appendix.

Theorem 1. *Let the reduced SVD of H_t be*

$$H_t = U_1 \Sigma_t V_1^T, \tag{4}$$

where $U_1 \in \mathbf{R}^{m\times\gamma}$ and $V_1 \in \mathbf{R}^{n\times\gamma}$ are column orthogonal, and $\Sigma_t \in \mathbf{R}^{\gamma\times\gamma}$ is diagonal and nonsingular with $\gamma = rank(H_t) = rank(S_t)$. Next, let the reduced SVD of $\Sigma_t^{-1}U_1^T H_b$ be

$$\Sigma_t^{-1} U_1^T H_b = P_1 \Sigma_b Q_1^T, \tag{5}$$

where $P_1 \in \mathbf{R}^{\gamma\times q}$, $Q_1 \in \mathbf{R}^{k\times q}$ are column orthogonal, $\Sigma_b \in \mathbf{R}^{q\times q}$ is diagonal and nonsingular. Then $q = rank(H_b) = rank(S_b)$, and G is a solution of the optimization problem (3) if and only if $q \le l \le \gamma$ and

$$G = \left(U_1\Sigma_t^{-1}\begin{bmatrix}P_1 & \mathcal{M}_1\end{bmatrix} + \mathcal{M}_2\right)\mathcal{Z}, \tag{6}$$

where $\mathcal{M}_1 \in \mathbf{R}^{\gamma\times(l-q)}$ is column orthogonal satisfying $\mathcal{M}_1^T P_1 = 0$, $\mathcal{M}_2 \in \mathbf{R}^{m\times l}$ is an arbitrary matrix satisfying $\mathcal{M}_2^T U_1 = 0$, and $\mathcal{Z} \in \mathbf{R}^{l\times l}$ is orthogonal.

A similar result as Theorem 1 has been established in (Chu et al., 2011), where the optimal solution to the optimization problem (3) is computed by means of economic QR decomposition with/without column pivoting.

When we compute the optimal linear transformation G^* of LDA for data dimensionality reduction, we prefer the dimension of the transformed space to be as small as possible. Hence, we parameterize all minimum dimension solutions of the optimization problem (3) in Corollary 2 which is a special case of Theorem (1) with $l = q$.

Corollary 2. *$G \in \mathbf{R}^{m\times l}$ is a minimum dimension solution of the optimization problem (3) if and only if $l = q$ and*

$$G = (U_1\Sigma_t^{-1}P_1 + \mathcal{M}_2)\mathcal{Z}, \tag{7}$$

where $\mathcal{M}_2 \in \mathbf{R}^{m\times q}$ is any matrix satisfying $\mathcal{M}_2^T U_1 = 0$ and $\mathcal{Z} \in \mathbf{R}^{q\times q}$ is orthogonal.

Another motivation for considering minimum dimension solutions of (3) is that results in (Chu et al., 2011) show that among all solutions of the optimization problem (3), minimum dimension solutions maximize the ratio $\text{Trace}(S_b^L)/\text{Trace}(S_w^L)$, which is also an important measure of class discrimination in LDA (Fukunaga, 1990).

From both equations (6) and (7), we see that the optimal solution G^* of generalized ULDA equals to the summation of two factors, $U_1\Sigma_t^{-1}\begin{bmatrix}P_1 & \mathcal{M}_1\end{bmatrix}\mathcal{Z}$ in the range space of S_t and $\mathcal{M}_2\mathcal{Z}$ in the null space of S_t. Since the factor $\mathcal{M}_2\mathcal{Z}$ belongs to $\text{null}(S_b) \cap \text{null}(S_w)$, it does not contain discriminative information. However, with the help of factor $\mathcal{M}_2\mathcal{Z}$ we can construct a sparse solution of ULDA in the next section.

4. Sparse ULDA

Note from Corallary 2 that G is a minimum dimension solution of the optimization problem (3) if and only if equality (7) holds, which is equivalent to

$$U_1^T G = \Sigma_t^{-1} P_1 \mathcal{Z},\ \ \mathcal{Z}^T\mathcal{Z} = I. \tag{8}$$

The main idea of our sparse ULDA algorithm is to find the sparsest solution of ULDA from all G satisfying (8). A natural way to do this is to find a matrix G that minimizes the ℓ_0-norm (cardinality). However, ℓ_0-norm is non-convex and NP-hard. Therefore, in our sparse ULDA, we replace the ℓ_0-norm with its convex relaxation ℓ_1-norm, which results in the following optimization problem

$$\begin{aligned} G^* = \arg\min_{G\in\mathbf{R}^{m\times q}} \quad & \|G\|_1 \\ s.t. \quad & U_1^T G = \Sigma_t^{-1}P_1\mathcal{Z},\ \mathcal{Z}^T\mathcal{Z} = I \end{aligned} \tag{9}$$

where $\|G\|_1$ is defined as $\|G\|_1 = \sum_{i=1}^m \sum_{j=1}^q |G_{ij}|$.

Note that $\mathcal{Z} \in \mathcal{R}^{q\times q}$ in (9) is orthogonal. However, on one hand, there still lack numerically efficient methods for solving optimization problems over the set of orthogonal matrices. On the other hand, it can introduce at most q^2 additional zeros in G by optimizing $\mathcal{Z}$ over all $q \times q$ orthogonal matrices assuming the the zero structure of the previous G is not destroyed; but, usually, $q < k << m$, so the number of the additional zeros in G introduced by optimizing $\mathcal{Z}$ is very small compared with mq. So it is acceptable that G^* in (9) is computed with a fixed $\mathcal{Z}$ ($\mathcal{Z} = I_q$ in our experiments).

When $q = 1$, the ℓ_1-minimization problem (9) is reduced to the basis pursuit problem

$$x_1^* = \arg\min\{\|x\|_1 : \ x \in \mathbf{R}^n, \ \mathcal{A}x = b\}, \qquad (10)$$

which has been studied extensively (Yin et al., 2008; Cai et al., 2009; Yin, 2010; Huang et al., 2011). This means that the numerical methods for solving (10) can be automatically extended to solve (9).

The linearized Bregman method (Yin et al., 2008; Cai et al., 2009; Yin, 2010) is considered as one of the most powerful methods for solving problem (10), and it has been accelerated in a recent study (Huang et al., 2011). The accelerated linearized Bregman method for (10) is:

$$\begin{cases} x^{k+1} = \delta \mathcal{S}_\mu(\tilde{v}^k), \\ v^{k+1} = \tilde{v}^k - \tau \mathcal{A}^T(\mathcal{A}x^{k+1} - b), \ k \geq 0, \\ \tilde{v}^{k+1} = \alpha_k v^{k+1} + (1 - \alpha_k) v^k, \end{cases} \qquad (11)$$

where $\tilde{v}^0 = v^0 = \tau \mathcal{A}^T b$, δ, μ and τ are positive parameters, $\alpha_k = \frac{2k+3}{k+3}$ and $\mathcal{S}_\mu(\cdot)$ is the componentwise soft-thresholding operator

$$\mathcal{S}_\mu(x) = \text{sign}(x) \odot \max\{|x| - \mu, 0\}.$$

Extending the accelerated linearized Bregman method (11) to the optimization problem (9), we get an analogue as follows:

$$\begin{cases} G^{k+1} = \delta \mathcal{S}_\mu(\tilde{V}^k), \\ V^{k+1} = \tilde{V}^k - \tau U_1(U_1^T G^{k+1} - \Sigma_t^{-1} P_1 \mathcal{Z}), \\ \tilde{V}^{k+1} = \alpha_k V^{k+1} + (1 - \alpha_k) V^k, \end{cases} \qquad (12)$$

where $\tilde{V}^0 = V^0 = \tau U_1 \Sigma_t^{-1} P_1 \mathcal{Z}$.

The convergence of the accelerated linear Bregman method is given in the following theorem which is a natural extension of results in (Cai et al., 2009; Yin, 2010; Huang et al., 2011).

Theorem 3. *Assume that* $0 < \delta < \frac{1}{\|(U_1^T)(U_1^T)^T\|_2} = 1$, $0 < \tau < \frac{1}{\delta}$, *and denote* G^* *as the solution of optimization problem* (9) *that has the minimal Frobenius norm among all solutions, that is*

$$G^* = \mathit{arg}\min\{\|G\|_F : G \ \textit{solves} \ (9)\}.$$

Then the sequence $\{G^k\}$ *generated by* (12) *converges to the unique solution of the optimization problem*

$$G_\mu^* = \mathit{arg}\min\{\mu\|G\|_1 + \frac{1}{2\delta}\|G\|_F^2 : \Sigma_t U_1^T G = P_1 \mathcal{Z}\}, \qquad (13)$$

that is,

$$\lim_{k \to \infty} \|G^k - G_\mu^*\|_F = 0.$$

Moreover, we have

$$\lim_{\mu \to \infty} \|G_\mu^* - G^*\|_F = 0.$$

In particular, for any fixed δ, *there exists a finite* μ^* *such that*

$$G_\mu^* = G^*, \ \forall \mu \geq \mu^*.$$

Algorithm 1 Sparse ULDA (SULDA)

```
**Input:** data $A \in \mathbf{R}^{m \times n}$ and tolerance $\epsilon > 0$
Compute the reduced SVDs (4) and (5)
Let $\mathcal{Z} = I_q$, $\tilde{V}^0 = V^0 = \tau U_1 \Sigma_t^{-1} P_1 \mathcal{Z}$
repeat
    Compute $G^{k+1}$ by (12)
    $error = \|U_1^T G^{k+1} - \Sigma_t^{-1} P_1 \mathcal{Z}\|_F$
until $error \leq \epsilon$
Output: $G^{k+1}$
```

Theorem 3 shows that the sequence $\{G^k\}$ generated by accelerated linearized Bregman method converges to the unique solution of (13). Moreover, if μ is large enough, then the limit point of $\{G^k\}$ is the solution of (9) that has the minimal Frobenius norm.

We are now ready to present our sparse ULDA algorithm, which is described in Algorithm 1. In Algorithm 1, we adopt the following stopping criterion

$$\|U_1^T G^k - \Sigma_t^{-1} P_1 \mathcal{Z}\|_F \leq \epsilon, \qquad (14)$$

where $\epsilon > 0$ is a tolerance parameter. Let

$$\Delta_k = U_1^T G^k - \Sigma_t^{-1} P_1 \mathcal{Z},$$

then $U_1^T G^k = \Delta_k + \Sigma_t^{-1} P_1 \mathcal{Z}$, $\|\Delta_k\|_F \leq \epsilon$, and

$$(G^k)^T S_t G^k = I_q + \Delta_k^T \Sigma_t P_1 \mathcal{Z} + \mathcal{Z}^T P_1^T \Sigma_t \Delta_k + \Delta_k^T \Sigma_t^2 \Delta_k.$$

The deviation of $G^T S_t G$ from I_q at the kth iteration can be measured by

$$\begin{aligned} & \frac{\|(G^k)^T S_t G^k - I_q\|_F}{\sqrt{q}} \\ \leq & \frac{\|\Delta_k^T \Sigma_t^2 \Delta_k\|_F + 2\|\mathcal{Z}^T P_1^T \Sigma_t \Delta_k\|_F}{\sqrt{q}} \\ \leq & \frac{\|\Sigma_t\|_2^2 \|\Delta_k\|_F^2 + 2\|\Sigma_t\|_2 \|\Delta_k\|_F}{\sqrt{q}} \\ = & \frac{\|\Sigma_t\|_2 \|\Delta_k\|_F (2 + \|\Sigma_t\|_2 \|\Delta_k\|_F)}{\sqrt{q}} \\ \leq & \frac{\|H_t\|_2 (2 + \|H_t\|_2 \epsilon) \epsilon}{\sqrt{q}} = \mathbf{O}(\epsilon). \end{aligned} \qquad (15)$$

Thus, it is expected that the extracted features are well uncorrelated if the tolerance ϵ is small enough.

5. Experimental Results

This section presents some experimental results comparing SULDA with other two sparse LDA algorithms, PLDA (Witten & Tibshirani, 2011) and SLDA (Clemmensen et al., 2011). PLDA applies a Lasso penalty in Fisher's LDA framework, where the within-class scatter matrix is approximated by its diagonal. SLDA applies an elastic net penalty to the Optimal Scoring problem. The codes of PLDA[1] and SLDA[2] are publicly avaiable from the authors in `R` and `MATLAB`, respectively.

In the implementation of SULDA, we selected tuning parameters $\delta = 0.9$, $\tau = 1$ ($< \frac{1}{\delta}$) and $\epsilon = 10^{-5}$. The tuning parameter of PLDA was selected by 10-fold cross-validation, while the tuning parameters of SLDA were selected so that the computed transformation has comparable spasity with SULDA.

We performed our experimental studies using five gene expression data sets: Colon, Leukemia, Prostate, SRBCT and Brain. Each data set was randomly split into training and test data using the following algorithm: within each class with n_i data, we randomly select $\lceil 0.5n_i \rceil$ of them as training data and the rest as testing data. The splitting was repeated 10 times. The statistics of the data sets are listed in Table 1.

Table 1. Data stuctures: data dimension (m), training size (n), the number of classes (k) and the number of testing data (# Test).

Data set	m	n	k	# Test
Colon	2000	31	2	31
Leukemia	3571	37	2	35
Prostate	6033	51	2	51
SRBCT	2308	32	4	31
Brain	5597	21	5	21

The results are summarized in Table 2, where 'Accuracy' denotes the classification accuracy obtained by 1NN classifier for SLDA and SULDA, 'Sparsity' denotes the percentage of zero entries in the solution G, that is,

$$\text{Sparsity} = \frac{\text{the number of zeros in } G}{m * q} \times 100\%,$$

'Orthogonality' measures the difference between $G^T S_t G$ and I_q which is defined as

$$\text{Orthogonality} = \frac{\|G^T S_t G - I_q\|_F}{\sqrt{q}},$$

'# Variable' denotes the number of selected variables.

We can observe from Table 2 that the overall performance of SULDA is better than the other two algorithms. In particular, SULDA achieves the highest classification accuracy for all data sets, followed by SLDA which achieves higher accuracy than PLDA. Regarding sparsity and the number of selected variables, SLDA and SULDA compute solutions with much higher sparsity than PLDA and select less variables, and SULDA achieves slightly higher sparsity than SLDA. An important advantage of SULDA over the other two algorithms is that the extracted features in the low-dimensional space are mutually uncorrelated. We see from the 'Orthogonality' column of Table 2 that for SULDA $\frac{\|G^T S_t G - I_q\|_F}{\sqrt{q}} = \mathbf{O}(\epsilon)$, which is consistent with error bound (15). However, for the other two algorithms $\frac{\|G^T S_t G - I_q\|_F}{\sqrt{q}}$ is relatively large, which implies that the extracted features are far away from uncorrelated.

A 2-dimensional visualization of the SRBCT data is shown in Figure 1, where the sample data were projected onto the first two sparse discriminant vectors (i.e., $l = 2$) computed by PLDA, SLDA and SULDA. We can see from Figure 1 that SULDA has the best class discrimination quality in the 2-dimensional space. For PLDA, class 1 and 4 intersect, while for SLDA, class 3 and 4 intersect. However, for SULDA, classes are well separated from each other and data points in the same class are close (training data from the same class were projected to the same point). In fact, when the training data consists of n linearly independent data, which is the case in our experiments, we can prove that $\gamma = \text{rank}(S_t) = n - 1$, $q = \text{rank}(S_b) = k - 1$ and $\text{rank}(S_t) = \text{rank}(S_b) + \text{rank}(S_w)$. In this case, the minimum dimension solutions G of ULDA belongs to $\text{range}(S_b) \cap \text{null}(S_w)$, which implies that $G^T H_w = 0$, and

$$G^T a_j = G^T c^{(i)} \quad \forall a_j \in \mathcal{A}_i \ (1 \le i \le k),$$

that is, all training data form the same class (class i) are projected to the same point ($G^T c^{(i)}$).

6. Conclution and Future Work

In this paper, we develop SULDA, an efficient algorithm that performs sparse uncorrelated LDA, based on the characterization of solutions of generalized ULDA. Specifically, we characterize all solutions of the optimization criterion (3) of generalized ULDA. Based on the characterization we incorporate sparsity into the transformation matrix by selecting the solution with minimum ℓ_1-norm from all minimum dimension solu-

[1] `http://cran.r-project.org/web/packages/penalizedLDA`

[2] `http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=5671`

Table 2. Numerical results for gene data over 10 training-test set splits: mean (and standard deviation) of classification accuracy, sparsity, orthogonality and the number of selected variables.

	Accuracy (%)	Sparsity (%)	Orthognality	# Variable
Colon				
PLDA	79.68 (4.82)	71.06 (8.41)	37.60 (8.86)	578.9 (168.2)
SLDA	80.97 (5.15)	97.94 (6.00e-4)	9.47 (1.64)	41.1 (1.2)
SULDA	83.87 (4.81)	98.49 (2.42e-4)	3.38e-6 (2.51e-6)	30.3 (0.5)
Leukemia				
PLDA	91.43 (9.62)	75.94 (8.50)	81.40 (29.24)	859.3 (303.4)
SLDA	94.00 (2.84)	98.95 (4.01e-4)	26.97 (6.35)	37.5 (1.4)
SULDA	94.86 (2.95)	98.99 (8.86e-5)	2.46e-6 (2.12e-6)	36.1 (0.3)
Prostate				
PLDA	76.67 (14.95)	83.24 (11.29)	66.57 (44.87)	1,011.4 (681.2)
SLDA	90.20 (3.20)	97.97 (2.10e-4)	15.42 (1.54)	122.5 (1.3)
SULDA	91.37 (3.94)	99.17 (1.17e-16)	4.69e-6 (4.27e-6)	50 (0)
SRBCT				
PLDA	95.48 (4.35)	82.71 (6.14)	30.65 (9.47)	962.8 (324.8)
SLDA	97.74 (2.18)	97.92 (2.26e-4)	36.34 (5.03)	139.8 (2.7)
SULDA	99.35 (1.36)	98.65 (7.61e-05)	3.91e-6 (1.59e-6)	79.6 (3.7)
Brain				
PLDA	45.24 (35.23)	86.33 (18.30)	36.96 (47.17)	1,762.7 (2292.7)
SLDA	79.05 (8.16)	98.99 (6.40e-5)	6.57 (0.49)	223.7 (2.3)
SULDA	80.00 (6.66)	99.64 (5.07e-5)	7.49e-6 (2.47e-6)	77.4 (2.6)

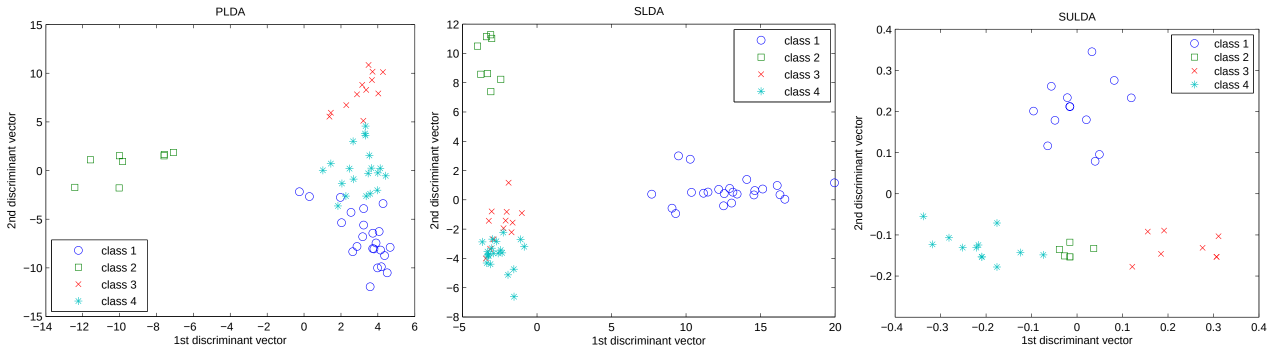

Figure 1. 2D visualization of the SRBCT data: the samples are projected onto the first two sparse discriminant vectors obtained by PLDA (left), SLDA (middle) and SULDA (right), respectively.

tions of ULDA. The resulting ℓ_1-minimization problem is solved by accelerated linearized Bregman method.

Different from existing sparse LDA algorithms, SULDA seeks a sparse solution directly from the solution set of ULDA. Thus, the computed sparse transformation is a solution of ULDA, instead of an approximation. This implies that features extracted by SULDA are mutually uncorrelated, which ensures minimum redundancy in the low-dimensional space.

Computationally, SULDA is easy to implement as it requires only matrix factorization (SVD) and multiplication. The effectiveness of SULDA is supported by experimental results using gene expression data. In our experiments, SULDA consistently outperformed its competitors in terms of both classification accuracy and interpretability (sparsity and number of used variables). The resulting sparse transformation can also be used to visualize observations and inspect class discrimination in the low-dimensional space.

In the derivation of SULDA, we fixed the orthogonal matrix $\mathcal{Z}$. One future focus is to consider arbitrary orthogonal $\mathcal{Z}$ and select the sparse solution of ULDA from a larger solution set. Another potential extension is to consider sparse kernel LDA using similar idea.

Moreover, it is well-known that LDA tends to give undesired results if samples in a class form several separate clusters (i.e., multimodal), and many extensions of LDA have been proposed to deal with this problem

(Hastie & Tibshirani, 1996; Sugiyama, 2007). In the future, we plan to extent the idea of SULDA to discriminant analysis approaches that can handle multimodal labeled data.

References

Cai, J., Osher, S., and Shen, Z. Convergence of the linearized bregman iteration for ℓ_1-norm minimization. *Mathematics of Computation*, 78(268):2127–2136, 2009.

Chu, D., Goh, S. T., and Hung, Y. S. Characterization of all solutions for undersampled uncorrelated linear discriminant analysis problems. *SIAM Journal on Matrix Analysis and Applications*, 32(3):820–844, 2011.

Clemmensen, L., Hastie, T., Wiiten, D., and Ersbøll, B. Sparse discriminant analysis. *Technometrics*, 53 (4):406–413, 2011.

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

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

Fukunaga, K. *Introduction to Statistical Pattern Recognition*. Academic Press, 2nd edition, 1990.

Fung, E. and Ng, M. K. On sparse fisher discriminant method for microarray data analysis. *Bioinformation*, 2(5):230–234, 2007.

Golub, G. and Loan, C. F. Van. *Matrix Computations*. The Johns Hopkins University Press, 3rd edition, 1996.

Hastie, T. and Tibshirani, R. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):155–176, 1996.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition, 2009.

Howland, P. and Park, H. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:995–1006, 2004.

Huang, B., Ma, S., and Goldfarb, D. Accelerated linearized bregman method. 2011. URL http://arxiv.org/abs/1106.5413.

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

Moghaddam, B., Weiss, Y., and Avidan, S. Generalized spectral bounds for sparse lda. In *Preceedings of the 23th International Conference on Machine Learning*, pp. 641–648, 2006.

Sugiyama, M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, 2007.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series* B, 58:15149–15154, 1996.

Witten, D. M. and Tibshirani, R. Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772, 2011.

Wu, M., Zhang, L., Wang, Z., Christiani, D., and Lin, X. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25 (9):1145–1151, 2009.

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

Ye, J. Least squares linear discriminant analysis. In *Preceedings of the 24th International Conference on Machine Learning*, pp. 1087–1094, 2007.

Ye, J., Li, T., Xiong, T., and Janardan, R. Using uncorrelated discriminant analysis for tissue classification with gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(4):181–190, 2004.

Ye, J., Janardan, R., Li, Q., and Park, H. Feature extraction via generalized uncorrelated linear discriminant analysis. *IEEE Transactions on Knoledge and Data Engineering*, 18(10):1312–1321, 2006.

Yin, W. Analysis and generalizations of the linearized bregman method. *SIAM Journal on Imaging Sciences*, 3(4):856–877, 2010.

Yin, W., Osher, S., Goldfarb, D., and Darbon, J. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008.

Appendix

A. Proof of Theorem 1

Proof. Let $U_2 \in \mathbf{R}^{m\times(m-\gamma)}$, $V_2 \in \mathbf{R}^{n\times(n-\gamma)}$, $P_2 \in \mathbf{R}^{\gamma\times(\gamma-q)}$ and $Q_2 \in \mathbf{R}^{k\times(k-q)}$ be column orthogonal matrices such that $U = \begin{bmatrix}U_1 & U_2\end{bmatrix}$, $V = \begin{bmatrix}V_1 & V_2\end{bmatrix}$, $P = \begin{bmatrix}P_1 & P_2\end{bmatrix}$ and $Q = \begin{bmatrix}Q_1 & Q_2\end{bmatrix}$ are orthogonal, respectively. Then, it is obvious that

$$H_t = \begin{bmatrix}U_1 & U_2\end{bmatrix}\begin{bmatrix}\Sigma_t & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}V_1 & V_2\end{bmatrix}^T,$$

and

$$S_t = H_tH_t^T = \begin{bmatrix}U_1\Sigma_t & U_2\end{bmatrix}\begin{bmatrix}I_\gamma & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}U_1\Sigma_t & U_2\end{bmatrix}^T.$$

Note that $S_t = S_b + S_w$, S_t, S_b and S_w are all symmetric and positive semi-definite, and $S_b = H_bH_b^T$, so $S_bU_2 = 0$ and it holds that

$$S_b = \begin{bmatrix}U_1\Sigma_t & U_2\end{bmatrix}\begin{bmatrix}\mathcal{S}_b & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}U_1\Sigma_t & U_2\end{bmatrix}^T,$$

where

$$\begin{aligned}\mathcal{S}_b &= (\Sigma_t^{-1}U_1^TH_b)(\Sigma_t^{-1}U_1^TH_b)^T\\ &= (P_1\Sigma_bQ_1^T)(P_1\Sigma_bQ_1^T)^T\\ &= \begin{bmatrix}P_1 & P_2\end{bmatrix}\begin{bmatrix}\Sigma_b^2 & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}P_1 & P_2\end{bmatrix}^T.\end{aligned}$$

So, let $\mathcal{Q} = \begin{bmatrix}U_1\Sigma_tP_1 & U_1\Sigma_tP_2 & U_2\end{bmatrix}$, we have

$$S_t = \mathcal{Q}\begin{bmatrix}I_q & 0 & 0\\ 0 & I_{\gamma-q} & 0\\ 0 & 0 & 0\end{bmatrix}\mathcal{Q}^T,\ S_b = \mathcal{Q}\begin{bmatrix}\Sigma_b^2 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\mathcal{Q}^T,$$

$$S_w = S_t - S_b = \mathcal{Q}\begin{bmatrix}I_q - \Sigma_b^2 & 0 & 0\\ 0 & I_{\gamma-q} & 0\\ 0 & 0 & 0\end{bmatrix}\mathcal{Q}^T,$$

which yield $q = \operatorname{rank}(S_b) = \operatorname{rank}(H_b)$, and

$$S_t^{(+)}S_b = \mathcal{Q}^{-T}\begin{bmatrix}\Sigma_b^2 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\mathcal{Q}^T.$$

For any $G \in \mathbf{R}^{m\times l}$, let $\mathcal{Q}^TG = \begin{bmatrix}G_1^T & G_2^T & G_3^T\end{bmatrix}^T$ where $G_1 \in \mathbf{R}^{q\times l}$, $G_2 \in \mathbf{R}^{(\gamma-q)\times l}$ and $G_3 \in \mathbf{R}^{(m-\gamma)\times l}$. We have

$$G^TS_tG = G_1^TG_1 + G_2^TG_2, \quad G^TS_bG = G_1^T\Sigma_b^2G_1.$$

Since it has been shown in (Ye, 2005) that

$$\max_G \operatorname{Trace}((S_t^L)^+S_b^L) = \operatorname{Trace}(S_t^{(+)}S_b) = \operatorname{Trace}(\Sigma_b^2), \tag{16}$$

we get that $G \in \mathbf{R}^{m\times l}$ is a solution of optimization problem (3) if and only if

$$G_1^TG_1 + G_2^TG_2 = I_l, \quad \operatorname{Trace}(G_1^T\Sigma_b^2G_1) = \operatorname{Trace}(\Sigma_b^2).$$

$G_1^TG_1 + G_2^TG_2 = I_l$ implies that $l \le \gamma$, and there exist $\mathcal{G}_1 \in \mathbf{R}^{q\times(\gamma-l)}$ and $\mathcal{G}_2 \in \mathbf{R}^{(\gamma-q)\times(\gamma-l)}$ such that $\begin{bmatrix}G_1 & \mathcal{G}_1\\ G_2 & \mathcal{G}_2\end{bmatrix}$ is orthogonal, which gives that $G_1G_1^T = I_q - \mathcal{G}_1\mathcal{G}_1^T$. Thus, we obtain

$$\begin{aligned}&\operatorname{Trace}(G_1^T\Sigma_b^2G_1) = \operatorname{Trace}(\Sigma_b^2)\\ \Leftrightarrow\ &\operatorname{Trace}(\Sigma_b^2G_1G_1^T) = \operatorname{Trace}(\Sigma_b^2)\\ \Leftrightarrow\ &\mathcal{G}_1 = 0\\ \Leftrightarrow\ &G_1G_1^T = I_q,\end{aligned}$$

which, in return, implies $q \le l \le \gamma$, and

$$\begin{aligned}&G \in \mathbf{R}^{m\times l}\text{is a solution of optimization problem (3)}\\ \Leftrightarrow\ &G = \mathcal{Q}^{-T}\begin{bmatrix}G_1\\ G_2\\ G_3\end{bmatrix},\ G_1G_1^T = I_q,\ \begin{bmatrix}G_1\\ G_2\end{bmatrix}^T\begin{bmatrix}G_1\\ G_2\end{bmatrix} = I_l\\ \Leftrightarrow\ &\begin{cases}G = \mathcal{Q}^{-T}\begin{bmatrix}G_1\\ G_2\\ G_3\end{bmatrix},\ \begin{bmatrix}G_1\\ G_2\end{bmatrix} = \begin{bmatrix}I_q & 0\\ 0 & \mathcal{G}_3\end{bmatrix}\mathcal{Z},\\ q \le l \le \gamma,\ \mathcal{G}_3 \in \mathbf{R}^{(\gamma-q)\times(l-q)}\text{is column}\\ \text{orthogonal and } \mathcal{Z} \in \mathbf{R}^{l\times l}\text{is orthogonal}\end{cases}\\ \Leftrightarrow\ &G = \left(U_1\Sigma_t^{-1}\begin{bmatrix}P_1 & P_2\mathcal{G}_3\end{bmatrix} + U_2G_3\right)\mathcal{Z},\end{aligned}$$

where in the last equality we used

$$\mathcal{Q}^{-T} = \begin{bmatrix}U_1\Sigma_t^{-1}P_1 & U_1\Sigma_t^{-1}P_2 & U_2\end{bmatrix}.$$

Since $\begin{bmatrix}P_1 & P_2\end{bmatrix}$ and $\begin{bmatrix}U_1 & U_2\end{bmatrix}$ are orthogonal, it follows that for any $\mathcal{M}_1 \in \mathbf{R}^{\gamma\times(l-q)}$ and $\mathcal{M}_2 \in \mathbf{R}^{m\times l}$

$$\begin{aligned}&\mathcal{M}_1 \text{ is column orthogonal, and } \mathcal{M}_1^TP_1 = 0\\ \Leftrightarrow\ &\mathcal{M}_1 = P_2\mathcal{G}_3,\ \text{ for some column orthogonal } \mathcal{G}_3,\end{aligned}$$

and

$$\mathcal{M}_2^TU_1 = 0 \Leftrightarrow \mathcal{M}_2 = U_2G_3,\ \text{ for some } G_3 \in \mathbf{R}^{(m-\gamma)\times l}.$$

Therefore, we have that $G \in \mathbf{R}^{m\times l}$ is a solution of optimization problem (3) if and only if $q \le l \le \gamma$ and

$$G = \left(U_1\Sigma_t^{-1}\begin{bmatrix}P_1 & \mathcal{M}_1\end{bmatrix} + \mathcal{M}_2\right)\mathcal{Z},$$

where $\mathcal{M}_1 \in \mathbf{R}^{\gamma\times(l-q)}$ is column orthogonal satisfying $\mathcal{M}_1^TP_1 = 0$, $\mathcal{M}_2 \in \mathbf{R}^{m\times l}$ is an arbitrary matrix satisfying $\mathcal{M}_2^TU_1 = 0$, and $\mathcal{Z} \in \mathbf{R}^{l\times l}$ is orthogonal. □