Kernel Alignment Inspired Linear Discriminant Analysis

Shuai Zheng and Chris Ding

Department of Computer Science and Engineering,
University of Texas at Arlington, TX, USA
zhengs123@gmail.com, chqding@uta.edu

Abstract. Kernel alignment measures the degree of similarity between two kernels. In this paper, inspired from kernel alignment, we propose a new Linear Discriminant Analysis (LDA) formulation, kernel alignment LDA (kaLDA). We first define two kernels, data kernel and class indicator kernel. The problem is to find a subspace to maximize the alignment between subspace-transformed data kernel and class indicator kernel. Surprisingly, the kernel alignment induced kaLDA objective function is very similar to classical LDA and can be expressed using between-class and total scatter matrices. This can be extended to multi-label data. We use a Stiefel-manifold gradient descent algorithm to solve this problem. We perform experiments on 8 single-label and 6 multi-label data sets. Results show that kaLDA has very good performance on many single-label and multi-label problems.

Keywords: Kernel Alignment, LDA.

1 Introduction

Kernel alignment [2] is a way to incorporate class label information into kernels which are traditionally directly constructed from data without using class labels. Kernel alignment can be viewed as a measurement of consistency between the similarity function (the kernel) and class structure in the data. Improving this consistency helps to enforce data become more separated when using the class label aligned kernel. Kernel alignment has been applied to pattern recognition and feature selection recently [3,28,10,11,4].

In this paper, we find that if we use the widely used linear kernel and a kernel built from class indicators, the resulting kernel alignment function is very similar to the widely used linear discriminant analysis (LDA), using the well-known between-class scatter matrix S_b and total scatter matrix S_t. We call this objective function as kernel alignment induced LDA (kaLDA). If we transform data into a linear subspace, the optimal solution is to maximize this kaLDA.

We further analyze this kaLDA and propose a Stiefel-manifold gradient descent algorithm to solve it. We also extend kaLDA to multi-label problems. Surprisingly, the scatter matrices arising in multi-label kernel alignment are identical those matrices developed in Multi-label LDA [21].

T. Calders et al. (Eds.): ECML PKDD 2014, Part III, LNCS 8726, pp. 401–416, 2014.

We perform extensive experiments by comparing kaLDA with other approaches on 8 single-label datasets and 6 multi-label data sets. Results show that kernel alignment LDA approach has good performance in terms of classification accuracy and F1 score.

2 From Kernel Alignment to LDA

Kernel Alignment is a similarity measurement between a kernel function and a target function. In other words, kernel alignment evaluates the degree of fitness between the data in kernel space and the target function. For this reason, we usually set the target function to be the class indicator function. The other kernel function is the data matrix. By measuring the similarity between data kernel and class indicator kernel, we can get a sense of how easily this data can be separated in kernel subspace. The alignment of two kernels $\mathcal{K}_1$ and $\mathcal{K}_2$ is given as [2]:

$$A(\mathcal{K}_1, \mathcal{K}_2) = \frac{\mathrm{Tr}(\mathcal{K}_1\mathcal{K}_2)}{\sqrt{\mathrm{Tr}(\mathcal{K}_1\mathcal{K}_1)}\sqrt{\mathrm{Tr}(\mathcal{K}_2\mathcal{K}_2)}}. \tag{1}$$

We first introduce some notations, and then present Theorem 1 and kernel alignment projective function.

Let data matrix be $X \in \Re^{p\times n}$ and $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, where p is data dimension, n is number of data points, $\mathbf{x}_i$ is a data point. Let normalized class indicator matrix be $Y \in \Re^{n\times K}$, which was used to prove the equivalence between PCA and K-means clustering [26,5], and

$$Y_{ik} = \begin{cases} \frac{1}{\sqrt{n_k}}, & \text{if point } i \text{ is in class } k. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where K is total class number, n_k is the number of data points in class k. Class mean is $\mathbf{m}_k = \sum_{\mathbf{x}_i\in k} \mathbf{x}_i/n_k$ and total mean of data is $\mathbf{m} = \sum_i \mathbf{x}_i/n$.

Theorem 1. *Define data kernel $\mathcal{K}_1$ and class label kernel $\mathcal{K}_2$ as follows:*

$$\mathcal{K}_1 = X^TX, \;\; \mathcal{K}_2 = YY^T, \tag{3}$$

we have

$$A(\mathcal{K}_1, \mathcal{K}_2) = c\frac{TrS_b}{\sqrt{TrS_t^2}} \tag{4}$$

where $c = 1/\sqrt{Tr(YY^T)^2}$ is a constant independent of X.

Furthermore, let $G \in \Re^{p\times k}$ be a linear transformation to a k-dimensional subspace

$$\widetilde{X} = G^TX, \;\; \widetilde{\mathcal{K}}_1 = \widetilde{X}^T\widetilde{X}, \tag{5}$$

we have

$$A(\widetilde{\mathcal{K}}_1, \mathcal{K}_2) = c\frac{Tr(G^T S_b G)}{\sqrt{Tr(G^T S_t G)^2}} \tag{6}$$

where

$$S_b = \sum_{k=1}^{K} n_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T, \tag{7}$$

$$S_t = \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T, \tag{8}$$

Theorem 1 shows that kernel alignment can be expressed using scatter matrices S_b and S_t. In applications, we adjust G such that kernel alignment is maximized, i.e., we solve the following problem:

$$\max_G \frac{\mathrm{Tr}(G^T S_b G)}{\sqrt{\mathrm{Tr}(G^T S_t G)^2}}. \tag{9}$$

In general, columns of G are assumed to be linearly independent.

A striking feature of this kernel alignment problem is that it is very similar to classic LDA.

2.1 Proof of Theorem 1 and Analysis

Here we note a useful lemma and then prove Theorem 1.

In most data analysis, data are centered, i.e., $\sum_i \mathbf{x}_i = \mathbf{0}$. Here we assume data is already centered. The following results remain correct if data is not centered. We have the following relations:

Lemma 1. *Scatter matrices S_b, S_t can be expressed as:*

$$S_b = XYY^T X^T, \tag{10}$$

$$S_t = XX^T. \tag{11}$$

These results are previously known, for example, Theorem 3 of [5].

Proof of Theorem 1. To prove Eq.(4), we substitute $\mathcal{K}_1, \mathcal{K}_2$ into Eq.(1) and obtain, noting $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$,

$$A(\mathcal{K}_1, \mathcal{K}_2) = \frac{\mathrm{Tr}(XYY^T X^T)}{\sqrt{\mathrm{Tr}(XX^T)^2}\sqrt{\mathrm{Tr}(YY^T)^2}} = c\frac{\mathrm{Tr}S_b}{\sqrt{\mathrm{Tr}S_t^2}}.$$

where we used Lemma 1. $c = 1/\sqrt{\mathrm{Tr}(YY^T)^2}$ is a constant independent of data X.

To prove Eq.(6),

$$A(\widetilde{\mathcal{K}}_1, \mathcal{K}_2) = c\frac{\mathrm{Tr}(G^T XYY^T X^T G)}{\sqrt{\mathrm{Tr}(G^T XX^T G)^2}} = c\frac{\mathrm{Tr}(G^T S_b G)}{\sqrt{\mathrm{Tr}(G^T S_t G)^2}},$$

thus we obtain Eq.(6) using Lemma 1.

2.2 Relation to Classical LDA

In classical LDA, the between-class scatter matrix S_b is defined as Eq.(7), and the within-class scatter matrix S_w and total scatter matrix S_t are defined as:

$$S_w = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T, \; S_t = S_b + S_w, \tag{12}$$

where $\mathbf{m}_k$ and $\mathbf{m}$ are class means. Classical LDA finds a projection matrix $G \in \Re^{p \times (K-1)}$ that minimizes S_w and maximizes S_b using the following objective:

$$\max_G \mathrm{Tr} \frac{G^T S_b G}{G^T S_w G}, \tag{13}$$

or

$$\max_G \frac{\mathrm{Tr}(G^T S_b G)}{\mathrm{Tr}(G^T S_w G)}. \tag{14}$$

Eq.(14) is also called trace ratio (TR) problem [22]. It is easy to see [1] that Eq.(14) can be expressed as

$$\max_G \frac{\mathrm{Tr}(G^T S_b G)}{\mathrm{Tr}(G^T S_t G)}. \tag{15}$$

As we can see, kernel alignment LDA objective function Eq.(9) is very similar to Eq.(15). Thus kernel alignment provides an interesting alternative explanation of LDA. In fact, we can similarly show that in Eq.(9), S_w is also maximized as in the standard LDA. First, Eq.(9) is equivalent to

$$\max_G \mathrm{Tr}(G^T S_b G) \;\; s.t. \;\; \mathrm{Tr}(G^T S_t G)^2 = \eta,$$

where η is a fixed-value. The precise value of η is unimportant, since the scale of G is undefined in LDA: if G^* is an optimal solution, and r is any real number, $G^{**} = rG^*$ is also an optimal solution with the same optimal objective function value. The above optimization is approximately equivalent to

$$\max_G \mathrm{Tr}(G^T S_b G) \;\; s.t. \;\; \mathrm{Tr}(G^T S_t G) = \eta,$$

This is same as

$$\max_G \mathrm{Tr}(G^T S_b G) \;\; s.t. \;\; \mathrm{Tr}(G^T S_w G) = \eta - \mathrm{Tr}(G^T S_b G),$$

In other words, S_b is maximized while S_w is minimized — recovering the LDA main theme.

[1] Eq.(14) is equivalent to $\min \frac{\mathrm{Tr}(G^T S_w G)}{\mathrm{Tr}(G^T S_b G)}$, which is $\min \left(\frac{\mathrm{Tr}(G^T S_w G)}{\mathrm{Tr}(G^T S_b G)} + 1 \right)$. Reversing to maximization and using $S_t = S_b + S_w$, we obtain Eq.(15).

3 Computational Algorithm

In this section, we develop efficient algorithm to solve kaLDA objective function Eq.(9):

$$\max_G J_1 = \frac{\mathrm{Tr}(G^T S_b G)}{\sqrt{\mathrm{Tr}(G^T S_t G)^2}}, \;\; s.t. \;\; G^T G = I. \tag{16}$$

The condition $G^T G = I$ ensures different columns of G mutually independent. The gradient of $J_1(G)$ is

$$\nabla J_1 \triangleq \frac{\partial J_1}{\partial G} = 2\frac{A}{\sqrt{\mathrm{Tr}D^2}} - 2\frac{\mathrm{Tr}B}{(\mathrm{Tr}D^2)^{\frac{3}{2}}}CD, \tag{17}$$

where $A = S_b G, \; B = G^T A, \; C = S_t G, \; D = G^T C$.

Constraint $G^T G = I$ enforces G on the Stiefel manifold. Variations of G on this manifold is parallel transport, which gives some restriction to the gradient. This has been been worked out in [6]. The gradient that reserves the manifold structure is

$$\nabla J_1 - G[\nabla J_1]^T G. \tag{18}$$

Thus the algorithm computes the new G is given as follows:

$$G \leftarrow G - \eta(\nabla J_1 - G[\nabla J_1]^T G). \tag{19}$$

The step size η is usually chosen as:

$$\eta = \tau \|G\|_1 / \|\nabla J_1 - G(\nabla J_1)^T G\|_1, \; \tau = 0.001 \sim 0.01. \tag{20}$$

where $\|G\|_1 = \sum_{ij} |G_{ij}|$.

Occasionally, due to the loss of numerical accuracy, we use projection $G \leftarrow G(G^T G)^{-\frac{1}{2}}$ to restore $G^T G = I$. Starting with the standard LDA solution of G, this algorithm is iterated until the algorithm converges to a local optimal solution. In fact, objective function will converge quickly when choosing η properly. Figure 1 shows that J_1 converges in about 200 iterations when $\tau = 0.001$, for datasets ATT, Binalpha, Mnist, and Umist (more details about the datasets will be introduced in experiment section). In summary, kernel alignment LDA (kaLDA) procedure is shown in Algorithm 1.

To show the effectiveness of proposed kaLDA, we visualize a real dataset in 2-D subspace in Figure 2. In this example, we take 3 classes of 644-dimension Umist data, 18 data points in each class. Figure 2a shows the original data projected in 2-D PCA subspace. Blue points are in class 1; red circle points are in class 2; black square points are in class 3. Data points from the three classes are mixed together in 2-D PCA subspace. It is difficult to find a linear boundary to separate points of different classes. Figure 2b shows the data in 2-D standard LDA subspace. We can see that data points in different classes have been projected into different clusters. Figure 2c shows the data projected in 2-D kaLDA subspace. Compared to Figure 2b, the within-class distance in Figure 2c is much smaller. The distance between different classes is larger.

Algorithm 1. $[G] = kaLDA(X, Y)$

Input: Data matrix $X \in \Re^{p \times n}$, class indicator matrix $Y \in \Re^{n \times K}$
Output: Projection matrix $G \in \Re^{p \times k}$
1: Compute S_b and S_t using Eq.(10) and Eq.(11)
2: Initialize G using classical LDA solution
3: **repeat**
4: Compute gradient using Eq.(17)
5: Update G using Eq.(19)
6: **until** J_1 Converges

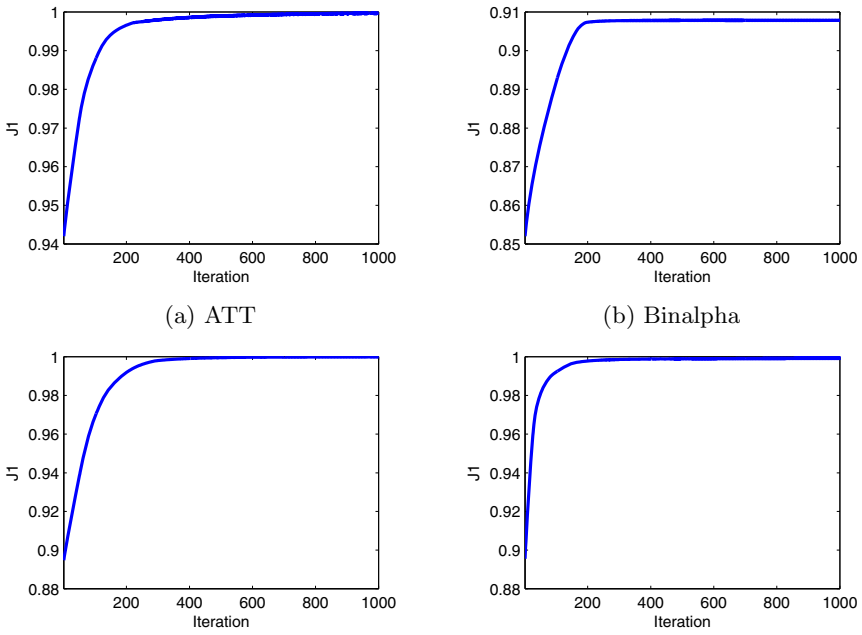
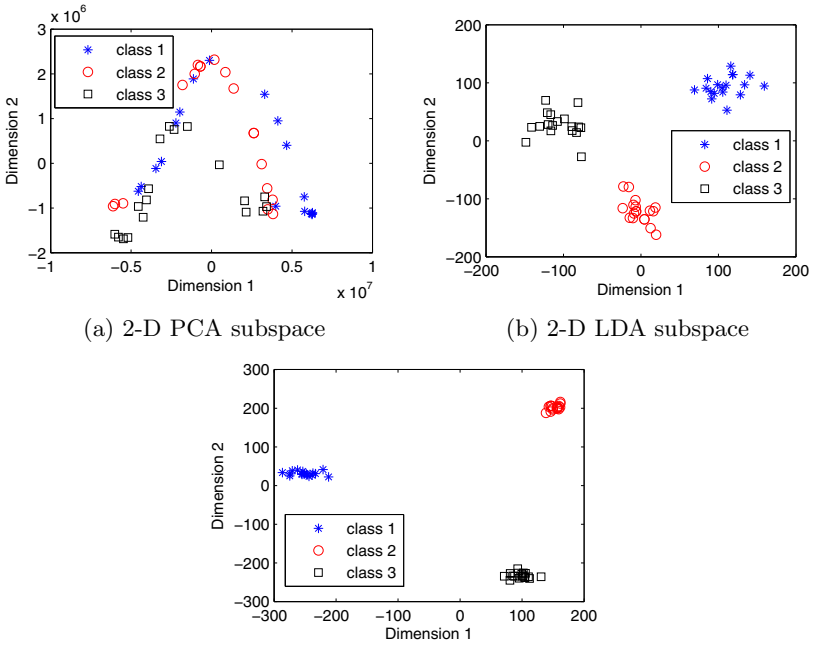


Fig. 1. Objective J1 converges using Stiefel-manifold gradient descent algorithm ($\tau = 0.001$)

4 Extension to Multi-label Data

Multi-label problem arises frequently in image and video annotations, multi-topic text categorization, music classification. etc.[21]. In multi-label data, a data point could have several class labels (belonging to several classes). For example, an image could have "cloud", "building", "tree" labels. This is different from the case of single-label problem, where one point can have only one class label. Multi-label is very natural and common in our everyday life. For example, a film can be simultaneously classified as "drama", "romance", "historic" (if it



(a) 2-D PCA subspace

(b) 2-D LDA subspace

(c) 2-D kaLDA subspace

Fig. 2. Visualization of Umist data in 2-D PCA, 2-D LDA and 2-D kaLDA subspace

is about a true story). A news article can have topic labels such as "economics", "sports", etc.

Kernel alignment approach can be easily and naturally extended to multi-label data, because the class label kernel can be clearly and unambiguously defined using class label matrix Z on both single label and multi-label data sets. The data kernel is defined as usual. In the following we further develop this approach.

One important result of our kernel alignment approach for single label data is that it has close relationship with LDA. For multi-label data, each data point could belong to several classes. The standard scatter matrices S_b, S_w are ambiguous, because S_b, S_w are only defined for single label data where each data point belongs to one class only. However, our kernel alignment approach on multi-label data leads to new definitions of scatter matrices and similar objective function; this can be viewed as the generalization of LDA from single-label data to multi-label data via kernel alignment approach.

Indeed, the new scatter matrices we obtained from kernel alignment approach are identical to the so-called "multi-label LDA" [21] developed from a class-separate, probabilistic point of view, very different from our point of view. The fact that these two approaches lead to the same set of scatter matrices show that the resulting multi-label LDA framework has a broad theoretical basis.

We first present some notations for multi-label data and then describe the kernel alignment approach for multi-label data in Theorem 2. The class label matrix $Z \in \Re^{n\times K}$ for data $X \in \Re^{p\times n}$ is given as:

$$Z_{ik} = \begin{cases} 1, & \text{if point } i \text{ is in class } k. \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

Let $\tilde{n}_k = \sum_{i=1}^{n} Z_{ik}$ be the number of data points in class k. Note that for multi-label data, $\sum_{k=1}^{K} \tilde{n}_k > n$. The **normalized** class indicator matrix $\widetilde{Y} \in \Re^{n\times K}$ is given as:

$$\widetilde{Y}_{ik} = \begin{cases} \frac{1}{\sqrt{\tilde{n}_k}}, & \text{if point } i \text{ is in class } k. \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

Let $\rho_i = \sum_{k=1}^{K} Z_{ik}$ be the number of classes that $\mathbf{x}_i$ belongs to. Thus ρ_i are the weights of $\mathbf{x}_i$. Define the diagonal weight matrix $\Omega = \mathrm{diag}(\rho_1, \cdots, \rho_n)$. The kernel alignment formulation for multi-label data can be stated as

Theorem 2. *For multi-label data X, let the data kernel and class label kernel be*

$$\mathcal{K}_1 = \Omega^{\frac{1}{2}} X^T X \Omega^{\frac{1}{2}}, \;\; \mathcal{K}_2 = \Omega^{-\frac{1}{2}} \widetilde{Y}\widetilde{Y}^T \Omega^{-\frac{1}{2}}. \tag{23}$$

We have the alignment

$$A(\mathcal{K}_1, \mathcal{K}_2) = c\frac{TrS_b}{\sqrt{TrS_t^2}} \tag{24}$$

where $c = 1/\sqrt{Tr(\Omega^{-1}\widetilde{Y}\widetilde{Y}^T)^2}$ is a constant independent of data X, and S_b, S_t are given in Eqs.(27, 28).

Furthermore, let $G \in \Re^{p\times k}$ be the linear transformation to a k-dimensional subspace,

$$\widetilde{X} = G^T X, \;\; \widetilde{\mathcal{K}}_1 = \Omega^{1/2} \widetilde{X}^T \widetilde{X} \Omega^{1/2}, \tag{25}$$

we have

$$A(\widetilde{\mathcal{K}}_1, \mathcal{K}_2) = c\frac{Tr(G^T S_b G)}{\sqrt{Tr(G^T S_t G)^2}} \tag{26}$$

The matrices S_b, S_t in Theorem 2 are defined as:

$$S_b = \sum_{k=1}^{K} \tilde{n}_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T, \tag{27}$$

$$S_t = \sum_{k=1}^{K} \sum_{i=1}^{n} Z_{ik} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T, \tag{28}$$

where $\mathbf{m}_k$ is the mean of class k and $\mathbf{m}$ is global mean, defined as:

$$\mathbf{m}_k = \frac{\sum_{i=1}^n Z_{ik}\mathbf{x}_i}{\tilde{n}_k}, \quad \mathbf{m} = \frac{\sum_{i=1}^n \rho_i \mathbf{x}_i}{\sum_{k=1}^K \tilde{n}_k}. \tag{29}$$

Therefore, we can seek an optimal subspace for multi-label data by solving Eq.(16) with S_b, S_t given in Eqs.(27,28)

4.1 Proof of Theorem 2 and Equivalence to Multi-label LDA

Here we note a useful lemma for multi-label data and then prove Theorem 2. We consider the case the data is centered, i.e., $\sum_{i=1}^n \rho_i \mathbf{x}_i = \mathbf{0}$. The results also hold when data is not centered, but the proofs are slightly complicated.

Lemma 2. *For multi-label data, S_b, S_t of Eqs.(27,28) can be expressed as*

$$S_b = X\widetilde{Y}\widetilde{Y}^T X^T \tag{30}$$

$$S_t = X\Omega X^T \tag{31}$$

Proof. From the definition of $\mathbf{m}_k$ and $\widetilde{Y}$ in multi-label data, we have

$$X\widetilde{Y} = (\mathbf{m}_1, \cdots, \mathbf{m}_K) \begin{pmatrix} \sqrt{\tilde{n}_1} & & \\ & \ddots & \\ & & \sqrt{\tilde{n}_K} \end{pmatrix}.$$

Thus $X\widetilde{Y}\widetilde{Y}^T X^T = \sum_{k=1}^K \tilde{n}_k \mathbf{m}_k \mathbf{m}_k^T$ recovers S_b of Eq.(27).

To prove Eq.(31), note that $X\Omega = (\rho_1 \mathbf{x}_1, \cdots, \rho_n \mathbf{x}_n)$, thus $X\Omega X^T = \sum_{i=1}^n \rho_i \mathbf{x}_i \mathbf{x}_i^T$.

Proof of Theorem 2. Using Lemma 2, to prove Eq.(24),

$$A(\mathcal{K}_1, \mathcal{K}_2) = c\frac{\mathrm{Tr}(X\widetilde{Y}\widetilde{Y}^T X^T)}{\sqrt{\mathrm{Tr}(X\Omega X^T)^2}} = c\frac{\mathrm{Tr}S_b}{\sqrt{\mathrm{Tr}S_t^2}},$$

where $c = 1/\sqrt{\mathrm{Tr}(\Omega^{-1}\widetilde{Y}\widetilde{Y}^T)^2}$ is independent of X.

To prove Eq.(26),

$$A(\widetilde{\mathcal{K}}_1, \mathcal{K}_2) = c\frac{\mathrm{Tr}(G^T X\widetilde{Y}\widetilde{Y}^T X^T G)}{\sqrt{\mathrm{Tr}(G^T X\Omega X^T G)^2}} = c\frac{\mathrm{Tr}(G^T S_b G)}{\sqrt{\mathrm{Tr}(G^T S_t G)^2}}.$$

For single-label data, $\rho_i = 1$, $\Omega = I$, $\tilde{n}_k = n_k$, Eqs.(30, 31) reduce to Eqs.(10, 11), and Theorem 2 reduces to Theorem 1.

As we can see, surprisingly, the scatter matrices S_b, S_t of Eqs.(27, 28) arising in Theorem 2 are identical to that in Multi-label LDA proposed in [21].

Table 1. Single-label datasets attributes

Data	n	p	k
Caltec07	210	432	7
Caltec20	1230	432	20
MSRC	210	432	7
ATT	400	644	40
Binalpha	1014	320	26
Mnist	150	784	10
Umist	360	644	20
Pie	680	1024	68

Table 2. Classification accuracy on Single-label datasets ($K-1$ dimension)

Data	kaLDA	LDA	TR	sdpLDA	MMC	RLDA	OCM
Caltec07	0.7524	0.6619	0.6762	0.5619	0.6000	**0.7952**	0.7619
Caltec20	**0.7068**	0.6320	0.4465	0.3386	0.5838	0.6812	0.6696
MSRC	**0.7762**	0.6857	0.5714	0.5952	0.5667	0.7333	0.7286
ATT	**0.9775**	0.9750	0.9675	0.9750	0.9750	0.9675	0.9675
Binalpha	0.7817	0.6078	0.4620	0.2507	0.7638	0.7983	**0.8204**
Mnist	**0.8800**	0.8733	0.8667	0.8467	0.8467	0.8667	0.8467
Umist	0.9900	0.9900	**0.9917**	0.9133	0.9633	0.9800	0.9783
Pie	0.8765	**0.8838**	0.8441	0.8632	0.8676	0.6515	0.6515

5 Related Work

Linear Discriminant Analysis (LDA) is a widely-used dimension reduction and subspace learning algorithm. There are many LDA reformulation publications in recent years. Trace Ratio problem is to find a subspace transformation matrix G such that the within-class distance is minimized and the between-class distance is maximized. Formally, Trace Ratio maximizes the ratio of two trace terms, $\max_G \mathrm{Tr}(G^T S_b G)/\mathrm{Tr}(G^T S_t G)$ [22,13], where S_t is total scatter matrix and S_b is between-class scatter matrix. Other popular LDA approach includes, regularized LDA(RLDA) [9], Orthogonal Centroid Method (OCM) [18], Uncorrelated LDA(ULDA) [23], Orthogonal LDA (OLDA) [23], etc.. These approaches mainly compute the eigendecomposition of matrix $S_t^{-1}S_b$, but use different formulation of total scatter matrix S_t [24].

Maximum Margin Criteria (MMC) [17] is a simpler and more efficient method. MMC finds a subspace projection matrix G to maximize $\mathrm{Tr}(G^T(S_b - S_w)G)$. Though in a different way, MMC also maximizes between-class distance while minimizing within-class distance. Semi-Definite Positive LDA (sdpLDA) [14]

Table 3. Multi-label datasets attributes

Data	n	p	k
MSRC-MOM	591	384	23
Barcelona	139	48	4
Emotion	593	72	6
Yeast	2,417	103	14
MSRC-SIFT	591	240	23
Scene	2,407	294	6

Table 4. Classification accuracy on Multi-label datasets ($K-1$ dimension)

Data	kaLDA	MLSI	MDDM	MLLS	MLDA
MSRC-MOM	**0.9150**	0.8962	0.9044	0.8994	0.9036
Barcelona	**0.6579**	0.6436	0.6470	0.6524	0.6290
Emotion	**0.7634**	0.7397	0.7540	0.7529	0.7619
Yeast	**0.7405**	0.7317	0.7371	0.7364	0.7368
MSRC-SIFT	0.8839	0.8762	0.8800	0.8807	**0.8858**
Scene	**0.8870**	0.8534	0.8713	0.8229	0.8771

solves the maximization of $\mathrm{Tr}(G^T(S_b - \lambda_1 S_w)G)$, where λ_1 is the largest eigenvalue of $S_w^{-1}S_b$. sdpLDA is derived from the maximum margin principle.

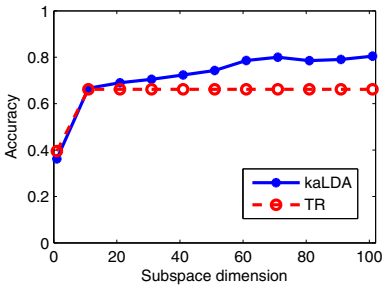
Multi-label problem arise frequently in image and video annotations and many other related applications, such as multi-topic text categorization [21]. There are many Multi-label dimension reduction approaches, such as Multi-label Linear Regression (MLR), Multi-label informed Latent Semantic Indexing (MLSI) [25], Multi-label Dimensionality reduction via Dependence Maximization (MDDM) [27], Multi-Label Least Square (MLLS) [12], Multi-label Linear Discriminant Analysis (MLDA) [21].

6 Experiments

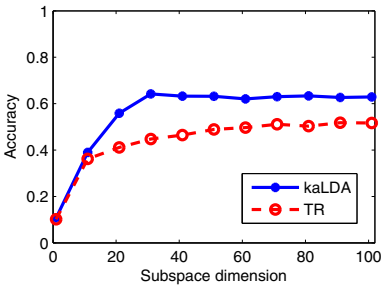
In this section, we first compare kernel alignment LDA (kaLDA) with other six different methods on 8 single label data sets and compare kaLDA multi-label version with four other methods on 6 multi-label data sets.

6.1 Comparison with Trace Ratio w.r.t. Subspace Dimension

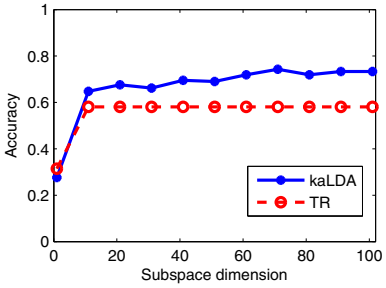
Eight single-label datasets are used in this experiment. These datasets come from different domains, such as image scene Caltec [8] and MSRC [16], face datasets ATT, Umist, Pie [19], and digit datasets Mnist [15] and Binalpha. Table 1 summarizes the attributes of those datasets.



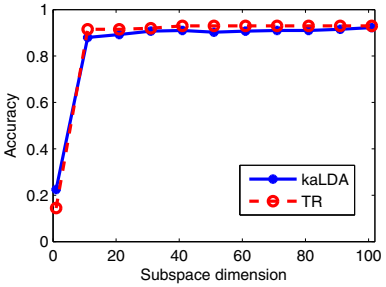
(a) Caltec07



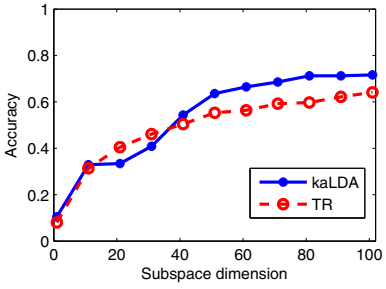
(b) Caltec20



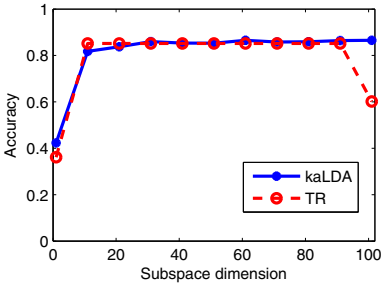
(c) MSRC



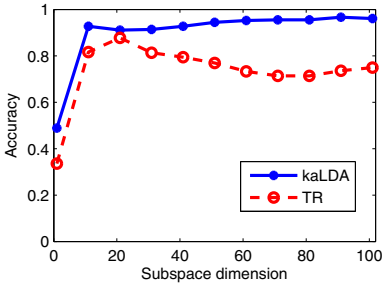
(d) ATT



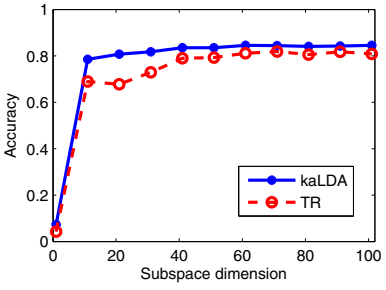
(e) Binalpha



(f) Mnist



(g) Umist



(h) Pie

Fig. 3. Classification accuracy w.r.t. dimension of the subspace

Table 5. Macro F1 score on Multi-label datasets ($K-1$ dimension)

Dataset	kaLDA	MLSI	MDDM	MLLS	MLDA
MSRC-MOM	**0.6104**	0.5244	0.5593	0.5426	0.5571
Barcelona	**0.7377**	0.7286	0.7301	0.7341	0.7169
Emotion	**0.6274**	0.5873	0.6101	0.6041	0.6200
Yeast	**0.5757**	0.5568	0.5696	0.5691	0.5693
MSRC-SIFT	0.4712	0.4334	0.4522	0.4544	**0.4773**
Scene	**0.6851**	0.5911	0.6411	0.5048	0.6568

Caltec07 and **Caltec20** are subsets of Caltech 101 data. Only the HOG feature is used in this paper.

MSRC is a image scene data, includes tree, building, plane, cow, face, car and so on. It has 210 images from 7 classes and each image has 432 dimension.

ATT data contains 400 images of 40 persons, with 10 images for each person. The images has been resized to 28×23.

Binalpha data contains 26 binary hand-written alphabets. It has 1014 images in total and each image has 320 dimension.

Mnist is a handwritten digits dataset. The digits have been size-normalized and centred. It has 10 classes and 150 images in total, with 784 dimension each image.

Umist is a face image dataset (Sheffield Face database) with 360 images from 20 individuals with mixed race, gender and appearance.

Pie is a face database collected by Carnegie Mellon Robotics Institute between October and December 2000. In total, it has 68 different persons.

In this part, we compare the classification accuracy of kaLDA and Trace Ratio [22] with respect to subspace dimension. The dimension of the subspace that kaLDA can find is not restricted to $K-1$. After subspace projection, KNN classifier ($knn = 3$) is applied to perform classification. Results are shown in Figure 3. Solid line denotes kaLDA accuracy and dashed line denotes Trace Ratio accuracy. As we can see, in Figures 3a, 3b, 3c, 3g, and 3h, kaLDA has higher accuracy than Trace Ratio when using the same number of reduced features. In Figures 3d, 3e, 3f, kaLDA has competitive classification accuracy with Trace Ratio. However, kaLDA is more stable than Trace Ratio. For example, in Figure 3f and 3g, we observe a decrease in accuracy when feature number increases using Trace Ratio.

6.2 Comparison with other LDA Methods

We compare kaLDA with six other different methods, including LDA, Trace Ratio (TR), spdLDA, Maximum Margin Criteria (MMC), regularized LDA (RLDA), and Orthogonal Centroid Method (OCM). All LDA will reduce data to $K-1$ dimension. KNN ($knn = 3$) will be applied to do the classification after data is

Table 6. Micro F1 score on Multi-label datasets ($K-1$ dimension)

Dataset	kaLDA	MLSI	MDDM	MLLS	MLDA
MSRC-MOM	**0.5138**	0.4064	0.4432	0.4370	0.4448
Barcelona	**0.6969**	0.6891	0.6861	0.6904	0.6772
Emotion	**0.6203**	0.5779	0.6030	0.5961	0.6151
Yeast	**0.4249**	0.4026	0.4205	0.4216	0.4213
MSRC-SIFT	0.3943	0.3510	0.3637	0.3667	**0.3959**
Scene	**0.6966**	0.6006	0.6493	0.5062	0.6643

projected into the selected subspace. The other algorithms have already been introduced in related work section. The final classification accuracy is the average of 5-fold cross validation, and is reported in Table 2. The first column "kaLDA" reports kaLDA classification accuracy. kaLDA has the highest accuracy on 4 out of 8 datasets, including Caltec20, MSRC-MOM, ATT and Mnist. For Umist and Pie, kaLDA results are very close to the highest accuracy. Overall, kaLDA performs better than all other methods.

6.3 Multi-label Classification

Six multi-label datasets are used in this part. These datasets include images features, music emotion and so on. Table 3 summarizes the attributes of those datasets.

MSRC-MOM and **MSRC-SIFT** data set is provided by Microsoft Research in Cambridge. It includes 591 images of 23 classes. **MSRC-MOM** is the Moment invariants (MOM) feature of images and each image has 384 dimensions. **MSRC-SIFT** is the SIFT feature and each image has 240 dimensions. About 80% of the images are annotated with at least one classes and about three classes per image on average.

Barcelona data set contains 139 images with 4 classes, i.e., "building", "flora", "people" and "sky". Each image has at least two labels.

Emotion [20] is a music emotion data, which comprises 593 songs with 6 emotions. The dimension of Emotion is 72.

Yeast [7] is a multi-label data set which contains functional classes of genes in the Yeast Saccharomyces cerevisiae.

Scene [1] contains images of still scenes with semantic indexing. It has 2407 images from 6 classes.

We use 5-fold cross validation to evaluate classification performance of different algorithms. K-Nearest Neighbour (KNN) classifier is used after the subspace projection. The algorithms we compared in this section includes Multi-label informed Latent Semantic Indexing (MLSI), Multi-label Dimensionality reduction via Dependence Maximization (MDDM), Multi-Label Least Square (MLLS), Multi-label Linear Discriminant Analysis (MLDA). These algorithms have been introduced in related work section.

We compare the performance of kaLDA and other algorithms using macro accuracy (Table 4), macro-averaged F1-score (Table 5) and micro-averaged (Table 6) F1-score. Accuracy and F1 score are computed using standard binary classification definitions. In multi-label classification, macro average is a standard class-wise average, and it is related to number of samples in each class. However, micro average gives equal weight to all classes [21]. kaLDA achieves highest classification accuracy on 5 out of 6 datasets. On the remaining MSRC-SIFT dataset, kaLDA result is very close to the best method MLDA and beat all rest methods. kaLDA achieves highest macro and micro F1 score on 5 out of 6 datasets. Furthermore, kaLDA has the second highest macro and micro F1 score on dataset MSRC-SIFT. Overall, kaLDA outperforms other multi-label algorithms in terms of classification accuracy and macro and micro F1 score.

7 Conclusions

In this paper, we propose a new kernel alignment induced LDA (kaLDA). The objective function of kaLDA is very similar to classical LDA objective. The Stifel-manifold gradient descent algorithm can solve kaLDA objective efficiently. We have also extended kaLDA to multi-label problems. Extensive experiments show the effectiveness of kaLDA in both single-label and multi-label problems.

Acknowledgment. This work is partially supported by US NSF CCF-0917274 and NSF DMS-0915228.

References

1. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition 37(9), 1757–1771 (2004)
2. Cristianini, N., Shawe-taylor, J., Elisseeff, A., Kandola, J.S.: On kernel target alignment. Advances in Neural Information Processing Systems 14, 367 (2002)
3. Cristianini, N., et al.: Method of using kernel alignment to extract significant features from a large dataset. US Patent 7,299,213 (2007)
4. Cuturi, M.: Fast global alignment kernels. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 929–936 (2011)
5. Ding, C., He, X.: K-means clustering via principal component analysis. In: Proc. of International Conference on Machine Learning, ICML 2004 (2004)
6. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM Journal on Matrix Analysis and Applications 20(2), 303–353 (1998)
7. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: NIPS, vol. 14, pp. 681–687 (2001)
8. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding 106(1), 59–70 (2007)
9. Guo, Y., Hastie, T., Tibshirani, R.: Regularized linear discriminant analysis and its application in microarrays. Biostatistics 8(1), 86–100 (2007)

10. Hoi, S.C., Lyu, M.R., Chang, E.Y.: Learning the unified kernel machines for classification. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 187–196. ACM (2006)
11. Howard, A., Jebara, T.: Transformation learning via kernel alignment. In: International Conference on Machine Learning and Applications, ICMLA 2009, pp. 301–308. IEEE (2009)
12. Ji, S., Tang, L., Yu, S., Ye, J.: Extracting shared subspace for multi-label classification. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 381–389. ACM (2008)
13. Jia, Y., Nie, F., Zhang, C.: Trace ratio problem revisited. IEEE Transactions on Neural Networks 20(4), 729–735 (2009)
14. Kong, D., Ding, C.: A semi-definite positive linear discriminant analysis and its applications. In: 2012 IEEE 12th International Conference on Data Mining (ICDM), pp. 942–947. IEEE (2012)
15. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
16. Lee, Y.J., Grauman, K.: Foreground focus: Unsupervised learning from partially matching images. International Journal of Computer Vision 85(2), 143–166 (2009)
17. Li, H., Jiang, T., Zhang, K.: Efficient and robust feature extraction by maximum margin criterion. IEEE Transactions on Neural Networks 17(1), 157–165 (2006)
18. Park, H., Jeon, M., Rosen, J.B.: Lower dimensional representation of text data based on centroids and least squares. BIT Numerical Mathematics 43(2), 427–448 (2003)
19. Sim, T., Baker, S., Bsat, M.: The cmu pose, illumination, and expression (pie) database of human faces. Tech. Rep. CMU-RI-TR-01-02, Robotics Institute, Pittsburgh, PA (January 2001)
20. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.P.: Multi-label classification of music into emotions. In: ISMIR, vol. 8, pp. 325–330 (2008)
21. Wang, H., Ding, C., Huang, H.: Multi-label linear discriminant analysis. In: Danilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 126–139. Springer, Heidelberg (2010)
22. Wang, H., Yan, S., Xu, D., Tang, X., Huang, T.: Trace ratio vs. ratio trace for dimensionality reduction. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8. IEEE (2007)
23. Ye, J.: Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. Journal of Machine Learning Research, 483–502 (2005)
24. Ye, J., Ji, S.: Discriminant analysis for dimensionality reduction: An overview of recent developments. Biometrics: Theory, Methods, and Applications. Wiley-IEEE Press, New York (2010)
25. Yu, K., Yu, S., Tresp, V.: Multi-label informed latent semantic indexing. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 258–265. ACM (2005)
26. Zha, H., Ding, C., Gu, M., He, X., Simon, H.: Spectral relaxation for K-means clustering. In: Advances in Neural Information Processing Systems 14 (NIPS 2001), pp. 1057–1064 (2001)
27. Zhang, Y., Zhou, Z.H.: Multilabel dimensionality reduction via dependence maximization. ACM Transactions on Knowledge Discovery from Data (TKDD) 4(3), 14 (2010)
28. Zhu, X., Kandola, J., Ghahramani, Z., Lafferty, J.D.: Nonparametric transforms of graph kernels for semi-supervised learning. In: Advances in Neural Information Processing Systems, pp. 1641–1648 (2004)