Multimodal Linear Discriminant Analysis via Structural Sparsity

Yu Zhang¹ and Yuan Jiang²

¹Department of Computer Science and Engineering, Hong Kong University of Science and Technology
²National Key Laboratory for Novel Software Technology, Nanjing University

zhangyu@cse.ust.hk, jiangy@lamda.nju.edu.cn

Abstract

Linear discriminant analysis (LDA) is a widely used supervised dimensionality reduction technique. Even though the LDA method has many real-world applications, it has some limitations such as the single-modal problem that each class follows a normal distribution. To solve this problem, we propose a method called multimodal linear discriminant analysis (MLDA). By generalizing the between-class and within-class scatter matrices, the MLDA model can allow each data point to have its own class mean which is called the instance-specific class mean. Then in each class, data points which share the same or similar instance-specific class means are considered to form one cluster or modal. In order to learn the instance-specific class means, we use the ratio of the proposed generalized between-class scatter measure over the proposed generalized within-class scatter measure, which encourages the class separability, as a criterion. The observation that each class will have a limited number of clusters inspires us to use a structural sparse regularizer to control the number of unique instance-specific class means in each class. Experiments on both synthetic and real-world datasets demonstrate the effectiveness of the proposed MLDA method.

1 Introduction

In the LDA model, each class is modeled by a Gaussian distribution, which is not capable of handling the multimodal structure contained in the data. The mixture discriminant analysis (MDA) [Hastie and Tibshirani, 1996] is one of the earliest multimodal extensions of the LDA model and it uses a Gaussian mixture model to describe the data in a class with the number of Gaussian components pre-defined and shared among different classes. Different from the MDA model, the LLDA method proposed in [Kim and Kittler, 2005] first groups the whole dataset by using the Gaussian mixture model or k-means clustering algorithm and then learns a LDA model on each cluster. The subclass discriminant analysis (SDA) method proposed in [Zhu and Martínez, 2006] aims at learning the number of clusters in each class based on the leave-one-out test criterion or a stability criterion proposed in [Martínez and Zhu, 2005]. The local Fisher discriminant analysis (LFDA) proposed in [Sugiyama, 2006; Sugiyama, 2007], which have similar ideas to nonparametric discriminant analysis [Kuo and Landgrebe, 2004; Li et al., 2009], conquers the multimodal problem by incorporating the local structure into the definitions of the within-class and between-class scatter matrices. One by-product of those multimodal LDA models is that the dimension of the embedding space is no longer limited to at most the number of classes minus 1 since the rank of the between-class scatter matrix becomes larger by exploiting the multimodal structure.

All the existing multimodal extensions of the LDA model have some limitations. For example, the MDA and LLDA models require that the number of components should be predefined, which is not an easy model selection problem. The assumptions of the SDA method that different classes have the same number of clusters and that clusters in each class have comparable numbers of data points seem a bit restricted. Moreover, the SDA method consists of two stages with the first stage learning the cluster structure while the second one learns the transformation, leading to a suboptimal learner. For the LFDA method, it is not easy to infer the cluster structure from data.

In this paper, we propose a method called multimodal linear discriminant analysis (MLDA) to alleviate those limitations in existing multimodal extensions of the LDA model. The MLDA model generalizes the definitions of the between-class and within-class scatter matrices by introducing the instance-specific class means. That is, different from the traditional LDA model which has only one class mean for each class, in the MLDA method each data point has its own class mean and data points, which share the same or similar class means, in each class are considered to belong to a cluster or modal. We reveal that the conventional scatter matrices are special cases of the generalized ones. Since the instance-specific class means are unknown for real-world problems, we utilize the ratio of the proposed generalized between-class scatter measure over the proposed generalized within-class scatter measure as an objective function to learn them. Usually each class has a limited number of clusters, implying that the number of unique instance-specific class means in each class is not very large, and this observation inspires us to use a structurally

*Corresponding Author
sparse regularizer (i.e., the $\ell_1,p$ norm) to enforce some pairs of instance-specific class means to be identical and in the meanwhile to control the number of unique instance-specific class means in each class. One advantage to use the structurally sparse regularizer is that we need not to specify the number of clusters and not to impose constraints on the cluster structure. We devise a proximal average method based on the GIST algorithm [Gong et al., 2013] to solve the resulting objective function with each subproblem having an analytical solution. Experiments on both synthetic and real-world datasets demonstrate the effectiveness of our proposed method.

2 The MLDA Model

In this section, we first review the conventional LDA model and then present the generalized scatter matrices, which set the stage for the introduction of the proposed MLDA model, as well as their properties.

2.1 LDA Revisited

The LDA model is a supervised dimensionality reduction method. Suppose that a training dataset consists of $n$ pairs of training samples $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^D$ denotes the $i$th data point and its class label is denoted by $y_i \in \{1, \ldots, c\}$, making the learning problem a multi-class classification problem with the number of classes being $c$. Let $n_i$ denote the number of data points in the $i$th class and so $n = \sum_{i=1}^c n_i$. Moreover, we define the overall mean $\bar{m}$ as the mean of all data points, i.e., $\bar{m} = \frac{1}{n} \sum_{i=1}^n x_i$, and the class mean for the $i$th class as $\bar{m}_i = \frac{1}{n_i} \sum_{y_i = 1}^n x_j$. Then the between-class and within-class scatter matrices are defined as

$$Q_b = \frac{1}{n} \sum_{i=1}^c n_i (\bar{m}_i - \bar{m})(\bar{m}_i - \bar{m})^T,$$

$$Q_w = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{m}_y_i)(x_i - \bar{m}_y_i)^T.$$

Then the LDA model seeks the optimal transformation matrix $W^*$ by solving the following problem

$$\max_{W \in \mathbb{R}^{D \times d}} \text{tr} \left( (W^T (Q_w + \alpha I_2) W)^{-1} W^T Q_b W \right),$$

where $d$ is the dimension of the reduced space, $I_2$ denotes an $a \times a$ identity matrix, the superscript $^{-1}$ denotes the matrix inverse, and $\alpha$ is a regularization parameter to guarantee the non-singularity.

2.2 Generalized Scatter Matrices

According to problem (3) and the definition of the within-class scatter matrix in Eq. (2), the LDA model enforces data points in one class to be close to the class mean, making it applicable to single-modal data. However, the data used in many applications exhibit multimodal structure, leading to unsatisfactory performance by using the LDA model. Here we propose the generalized within-class and between-class scatter matrices as

$$S_b = \frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n \delta(y_i \neq y_j)(m_i - m_j)(m_i - m_j)^T \quad (4)$$

$$S_w = \frac{1}{n} \sum_{i=1}^n (x_i - m_i)(x_i - m_i)^T,$$

where $\delta(z)$ returns 1 when $z$ holds and otherwise 0, and $m_i$ represents the instance-specific class mean for $x_i$. Suppose that the $i$th class has $l_i$ unique instance-specific class means, i.e., the set $\{m_{i,j} | y_j = i\}$ having $l_i$ elements $\{m_{i,j} \}$ for $k = 1, \ldots, l_i$. Then we can rewrite the generalized within-class scatter matrix defined in Eq. (5) as

$$S_w = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{l_i} \sum_{k \in C_{i,j}} (m_k - m_{i,j}^T)(m_k - m_{i,j}^T),$$

where $C_{i,j} = \{k|m_k = m_{i,j}^T\}$ denotes the set of the indices corresponding to the data points belonging to the $j$th modal (or cluster) of the $i$th class. By using the generalized scatter matrices, we can easily model the multimodal structure contained in the data.

2.3 Properties

The newly defined scatter matrices in Eqs. (4) and (5) are called the generalized between-class and within-class scatter matrices since the conventional scatter matrices are special cases of them, and the relation is revealed in the following theorem.

Theorem 1 By setting $m_i$ to be $\bar{m}_i$ when $x_i$ belongs to the $j$th class, we have $S_b = Q_b$ and $S_w = Q_w$.

Recall that the conventional between-class and within-class scatter matrices have one property that $Q_b + Q_w = Q_i$, where $Q_i = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{m})(x_i - \bar{m})^T$ is the total scatter matrix. Similar to that, the following theorem shows that the generalized scatter matrices have a similar property.

Theorem 2 For $S_b$ and $S_w$, we have

$$S_b + S_w = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{l_i} \sum_{k \in C_{i,j}} (m_k - \bar{m})(m_k - \bar{m})^T$$

$$- \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{l_i} \sum_{k \in C_{i,j}} (m_k - m_{i,j}^T)(m_{i,j}^T - \bar{m})^T$$

$$- \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{l_i} \sum_{k \in C_{i,j}} (m_{i,j}^T - \bar{m})(m_{i,j} - m_{i,j}^T)^T$$

$$- \frac{1}{n^2} \sum_{i=1}^c \sum_{j=1}^{l_i} \sum_{k \in C_{i,j}} n_{i,j}(m_{i,j} - \bar{m})(m_{i,j} - \bar{m})^T,$$

where $n_{i,j}$ is the cardinality of $C_{i,j}$ implying that the $j$th cluster of the $i$th class has $n_{i,j}$ data points, $\bar{m}_i = \frac{1}{l_i} \sum_{j=1}^{l_i} m_{i,j}$, and $\bar{m} = \frac{1}{n} \sum_{i=1}^n m_i$. Moreover, when $m_{i,j}^T$ is set to be the average of all the data points in the $j$th cluster of the $i$th class, we have

$$S_b + S_w = Q_i - \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{l_i} n_{i,j}(m_{i,j}^T - \bar{m}_i)(m_{i,j}^T - \bar{m}_i)^T.$$

3449
According to Theorem 2, we can see that when \( \{m_i\} \) take appropriate values, the sum of the generalized between-class and within-class scatter matrices equals the total scatter matrix minus a matrix called the average between-cluster scatter matrix whose formulation is just the second term in the right-hand side of Eq. (6). The average between-cluster scatter matrix measures the separability between different clusters in the same class and is not useful for discriminating different classes.

### 2.4 The Model

When given \( M = (m_1, \ldots, m_n) \), we formulate the objective function of the MLDA model in a similar way to the conventional LDA model as

\[
\max_{W \in \mathbb{R}^{d \times d}} \text{tr} \left( (W^T(S_w + \alpha I_D)W)^{-1} W^T S_b W \right).
\]

(7)

A by-product of this formulation is that if the unique instance-specific class means are linearly independent, the rank of \( S_b \) is \( \sum_{i=1}^n t_i - 1 \) which is larger than \( c - 1 \) and so the dimension of the reduced space \( d \) can be larger than \( c - 1 \), which overcomes a limitation of the conventional LDA model that \( d \) is at most \( c - 1 \). The solution of problem (7) can be obtained by solving a generalized eigen-decomposition problem as

\[
S_b W = (S_w + \alpha I_D)W \Lambda,
\]

where \( \Lambda \) is a diagonal matrix containing the largest \( d \) eigenvalues.

### 2.5 Learning M

When \( M \) is given, we can compute the transformation matrix \( W \). However, in most applications, \( M \) is unknown and we need to learn it from data automatically.

We propose to use the objective function of the MLDA model, which encourages the class separability, to learn \( M \), that is, maximizing

\[
\text{tr} \left( (W^T(S_w + \alpha I_D)W)^{-1} W^T S_b W \right)
\]

with respect to \( W \) and \( M \). However, this problem contains two variables, making it not very easy to be solved. In the following theorem, we introduce its upper bound which can be used as a surrogate function.

**Theorem 3**

\[
\max_{W \in \mathbb{R}^{d \times d}} \text{tr} \left( (W^T(S_w + \alpha I_D)W)^{-1} W^T S_b W \right) \leq \text{tr} \left( (S_w + \alpha I_D)^{-1} S_b \right).
\]

(8)

By using Theorem 3, we can use the upper bound as a surrogate function to learn \( M \) as

\[
\max_M \text{tr} \left( (S_w + \alpha I_D)^{-1} S_b \right).
\]

If directly solving this problem, we may get a trivial solution such as \( m_i = x_i \). In order to avoid this situation, we can utilize the structural sparsity among all \( m_i \)'s that each class has a limited number of clusters, which corresponds to the situation that the number of unique instance-specific class means for each class is not very large or equivalently that many elements in \( \{m_i - m_j\} | y_i = y_j \} \) are zero vectors. In the next section, we discuss how to learn \( M \) via the structurally sparse regularization without knowing the number of clusters in each class.

### 3 Learning M via Structural Sparsity

When there is no information available about \( M \), we need to learn \( M \) from data automatically. In this case, we utilize the structurally sparse regularization \cite{Hocking et al., 2011; Bach et al., 2012b; Bach et al., 2012a} to enforce the sparsity in the set \( \{m_i - m_j\} | y_i = y_j \} \).

The objective function to learn \( M \) is formulated as

\[
\min_M \beta \sum_{i<j, y_i = y_j} \gamma_{ij} ||m_i - m_j||_p - \text{tr} \left( (S_w + \alpha I_D)^{-1} S_b \right),
\]

(9)

where \( S_b \) and \( S_w \) are the generalized between-class and within-class scatter matrices defined in Eqs. (4) and (5), \( \beta \) is a positive regularization parameter, \( \| \cdot \|_p \) denotes the \( \ell_p \) norm of a vector, and \( \gamma_{ij} \) denotes the nonnegative similarity between \( x_i \) and \( x_j \). \( \gamma_{ij} \) can be set to 1 by assuming that each pair of data points are equal of similarity or to be \( \exp(-\frac{||x_i - x_j||^2}{2\sigma}) \) to incorporate the local density into the consideration.

Similar to \cite{Hocking et al., 2011}, the regularization term in problem (9) (i.e., the first term) enforces the difference between some pairs of the instance-specific class means corresponding to data points in the same class to be zero, leading to a limited number of unique instance-specific class means in each class. In order to make the subproblem in the proximal method convex as we will see later, \( p \) is assumed to be no less than 1, i.e., \( p \geq 1 \). Moreover, when \( p \) equals 1, the regularization term only enforces some corresponding entries in some pairs of columns in \( M \) but not the entire columns to be identical, which does not match our expectation. Hence \( p \) is assumed to be larger than 1, i.e., \( p > 1 \).

Obviously problem (9) is non-convex due to the non-convexity of the second term in problem (9). To solve problem (9), we use a proximal method for non-convex optimization problems called the GIST algorithm \cite{Gong et al., 2013}. To apply the GIST algorithm to problem (9), we define \( f(M) \) and \( g(M) \) as \( f(M) = -\text{tr} \left( (S_w + \alpha I_D)^{-1} S_b \right) \) and \( g(M) = \beta \sum_{i<j, y_i = y_j} \gamma_{ij} ||m_i - m_j||_p \). We rewrite \( S_b \) and \( S_w \) in Eqs. (4) and (5) as

\[
S_b = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T \quad \text{and} \quad S_w = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T,
\]

where \( \bar{x} \) is the mean vector of \( X \) and \( G \) is the Laplacian matrix of \( G \). Then we can compute the gradient \( \nabla g(M) \) as

\[
\nabla g(M) = \frac{2}{n} S_b^{-1} - \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T S_b^{-1} (X - \bar{x}) - \frac{2}{n} S_b^{-1} ML,
\]

where \( L \) is the Laplacian matrix of \( G \). In the GIST algorithm, we only need to solve the following problem:

\[
\min_M \frac{1}{2} ||M - \hat{M}(k)||_F^2 + \beta \sum_{i<j, y_i = y_j} \gamma_{ij} ||m_i - m_j||_p
\]

where \( \hat{M}(k) = M(k) - \frac{1}{2\ell_k} \nabla f(M(k)) \). Note that in the above problem, two submatrices of \( M \), \( M_i \) and \( M_j \) for any \( i \neq j \), where matrix \( M_i \) consists of all the instance-specific class means for the \( i \)th class as its columns, are decoupled in the two terms of the objective function and so the above problem can be decomposed into \( c \) independent problems with the \( i \)th one formulated as

\[
\min_{M_i} \frac{1}{2} ||M_i - \hat{M}_i(k)||_F^2 + \beta \sum_{j,k} \gamma_{jk} ||m_j - m_k||_p
\]

(10)
where $R_i = \{j | y_j = i\}$. By this decomposition, we not only reduce the size of the problem to be optimized but also enable the parallel optimization for the $c$ independent problems. In the following section, we discuss the optimization procedure to solve problem (10) in details.

### 3.1 Optimization Procedure for Problem (10)

In this section, we discuss how to solve problem (10). To simplify the presentation, we use slightly different notations from problem (10) to formulate it as

$$
\min_{Z \in \mathbb{R}^{m \times n}} \frac{t}{2} \|Z - A\|_F^2 + \frac{1}{t} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \theta_{ij} \|z_i - z_j\|_p, \tag{11}
$$

where $z_i$ is the $i$th column of $Z$. Note that problems (11) and (10) are equivalent. Here we investigate a case that $p = 2$ and other cases will be left for future study. Problem (11) is a convex proximal problem and due to the complex regularizer on $Z$, it has no closed-form solution. Here we use the proximal average technique [Yu, 2013] to solve it. The proximal average method approximates the solution of a proximal problem with a convex combination of multiple regularizers by a convex combination of the solutions of proximal problems with each regularizer individually. One advantage of the proximal average method is that each simple proximal problem can have an analytical solution and hence the approximated solution also has, leading to a low complexity for solving problem (11).

For problem (11), without loss of generality, we assume that $\sum_{i,j} \theta_{ij} = 1$. Then we can see that the regularizer in problem (11) is a convex combination of simple regularizers $\|z_i - z_j\|_p$. Suppose $Z^{(i,j)}$ is the solution of the following problem as

$$
\min_{Z \in \mathbb{R}^{m \times n}} \frac{t}{2} \|Z - A\|_F^2 + \frac{1}{t} \|z_i - z_j\|_p. \tag{12}
$$

Then the proximal average method approximate the solution of problem (11) by $\sum_{i,j} \theta_{ij} Z^{(i,j)}$. For problem (12), it has an analytical solution as shown in the following theorem.

**Theorem 4** When $p = 2$, problem (12) has an analytical solution as

$$
z_k = a_k \forall k \neq i, j$$

$$
z_i = \begin{cases} \frac{1}{2} (a_i + a_j) & \text{if } \|a_i - a_j\|_2 \leq 2/t \\ a_i - \frac{a_i - a_j}{\|a_i - a_j\|_2} & \text{otherwise} \end{cases},$$

$$
z_j = \begin{cases} \frac{1}{2} (a_i + a_j) & \text{if } \|a_i - a_j\|_2 \leq 2/t \\ a_j - \frac{a_i - a_j}{\|a_i - a_j\|_2} & \text{otherwise} \end{cases},$$

where $a_i$ is the $i$th column of $A$.

Moreover, we can divide the set of index pairs $\{(i, j) | i < j\}$ into several subsets $S_1, \ldots, S_l$ so that in each subset $S_l$, the indices in different pairs are non-overlapping, i.e., for any two different pairs $(i_1, j_1)$ and $(i_2, j_2)$ in $S_l$, $i_1, j_1, i_2, j_2$ are different from each other. Suppose $Z^{(i)}$ is the solution of the following problem as

$$
\min_{Z \in \mathbb{R}^{m \times n}} \frac{t}{2} \|Z - A\|_F^2 + \frac{1}{\theta(i)} \sum_{(i, j) \in S_l} \theta_{ij} \|z_i - z_j\|_p, \tag{13}
$$

where $\theta(l) = \sum_{(i, j) \in S_l} \theta_{ij}$. Then the proximal average method approximates the solution of problem (11) by $\sum_{l} \theta(l) Z^{(l)}$. Similar to problem (12), problem (13) has an analytical solution by using a similar analysis and we omit the details. For the construction of the subsets, we can adopt a greedy approach which $S_l$ choose successive indices as a pair, $S_2$ choose the indices with distance 2 as a pair, and so on. It is easy to show that $t = O(m)$.

For the two proximal average variants, we can compute the solutions of all the subproblems (i.e., problems (12) and (13)) in a parallel way, which can accelerate the training process. It is easy to show that problem (12) needs to be solved for $O(n^2)$ times but problem (13) is only needed for $O(m)$ times. So for complexity consideration, we are more preferable to the second approach even though the analysis in the first approach is the footstone of the second one.

For the whole GIST algorithm where the proximal subproblem is solved by the proximal average method approximately, by following [Zhong and Kwok, 2014] we can prove that the algorithm can converge to a critical point, which can be a local optimum for problem (9).

### 3.2 Discussion

To see which condition leads to different instance-specific class means of the same value, we investigate a special case of problem (9) that $p$ equals 2 and $\gamma_{ij}$ is equal to $\gamma$ for all pairs $(i, j)$ where $\gamma$ is a constant. Suppose in a class there exists a cluster consisting of two data points $x_i$ and $x_j$ only with their instance-specific class means denoted by $\mathbf{m}_i$ and $\mathbf{m}_j$. By setting the derivative of the objective function in problem (9) with respect to $\mathbf{m}_i$ and $\mathbf{m}_j$ to zero respectively, we can get the stationary condition as

$$
\sum_{k \neq i, j} \frac{\partial}{\partial \mathbf{m}_i} \|\mathbf{m}_i - \mathbf{m}_k\|_2 + \frac{\partial f(M)}{\partial \mathbf{m}_i} + \frac{\partial f(M)}{\partial \mathbf{m}_j} = 0, \tag{14}
$$

$$
\sum_{k \neq i, j} \frac{\partial}{\partial \mathbf{m}_j} \|\mathbf{m}_j - \mathbf{m}_k\|_2 + \frac{\partial f(M)}{\partial \mathbf{m}_i} + \frac{\partial f(M)}{\partial \mathbf{m}_j} = 0. \tag{15}
$$

where $f(M) = -\text{tr} \left( (S_w + \alpha \mathbf{I}_p)^{-1} S_b \right)$ and $\frac{\partial g(x)}{\partial x}$ denotes the (sub)gradient of a function $g(\cdot)$ with respect to a variable $x$. Since $x_i$ and $x_j$ compose a cluster which implies that $\mathbf{m}_i$ is equal to $\mathbf{m}_j$ and $\mathbf{m}_j$ not equal to $\mathbf{m}_k$ for all other $k$'s in the same class, then the first terms in Eqs. (14) and (15) are equal to each other and based on the chain rule, the last terms in those two equations are equal to the subgradients $\partial f(0)/\partial \mathbf{m}_i$ and $-\partial f(0)/\partial \mathbf{m}_j$ which satisfy that their $l_2$ norms are no larger than 1. By computing the difference between Eqs. (14) and (15), we can have $\left\| \frac{\partial f(M)}{\partial \mathbf{m}_i} - \frac{\partial f(M)}{\partial \mathbf{m}_j} \right\|_2 \leq 2\beta\gamma$, which implies that data points with similar gradients of the function $f(\cdot)$ are more likely to share the same instance-specific class mean. So for nearby data points in one class, the difference between their gradients can be small and so they tend to lie in a cluster.

### 4 Experiments

In this section, we empirically test the performance of the MLDA model.
The methods in comparison include the LDA method [Bel-
humeur et al., 1997], the LLDA method [Kim and Kittler,
2005], the SDA method [Zhu and Martínez, 2006], and the
LFDA method [Sugiyama, 2007]. As discussed in the intro-
duction, the LLDA method is a two-stage method which first
clusters the whole dataset via the $k$-means clustering algorithm
and then learns a LDA model on each cluster, the SDA method
can learn the number of clusters, and the LFDA method uti-
lizes the local density to model the multimodal structure.

Recall that the proposed objective function to learn $M$ in
problem (9) is non-convex, making it sensitive to the initial val-
ue of $M$. In the following experiments, the initial value for $m_i$
is set to be $x_1$. The parameters (i.e., $\eta, \sigma,$ and $t_0$) in the GIST
algorithm are set to be 2, 0.2 and 1 respectively. The reg-
ularization parameters $\alpha$ and $\beta$ are selected via the 10-fold
cross validation method from the candidate set $\{0.001, 0.01, 0.1, 1\}$
and the hyperparameters of the baselines in comparison are also
selected via the 10-fold cross validation. For fair comparison
with the conventional LDA method, we set the reduced
dimensions of all the methods in comparison to $c − 1$ where $c$
is the number of classes. After learning the transformations in
all the methods, we use the nearest neighbor classifier to make prediction.

4.1 Experiments on Synthetic Data

The first synthetic data is generated as follows. There are
two classes in this dataset. The first class contains two clus-
ters with the first cluster generated from a Gaussian distribu-
tion $\mathcal{N}(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, I_2)$ and the second one following $\mathcal{N}(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, I_2)$. The second class has three clusters with them sampled from
$\mathcal{N}(\begin{bmatrix} 10 \\ 0 \end{bmatrix}, I_2)$, $\mathcal{N}(\begin{bmatrix} 10 \\ 10 \end{bmatrix}, I_2)$, and $\mathcal{N}(\begin{bmatrix} 0 \\ -10 \end{bmatrix}, I_2)$ respectively.
For each cluster in those two classes, we randomly generate
100 data points respectively to form the training set and the
data set is plotted in Figure 1(a). The settings of the second
synthetic data keep almost the same as those of the first one
with the difference lying in the setting of the covariance ma-
trices in different clusters. Here different covariance matrices
are used for different classes, i.e., the two clusters in the first
class use $\begin{bmatrix} 1.5 & 0 \\ 0 & 1 \end{bmatrix}$ as the covariance matrix and all the co-
variance matrices of the three clusters in the second class equal
$\begin{bmatrix} 1 & 0 \\ 0 & 1.5 \end{bmatrix}$.

The one-dimensional transformations of the LDA, SDA, LFDA,
and MLDA methods for the two synthetic datasets are plotted in Figure 1. Here the LLDA method is not in-
cluded since it cannot discover the clusters within each class.
Based on the data distributions of two synthetic data, the ide-
al transformation is the horizontal line and we can compare
with the transformations produced by the methods in compari-
son. From the results, we can see the discriminative ability of
the transformation produced by the SDA model is the worst
since some clusters in different classes are overlapped after
the dimensionality reduction. One reason is that the data dis-
tribution violates an assumption of the SDA model that the
numbers of clusters in different classes are equal to each other.
The LDA model is better than the SDA model but the data
points in different classes still have some overlap after the
projection, leading to possible classification error for testing.

4.2 Experiments on Real-World Datasets

Six real-world datasets are used in our experiments, including
the ETH-80, COIL-20, COIL-100, AR, UMIST, and MNIST
databases. The ETH-80 dataset [Leibe and Schiele, 2003]
contains images of the following categories: apples, pears,
cars, cows, horses, dogs, tomatoes, and cups, and each category includes the images of 10 objects taken at 41 orientations, which give us a total of 410 images per category. The COIL-20 dataset contains 1,440 gray-scale images for 20 objects and each object has 72 images of size $16 \times 16$ with the difference of two successive viewpoint as 5 degrees. The images in the COIL-100 dataset are taken in a similar way to the COIL-20 dataset and it contains 7,200 gray-scale images for 100 objects. The AR dataset [Martínez and Benavente, 1998] contains frontal face images of 100 persons (50 men and 50 women) with different expressions, illuminations, and occlusions and there are 26 images for each person taken in two sessions, each having 13 images. The UMIST dataset [Graham and Allinson, 1998] is a multi-view dataset consisting of 575 gray-scale images of 20 people (subject) with each covering a wide range of poses from profile to frontal views. The MNIST dataset [LeCun and Cortes, 1998] contains 70,000 handwritten digits ranging from 0 to 9 with each one having about 7,000 samples where each image is normalized to a gray-level image with the size as $14 \times 14$.

In the proposed MLDA method, $\gamma_{ij}$ is set to be $\exp\left\{ -\frac{1}{\sigma_i^2} \| x_i - x_j \|_2^2 \right\}$ when $y_i = y_j$ and 0 otherwise, where $\sigma_i$ denotes the average pairwise Euclidean distance in the $i$th class. In order to investigate the effect of varying the size of the training set, we randomly sample 30%, 50%, and 70% of the total data as the training set and the rest forms the test set. Each configuration repeats for 10 times and the average results as well as the standard deviations are plotted in Figure 2. From the results, we can see that sometimes the performance of the LLDA, SDA and LFDA models is better than that of the LDA, for example, on the COIL-20 and COIL-100 datasets, but in other cases, they perform only comparably to the LDA model. The proposed MLDA method has the best performance in all settings. In the ETH-80 and AR datasets, the performance of the proposed MLDA method using 30% data for training is comparable to that of the SDA or LFDA method with the training percentage as 70%, which in some aspect shows the effectiveness of the MLDA method.

5 Conclusion

In this paper, we propose a multimodal extension of the LDA model by generalizing the between-class and within-class scatter matrices based on the instance-specific class means assigned to each data point. The learning of the instance-specific class means is accomplished by maximizing a surrogate function regularized by the structural sparsity and the resultant objective function is solved by a proximal algorithm.

The objective function in problem (8) can be viewed as a loss function, which encourages the class separability, for $M$, but it is not the only one. We will try other criterions such as the stability criterion proposed in [Martínez and Zhu, 2005] and the implementation can be reused since we only need to modify the definition of the function $f(\cdot)$ and the calculation of its gradient. Moreover, the proposed MLDA model is a linear model and we are interested in investigating its nonlinear extension.

Acknowledgments

This work is supported by the NSFC (61473087, 61673201, 61673202), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Natural Science Foundation of Jiangsu Province (BK20141340).

References

[Bach et al., 2012a] Francis R. Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization


