Multimodal Linear Discriminant Analysis via Structural Sparsity

Yu Zhang[1] **and Yuan Jiang**[2*]
[1]Department of Computer Science and Engineering, Hong Kong University of Science and Technology
[2]National 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 regularizor 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 $\{\mathbf{x}_i, y_i\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^D$ denotes the ith 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 ith class and so $n = \sum_{i=1}^c n_i$. Moreover, we define the overall mean $\bar{\mathbf{m}}$ as the mean of all data points, i.e., $\bar{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$, and the class mean for the ith class as $\bar{\mathbf{m}}_i = \frac{1}{n_i}\sum_{y_j=i} \mathbf{x}_j$. Then the between-class and within-class scatter matrices are defined as

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

$$\mathbf{Q}_w = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{m}}_{y_i})(\mathbf{x}_i - \bar{\mathbf{m}}_{y_i})^T. \quad (2)$$

Then the LDA model seeks the optimal transformation matrix $\mathbf{W}^\star$ by solving the following problem

$$\max_{\mathbf{W}\in\mathbb{R}^{D\times d}} \operatorname{tr}\left(\left(\mathbf{W}^T(\mathbf{Q}_w + \alpha\mathbf{I}_D)\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{Q}_b\mathbf{W}\right), \quad (3)$$

where d is the dimension of the reduced space, $\mathbf{I}_a$ denotes an $a \times a$ identity matrix, the superscript $^{-1}$ denotes the matrix inverse, and α 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

$$\mathbf{S}_b = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta(y_i \neq y_j)(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \quad (4)$$

$$\mathbf{S}_w = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m}_i)(\mathbf{x}_i - \mathbf{m}_i)^T, \quad (5)$$

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

$$\mathbf{S}_w = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{l_i}\sum_{k\in\mathcal{C}_{i,j}}(\mathbf{x}_k - \mathbf{m}_j^i)(\mathbf{x}_k - \mathbf{m}_j^i)^T,$$

where $\mathcal{C}_{i,j} = \{k | \mathbf{m}_k = \mathbf{m}_j^i\}$ denotes the set of the indices corresponding to the data points belonging to the jth modal (or cluster) of the ith 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 $\mathbf{m}_i$ to be $\bar{\mathbf{m}}_j$ when $\mathbf{x}_i$ belongs to the jth class, we have $\mathbf{S}_b = \mathbf{Q}_b$ and $\mathbf{S}_w = \mathbf{Q}_w$.*

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

Theorem 2 *For $\mathbf{S}_b$ and $\mathbf{S}_w$, we have*

$$\begin{aligned}\mathbf{S}_b + \mathbf{S}_w =& \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{l_i}\sum_{k\in\mathcal{C}_{i,j}}(\mathbf{x}_k - \tilde{\mathbf{m}})(\mathbf{x}_k - \tilde{\mathbf{m}})^T \\ &- \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{l_i}\sum_{k\in\mathcal{C}_{i,j}}(\mathbf{x}_k - \mathbf{m}_j^i)(\mathbf{m}_j^i - \tilde{\mathbf{m}})^T \\ &- \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{l_i}\sum_{k\in\mathcal{C}_{i,j}}(\mathbf{m}_j^i - \tilde{\mathbf{m}})(\mathbf{x}_k - \mathbf{m}_j^i)^T \\ &- \frac{1}{n^2}\sum_{i=1}^{c} n_i\sum_{j=1}^{l_i} n_{i,j}(\mathbf{m}_j^i - \tilde{\mathbf{m}}_i)(\mathbf{m}_j^i - \tilde{\mathbf{m}}_i)^T,\end{aligned}$$

where $n_{i,j}$ is the cardinality of $\mathcal{C}_{i,j}$ implying that the jth cluster of the ith class has $n_{i,j}$ data points, $\tilde{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^n \mathbf{m}_i$, and $\tilde{\mathbf{m}}_i = \frac{1}{n_i}\sum_{y_j=i}\mathbf{m}_j$. Moreover, when $\mathbf{m}_j^i$ is set to be the average of all the data points in the jth cluster of the ith class, we have

$$\mathbf{S}_b + \mathbf{S}_w = \mathbf{Q}_t - \sum_{i=1}^{c}\frac{n_i}{n^2}\sum_{j=1}^{l_i} n_{i,j}(\mathbf{m}_j^i - \bar{\mathbf{m}}_i)(\mathbf{m}_j^i - \bar{\mathbf{m}}_i)^T. \quad (6)$$

According to Theorem 2, we can see that when $\{\mathbf{m}_i\}$ 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 $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_n)$, we formulate the objective function of the MLDA model in a similar way to the conventional LDA model as

$$\max_{\mathbf{W}\in\mathbb{R}^{D\times d}} \operatorname{tr}\left(\left(\mathbf{W}^T(\mathbf{S}_w + \alpha\mathbf{I}_D)\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right). \quad (7)$$

A by-product of this formulation is that if the unique instance-specific class means are linearly independent, the rank of $\mathbf{S}_b$ is $\sum_{i=1}^{c} l_i - 1$ 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

$$\mathbf{S}_b\mathbf{W} = (\mathbf{S}_w + \alpha\mathbf{I}_D)\mathbf{W}\mathbf{\Lambda},$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the largest d eigenvalues.

2.5 Learning $\mathbf{M}$

When $\mathbf{M}$ is given, we can compute the transformation matrix $\mathbf{W}$. However, in most applications, $\mathbf{M}$ 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 $\mathbf{M}$, that is, maximizing $\operatorname{tr}\left(\left(\mathbf{W}^T(\mathbf{S}_w + \alpha\mathbf{I}_D)\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)$ with respect to $\mathbf{W}$ and $\mathbf{M}$. 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

$$\max_{\mathbf{W}\in\mathbb{R}^{D\times d}} \operatorname{tr}\left(\left(\mathbf{W}^T(\mathbf{S}_w + \alpha\mathbf{I}_D)\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right) \leq \operatorname{tr}\left((\mathbf{S}_w + \alpha\mathbf{I}_D)^{-1}\mathbf{S}_b\right).$$

By using Theorem 3, we can use the upper bound as a surrogate function to learn $\mathbf{M}$ as

$$\max_{\mathbf{M}} \operatorname{tr}\left((\mathbf{S}_w + \alpha\mathbf{I}_D)^{-1}\mathbf{S}_b\right). \quad (8)$$

If directly solving this problem, we may get a trivial solution such as $\mathbf{m}_i = \mathbf{x}_i$. In order to avoid this situation, we can utilize the structural sparsity among all $\mathbf{m}_i$'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 $\{\mathbf{m}_i - \mathbf{m}_j | y_i = y_j\}$ are zero vectors. In the next section, we discuss how to learn $\mathbf{M}$ via the structurally sparse regularization without knowing the number of clusters in each class.

3 Learning $\mathbf{M}$ via Structural Sparsity

When there is no information available about $\mathbf{M}$, we need to learn $\mathbf{M}$ from data automatically. In this case, we utilize the structurally sparse regularization [Hocking *et al.*, 2011; Bach *et al.*, 2012b; Bach *et al.*, 2012a] to enforce the sparsity in the set $\{\mathbf{m}_i - \mathbf{m}_j | y_i = y_j\}$.

The objective function to learn $\mathbf{M}$ is formulated as

$$\min_{\mathbf{M}} \beta \sum_{i<j, y_i=y_j} \gamma_{ij}\|\mathbf{m}_i - \mathbf{m}_j\|_p - \operatorname{tr}\left((\mathbf{S}_w + \alpha\mathbf{I}_D)^{-1}\mathbf{S}_b\right), \quad (9)$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are the generalized between-class and within-class scatter matrices defined in Eqs. (4) and (5), β is a positive regularization parameter, $\|\cdot\|_p$ denotes the ℓ_p norm of a vector, and γ_{ij} denotes the nonnegative similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. γ_{ij} can be set to 1 by assuming that each pair of data points are equal of similarity or to be $\exp\{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\omega^2}\}$ to incorporate the local density into the consideration.

Similar to [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 $\mathbf{M}$ but not the entire columns to be identical, which does not match our expectation. Hence p is assumed to be large 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 [Gong *et al.*, 2013]. To apply the GIST algorithm to problem (9), we define $f(\mathbf{M})$ and $g(\mathbf{M})$ as $f(\mathbf{M}) = -\operatorname{tr}\left((\mathbf{S}_w + \alpha\mathbf{I}_D)^{-1}\mathbf{S}_b\right)$ and $g(\mathbf{M}) = \beta\sum_{\substack{i<j \\ y_i=y_j}} \gamma_{ij}\|\mathbf{m}_i - \mathbf{m}_j\|_p$. We rewrite $\mathbf{S}_b$ and $\mathbf{S}_w$ in Eqs. (4) and (5) as $\mathbf{S}_b = \frac{1}{n^2}\mathbf{M}\mathbf{L}\mathbf{M}^T$ and $\mathbf{S}_w = \frac{1}{n}(\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})^T$, where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{G}$ is an $n \times n$ matrix with the (i,j)th element being 1 when $y_i \neq y_j$ and 0 otherwise, and $\mathbf{L}$ is the Laplacian matrix of $\mathbf{G}$. Then we can compute the gradient $\nabla f(\mathbf{M})$ as $\nabla f(\mathbf{M}) = \frac{2}{n^3}\tilde{\mathbf{S}}_w^{-1}\mathbf{M}\mathbf{L}\mathbf{M}^T\tilde{\mathbf{S}}_w^{-1}(\mathbf{M}-\mathbf{X}) - \frac{2}{n^2}\tilde{\mathbf{S}}_w^{-1}\mathbf{M}\mathbf{L}$, where $\tilde{\mathbf{S}}_w = \mathbf{S}_w + \alpha\mathbf{I}_D$.

In the GIST algorithm, we only need to solve the following problem:

$$\min_{\mathbf{M}} \frac{t_k}{2}\|\mathbf{M} - \hat{\mathbf{M}}^{(k)}\|_F^2 + \beta \sum_{i<j, y_i=y_j}^{n} \gamma_{ij}\|\mathbf{m}_i - \mathbf{m}_j\|_p$$

where $\hat{\mathbf{M}}^{(k)} = \mathbf{M}^{(k)} - \frac{1}{t_k}\nabla f(\mathbf{M}^{(k)})$. Note that in the above problem, two submatrices of $\mathbf{M}$, $\mathbf{M}_i$ and $\mathbf{M}_j$ for any $i \neq j$, where matrix $\mathbf{M}_i$ consists of all the instance-specific class means for the ith 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 ith one formulated as

$$\min_{\mathbf{M}_i} \frac{t_k}{2}\|\mathbf{M}_i - \hat{\mathbf{M}}_i^{(k)}\|_F^2 + \beta \sum_{j,k\in\mathcal{R}_i, j<k} \gamma_{jk}\|\mathbf{m}_j - \mathbf{m}_k\|_p, \quad (10)$$

where $\mathcal{R}_i = \{j|y_j = i\}$. 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

$$\min_{\mathbf{Z}\in\mathbb{R}^{D\times m}} \frac{t}{2}\|\mathbf{Z}-\mathbf{A}\|_F^2 + \sum_{i=1}^{m}\sum_{j=1,j>i}^{m}\theta_{ij}\|\mathbf{z}_i-\mathbf{z}_j\|_p, \quad (11)$$

where $\mathbf{z}_i$ is the ith column of $\mathbf{Z}$. 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 $\mathbf{Z}$, 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 $\|\mathbf{z}_i - \mathbf{z}_j\|_p$. Suppose $\mathbf{Z}^{(i,j)}$ is the solution of the following problem as

$$\min_{\mathbf{Z}} \frac{t}{2}\|\mathbf{Z}-\mathbf{A}\|_F^2 + \|\mathbf{z}_i-\mathbf{z}_j\|_p. \quad (12)$$

Then the proximal average method approximate the solution of problem (11) by $\sum_{j>i}\theta_{ij}\mathbf{Z}^{(i,j)}$. 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*

$$\mathbf{z}_k = \mathbf{a}_k \forall k \neq i, j$$

$$\mathbf{z}_i = \begin{cases} \frac{1}{2}(\mathbf{a}_i+\mathbf{a}_j) & \text{if } \|\mathbf{a}_i-\mathbf{a}_j\|_2 \leq 2/t \\ \mathbf{a}_i - \frac{\mathbf{a}_i-\mathbf{a}_j}{t\|\mathbf{a}_i-\mathbf{a}_j\|_2} & \text{otherwise} \end{cases},$$

$$\mathbf{z}_j = \begin{cases} \frac{1}{2}(\mathbf{a}_i+\mathbf{a}_j) & \text{if } \|\mathbf{a}_i-\mathbf{a}_j\|_2 \leq 2/t \\ \mathbf{a}_j - \frac{\mathbf{a}_j-\mathbf{a}_i}{t\|\mathbf{a}_i-\mathbf{a}_j\|_2} & \text{otherwise} \end{cases}$$

where $\mathbf{a}_i$ is the ith column of $\mathbf{A}$.

Moreover, we can divide the set of index pairs $\{(i,j)|i<j\}$ into several subsets $\mathcal{S}_1,\ldots,\mathcal{S}_t$ so that in each subset $\mathcal{S}_l$, 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 $\mathcal{S}_l$, i_1,j_1,i_2,j_2 are different from each other. Suppose $\mathbf{Z}^{(l)}$ is the solution of the following problem as

$$\min_{\mathbf{Z}} \frac{t}{2}\|\mathbf{Z}-\mathbf{A}\|_F^2 + \frac{1}{\theta_{(l)}}\sum_{(i,j)\in\mathcal{S}_l}\theta_{ij}\|\mathbf{z}_i-\mathbf{z}_j\|_p, \quad (13)$$

where $\theta_{(l)} = \sum_{(i,j)\in\mathcal{S}_l}\theta_{ij}$. Then the proximal average method approximates the solution of problem (11) by $\sum_l \theta_{(l)}\mathbf{Z}^{(l)}$. 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 $\mathcal{S}_1$ choose successive indices as a pair, $\mathcal{S}_2$ 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 $O(m^2)$ 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 γ_{ij} is equal to γ for all pairs (i,j) where γ is a constant. Suppose in a class there exists a cluster consisting of two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ 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

$$\sum_{\substack{k\neq j \\ y_k=y_i}} \frac{\partial\|\mathbf{m}_i-\mathbf{m}_k\|_2}{\partial\mathbf{m}_i} + \frac{\frac{\partial f(\mathbf{M})}{\partial\mathbf{m}_i}}{\beta\gamma} + \frac{\partial\|\mathbf{m}_i-\mathbf{m}_j\|_2}{\partial\mathbf{m}_i} = \mathbf{0} \quad (14)$$

$$\sum_{\substack{k\neq i \\ y_k=y_j}} \frac{\partial\|\mathbf{m}_j-\mathbf{m}_k\|_2}{\partial\mathbf{m}_j} + \frac{\frac{\partial f(\mathbf{M})}{\partial\mathbf{m}_j}}{\beta\gamma} + \frac{\partial\|\mathbf{m}_i-\mathbf{m}_j\|_2}{\partial\mathbf{m}_j} = \mathbf{0}, \quad (15)$$

where $f(\mathbf{M}) = -\mathrm{tr}\left((\mathbf{S}_w+\alpha\mathbf{I}_D)^{-1}\mathbf{S}_b\right)$ and $\frac{\partial g(x)}{\partial x}$ denotes the (sub)gradient of a function $g(\cdot)$ with respect to a variable x. Since $\mathbf{x}_i$ and $\mathbf{x}_j$ compose a cluster which implies that $\mathbf{m}_i$ is equal to $\mathbf{m}_j$ and $\mathbf{m}_k$ is not equal to $\mathbf{m}_i$ 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 $\partial\|\mathbf{0}\|_2$ and $-\partial\|\mathbf{0}\|_2$ which satisfy that their ℓ_2 norms are no larger than 1. By computing the difference between Eqs. (14) and (15), we can have $\left\|\frac{\partial f(\mathbf{M})}{\partial\mathbf{m}_i} - \frac{\partial f(\mathbf{M})}{\partial\mathbf{m}_j}\right\|_2 \leq 2\beta\gamma$, 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.

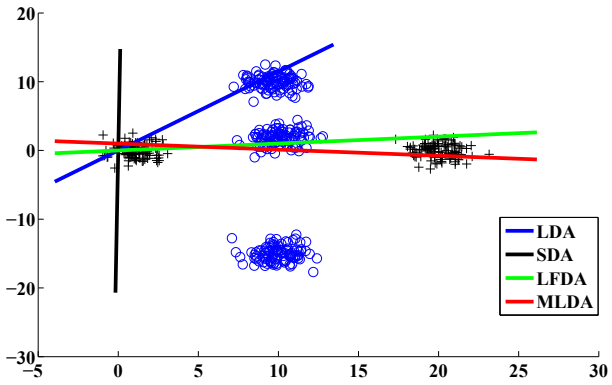
The methods in comparison include the LDA method [Belhumeur *et al.*, 1997], the LLDA method [Kim and Kittler, 2005], the SDA method [Zhu and Martínez, 2006], and the LFDA method [Sugiyama, 2007]. As discussed in the introduction, the LLDA method is a two-stage method which first clusters the whole dataset 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 utilizes the local density to model the multimodal structure.

Recall that the proposed objective function to learn $\mathbf{M}$ in problem (9) is non-convex, making it sensitive to the initial value of $\mathbf{M}$. In the following experiments, the initial value for $\mathbf{m}_i$ is set to be $\mathbf{x}_i$. The parameters (i.e., η, σ, and t_0) in the GIST algorithm are set to be 2, 0.2 and 1 respectively. The regularization parameters α and β 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.

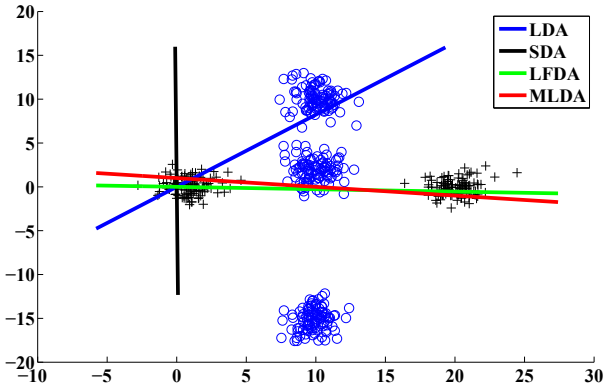
4.1 Experiments on Synthetic Data

The first synthetic data is generated as follows. There are two classes in this dataset. The first class contains two clusters with the first cluster generated from a Gaussian distribution $\mathcal{N}(\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right], \mathbf{I}_2)$ and the second one following $\mathcal{N}(\left[\begin{smallmatrix}20\\0\end{smallmatrix}\right], \mathbf{I}_2)$. The second class has three clusters with them sampled from $\mathcal{N}(\left[\begin{smallmatrix}10\\10\end{smallmatrix}\right], \mathbf{I}_2)$, $\mathcal{N}(\left[\begin{smallmatrix}10\\2\end{smallmatrix}\right], \mathbf{I}_2)$, and $\mathcal{N}(\left[\begin{smallmatrix}10\\-15\end{smallmatrix}\right], \mathbf{I}_2)$ 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 matrices in different clusters. Here different covariance matrices are used for different classes, i.e., the two clusters in the first class use $\left[\begin{smallmatrix}1.5 & 0\\0 & 1\end{smallmatrix}\right]$ as the covariance matrix and all the covariance matrices of the three clusters in the second class equal $\left[\begin{smallmatrix}1 & 0\\0 & 1.5\end{smallmatrix}\right]$.

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 included since it cannot discover the clusters within each class. Based on the data distributions of two synthetic data, the ideal transformation is the horizontal line and we can compare with the transformations produced by the methods in comparison. 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 distribution 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.



(a) Synthetic Data I



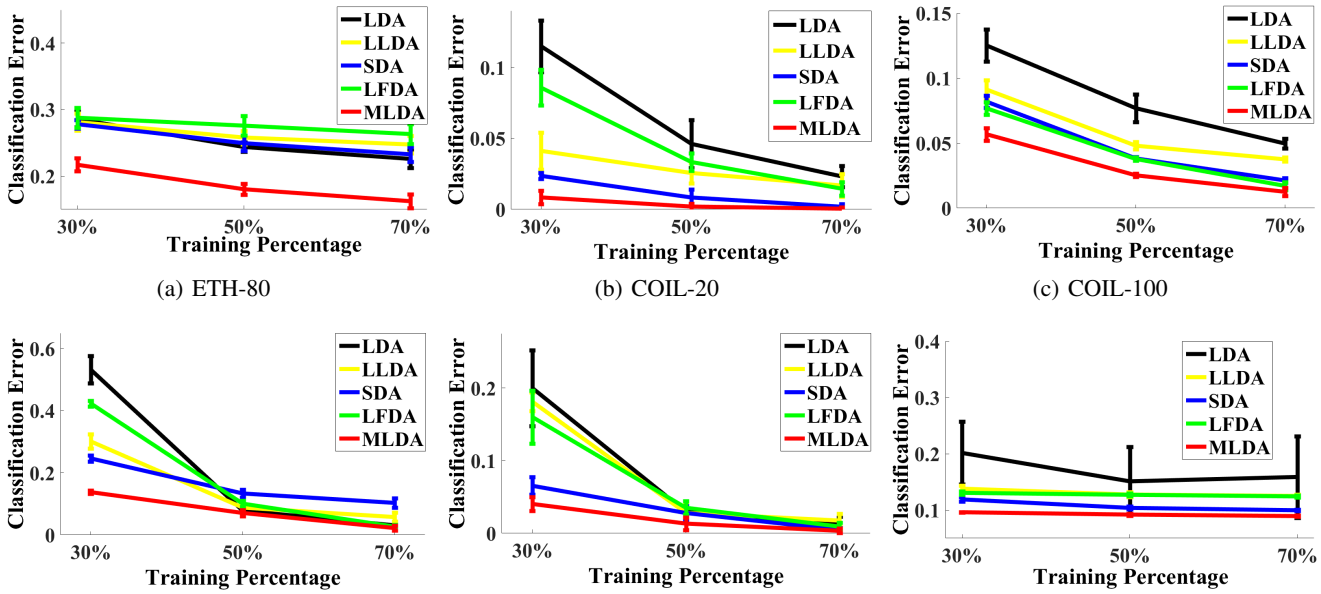
(b) Synthetic Data II

Figure 1: Experiments on two synthetic datasets. Blue and black points are from the first and second classes.

The LFDA and MLDA models have the best results since the learned transformations are almost horizontal. Moreover, one advantage of the proposed MLDA model over the LFDA model is that the MLDA model can identify the clusters in each class via the learned $\mathbf{M}$. According to $\mathbf{M}$, we find the clusters found by the MLDA model are exactly the same as the ground truth on those two synthetic datasets, which demonstrates the cluster-finding ability of the MLDA model. The five respective instance-specific class means corresponding to the five clusters in the first dataset are $\left[\begin{smallmatrix}0.8864\\-0.0386\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}20.3167\\-0.0110\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}9.6911\\10.0980\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}10.0881\\1.8842\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}9.9684\\-15.2609\end{smallmatrix}\right]$, and those in the second dataset are $\left[\begin{smallmatrix}0.6381\\0.0925\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}20.3047\\-0.0310\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}10.0538\\9.8551\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}9.9588\\2.2973\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}9.8401\\-15.4004\end{smallmatrix}\right]$. We can see that the unique instance-specific class means are very close to the means of the underlying Gaussian distributions, which demonstrates the effectiveness of the MLDA model on these two datasets.

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,



(a) ETH-80 (b) COIL-20 (c) COIL-100

(d) AR (e) UMIST (f) MNIST

Figure 2: Test errors on real-world datasets when varying the size of the training set.

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 16×16 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 14×14.

In the proposed MLDA method, γ_{ij} is set to be $\exp\{-\frac{1}{\sigma_{y_i}^2}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2\}$ when $y_i = y_j$ and 0 otherwise, where σ_i denotes the average pairwise Euclidean distance in the ith 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 $\mathbf{M}$, 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

with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

[Bach *et al.*, 2012b] Francis R. Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

[Belhumeur *et al.*, 1997] Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[Gong *et al.*, 2013] Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *Proceedings of the 30th International Conference on Machine Learning*, pages 37–45, 2013.

[Graham and Allinson, 1998] Daniel B. Graham and Nigel Allinson. Characterizing virtual eigensignatures for general purpose face recognition. In *Face Recognition: From Theory to Applications*, volume 163, pages 446–456. Springer, 1998.

[Hastie and Tibshirani, 1996] Trevor Hastie and Robert Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, 58(1):155–176, 1996.

[Hocking *et al.*, 2011] Toby Hocking, Jean-Philippe Vert, Francis R. Bach, and Armand Joulin. Clusterpath: An algorithm for clustering using convex fusion penalties. In *Proceedings of the 28th International Conference on Machine Learning*, pages 745–752, 2011.

[Kim and Kittler, 2005] Tae-Kyun Kim and Josef Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):318–327, 2005.

[Kuo and Landgrebe, 2004] B.-C. Kuo and D. A. Landgrebe. Nonparametric weighted feature extraction for classification. *IEEE Transactions on Geoscience and Remote Sensing*, 42(5):1096–1105, 2004.

[LeCun and Cortes, 1998] Yann LeCun and Corinna Cortes. The MNIST database of handwritten digits, 1998.

[Leibe and Schiele, 2003] Bastian Leibe and Bernt Schiele. Analyzing appearance and contour based methods for object categorization. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 409–415, 2003.

[Li *et al.*, 2009] Zhifeng Li, Dahua Lin, and Xiaoou Tang. Nonparametric discriminant analysis for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):755–761, 2009.

[Martínez and Benavente, 1998] Aleix M. Martínez and Robert Benavente. The AR-face database. Technical Report 24, CVC, 1998.

[Martínez and Zhu, 2005] Aleix M. Martínez and Manli Zhu. Where are linear feature extraction methods applicable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1934–1944, 2005.

[Sugiyama, 2006] Masashi Sugiyama. Local Fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 905–912, 2006.

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

[Yu, 2013] Yaoliang Yu. Better approximation and faster algorithm using the proximal average. In *Advances in Neural Information Processing Systems 26*, pages 458–466, 2013.

[Zhong and Kwok, 2014] Wenliang Zhong and James T. Kwok. Gradient descent with proximal average for non-convex and composite regularization. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 2206–2212, 2014.

[Zhu and Martínez, 2006] Manli Zhu and Aleix M. Martínez. Subclass discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1274–1286, 2006.