Heteroscedastic Max-min Distance Analysis

Bing Su[1], Xiaoqing Ding[1], Changsong Liu[1], Ying Wu[2]
[1]Tsinghua University, Beijing, 100084, China
subingats@gmail.com,{dxq,lcs}@ocrserv.ee.tsinghua.edu.cn
[2] Northwestern University, Evanston, IL, 60208, USA
yingwu@eecs.northwestern.edu

Abstract

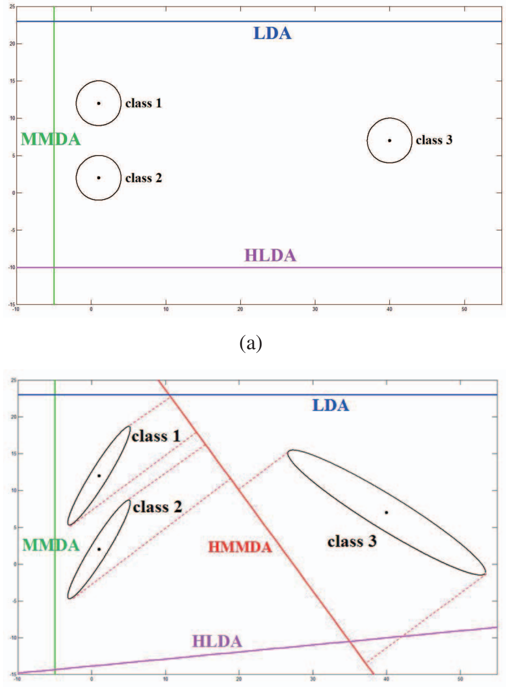
Many discriminant analysis methods such as LDA and HLDA actually maximize the average pairwise distances between classes, which often causes the class separation problem. Max-min distance analysis (MMDA) addresses this problem by maximizing the minimum pairwise distance in the latent subspace, but it is developed under the homoscedastic assumption. This paper proposes Heteroscedastic MMDA (HMMDA) methods that explore the discriminative information in the difference of intra-class scatters for dimensionality reduction. WHMMDA maximizes the minimal pairwise Chenoff distance in the whitened space. OHMMDA incorporates this objective and the minimization of class compactness into a trace quotient formulation and imposes an orthogonal constraint to the final transformation, which can be solved by a bisection search algorithm. Two variants of OHMMDA are further proposed to encode the margin information. Experiments on several UCI Machine Learning datasets and the Yale Face database demonstrate the effectiveness of the proposed HMMDA methods.

1. Introduction

Dimensionality reduction (DR) has become a ubiquitous procedure in many pattern recognition and machine learning applications. Among a number of DR approaches, a class of linear supervised techniques referred to as *discriminant analysis (DA)* has received a lot of attention, which maximizes the separability of classes. Various criteria have been proposed based on different definitions of separability in the literature. Linear discriminant analysis (LDA) is probably the most widely used method, which was first proposed for two-class problems by Fisher in [6] and extended to general multi-class problems by Rao in [16]. LDA optimizes the so-called Fisher criterion by maximizing the ratio of between-class scatter over within-class scatter under the homoscedastic Gaussian assumption. The Fisher criterion is extended to handle tensor data in [27] and sequence data in [19].

However, in practice, the distributions of classes are often non-Gaussian, and the covariances of different classes are not equal. These situations have been studied in extensive literatures. Subclass discriminant analysis [24, 30] considers the general distribution types by dividing each class into several subclasses each described by one Gaussian distribution. Marginal Fisher Analysis [25] uses marginal and neighboring points to construct inter-class separability and intra-class compactness. Locality preserving property is combined with LDA in [10] to handle multimodal data. The optimal class representation is determined in [20] to replace the mean class vector. The Bayes error is directly minimized in [7]. The heteroscedastic Gaussian model parameters are jointly estimated with DR in the maximum-likelihood framework in [11]. Heteroscedastic LDA (HLDA) [12] extends LDA to heteroscedastic cases by utilizing the Chernoff criterion instead of the Fisher criterion, where the Chernoff distance is employed to generalize the between-class scatter. A theoretical analysis of HLDA is presented in [15]. In [17], the Chernoff distance is maximized in the transformed space by a gradient-based algorithm, and its performance compared with LDA and HLDA is evaluated in [1].

These methods actually maximize the average of all pairwise distances between classes due to the definition of between-class scatter. This will cause the so-called "class separation" problem [13]. Specifically, these methods tend to pay close attention to classes with larger distances, but ignore those with smaller distances, resulting in the overlap of "neighbouring" classes in the projected subspace on the basis of a specific distance measure. An example to illustrate the class separation problem is shown in Fig. 1(a), where class 1 and class 2 locate closely to each other while class 3 is far away from them. All classes have the same unit covariance. (In this case, HLDA degenerates into LDA.) The average of pairwise distances between classes is maximized when the dominant large distances between class 3 and the remainders are preserved. Thus the direction that separates class 3 from the remaining classes as much as possible is determined as the projection axis of LDA and HLDA, resulting in a complete confusion of class 1 and class 2.



(a)

(b)

Figure 1. (a) An illustration of the class separation problem. (b) An illustration of the influence of heteroscedastic data. The means of the three classes in (b) are the same as those in (a), while the covariances are different. MMDA can address the class separation problem as shown in (a) but fails in heteroscedastic case as shown in (b). The proposed HMMDA can successfully separate the three classes in (b).

To address this problem, some methods such as aPAC [13] and weighted Fisher criteria [28] impose larger weights to similar class pairs, but the optimal weighting function is often ad-hoc and thus difficult to choose. Alternatively, several *max-min distance analysis (MMDA)* approaches have been proposed by assuming homoscedastic Gaussian distributions in [3, 29, 26], which maximize the minimum pairwise separability between classes in the latent subspace. The three classes are better separated in Fig. 1(a) when they are projected to the subspace (the vertical line) determined by MMDA. However, in practice classes usually do not satisfy homoscedastic Gaussian assumption, so difference of class means can not fully reflect the gap between classes. Existing MMDA algorithms simply try to separate the nearest class means in the latent subspace as much as possible without considering the difference of class covariances. An example is shown in Fig. 1(b), where the means of the three classes remain unchanged with those in Fig. 1(a), but the covariances differ in different classes. The sum of these covariance is designed to be multiples of unit matrix, thus whitening has no effect on the final projection. Hence the resulted projections of LDA and MMDA also remain unchanged. It can be seen that LDA and HLDA still suffer from the class separation problem. For MMDA, due to the impact of heteroscedasticity, the three classes become almost inseparable after projection.

In this paper, we propose *heteroscedastic max-min distance analysis (HMMDA)* methods for DR to exploit discriminative information presented in the difference of class covariances: *whitened HMMDA (WHMMDA)* and *orthogonal HMMDA (OHMMDA)* along with its two variants. In respect of tackling within-class scatter, WHMMDA performs a whitening preprocessing thus can be considered as a direct extension of MMDA, while OHMMDA explicitly incorporates the minimization of the closeness of classes into the objective function and applies an orthogonal constraint. Variants of HMMDA further capture the margin information by utilizing the pairwise support scatters when constructing the second-order separability. All these methods can effectively deal with the class separation problem in heteroscedastic situations. In the previous example, the slant (red) line represents the projection axis of the HMMDA methods[1]. It is clear that the three classes can be completely separated when projected to this direction.

The rest of this paper is organized as follows: Section 2 and 3 presents the proposed WHMMDA and OHMMDA together with its two variants, respectively; Section 4 reports experiments and results; Section 5 draws the conclusions.

2. Whitened HMMDA

In this section, we generalize MMDA to WHMMDA following a two-step framework. That is, a whitening preprocessing is first implemented, separability is then maximized in the whitened space. Assume that d-dimensional data belong to C classes. Let $\mathbf{m}_i$, $\mathbf{\Sigma}_i$ and p_i denote the mean, covariance and prior probability of class i, respectively, let $\mathbf{S}_w$ be the within-class scatter defined as: $\mathbf{S}_w = \sum_{i=1}^{C} p_i \mathbf{\Sigma}_i$, then the whitening transformation is determined as: $\mathbf{W}_1 = \mathbf{S}_w{}^{-1/2} \in \mathbb{R}^{d \times d}$.

Although $\mathbf{S}_w$ is transformed into an identity matrix, the covariances of different classes remain different. Hence difference between class means cannot fully reflect the separability of classes. Under the assumption that each class obeys a Gaussian distribution with different mean and covariance from other classes, Chernoff distance d_{Cij} between classes i and j takes the discriminative information within difference of covariances into account and thus has the ability to

[1] They produce the same projection in this case.

better describe the entanglement of classes.

$$\begin{aligned} d_{Cij} &= (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T \hat{\mathbf{\Sigma}}_{ij}^{-1} (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j) \\ &+ \tfrac{1}{\alpha_{ij}(1-\alpha_{ij})} \log \frac{\left|\hat{\mathbf{\Sigma}}_{ij}\right|}{\left|\hat{\mathbf{\Sigma}}_i\right|^{\alpha_{ij}} \left|\hat{\mathbf{\Sigma}}_j\right|^{1-\alpha_{ij}}} \end{aligned} \tag{1}$$

where $\alpha_{ij} = p_i/(p_i + p_j)$, $\hat{\mathbf{m}}_i = {\mathbf{W}_1}^T \mathbf{m}_i$, $\hat{\mathbf{\Sigma}}_i = {\mathbf{W}_1}^T \mathbf{\Sigma}_i \mathbf{W}_1$ are the mean and variance of class i in the whitened space, and $\hat{\mathbf{\Sigma}}_{ij} = \alpha_{ij}\hat{\mathbf{\Sigma}}_i + (1-\alpha_{ij})\hat{\mathbf{\Sigma}}_j$. It has been shown in [12] that d_{Cij} can be obtained as the trace of a positive semi-definite matrix $\mathbf{S}_{Cij}$ ($d_{Cij} = tr(\mathbf{S}_{Cij})$).

$$\begin{aligned} &\mathbf{S}_{Cij} = \hat{\mathbf{\Sigma}}_{ij}^{-1/2} (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T \hat{\mathbf{\Sigma}}_{ij}^{-1/2} \\ &+ \tfrac{1}{\alpha_{ij}(1-\alpha_{ij})} (\log \hat{\mathbf{\Sigma}}_{ij} - \alpha_{ij} \log \hat{\mathbf{\Sigma}}_i - (1-\alpha_{ij}) \log \hat{\mathbf{\Sigma}}_j) \end{aligned} \tag{2}$$

The objective of WHMMDA is to maximize the minimal pairwise Chernoff distance in the latent subspace:

$$\begin{aligned} &\max_{\mathbf{W}_2} \min_{1\le i<j\le C} (p_i p_j)^{-1} tr({\mathbf{W}_2}^T \mathbf{S}_{Cij} \mathbf{W}_2) \\ &= \max_{\mathbf{W}_2} \min_{1\le i<j\le C} tr({\mathbf{W}_2}^T \tilde{\mathbf{S}}_{Cij} \mathbf{W}_2) \\ &s.t. \ {\mathbf{W}_2}^T \mathbf{W}_2 {=} \mathbf{I}_{d'} \end{aligned} \tag{3}$$

where $\tilde{\mathbf{S}}_{Cij} = (p_i p_j)^{-1} \mathbf{S}_{Cij}$ is the scatter between classes i and j weighted by the inverse factor of prior probabilities.

We use the method proposed in [3] to optimize (3). Specifically, by introducing two new variables t and $\mathbf{Z} = \mathbf{W}_2 {\mathbf{W}_2}^T$, and relaxing the feasible set of $\mathbf{Z}$: $\left\{\mathbf{Z} | \mathbf{Z} = \mathbf{W}_2{\mathbf{W}_2}^T, {\mathbf{W}_2}^T\mathbf{W}_2 {=} \mathbf{I}_{d'}\right\}$ to its convex hull $\{\mathbf{Z} |\, tr(\mathbf{Z}) = d', \mathbf{0} \le \mathbf{Z} \le \mathbf{I}_d\}$ [14], the non-convex problem (3) is transformed into the following Semi-Definite Programming (SDP) problem:

$$\begin{aligned} &\max \ t \\ &s.t. \ tr(\tilde{\mathbf{S}}_{Cij}\mathbf{Z}) \ge t, \ 1 \le i < j \le C \\ &\quad \mathbf{0} \le \mathbf{Z} \le \mathbf{I}_d \\ &\quad tr(\mathbf{Z}) = d' \end{aligned} \tag{4}$$

This relaxation is called the global relaxation in [3]. Let $\mathbf{Z}_{glo}$ denotes the optimal solution of (4).

The relaxation is loose for only extreme points of the convex hull are contained in the original feasible set of $\mathbf{Z}$. It is proved in [3] that the convex hull is equivalent to the original feasible set if an additional condition $\det(\mathbf{Z} + \xi\mathbf{I}_d) = (1+\xi)^{d'}\xi^{d-d'}, \forall \xi > 0$ is added. Hence by using $\mathbf{Z}_{glo}$ as the initial $\mathbf{Z}_0$, a more precise solution can be obtained by iteratively solving the following sequential SDP relaxation using the solution of the previous iteration as $\mathbf{Z}_0$:

$$\begin{aligned} &\max \ t \\ &s.t. \ tr(\tilde{\mathbf{S}}_{Cij}\mathbf{Z}) \ge t, \ 1 \le i < j \le C \\ &\quad \mathbf{0} \le \mathbf{Z} \le \mathbf{I}_d \\ &\quad tr(\mathbf{Z}) = d' \\ &\quad tr((\mathbf{Z}_0 + \mathbf{I}_d)^{-1}\mathbf{Z}) \le (1+\lambda) tr((\mathbf{Z}_0 + \mathbf{I}_d)^{-1}\mathbf{Z}_0) \\ &\quad tr((\mathbf{Z}_0 + \mathbf{I}_d)^{-1}\mathbf{Z}) \ge (1-\lambda) tr((\mathbf{Z}_0 + \mathbf{I}_d)^{-1}\mathbf{Z}_0) \end{aligned} \tag{5}$$

where λ is a parameter decreasing from 10^{-2} to 10^{-6} for 20 iterations. We use SDPT3 [21, 22] to solve problems (4) and (5). Let $\mathbf{Z}_{loc}$ denotes the final solution of (5). $\mathbf{W}_2$ is recovered by decomposing $\mathbf{Z}_{loc}$ and using the eigenvectors corresponding to the d' largest eigenvalues as columns. The final transformation matrix of WHMMDA is: $\mathbf{W}_{WH} = \mathbf{W}_1\mathbf{W}_2$. Different from LDA, the maximal dimensionality of reduced subspace generated by WHMMDA is not limited by $C-1$, because the transformation is obtained via solving SDP rather than algebraic eigen-decomposition.

3. Orthogonal HMMDA

3.1. Formulation and solution of OHMMDA

The transformation matrix resulted by WHMMDA is not orthogonal, thus the features produced may be correlated. It has been shown that in some applications [5, 9, 18] orthogonal basis transformations may have better discriminating power. Furthermore, whitening the within-class scatter cannot directly strengthen the closeness of data in the same class. In this section, we develop an algorithm in the original space that can explicitly obtain a joint optimization of maximizing the minimum pairwise Chernoff distance between classes and minimizing the diffusion within the same classes with the orthogonal constraint in a trace quotient fashion, which we call *orthogonal HMMDA (OHMMDA)*.

We first compute the pairwise between-class scatter matrix $\mathbf{S}_{OCij}$ directly in the original space:

$$\begin{aligned} &\mathbf{S}_{Cij} = {\mathbf{\Sigma}_{ij}}^{-1/2} (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T {\mathbf{\Sigma}_{ij}}^{-1/2} \\ &+ \tfrac{1}{\alpha_{ij}(1-\alpha_{ij})} (\log \mathbf{\Sigma}_{ij} - \alpha_{ij} \log \mathbf{\Sigma}_i - (1-\alpha_{ij}) \log \mathbf{\Sigma}_j) \end{aligned} \tag{6}$$

The objective of O-HMMDA can then be expressed as the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{W}} \frac{\min\limits_{1\le i<j\le C} (p_i p_j)^{-1} tr(\mathbf{W}^T \mathbf{S}_{OCij} \mathbf{W})}{tr(\mathbf{W}^T \mathbf{S}_w \mathbf{W})} \\ &= \max_{\mathbf{W}} \min_{1\le i<j\le C} \frac{tr(\mathbf{W}^T \tilde{\mathbf{S}}_{OCij} \mathbf{W})}{tr(\mathbf{W}^T \mathbf{S}_w \mathbf{W})} \\ &s.t. \ \mathbf{W}^T\mathbf{W} = \mathbf{I}_{d'} \end{aligned} \tag{7}$$

Similarly, $\tilde{\mathbf{S}}_{OCij} = (p_i p_j)^{-1}\mathbf{S}_{OCij}$.

Problem (7) is equivalent to the following problem by introducing an auxiliary variables t:

$$\begin{aligned} &\max \ \ t \\ s.t.\ \ &tr(\mathbf{W}^T\tilde{\mathbf{S}}_{OCij}\mathbf{W}) \geq t \cdot tr(\mathbf{W}^T\mathbf{S}_w\mathbf{W}), \\ &\qquad 1 \leq i < j \leq C \\ &\mathbf{W}^T\mathbf{W} = \mathbf{I}_{d'} \end{aligned} \tag{8}$$

By introducing variable $\mathbf{Z} = \mathbf{W}\mathbf{W}^T$, and relaxing the feasible set of $\mathbf{Z}$ to its convex hull again as in WHMMDA, problem (8) is transformed into:

$$\begin{aligned} &\max \ \ t \\ s.t.\ \ &tr((\tilde{\mathbf{S}}_{OCij} - t \cdot \mathbf{S}_w)\mathbf{Z}) \geq 0, \\ &\qquad 1 \leq i < j \leq C \\ &\mathbf{0} \leq \mathbf{Z} \leq \mathbf{I}_d \\ &tr(\mathbf{Z}) = d' \end{aligned} \tag{9}$$

This problem can be solved by the bisection search strategy proposed in [18]. Assume that t^* is the unknown optimal value of (9). For a given $\hat{t} \in \mathbb{R}$, if the following SDP problem (10) is feasible, then it means that $\hat{t}$ is a feasible solution of (9), thus the optimal solution t^* of (9) should satisfy: $t^* \geq \hat{t}$ since t^* is the maximum of all feasible t. Otherwise, if (10) is infeasible, then it means $\hat{t}$ exceeds the maximum value of all feasible t and thus $t^* < \hat{t}$.

$$\begin{aligned} &find\ \ \mathbf{Z} \\ s.t.\ \ &tr((\tilde{\mathbf{S}}_{OCij} - \hat{t} \cdot \mathbf{S}_w)\mathbf{Z}) \geq 0, \\ &\qquad 1 \leq i < j \leq C \\ &\mathbf{0} \leq \mathbf{Z} \leq \mathbf{I}_d \\ &tr(\mathbf{Z}) = d' \end{aligned} \tag{10}$$

Based on this inference, by determining a lower bound t_L and an upper bound t_U of t^*, a recursive algorithm can be developed to gradually narrow down the feasible region until the position of t^* is located. Specifically, we first initialize the value of t as: $t = (t_L + t_U)/2$, then we check if (10) is feasible. If it is feasible, then we update the lower bound as: $t_L = t$; otherwise we update the upper bound as: $t_U = t$. Thereupon the value of $t = (t_L + t_U)/2$ is changed and (10) is checked again. This process is repeated until the difference of t_U and t_L is below a preset threshold ξ: $t_U - t_L \leq \xi$.

The initial lower bound t_L and upper bound t_U can be estimated following the theorem from [14] and [18]:

Theorem. If $\mathbf{S} \in \mathbb{R}^{d\times d}$ is a symmetric matrix and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are the eigenvalues of $\mathbf{S}$ sorted in descending order. Then for any $\mathbf{W} \in \mathbb{R}^{d\times d'}$ that satisfies the condition: $\mathbf{W}^T\mathbf{W} = \mathbf{I}_{d'}$, we have: $\max\limits_{\mathbf{W}}\ tr(\mathbf{W}^T\mathbf{S}\mathbf{W}) = \sum\limits_{i=1}^{d'} \lambda_i$ and $\min\limits_{\mathbf{W}}\ tr(\mathbf{W}^T\mathbf{S}\mathbf{W}) = \sum\limits_{i=d-d'+1}^{d} \lambda_i$.

$\tilde{\mathbf{S}}_{OCij}, 1 \leq i < j \leq C$ and $\mathbf{S}_w$ are all symmetric and positive semi-definite matrixes, then their corresponding sorted eigenvalues satisfy: $\lambda_{OCij1} \geq \lambda_{OCij2} \geq \cdots \geq \lambda_{OCijd} \geq 0, 1 \leq i < j \leq C$ and $\lambda_{w1} \geq \lambda_{w2} \geq \cdots \geq \lambda_{wd} \geq 0$. Then by amplifying and magnifying the objective of (7), the upper and lower bound of t can be estimated as:

$$t_U = \max_{1\leq i<j\leq C} \frac{\sum\limits_{k=1}^{d'} \lambda_{OCijk}}{\sum\limits_{k=d-d'+1}^{d} \lambda_{wk}} \tag{11}$$

$$t_L = \min_{1\leq i<j\leq C} \frac{\sum\limits_{k=d-d'+1}^{d} \lambda_{OCijk}}{\sum\limits_{k=1}^{d'} \lambda_{wk}} \tag{12}$$

respectively. When the smallest eigenvalues of $\mathbf{S}_w$ are all zeros, the denominator of (11) is also zero. This happens when there are too few training samples and $\mathbf{S}_w$ is not full rank. This problem can be solved by either using PCA to remove the null space of $\mathbf{S}_w$ [18] or by regularization.

The bisection algorithm is guaranteed to converge to the global optimum within $\log_2((t_U - t_L)/\xi)$ iterations [18]. We also use SDPT3 [21, 22] to solve problem (10). After obtaining the optimal solution of $\mathbf{Z}$, the final transformation $\mathbf{W}$ is constructed by eigen-decomposing $\mathbf{Z}$ and using the eigenvectors corresponding to the d' largest eigenvalues as columns as before.

3.2. Two variants of OHMMDA

Since the ratio of the minimum pairwise between-class scatter $\mathbf{S}_{OCij}$ over the within-class scatter $\mathbf{S}_w$ is explicitly maximized in a trace quotient fashion (7) instead of first whitening $\mathbf{S}_w$, both $\mathbf{S}_{OCij}$ and $\mathbf{S}_w$ are not necessarily restricted to those used in (6) and (7). Different formulations of $\mathbf{S}_{OCij}$ and $\mathbf{S}_w$ represent the dissimilarity between classes and the compactness of classes in different aspects. Generally, using the sum of variances as the within-class scatter tends to aggregate all the samples, which is indeed a strict requirement. An alternative choice is to force adjacent samples to be close. Similarly, the pairwise between-class scatter can be built using only those neighboring samples distributed in the border of the two classes to reflect their separability. Encoding such margin information is a commonly used strategy [25, 18]. In [18], the intra-class scatter for class i is defined as:

$$\mathbf{\Sigma}_i^m = \sum_{\mathbf{x}_p \in class\, i} \sum_{\mathbf{x}_q \in N_k^i(\mathbf{x}_p)} (\mathbf{x}_p - \mathbf{x}_q)(\mathbf{x}_p - \mathbf{x}_q)^T \tag{13}$$

where $N_k^i(\mathbf{x}_p)$ denotes the set of k-nearest neighbors in class i. That is, for each sample in class i, only its k nearest neighbors in the same class are used to construct $\mathbf{\Sigma}_i^m$.

The value of k is set to be 3 in our experiments. The overall intra-class scatter of margin is the weighted sum of the intra-class scatter of all classes:

$$\mathbf{S}_w^m = \sum_{i=1}^{C} p_i \mathbf{\Sigma}_i^m \tag{14}$$

The pairwise inter-class scatter of margin $\mathbf{S}_{Cij}^m$ is accordingly defined by replacing $\mathbf{\Sigma}_i$ with $\mathbf{\Sigma}_i^m$ in (6), which considers the difference between locally similarity within classes at the second-order level.

$$\begin{aligned}\mathbf{S}_{Cij}^m &= (\mathbf{\Sigma}_{ij}^m)^{-1/2}(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T(\mathbf{\Sigma}_{ij}^m)^{-1/2} \\ &+ \tfrac{1}{\alpha_{ij}(1-\alpha_{ij})}(\log \mathbf{\Sigma}_{ij}^m - \alpha_{ij}\log \mathbf{\Sigma}_i^m - (1-\alpha_{ij})\log \mathbf{\Sigma}_j^m)\end{aligned} \tag{15}$$

where $\mathbf{\Sigma}_{ij}^m = \alpha_{ij}\mathbf{\Sigma}_i^m + (1-\alpha_{ij})\mathbf{\Sigma}_j^m$.

The pairwise between-class scatter can also incorporate such margin information by exploiting those "support points" distributed in the boundary region of two classes, which we call the support pairwise inter-class scatter and denote it by $\mathbf{S}_{Cij}^s$. We compute all the pairwise Euclidean distances between the N_i samples in class i and the N_j samples in class j, and collect the sample pairs from different classes that belong to the k' nearest pairs into a set Ψ_{ij}, $\Psi_{ij} = \{(\mathbf{x}_p, \mathbf{x}_q)|\mathbf{x}_p \in class\ i, \mathbf{x}_q \in class\ j, (\mathbf{x}_p, \mathbf{x}_q) \in P_{k'}(i,j)\}$, $i \neq j$, where $P_{k'}(i,j)$ denotes the set of k' nearest sample pairs between class i and j. Instead of purely using the difference between means of the two classes at the first-order level, we utilize these "support point pairs" to represent the first-order separability and replace the term $(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T$ in $\mathbf{S}_{Cij}^m$ by $\sum\limits_{(\mathbf{x}_p,\mathbf{x}_q)\in\Psi_{ij}} (\mathbf{x}_p - \mathbf{x}_q)(\mathbf{x}_p - \mathbf{x}_q)^T$ to construct $\mathbf{S}_{Cij}^s$.

In this paper, we refer to the original OHMMDA algorithm that uses $\mathbf{S}_{OCij}$ and $\mathbf{S}_w$ as pairwise between-class and within-class scatters as OHMMDA, and use mar-OHMMDA and sup-OHMMDA to denote the algorithms that use $\mathbf{S}_{Cij}^m$, $\mathbf{S}_w^m$ and $\mathbf{S}_{Cij}^s$, $\mathbf{S}_w^m$ as pairwise inter-class and intra-class scatters, respectively. Similar to WHMMDA, the dimensionality reduced by OHMMDA, mar-OHMMDA or sup-OHMMDA is also not limited by $C-1$.

3.3. Complexity

The most time-consuming part is to solve SDPs (5) and (10). The number of variables and constraints are $O(d^2)$ and $O(d + C^2)$, thus the worst-case complexity of interior-point method is $O(d^4(d + C^2)^2)$. WHMMDA takes a fixed number of iterations, and OHMMDAs take $log_2((t_U - t_L)/e)$ iterations to converge. The space complexity for all variants is $O(C^2d^2)$. Any scalable distributed technique such as alternating direction method of multipliers can be used to cope with large scale SDPs.

4. Experiments and results

4.1. Experiments on UCI datasets

Datasets. We evaluate the performances of the proposed HMMDA algorithms on five real-world datasets from the UCI Machine Learning Repository [2]. These datasets are briefly described as follows: (1). The "Breast Cancer Wisconsin (Diagnostic)" (BCWD) dataset contains 569 instances from two classes indicating whether the diagnosis is malignant or benign. Each instance is represented by 32 features extracted form a digitized image of a fine needle aspirate of a breast mass. (2). The "Iris" dataset consists of 150 instances with a dimension of 4 from three classes of 50 instances each, where each class refers to a type of iris plant. (3). The "Glass Identification" dataset consists of 214 instances with a dimension of 10 from six glass classes. (4). The "Landsat Satellite" dataset has six decision classes with a dimension of 36 for each example. There are 4,435 examples in the training set and 2,000 examples in the test set. We merged the training set and test set into a single set. (5). The "Multiple Features" dataset consists of features of ten handwritten numeral classes. There are 2,000 instances in total with 200 instances for each class. Every instance is represented in terms of six feature sets. We use the set that contains 47 Zernike moments in our experiments.

Experimental setup. For each dataset, we divide the whole set into five subsets, of which four are used as training set and the left one is used for test. We evaluate the effectiveness of the proposed WHMMDA, OHMMDA, mar-OHMMDA and sup-OHMMDA methods on such fivefold cross validation. Four widely used discriminant analysis approaches, including LDA, HLDA, aPAC and MMDA, are also evaluated for comparison. For all these datasets, the original dimensionality d is reduced to all possible values. For LDA and aPAC, the maximum dimensionality of the selected subspace is limited to $C-1$; while other methods are able to preserve a dimensionality higher than $C-1$, so we reduced the dimensionality of data to from 1 to $d-1$. Three classifiers including the nearest neighbor classifier (1-NN), the nearest mean classifier (NM) and the quadratic classifier are used to perform classification in the low-dimensional subspaces. The quadratic classifier uses the quadratic discriminant function (QDF) as the decision function:

$$\mathbf{x} \in \arg \min_{i=1,\cdots,C} (\mathbf{x} - \mathbf{m}_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \mathbf{m}_i) + \log|\mathbf{\Sigma}_i|$$

Results and analysis. Table 1(a) to 1(e) show the optimal average classification error rates among all possible reduced dimensionalities together with the corresponding standard deviations and dimensions of different DR methods on the six datasets, respectively.

For the BCWD dataset, there are only two classes, in this case the max-min criterion is equivalent to the max-average criterion. Hence WHMMDA boils down to HLDA. They

Table 1. Optimal average classification error rates by the three classifiers of different DR methods on the six datasets. The two values in each bracket are the corresponding dimensions and standard deviations. The best result achieved in each dataset among all methods by each classifier is shown in bold.

(a) Results on the BCWD dataset

Classifier	LDA	HLDA	aPAC	MMDA	WHMMDA	OHMMDA	mar- OHMMDA	sup- OHMMDA
1-NN	0.0473 (1,0.0177)	**0.0298** (2,0.0151)	0.0473 (1,0.0177)	0.0438 (2,0.0180)	**0.0298** (2,0.0151)	0.0651 (29,0.0075)	0.0704 (29,0.0204)	0.0474 (20,0.0213)
NM	0.0421 (1,0.0101)	**0.0404** (16,0.0118)	0.0421 (1,0.0101)	0.0421 (1,0.0101)	**0.0404** (16,0.0118)	0.0897 (28,0.0132)	0.1073 (21,0.0289)	0.1004 (20,0.0378)
QDF	0.0316 (1,0.0089)	**0.0298** (1,0.0088)	0.0316 (1,0.0089)	0.0316 (1,0.0089)	**0.0298** (1,0.0088)	0.0405 (26,0.0206)	0.0387 (21,0.0183)	0.0405 (23,0.0183)

(b) Results on the Iris dataset

Classifier	LDA	HLDA	aPAC	MMDA	WHMMDA	OHMMDA	mar- OHMMDA	sup- OHMMDA
1-NN	0.0333 (1,0.0365)	**0.0267** (2,0.0249)	0.0400 (1,0.0327)	0.0400 (3,0.0249)	0.0333 (2,0.0211)	0.0333 (3,0.0211)	0.0600 (3,0.0389)	0.0600 (3,0.0249)
NM	**0.0200** (1,0.0163)	**0.0200** (1,0.0163)	**0.0200** (1,0.0163)	**0.0200** (1,0.0267)	**0.0200** (1,0.0267)	0.0267 (3,0.0133)	0.0533 (1,0.0452)	0.0400 (2,0.0327)
QDF	0.0200 (1,0.0163)	0.0200 (1,0.0163)	0.0200 (1,0.0163)	0.0267 (1,0.0249)	**0.0133** (2,0.0163)	0.0400 (3,0.0133)	0.0333 (3,0.0516)	0.0333 (3,0.0211)

(c) Results on the Glass dataset

Classifier	LDA	HLDA	aPAC	MMDA	WHMMDA	OHMMDA	mar- OHMMDA	sup- OHMMDA
1-NN	0.3764 (5,0.0617)	0.3398 (8,0.0718)	0.3764 (5,0.0617)	0.3168 (8,0.0596)	0.3302 (6,0.0520)	0.3154 (8,0.0670)	**0.2793** (8,0.0722)	0.2877 (8,0.0742)
NM	0.4147 (5,0.0597)	0.4286 (8,0.0432)	0.4147 (5,0.0597)	0.4101 (8,0.0591)	**0.4087** (8,0.0641)	0.4340 (7,0.0410)	0.5651 (8,0.0462)	0.5740 (8,0.0518)
QDF	0.4603 (2,0.0831)	0.4302 (5,0.0424)	0.4240 (1,0.0676)	**0.3729** (2,0.0430)	0.4201 (5,0.0294)	0.3920 (5,0.0587)	0.3922 (5,0.0258)	0.3969 (4,0.0848)

(d) Results on the Landset dataset

Classifier	LDA	HLDA	aPAC	MMDA	WHMMDA	OHMMDA	mar- OHMMDA	sup- OHMMDA
1-NN	0.1442 (5,0.0036)	0.1343 (18,0.0082)	0.1442 (5,0.0036)	0.1442 (5,0.0033)	0.1442 (10,0.0159)	**0.0892** (18,0.0070)	0.0908 (33,0.0080)	0.0928 (35,0.0046)
NM	0.1573 (5,0.0044)	**0.1556** (23,0.0050)	0.1573 (5,0.0044)	0.1573 (6,0.0044)	0.1568 (35,0.0030)	0.2003 (35,0.0108)	0.1893 (5,0.0108)	0.2152 (16,0.0138)
QDF	0.1419 (5,0.0029)	0.1419 (25,0.0065)	0.1417 (4,0.0042)	**0.1358** (11,0.0070)	0.1434 (27,0.0041)	0.1413 (21,0.0034)	0.1378 (23,0.0081)	0.1444 (35,0.0052)

(e) Results on the Multi-feature Digit dataset

Classifier	LDA	HLDA	aPAC	MMDA	WHMMDA	OHMMDA	mar- OHMMDA	sup- OHMMDA
1-NN	0.2050 (9,0.0134)	0.1895 (22,0.0100)	0.2050 (9,0.0134)	0.1850 (20,0.0106)	0.1990 (41,0.0099)	0.1870 (42,0.0108)	0.1900 (46,0.0143)	**0.1845** (46,0.0144)
NM	0.1800 (9,0.0101)	**0.1770** (31,0.0107)	0.1800 (9,0.0101)	0.1800 (9,0.0101)	0.1825 (46,0.0104)	0.2370 (37,0.0196)	0.2840 (46,0.0218)	0.2440 (27,0.0197)
QDF	0.1730 (8,0.0114)	0.1745 (31,0.0149)	0.1670 (7,0.0149)	0.1690 (13,0.0129)	0.1910 (45,0.0124)	0.1680 (28,0.0112)	0.1640 (23,0.0112)	**0.1590** (20,0.0142)

achieve the best results among all DR methods for all classifiers, owing to the exploitation of second-order Chernoff distance. For Other datasets which contain more than two classes, the performances of HMMDA algorithms are quite satisfactory. This is reflected in several respects. First, the best results among all DR methods and all classifiers on all these datasets are always achieved by one variant of HMMDA algorithms. Second, for both the two non-linear classifiers (1-NN and QDF), HMMDA variants obtain the best results or comparable results with the best ones on maximum number of datasets. Third, for all classifiers, on most datasets where the results of HMMDA algorithms are not the best ones, they are also not far from the best results.

It can be observed that non-linear classifiers perform better than the NM classifier on nearly all datasets. This means that the distributions of classes are generally not homoscedastic Gaussian distributions even not Gaussian, which is the common case in real world applications. HMMDA algorithms always perform better with lower error rates and lower or comparable standard deviations than competitive DR methods by non-linear classifiers. This indicates that through successfully combining the advantages of max-min criterion and the Chernoff criterion, HMMDA algorithms are able to provide a constrained subspace in which the closest two classes are projected as far as possible hence all classes are better separated. In this sense "far" means that two classes have a larger Chenoff distance thus the discriminative information contained in the differences of secondary moments under different definitions can be captured. On the other hand, the performances of HMMDA algorithms when using the NM classifier are inferior to those using non-linear classifiers. The reason could be that data after HMMDA transformation are not linearly separable even they may have a better nonlinear separability.

For the two non-linear classifiers, WHMMDA generally performs the best on the BCWD dataset and the Iris dataset. The scales of the two datasets are quite small, hence less information is presented. WHMMDA without the orthogonal constraint is able to project data to subspaces with few less relevant dimensionalities that have good discriminative powers. For the other three datasets of the ascending scale, OHMMDA, mar-OHMMDA and sup-OHMMDA generally rank inside the top three. The explicit trace quotient formulation together with the orthogonal constraint and the effective solution make OHMMDA varietal algorithms fully functional.

Among the proposed HMMDA methods, no one variant can outperform all other variants on all datasets. The reason is that different variants characterize separability from different perspectives, and thus which variant is more effective highly depends on the data. As the scales and domains of these benchmark datasets vary a lot, no variant can win in all cases. When there are only a few training samples, mar-OHMMDA is preferred since the intra-class scatter of margin (15) can be better utilized to reflect the margin information. If the data distribution is highly non-Gaussian, sup-OHMMDA is preferred as support inter-class scatter is able to capture the marginal separability between classes. If sufficient samples are available and the distribution is not very singular, the original OHMMDA is suitable, since the estimations of class variances are accurate and can provide the overall dispersion of classes. Nevertheless, the experiments show that any of the variants is indeed able to achieve superior results.

4.2. Experiments on the Yale Face database

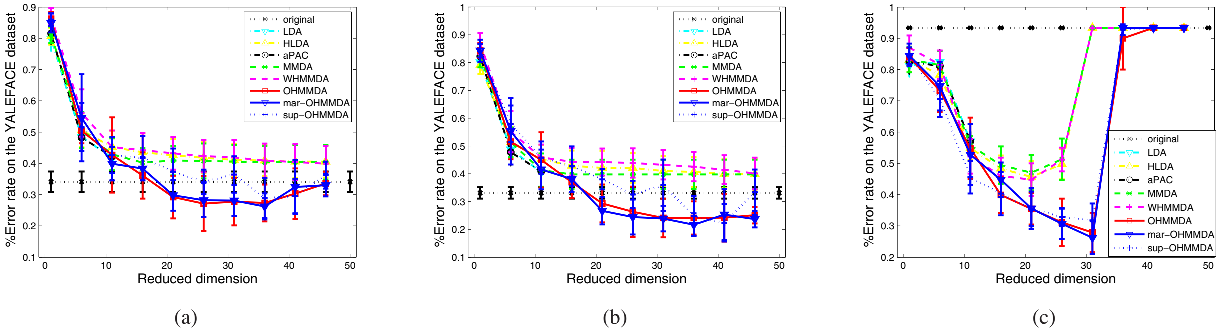
We have experimentally demonstrated that HMMDA algorithms perform quite well on real world datasets where both whether the distribution of different classes is imbalanced and whether the distribution of each class is Gaussian are unknown. In this subsection, we further evaluate the effectiveness of the proposed HMMDA algorithms in face recognition application on the Yale Face database [23].

Database. The original Yale Face database contains 165 gray-scale images in GIF format of 15 individuals. There are 11 images per individual, one per different facial expression, lighting condition, with/without glasses or other configuration. In [8, 4], these face images were normalized in scale and orientation, aligned such that the two eyes were at the same position, and finally cropped into 32×32 gray pixels such that only the facial areas were preserved, hence each image was represented by a 1,024-dimensional vector. We use these processed data [4] as input.

Experimental setup. We first use PCA to reduce the dimensionality of all sample vectors from 1,024 to 50, and this preserves more than 98% of the total energy. The 11 samples of each individual are randomly divided into training and test sets, that is, p samples per individual are randomly selected for training and the rest of the $11-p$ samples are for test. In our experiments, we set $p = 5, 6$. For each given p, we randomly split the whole database to generate the training and test sets 10 times, and take the average error rates and standard deviations as evaluation indicators. We compare the proposed HMMDA algorithms again with LDA, HLDA, aPAC and MMDA. The original dimensionality 50 is reduced to all possible values in the range of 1 to $C - 1$ with interval of 5 by LDA and aPAC, and to all possible values in the range of 1 to $50 - 1$ with interval of 5 by the other DR methods. The three classifiers (1-NN, NM and QDF) are used to perform classification in the low-dimensional subspaces. The error rates of the three classifiers on original data are also evaluated as references.

Results and analysis. The results on the Yale Face database by using $p = 5, 6$ samples per class for training are shown in Fig. 2 and Fig. 3, respectively. We can find that in all cases, when more than a third of the original dimensionality is preserved, the average error rates in the subspace projected by nearly all these DR methods are lower than those in the original space with smaller or comparable standard deviations. Hence no matter how many samples per class are used for training and no matter which classifier is used, these DR methods, especially the proposed HMMDA algorithms, indeed can improve the performance.

LDA and aPAC can preserve at most $C - 1$ dimensionalities, but it is hard to achieve sufficiently good performance within such a small number of dimensionalities. However, in many cases the proposed HMMDA algorithms outperform the two methods even when projecting to subspaces with a same small dimensionality, and when more than $C - 1$ dimensionalities are retained, the error rates of HMMDAs continue to drop. For almost all these cases, the performances of OHMMDA, mar-OHMMDA and sup-OHMMDA are always the best three, even when few samples per class are available for training and when the linear NM classifier is used for classification. The quadric classifier performs poorly in the original space, which may indicate that the second order moments of the class distributions seem rather unsystematic and cannot directly contribute to the classification. Even so, the proposed HMMDA algorithms can still identify the separation information potentially presented in them and successfully utilize them to greatly improve the classification performance.

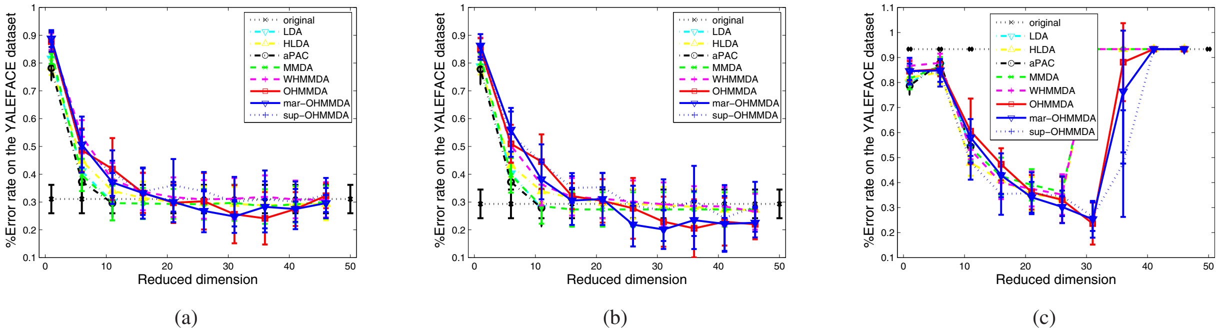


(a)

(b)

(c)

Figure 2. Error rates of different dimensionality reduction methods on the Yale Face database by (a) the 1-NN classifiers; (b) the NM classifier and (c) the QDF classifier when $p = 5$.



(a)

(b)

(c)

Figure 3. Error rates of different dimensionality reduction methods on the Yale Face database by (a) the 1-NN classifiers; (b) the NM classifier and (c) the QDF classifier when $p = 6$.

5. Conclusions

In this paper, we have proposed several HMMDA methods for DR, which can both address the class separation problem and take the heteroscedasticity of data into account. WHMMDA first performs a whitening preprocessing and then maximizes the minimal pairwise Chenoff distance in the latent subspace, while OHMMDA explicitly incorporates this objective with minimizing the closeness of classes into a trace quotient formulation and imposes an orthogonal constraint. Two variants of OHMMDA, namely mar-OHMMDA and sup-OHMMDA, are developed to better handle general class distributions and encode the margin information. HMMDA algorithms achieve superior performances compared against several popular DR approaches on five UCI datasets and the Yale Face database.

Although these HMMDA algorithms are developed under the Gaussian assumption, they work quite well in practice and the consideration of heteroscedasticity still makes sense. Following the idea of [30], since a large majority of distributions can be approximated by a mixture of Gaussians, we can first decompose the general underlying distribution of each class into a set of heteroscedastic Gaussians, and then view each Gaussian component as a single subclass. Under the max-min distance criterion, we can put lower weights or apply constraints to the pairwise within-class scatter between two subclasses from the same class. Hence the proposed HMMDA algorithms can be extended to tackle general distributions. Our future work involves progressing towards such extensions and exploring nonlinear extensions of HMMDA by kernelization.

Acknowledgements

This work was supported by the National Basic Research Program of China (973 program) under Grant No. 2013CB329403, the National Natural Science Foundation of China under Grant No. 61032008, 61471214 and National Science Foundation grant IIS-0916607, IIS-1217302.

References

[1] M. L. Ali, L. Rueda, and M. Herrera. On the performance of chernoff-distance-based linear dimensionality reduction techniques. In *Advances in Artificial Intelligence*, pages 467–478. 2006.

[2] K. Bache and M. Lichman. *UCI Machine Learning Repository*. http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences, 2013.

[3] W. Bian and D. Tao. Max-min distance analysis by using sequential sdp relaxation for dimension reduction. *PAMI*, 33(5):1037–1050, 2011.

[4] D. Cai. Popular face databases in matlab format. http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html, 2014.

[5] D. Cai, X. He, J. Han, and H.-J. Zhang. Orthogonal laplacianfaces for face recognition. *TIP*, 15(11):3608–3614, 2006.

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

[7] O. Hamsici and A. Martinez. Bayes optimality in linear discriminant analysis. *PAMI*, 30(4):647–657, 2008.

[8] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Graph embedding and extensions: a general framework for dimensionality reduction. *PAMI*, 27(3):328–340, 2005.

[9] G. Hua, P. A. Viola, and S. M. Drucker. Face recognition using discriminatively trained orthogonal rank one tensor projections. In *CVPR*, 2007.

[10] A. Iosifidis, A. Tefas, and I. Pitas. On the optimal class representation in linear discriminant analysis. *IEEE Trans. Neural Networks and Learning Systems*, 24(9):1491–1497, 2013.

[11] N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech communication*, 26(4):283–297, 1998.

[12] M. Loog and R. Duin. Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. *PAMI*, 26(6):732–739, 2004.

[13] M. Loog, R. P. W. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise fisher criteria. *PAMI*, 23(7):762–766, 2001.

[14] M. L. Overton and R. S. Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45, 1992.

[15] J. Peng, G. Seetharaman, S. A. Robila, A. S. Varde, and W. Fan. Chernoff dimensionality reduction–where fisher meets fkt. In *SDM*.

[16] C. Rao. The utilization of multiple measurements in problems of biological classification. *J. Royal Statistical Soc., Series B*, 10:159–203, 1948.

[17] L. Rueda and M. Herrera. Linear dimensionality reduction by maximizing the chernoff distance in the transformed space. *PR*, 41(10):3138–3152, 2008.

[18] C. Shen, H. Li, and M. Brooks. Supervised dimensionality reduction via sequential semidefinite programming. *PR*, 41(12):3644–3652, 2008.

[19] B. Su and X. Ding. Linear sequence discriminant analysis: a model-based dimensionality reduction method for vector sequences. In *ICCV*, 2013.

[20] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *JMLR*, 8:1027–1061, 2007.

[21] K. Toh, M. Todd, and R. Tutuncu. Sdpt3 — a matlab software package for semidefinite programming. *Optimization Methods and Software*, 11:545–581, 1999.

[22] R. Tutuncu, K. Toh, and M. Todd. Solving semidefinite-quadratic-linear programs using sdpt3. *Mathematical Programming Ser. B*, 95:189–217, 2003.

[23] Y. Univ. Yale face database. http://cvc.yale.edu/projects/yalefaces/yalefaces.html, 2002.

[24] X. wen Chen and T. Huang. Facial expression recognition: A clustering-based approach. *PRL*, 24(9):1295–1302, 2003.

[25] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *PAMI*, 29(1):40–51, 2007.

[26] Y. Yu, J. Jiang, and L. Zhang. Distance metric learning by minimal distance maximization. *PR*, 44(3):639–649, 2011.

[27] W. Zhang, Z. Lin, and X. Tang. Tensor linear laplacian discrimination (tlld) for feature extraction. *PR*, 42(9):1941–1948, 2009.

[28] X.-Y. Zhang and C.-L. Liu. Evaluation of weighted fisher criteria for large category dimensionality reduction in application to chinese handwriting recognition. *PR*, 46(9):2599–2611, 2013.

[29] Y. Zhang and D.-Y. Yeung. Worst-case linear discriminant analysis. In *NIPS*, 2010.

[30] M. Zhu and A. Martinez. Subclass discriminant analysis. *PAMI*, 28(8):1274–1286, 2006.