Pattern Recognition 44 (2011) 639–649

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Distance metric learning by minimal distance maximization

Yaoliang Yu [a,*,1], Jiayan Jiang [b], Liming Zhang [b]

[a] Department of Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8
[b] Department of Electronic Engineering, Fudan University, Shanghai 200433, PR China

ARTICLE INFO

Article history:
Received 22 July 2008
Received in revised form
25 November 2009
Accepted 29 September 2010

Keywords:
Linear dimensionality reduction (LDR)
Metric learning
Convex optimization
Minimal distance maximization

ABSTRACT

Classic linear dimensionality reduction (LDR) methods, such as principal component analysis (PCA) and linear discriminant analysis (LDA), are known not to be robust against *outliers*. Following a systematic analysis of the multi-class LDR problem in a unified framework, we propose a new algorithm, called minimal distance maximization (MDM), to address the non-robustness issue. The principle behind MDM is to maximize the *minimal* between-class distance in the *output* space. MDM is formulated as a semi-definite program (SDP), and its dual problem reveals a close connection to "weighted" LDR methods. A soft version of MDM, in which LDA is subsumed as a special case, is also developed to deal with overlapping centroids. Finally, we drop the homoscedastic Gaussian assumption made in MDM by extending it in a non-parametric way, along with a gradient-based convex approximation algorithm to significantly reduce the complexity of the original SDP. The effectiveness of our proposed methods are validated on two UCI datasets and two face datasets.

1. Introduction

Many researchers have pointed out that principal component analysis (PCA) is not robust against outliers [1–4]. Similar non-robustness issue has been confirmed to exist in linear discriminant analysis (LDA) [5–8]. For the latter, when the reduced dimension is strictly less than $c-1$ (c is the number of classes), this non-robustness issue becomes much more severe, causing LDA being far from (Bayes) optimal [8]. The reason, as we discuss in details later, is that both PCA and LDA try to maximize the *sum* of squared (pairwise) distances, which inevitably make the "outliers" unnecessarily important. Various algorithms have been proposed to address this issue [1–12], and we postpone the detailed discussions about them into Section 2.

In this paper, under the homoscedastic Gaussian assumption (see Assumption 1), we first provide a systematic analysis for the multi-class LDR problem in a unified framework. After exposing the deficiencies of current methods, a new algorithm, called minimal distance maximization (MDM), is then proposed. MDM tries to maximize the *minimal* between-class distance in the *output* space. This makes sense both intuitively and theoretically since the minimal distance is the bottleneck and is not affected by distant class centroids (*i.e.*, "outliers"). MDM is formulated as a Semi-Definite Program (SDP) [13], which is immune to local minima. Interestingly, we show that the dual problem of MDM reveals a close connection to "weighted" LDR methods, for example, aPAC in [5]: It tries to *learn* the weights by minimizing the sum of the largest eigenvalues of the "weighted" between-class scatter matrix. However, the key difference is that the dual of MDM dynamically adjusts the weights in the *output* space while previous work such as aPAC empirically determines the weights in the *input* space beforehand. A sufficient condition is also given to determine the optimality of MDM. When two (or more) centroids overlap, MDM will output randomly because the minimal between-class distance is trivially 0. To tackle this problem, a "soft" version of MDM, in which LDA is subsumed as a special case, is developed correspondingly.

Since it is not easy in practice to test if the homoscedastic Gaussian assumption applies, we drop it in Section 4 by extending MDM in a non-parametric way. Following the same principle as behind MDM, we argue that the minimal distance of differently labeled samples well assesses the separability of all classes, while the maximal distance of each class evaluates its compactness. We propose to Simultaneously Maximize the minimal distance of differently labeled samples and Minimize the sum of maximal distances of each class (SMM). After examining the dual problem, a similar relationship to non-parametric "weighted" LDR methods [9–12] can also be established. However, the difference again lies in how the weights are determined. Furthermore, to make the algorithm scalable to large-size problems, the "softmax" inequality is employed to give a smooth convex approximation of the original SDP formulation in SMM. Projected gradient descent is then used to optimize the smooth convex approximation.

* Corresponding author.
E-mail addresses: yaoliang@cs.ualberta.ca, ever4ys@gmail.com (Y. Yu), 041021021@fudan.edu.cn (J. Jiang), lmzhang@fudan.edu.cn (L. Zhang).
[1] Work was done while the author was at Department of Electronic Engineering, Fudan University.

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.09.019

The effectiveness of our proposed methods (MDM and aSMM) is evaluated in two databases of the UCI repository and two face databases, i.e., the ORL face database and the YALE face database, comparing with state-of-art LDR methods and metric learning (ML) algorithms. In most cases, our methods perform promisingly, especially when the reduced dimension is fairly small.

This paper is organized as follows: We briefly review related work in Section 2. The unified framework is provided and analyzed in Section 3, as well as our new algorithm, discussion and extension. To drop the homoscedastic Gaussian assumption, Section 4 extends MDM non-parametrically. A gradient-based convex approximation algorithm is also proposed to speed up SMM. We present experimental results in Section 5, and finally conclude in the last section.

2. Related work

There have been two main streams of works relevant to us. One is metric learning (ML), which learns the Mahalanobis distance, parameterized by a semi-definite matrix M, directly from data; the other being LDR methods, which learn a linear transformation W instead. Note that the two are closely related by the equation $M=WW^T$.

Among the first category, the pioneering work [14] maximizes the sum of (square) distances between dissimilar samples while constraining the sum of distances between similar samples to be less than a threshold. An encouraging improvement in semi-supervised clustering by employing the learned Mahalanobis distance instead of the Euclidean distance is shown in [14]. Recently, two fully supervised ML algorithms are proposed in [15,16]. Based on large margin theory, large margin nearest neighbor (LMNN) [15] pulls the within-class nearest neighbors while pushes between-class points out of a margin. Maximally collapsing metric learning (MCML) [16] tries to collapse data from the same class to a point while separates data from different class as much as possible. Actually, MCML can be treated as a linearization of stochastic neighborhood embedding (SNE) [17]. Results of [15,16] show significant improvement in classification problems when using the learned Mahalanobis distance. Other notable work include metric learning by kernel alignment [18], kernel classification rule with the data-dependent Mahalanobis distance metric [19], and incremental ML methods [20,21], etc.

In dealing with W, the most famous instances are arguably PCA [22,23] and LDA [24–26]. PCA, which works unsupervisedly, tries to minimize the reconstruction error or maximize the variance, equivalently. Several researchers have pointed out that PCA is not robust against outliers [1–4]. This non-robustness of PCA is generally attributed to the L_2 norm utilized in its criterion and L_1 norm is favored instead in [1–3]. By adopting a probabilistic model, the univariate Laplace distribution, which is more heavy-tailed than the Gaussian distribution, is chosen in [4]. However, none of [1–4] belongs to convex optimization, thus they all suffer from poor local minima. LDA, closely related to PCA, has been confirmed to have similar non-robustness issues in [5–8]. Considerable efforts have been put in improving LDA. Approximate pairwise accuracy criterion (aPAC) [5] proposes to weigh all pairs of class centroids according to their distances to each other. A similar "weighting" strategy is utilized in [6], which also reduces the dimensionality incrementally. In [7], the multi-class problem is decomposed into a series of pairwise problems and an approximation of the (optimal) Bayes error is directly minimized. Bayes LDA [8] proves that when reduced to one-dimensional space, the optimal Bayes decision plane can be found by solving $c!/2$ convex sub-problems (c is the number of classes), under homoscedastic Gaussian assumption. However, the optimality of Bayes LDA can only be guaranteed in one-dimensional space and it adopts a similar "sub-optimal" strategy as in [6] to successively reduce the dimensionality. Besides, both [7] and [8] lack a straightforward non-parametric extension. Recently, inspired by manifold learning algorithms, much work has been focused on "weighted" discriminant analysis [9–12], which can be regraded as non-parametric extensions of aPAC [5]. All of these "weighted" LDR methods end up with (generalized) eigen-decomposition, however, weights in these methods are generally computed in the *input* space and never change once determined.

3. Minimal distance maximization

In this section, we propose a new method (MDM), which maximizes the *minimal* distance in the *output* space and is immune to local minima, to address the non-robustness issue of LDA. We formalize the problem in Section 3.1 and then lay our homoscedastic Gaussian assumption in Section 3.2. A unified framework is proposed next, followed by our new algorithm, discussion, and extension. Note that we will drop the homoscedastic Gaussian assumption in Section 4.

3.1. Problem setup

Let training data be $\{x_i,y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^m$ is associated with its class label $y_i \in \{1,2,\ldots,c\}$. Training data is assembled in $X=[x_1,x_2,\ldots,x_n] \in \mathbb{R}^{m\times n}$, $y=[y_1,y_2,\ldots,y_n] \in \mathbb{R}^n$, and C_g is a set defined as $\{h|y_h=g\}$, $n_g=|C_g|$, $\mu_g=(1/n_g)\sum_{h\in C_g} x_h$.

The Mahalanobis (square) distance between x_i and x_j is defined as

$$d_{ij}^M=(x_i-x_j)^T M(x_i-x_j), \tag{1}$$

where $M\preceq 0$ is a positive semi-definite matrix.[2] The objective of most metric learning algorithms is to learn M in some appropriate manner such that the classification error is minimized.

3.2. Homoscedastic Gaussians

Like LDA and many previous works [5–8], our main assumption about the data in this section is:

Assumption 1 (*Homoscedastic Gaussian*). Each class follows the Gaussian distribution with common covariance matrix but different means.

Although it might seem rare to meet this assumption in practice, we emphasize that even in this ideal setting, there is no effective algorithm capable of finding the optimal Bayes plane in polynomial time, except in trivial cases such as when the reduced dimension is *exactly* $c-1$. Note however that it is necessary to reduce the dimension to less than $c-1$ in many applications arising in biology, medicine, psychology, and anthropology [8].

Our intention here is to clearly expose our idea under the ideal assumption, while in the next section this assumption will be dropped by working fully non-parametrically. On the other hand, albeit the questionability of their assumptions, parametric methods like LDA are still widely used, mainly due to their simplicity, efficiency and mysterious effectiveness. The latter is also confirmed in our experiments. Notice that the Gaussian assumption is partly responsible for the quadratic form of the Mahalanobis distance in (1).

It is well known that, for two homoscedastic Gaussian distributions, the hyperplane perpendicular to the line connecting the two

[2] Without loss of generality, M is assumed to be symmetric throughout this paper.

means *in the whitened space* is the optimal Bayes decision plane. Not surprisingly, one natural strategy to deal with the general problem is then to decouple the multi-class problem into a series of pairwise subproblems [7]. Of course, in which way these separate pairwise subproblems are integrated becomes particularly relevant.

3.3. A unified framework

We first propose the following unified framework for general linear dimensionality reduction problems:

$$\begin{aligned}\max_{W}\quad & E'(W)=\left(\sum_{g>h} d_{gh}^{p}\right)^{2/p}\\ \text{s.t.}\quad & d_{gh}^{2}=(\mu_g-\mu_h)^T WW^T(\mu_g-\mu_h)\\ & W^T S_w W=\mathbb{I}.\end{aligned}\tag{2}$$

Here p is a constant, $W\in R^{m\times r}$ reduces the dimension from m to r and $\mathbb{I}$ denotes the identity matrix with appropriate sizes throughout the paper. S_w is the within-class scatter matrix defined as

$$S_w=\sum_{g}\sum_{h\in C_g}(x_h-\mu_g)(x_h-\mu_g)^T.$$

Framework (2) can be regarded as integrating pairwise subproblems in an $L_{p/2}$-norm way. Clearly, LDA is a special case ($p=2$) of our framework.

When S_w is non-singular, we can first whiten[3] the data: $\tilde{x}_i=S_w^{-1/2}x_i$. This reduces (2) to

$$\begin{aligned}\max_{\tilde{W}}\quad & E(\tilde{W})=\left(\sum_{g>h}\tilde{d}_{gh}^{p}\right)^{2/p}\\ \text{s.t.}\quad & \tilde{d}_{gh}^{2}=(\tilde{\mu}_g-\tilde{\mu}_h)^T\tilde{W}\tilde{W}^T(\tilde{\mu}_g-\tilde{\mu}_h)\\ & \tilde{W}^T\tilde{W}=\mathbb{I}.\end{aligned}\tag{3}$$

To avoid ambiguity, the superscript $\tilde{}$ is adopted to denote variables in the whitened space, for example, $\tilde{\mu}_g$ means the centroid of the g-th class in the whitened space.

Unfortunately, (3) is not convex. We next introduce the symmetric positive semi-definite matrix $M=WW^T$, and transform (3) to

$$\begin{aligned}\max_{\tilde{M}}\quad & E(\tilde{M})=\left(\sum_{g>h}\tilde{d}_{gh}^{p}\right)^{2/p}\\ \text{s.t.}\quad & \tilde{d}_{gh}^{2}=(\tilde{\mu}_g-\tilde{\mu}_h)^T\tilde{M}(\tilde{\mu}_g-\tilde{\mu}_h)\\ & \tilde{M}^2=\tilde{M},\quad \operatorname{rank}(\tilde{M})=r.\end{aligned}\tag{4}$$

Note that (4) is equivalent to (3), though not convex either. Since $\tilde{M}$ is constrained to be idempotent, its rank (which is not a convex function) can be equivalently replaced by its trace (which is a convex function). Thus the only "unpleasant" constraint so far preventing us from "seemingly convex" is $\tilde{M}^2=\tilde{M}$. After relaxing it to $0\preceq\tilde{M}\preceq\mathbb{I}$, we arrive at

$$\begin{aligned}\max_{\tilde{M}}\quad & E(\tilde{M})=\left(\sum_{g>h}\tilde{d}_{gh}^{p}\right)^{2/p}\\ \text{s.t.}\quad & \tilde{d}_{gh}^{2}=(\tilde{\mu}_g-\tilde{\mu}_h)^T\tilde{M}(\tilde{\mu}_g-\tilde{\mu}_h)\\ & 0\preceq\tilde{M}\preceq\mathbb{I},\quad \operatorname{Tr}(\tilde{M})=r.\end{aligned}\tag{5}$$

Now, (5) formally becomes convex according to the following theorem:

Theorem 1. *Problem* (5) *is a convex program, when* $p\le 2$ *and* $p\neq 0$.

Proof. Since $E(\tilde{M})$ is the $L_{p/2}$ norm[4] of $\tilde{d}_{gh}^2$, which itself is a linear function of $\tilde{M}$. According to the concavity of L_q norm, we know that $E(\tilde{M})$ is concave when $p\le 2$ and $p\neq 0$ (note that we are *maximizing* a *concave* function). Besides, the constraints $0\preceq\tilde{M}\preceq\mathbb{I}$ and $\operatorname{Tr}(\tilde{M})=r$ are easily verified to be convex and linear, respectively. □

Before going any further, we would like to introduce a technical lemma which will be used subsequently:

Lemma 1 (*Alizadeh [27, Theorems 4.1 and 4.3]*). *For the sum of the largest* r *eigenvalues of a symmetric matrix* A, *the following SDP characterization holds*:

$$\begin{aligned}\operatorname{MaxEigSum}_r(A)=\max_{M}\quad & \operatorname{Tr}(AM)\\ \text{s.t.}\quad & 0\preceq M\preceq\mathbb{I},\quad \operatorname{Tr}(M)=r,\end{aligned}$$

and its dual

$$\begin{aligned}\operatorname{MaxEigSum}_r(A)=\min_{b,\Sigma}\quad & br+\operatorname{Tr}(\Sigma)\\ \text{s.t.}\quad & b\mathbb{I}+\Sigma\succeq A,\quad \Sigma\succeq 0.\end{aligned}$$

Here we use $\operatorname{MaxEigSum}_r(A)$ to denote the sum of the largest r eigenvalues of matrix A.

The following observation is a simple consequence of the previous lemma:

Corollary 1. *When* $p=2$, *problem* (5) *degenerates to LDA.*

Proof. $p=2$ simplifies (5) to

$$\begin{aligned}\max_{\tilde{M}}\quad & E(\tilde{M})=\sum_{g>h}(\tilde{\mu}_g-\tilde{\mu}_h)^T\tilde{M}(\tilde{\mu}_g-\tilde{\mu}_h)\\ & \propto\operatorname{Tr}[\tilde{S}_b\cdot\tilde{M}]\\ \text{s.t.}\quad & 0\preceq\tilde{M}\preceq\mathbb{I},\operatorname{Tr}(\tilde{M})=r.\end{aligned}\tag{6}$$

Here $\tilde{S}_b\coloneqq\sum_g(\tilde{\mu}_g-\tilde{\mu})(\tilde{\mu}_g-\tilde{\mu})^T$ is the between-class scatter matrix (in the whitened space) while $\tilde{\mu}\coloneqq(1/c)\sum_g\tilde{\mu}_g$ is the overall mean.

By Lemma 1, (6) is equivalent to maximize the sum of the largest r eigenvalues of $\tilde{S}_b$. Since $\tilde{S}_b$ is a constant matrix, one can then recognize that $M^\star=VV^T$ is the optimal solution of (6), where V is an orthogonal base of $\tilde{S}_b$'s principal eigenspace (corresponding to the first r largest eigenvalues). □

The importance of our unified framework is to provide a platform for proposing new LDR or ML algorithms: one has the freedom of choosing the value of p. One natural question arising is what can we say about p in a certain optimal sense? Let us first check this with $p=2$, a case corresponding to the classic LDA (see Corollary 1 above).

As mentioned before, LDA is not robust against "outliers". In the whitened space, LDA is shown to maximize the *sum* of pairwise distances between centroids [5]. This *sum* nature of LDA's criterion makes it aggressive in pushing distant pairs even more far apart but reluctant in pushing nearby pairs to some extent, i.e., the result of LDA is dominated by some particular pairs of centroids (which have large distances to each other). Besides,

[3] In this paper, unless otherwise stated, by whiten we always mean whitened by the within-class scatter matrix S_w.

[4] The L_q norm is defined as: $\|\alpha\|_{L_q}=(\sum_i|\alpha_i|^q)^{1/q}$. When $q=0$, $\|\alpha\|_{L_0}$ denotes the number of non-zero elements of α.

the *square* distance employed in LDA's criterion aggravates the problem and this disadvantage becomes much more severe when the reduced dimension is lower than $c-1$, as shown in the following example.

In Fig. 1(a), three classes of data are sampled from three homoscedastic Gaussian distributions. After reducing the dimension from two to one, two classes are heavily overlapped in LDA's result, see Fig. 1(b). The reason becomes apparent by recalling that LDA tries to maximize the *sum* of all pairwise (square) distances.

Among others, there have been two attempts (relevant to our work) to improve LDA. One popular way, pioneered in aPAC [5], is to adopt a weighted sum criterion to alleviate LDA's dominating problem. The general principle behind is to assign smaller weights to distant pairs and bigger weights to nearby ones. However, the weights in general are computed in the *input* space and never change. In Fig. 1(c), aPAC fails like LDA, which clearly illustrates that a static weighting strategy is not adequate. Interestingly, the dual problem of our new algorithm is to adjust the weights dynamically in the *output* space (see Section 3.5). On the other hand, one can try to change $p=2$ to some smaller value, say, $p=1$ in our unified framework (5) [14]. However, as can be argued, changing p to any *finite* value can not help much in addressing LDA's non-robustness issue. To see this, suppose $p>0$ ($p<0$ respectively), the effect of reducing p can be simply neutralized by scaling up (down respectively) the data. The only option remains to try is thus $p=-\infty$, which we will discuss in the next subsection.

3.4. Primal formulation

We propose to "exaggeratively" put $p=-\infty$ in the unified framework (5). After some simple algebra, we get

$$\max_{\tilde{M}} \ \min_{g>h}(\tilde{\mu}_g-\tilde{\mu}_h)^T\tilde{M}(\tilde{\mu}_g-\tilde{\mu}_h)$$
$$\text{s.t.} \quad 0\preceq\tilde{M}\preceq\mathbb{I}, \quad \mathrm{Tr}(\tilde{M})=r. \tag{7}$$

The objective of (7) tries to maximize the minimal distance among all pairwise distances in the *output* space. This is meaningful since the minimal distance (in the *output* space) is the bottleneck. We can also see that this choice of p leads to the most robust formulation against outliers in the family (5). Fig. 2 schematically illustrates the proposed new method, which we call minimal distance maximization (MDM). Note that (7) has completely avoided the *sum*-type criterion of LDA.

Eq. (7) is easily recognized as an instance of SDP, thus can be solved by many existing numerical packages such as SeDuMi [28] or CSDP [29]. The result of MDM is also shown in Fig. 1(d). It is clear that three classes are well separated in the reduced space.

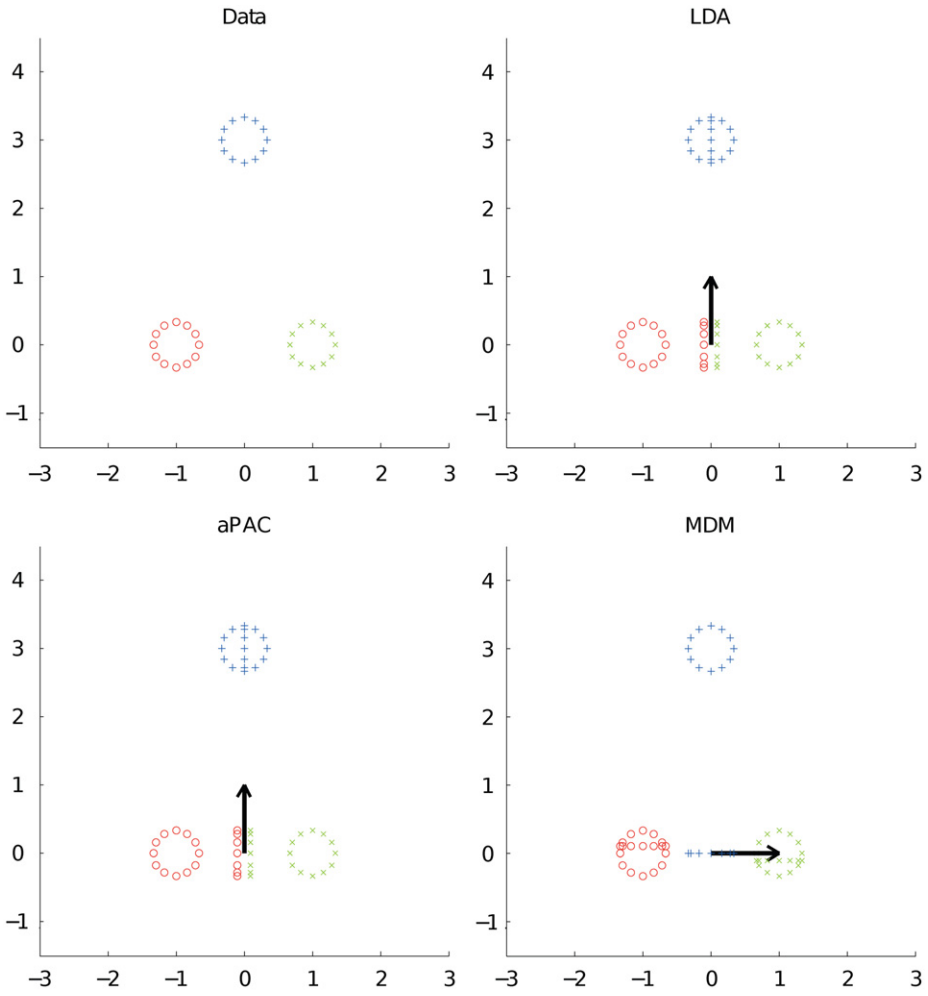

Fig. 1. A toy example. Data are sampled from three homoscedastic Gaussian distributions. After reducing to one-dimensional space, two classes are heavily overlapped in the results produced by both LDA and aPAC, while all three classes are well separated in the result produced by the proposed method MDM.

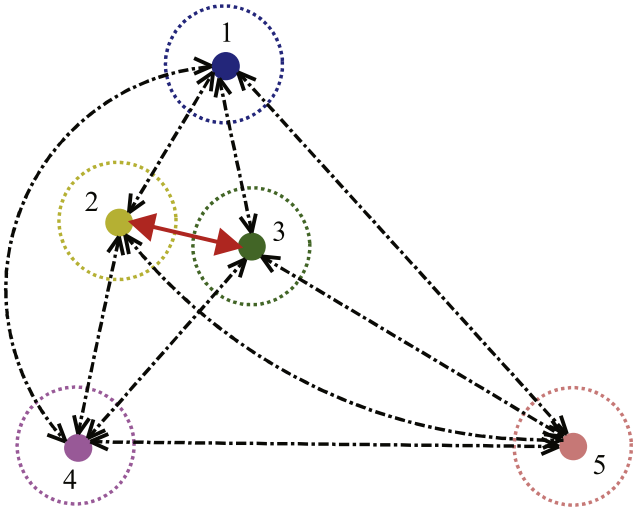

Fig. 2. Schematic illustration of MDM. Among all pairwise distances of class centroids (dashed line), the minimal one (solid line) is the bottleneck.

3.5. Dual formulation

Dual problems usually play an important role in deeply understanding the primal ones. The dual problem of (7) can be derived as

$$\begin{aligned}\min_{\alpha,b,\Sigma}\quad & br+\mathrm{Tr}(\Sigma)\\ \text{s.t.}\quad & \tilde{S}_b(\alpha)\preceq b\mathbb{I}+\Sigma\\ & \alpha_{gh}\geq 0,\quad \sum_{g>h}\alpha_{gh}=1\\ & \Sigma\preceq 0,\end{aligned}\tag{8}$$

where $\tilde{S}_b(\alpha):=\sum_{g>h}\alpha_{gh}(\tilde{\mu}_g-\tilde{\mu}_h)(\tilde{\mu}_g-\tilde{\mu}_h)^T$. By Lemma 1, we can identify (8) as an eigenvalue optimization problem:

$$\begin{aligned}\min_{\alpha}\quad & \mathrm{MaxEigSum}_r[\tilde{S}_b(\alpha)]\\ \text{s.t.}\quad & \alpha_{gh}\geq 0,\quad \sum_{g>h}\alpha_{gh}=1.\end{aligned}\tag{9}$$

Clearly, (9) can be interpreted as dynamically learning a set of (normalized) weights for all pairs of centroids. Note that $\tilde{S}_b(\alpha)$ has the form of the "weighted" between-class scatter matrix proposed in [5]. However, the weights α_{gh} in our formulation is obtained through optimization in the *output* space other than empirically or statically determined in the *input* space.

The optima of primal problem (7) and dual problem (8) should satisfy the Karush–Kuhn–Tucker (KKT) conditions [13]:

$$\alpha^\star_{gh}\cdot[(\tilde{\mu}_g-\tilde{\mu}_h)^T\tilde{M}^\star(\tilde{\mu}_g-\tilde{\mu}_h)-t^\star]=0,\tag{10}$$

$$[\tilde{S}_b(\alpha^\star)-b^\star\mathbb{I}-\Sigma^\star]\cdot\tilde{M}^\star=0,\tag{11}$$

$$(\mathbb{I}-\tilde{M}^\star)\cdot\Sigma^\star=0,\tag{12}$$

where $t^\star=\min_{g>h}(\tilde{\mu}_g-\tilde{\mu}_h)^T\tilde{M}^\star(\tilde{\mu}_g-\tilde{\mu}_h)$. We have used the superscript * to denote the optima of (7) and (8).

Now we can see a good sparse property of the dual variable α: α_{gh} is non-zero only when the pair (g,h) achieves the minimal (Mahalanobis) distance (in the *output* space). We call these pairs, which have a non-zero weight α_{gh}, *support centroids*.

3.6. Optimality

The KKT conditions (11)–(12), along with the strong duality of primal–dual convex problems, also reveal a close relation between our convex relaxation (5) and the original non-convex problem (4).[5] By exploiting these KKT conditions, we are sometimes able to recover the globally optimal solution of the non-convex problem (4) by solving its convex relaxation (5).

To see that, we first derive the following equation from (11)–(12):

$$\mathrm{Tr}[\tilde{S}_b(\alpha^\star)\tilde{M}^\star]=b^\star r+\mathrm{Tr}(\Sigma^\star)=\mathrm{MaxEigSum}_r[\tilde{S}_b(\alpha^\star)],$$

where the last equality follows from the strong duality between (8) and (9). From this equation, we conclude that $\tilde{M}^\star$ is an optimum of the following eigenvalue optimization problem (see Lemma 1):

$$\begin{aligned}\max_{\tilde{M}}\quad & \mathrm{Tr}[\tilde{S}_b(\alpha^\star)\tilde{M}]\\ \text{s.t.}\quad & 0\preceq\tilde{M}\preceq\mathbb{I},\quad \mathrm{Tr}(\tilde{M})=r.\end{aligned}\tag{13}$$

Another lemma will be needed in our next theorem:

Lemma 2 (*Alizadeh [27], Lemma 4.2*). *Let S_1 be the convex hull of the set $\{\tilde{W}\tilde{W}^T:\tilde{W}\in\mathbb{R}^{m\times r},\tilde{W}^T\tilde{W}=\mathbb{I}_r\}$, and S_2 be defined as $\{\tilde{M}:\tilde{M}=\tilde{M}^T\in\mathbb{R}^{m\times m},\mathrm{Tr}(\tilde{M})=r,0\preceq\tilde{M}\preceq\mathbb{I}\}$. Then $S_1=S_2$.*

Theorem 2 (*Sufficient condition*). *If the largest $r+1$ eigenvalues of $\tilde{S}_b(\alpha^\star)$ are all different, then $\tilde{M}^\star=\tilde{W}^\star\tilde{W}^{\star T}$, where $\tilde{W}^\star$ assembles eigenvectors of $\tilde{S}_b(\alpha^\star)$ associated with its largest r eigenvalues.*

Proof. Through Lemma 2 it is easy to verify that $\tilde{M}^\star$, a maximizer of problem (13), is a convex combination of $\tilde{W}^\star_i\tilde{W}^{\star T}_i$, where $\tilde{W}^\star_i$ is constituted by an orthonormal base of the eigenspace of $\tilde{S}_b(\alpha^\star)$, associated with its largest r eigenvalues. If the largest $r+1$ eigenvalues of $\tilde{S}_b(\alpha^\star)$ are all different, then $\forall i\neq j,\tilde{W}^\star_i\tilde{W}^{\star T}_i=\tilde{W}^\star_j\tilde{W}^{\star T}_j$ (the eigenspace is uniquely defined). This concludes the proof. □

Note that when the condition in Theorem 2 is satisfied, $\tilde{M}^{\star 2}=\tilde{M}^\star$ and the rank of $\tilde{M}^\star$ is exactly r, hence the global optimality of $\tilde{M}^\star$ to the non-convex problem (4).

3.7. L_1-norm soft MDM

One obvious drawback of MDM is when two or more centroids in the whitened space overlap, MDM will output a trivial solution since the minimal (Mahalanobis) distance will always be zero. By introducing slack variables, we get the following L_1-norm soft MDM to overcome this deficiency:

$$\begin{aligned}\max_{\tilde{M},\xi}\quad & \min_{g>h}[(\tilde{\mu}_g-\tilde{\mu}_h)^T\tilde{M}(\tilde{\mu}_g-\tilde{\mu}_h)+\xi_{gh}]-C\cdot\sum_{g>h}\xi_{gh}\\ \text{s.t.}\quad & 0\preceq\tilde{M}\preceq\mathbb{I},\quad \mathrm{Tr}(\tilde{M})=r\\ & \forall g>h,\quad \xi_{gh}\geq 0.\end{aligned}$$

Here ξ_{gh} are the introduced non-negative slack variables. C is a tuning parameter to balance the two terms, and its meaning will become more clear by exploring the dual eigenvalue optimization

[5] We have set $p=-\infty$ throughout this subsection.

problem:

$$\min_{\alpha} \quad \mathrm{MaxEigSum}_r[\tilde{S}_b(\alpha)]$$
$$\text{s.t.} \quad C \geq \alpha_{gh} \geq 0, \quad \sum_{g>h} \alpha_{gh} = 1. \tag{14}$$

Through constraints on α in (14), we immediately realize that C should lie in the range $[2/c(c-1),1]$. Recall that c is the number of classes and we have $c(c-1)/2$ different pairs of centroids. Interestingly, at the two extreme points, $C=1$ degenerates L_1-norm soft MDM to MDM while $C=2/c(c-1)$ degenerates L_1-norm soft MDM to LDA, again.

Similar as in support vector machines (SVM) [30], it is also possible to investigate the L_2-norm soft MDM, which penalizes the slack variables ξ in its L_2 norm. It turns out that L_2-norm soft MDM prefers a non-sparse solution of the dual variable α. Since it is not clear to us what benefits could be brought in by non-sparsity, we will not pursue L_2-norm soft MDM any further in this paper.

4. Non-parametric extension

We now present our non-parametric extension of MDM, to drop the homoscedastic Gaussian assumption. A gradient-based convex approximation algorithm is also proposed to reduce the time complexity.

4.1. Minimal distance of differently labeled samples

We use the minimal distance between x_i and any other differently labeled samples to assess their separation:

$$MDDLS_i \coloneqq \min_{j|y_j \neq y_i} (x_i - x_j)^T M (x_i - x_j).$$

Again, to avoid being dominated by some particular instances, we will use the minimal one among all $MDDLS_i$ to represent the separation of all differently labeled data:

$$MDDLS \coloneqq \min_i MDDLS_i.$$

Note that *MDDLS* is itself a function of M. The solid line in the left panel of Fig. 3 illustrates the concept of *MDDLS*.

4.2. Maximal distance of each class

Likewise, the compactness of the g-th class is defined as its maximal within-class distance:

$$MDEC_g \coloneqq \max_{i,j|y_j = y_i \in C_g} (x_i - x_j)^T M (x_i - x_j).$$

We sum up all $MDEC_g$ to represent the compactness of the whole data:

$$MDEC \coloneqq \sum_g MDEC_g.$$

Note also that *MDEC* is itself a function of M. The solid line in the right panel of Fig. 3 illustrates the concept of *MDEC*.

4.3. A naive formulation

After defining *MDDLS* and *MDEC*, our objective now simply converts to Simultaneously Maximize *MDDLS* and Minimize *MDEC* (SMM), formulated as

$$\min_M \quad \sum_g \max_{i,j|y_i = y_j \in C_g} d_{ij}^M - \rho \cdot \min_{i,j|y_i \neq y_j} d_{ij}^M$$
$$\text{s.t.} \quad 0 \preceq M \preceq \mathbb{I}, \quad \mathrm{Tr}(M) = r. \tag{15}$$

Here d_{ij}^M is defined in (1) and ρ is a tuning parameter. Fig. 3 shows the main idea of SMM (15).

Eq. (15) is easily recognized as an instance of SDP and its dual problem can be derived as

$$\min_{\alpha,\beta,b} \quad br + \mathrm{Tr}(\Sigma)$$
$$\text{s.t.} \quad L(\alpha,\beta) \coloneqq L_1(\alpha) - L_2(\beta) \preceq b\mathbb{I} + \Sigma$$
$$\beta_{ij} \geq 0, \quad \sum_{i,j|y_i = y_j \in C_g} \beta_{ij} = 1$$
$$\alpha_{ij} \geq 0, \quad \sum_{i,j|y_i \neq y_j} \alpha_{ij} = \rho$$
$$\Sigma \preceq 0, \tag{16}$$

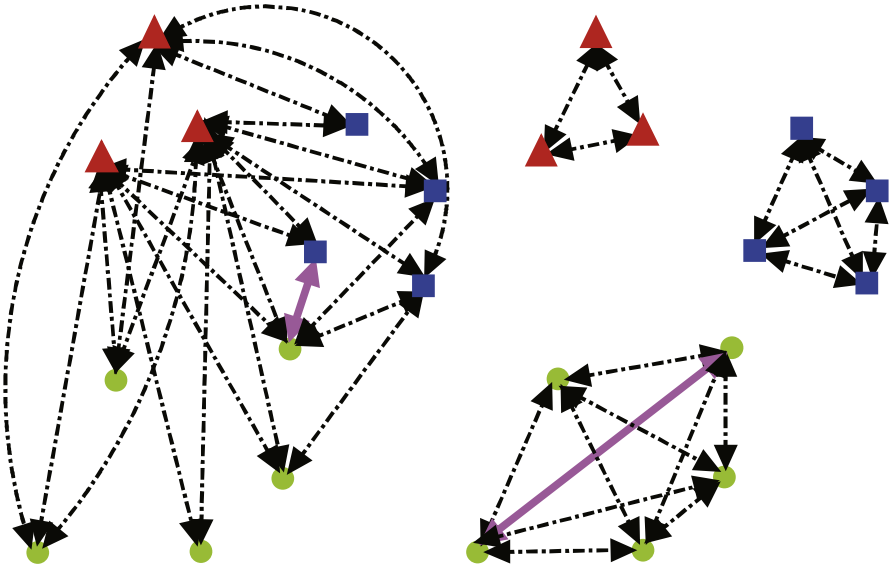

Fig. 3. In the left part, dashed lines denote the pairwise distances between all differently labeled samples, while the solid line shows the minimal one among them. In the right part, dashed lines denote the pairwise distances among each class, while the solid lines show the maximal one in each class.

where

$$L_1(\alpha) := \sum_{i,j|y_i \neq y_j} \alpha_{ij}(x_i - x_j)(x_i - x_j)^T,$$
$$L_2(\beta) := \sum_{g} \sum_{i,j|y_i = y_j \in C_g} \beta_{ij}(x_i - x_j)(x_i - x_j)^T.$$

Note that $L_1(\alpha)$ and $L_2(\beta)$ have the form of the "weighted" between-class scatter matrix and the "weighted" within-class scatter matrix [11,12].

Eq. (16) can also be identified as an eigenvalue optimization problem:

$$\begin{aligned} \min_{\alpha,\beta} \quad & \mathrm{MaxEigSum}_r[L(\alpha,\beta)] \\ \text{s.t.} \quad & \beta_{ij} \geq 0, \quad \sum_{i,j|y_i = y_j \in C_g} \beta_{ij} = 1 \\ & \alpha_{ij} \geq 0, \quad \sum_{i,j|y_i \neq y_j} \alpha_{ij} = \rho. \end{aligned} \tag{17}$$

Again, we see that the dual problem of SMM adjusts the weights α,β for the "weighted" scatter matrices dynamically in the output space, instead of setting them empirically and statically in the input space. This not only makes our algorithm quite different from recent "weighted" LDR methods [9–12], but also provides an intuitive way to set the weights: After solving the dual problem (16) to obtain the weights (α,β), one could also apply them in those "weighted" LDR methods [9–12]. Note that in this case, we do not need to empirically build an adjacency graph, which is usually needed in [10–12]. Moreover, as in the parametric case, by writing out the KKT conditions (omitted), we can see that the dual variables α,β will be sparse, i.e., only a small portion of them will be non-zero.

Following the same procedure as in MDM, it is also possible to develop soft versions of SMM by introducing slack variables. The detailed analysis has been omitted due to similarity. On the other hand, since the time complexity of (16) is as high as $O([n^2+m^2]^2 m^{2.5})$, we will focus on reducing the computational burden in the next subsection.

4.4. Convex approximation

Although one can use interior-point-based numerical packages like SeDuMi or CSDP to solve the SDP problem in (15), the time complexity of these SDP routines does not scale very well with the size of practical problems. On the other hand, gradient-based algorithms, though typically require more iterations, do scale very well due to the low cost of each iteration step.

To address the non-smoothness of the objective function in (15), the well-known softmax inequality [13]:

$$\max\{a_1,\ldots,a_n\} \leq \log \sum_{i=1}^{n} e^{a_i} \leq \max\{a_1,\ldots,a_n\} + \log n$$

is employed to approximate the original problem (15):

$$\begin{aligned} \min_{M} \quad & \sum_{g} \log \sum_{i,j|y_i = y_j \in C_g} e^{d_{ij}^M} + \rho \cdot \log \sum_{i,j|y_i \neq y_j} e^{-d_{ij}^M} \\ \text{s.t.} \quad & 0 \preceq M \preceq \mathbb{I}, \quad \mathrm{Tr}(M) = r. \end{aligned} \tag{18}$$

We call (18) approximate SMM (aSMM). Note that recently the softmax inequality was also used in [31] to learn a large margin hidden Markov model.

An important fact of (18) is that it remains to be convex, therefore we can safely use projected gradient descent to optimize it. In each step, after updating M by a fraction of the negative gradient of the objective function, we alternatively project it back to the two convex constraints $\mathbb{A} = \{M : 0 \preceq M \preceq \mathbb{I}\}$ and $\mathbb{B} = \{M : \mathrm{Tr}(M) = r\}$, until convergence. Since $\mathbb{A} \cap \mathbb{B} \neq \emptyset$, it is guaranteed to find a (unique) point in the intersection of the two convex constraints.

Let us write out the gradient of (18):

$$\nabla_M = \sum_{g} \sum_{i,j|y_i = y_j \in C_g} p_{ij} x_{ij} x_{ij}^T - \rho \cdot \sum_{i,j|y_i \neq y_j} q_{ij} x_{ij} x_{ij}^T, \tag{19}$$

where $x_{ij} := x_i - x_j$, and

$$p_{ij} = \frac{e^{d_{ij}^M}}{\sum_{i,j|y_i = y_j \in C_g} e^{d_{ij}^M}}, \quad q_{ij} = \frac{e^{-d_{ij}^M}}{\sum_{i,j|y_i \neq y_j} e^{-d_{ij}^M}}. \tag{20}$$

Note that here p_{ij} indicates the *soft* probability of pair (i,j) being $MDEC_g$, and q_{ij} indicates the *soft* probability of pair (i,j) being *MDDLS*. One can also rewrite (19) in a more compact form $\nabla_M = XBX^T$. The detailed description of the Laplacian matrix B is postponed to Algorithm 1.

The gradient (19) is very similar to the "weighted" scatter matrices in [9–12]. However, we emphasize that the weights (p_{ij},q_{ij}) are themselves a function of M, meaning that they are varying from step to step, rather than empirically or statically determined beforehand.

Algorithm 1 gives the pseudo-code of aSMM. The time and space complexity have been reduced to $O(\tau_1[m^2 n + \tau_2 m^3])$ and $O(m^2+n^2)$, respectively. Here τ_1,τ_2 are two iteration steps needed by the algorithm for convergence. Note that the same approximation can also be made in MDM. In fact, MDM can be treated as a special case of SMM where each class only has one sample.

5. Experiments

In this section, we will compare MDM and aSMM to state-of-art LDR algorithms such as PCA [22,23], LDA [24–26], aPAC [5], BLDA [8], MFA [11] and ML algorithms such as LMNN [15] and MCML [16] in two databases of the UCI repository and two face databases. For clarity, these acronyms are summarized in Table 1. Note that the two databases of the UCI repository we chose in this paper have already been split into training and testing set so that experimental results can be easily repeated. We divide the above-mentioned methods into two categories: parametric methods including LDA [24–26], aPAC [5], BLDA [8] and L_1-MDM; non-parametric ones including PCA [22,23], MFA [11], LMNN [15], MCML [16] and aSMM.

For LMNN, MCML, L_1-MDM and aSMM, we perform singular value decomposition (SVD) on the learned Mahalanobis metric M and truncate its eigenvectors corresponding to the largest r eigenvalues when reducing to the low r-dimensional space. We did not test SMM because its computational cost is prohibitive even in moderate-sized database. Note that nearest centroid classifier is employed for parametric methods when classifying a test sample while nearest neighbor classifier is used for non-parametric methods.

Table 1
Acronyms for different methods compared in our experiments.

PCA	Principal component analysis [22,23]
LDA	Linear discriminant analysis [24–26]
BLDA	Bayes linear discriminant analysis [8]
aPAC	Approximate pairwise accurate criterion [5]
MFA	Marginal Fisher analysis [11]
LMNN	Large margin nearest neighbor [15]
MCML	Maximally collapsing metric learning [16]
MDM	Minimal distance maximization
L_1-MDM	L_1-norm minimal distance maximization
aSMM	Approximate simultaneous maximization and minimization

Algorithm 1. Approximate SMM

Input: X: input data matrix, y: corresponding label, M_0: initial guess, ρ: tuning parameter, r: reduced dimensionality
Output: M: learned Mahalanobis matrix

$M \leftarrow (M_0 + M_0^T)/2$
repeat
  for all (i,j) **do**
    if $y_i = y_j \in C_g$ **then**
      $A_{ij} \leftarrow \dfrac{e^{d_{ij}^M}}{\sum_{i,j|y_i=y_j\in C_g} e^{d_{ij}^M}}$
    else if $y_i \neq y_j$ **then**
      $A_{ij} \leftarrow -\rho \cdot \dfrac{e^{-d_{ij}^M}}{\sum_{i,j|y_i\neq y_j} e^{-d_{ij}^M}}$
    end if
  end for
  $A \leftarrow A + A^T$
  $B \leftarrow \mathrm{diag}(\mathrm{sum}(A)) - A$
  line search stepsize η
  $M \leftarrow M - \eta \cdot XBX^T$
  repeat
    $[V,D] = \mathrm{eig}(M)$
    $d = \mathrm{diag}(D)$
    $d \leftarrow \min(\max(d,0),1)$
    $M \leftarrow V \cdot \mathrm{diag}(d) \cdot V^T$
    $M \leftarrow M - \dfrac{\mathrm{Tr}(M)-r}{\mathrm{size}(M,1)} \cdot \mathbb{I}$
  until Convergence
until Convergence

5.1. Image segmentation database

The image segmentation database in the UCI repository [32] has seven classes and 2310 samples, of which 210 instances (30 samples per class) are divided into the training set while the rest 2100 instances (300 per class) constitute the testing set. Each sample has 19 attributes, which are extracted randomly from a 3×3 region of seven different outdoor images (brickface, sky, foliage, cement, window, path, and grass). To avoid singularity of scatter matrices, we have eliminated the 3rd attribute since it is constantly 9, thus each sample is left with 18 attributes. Note that the second class and the seventh class can be regarded as "outliers" in this dataset since their average distance (in the whitened space) to other classes are 3–4 times larger.

Table 2 gives the classification accuracies of all relevant parametric methods. From the table, we can clearly see that L_1-MDM is much better than LDA and aPAC in most dimensions. Compared with the very latest work BLDA [8], L_1-MDM wins in all reduced dimensions but the first and the last dimensions. Note that it is expected that BLDA wins in the first dimension, since it is guaranteed to be optimal in one-dimensional space (although achieved by solving $c!/2$ convex sub-problems and under the homoscedastic Gaussian assumption). However, when the reduced dimension exceeds one, the optimality of BLDA is lost due to its successive procedure. This is indeed verified by the results in Table 2. Another potential drawback of BLDA is its computational cost since it needs to solve $c!/2$ convex sub-problems in reducing the dimension only by one.

Since the training set and testing set are fixed in this database, we can further explore some interesting properties of L_1-MDM. First, we check the sparsity of the dual variable α, which has 21 different entries in this database. The number of non-zero elements in α ($C = \frac{4}{21}$) is 9 ($r=1,2$), 7 ($r=3,4$), 6 ($r=5,6$) respectively, showing that almost $\frac{2}{3}$ of α are zeros. Next we check the optimality of the relaxed framework (5). Recall that in (5) we have relaxed the constraints $\tilde{M}^2 = \tilde{M}$ and $\mathrm{rank}(\tilde{M}) = r$ to $0 \preceq \tilde{M} \preceq \mathbb{I}$ and $\mathrm{Tr}(\tilde{M}) = r$. In three different dimensions ($r=2,3,4$), the optimal solution $\tilde{M}^*$ of (5) is found to be idempotent and its rank exactly equals r, which means the convex relaxation (5) finds the (globally) optimal solution of the non-convex problem (4). This conclusion can also be drawn through the sufficient condition stated in Theorem 2. For example, when $r=2$, the largest three eigenvalues of $\tilde{S}_b(\alpha^*)$ are 7.4299, 5.3025 and 5.1666. Apparently, they are all different to each other. However, in other dimensions ($r=1,5,6$), by checking the rank and idempotentness of M, the relaxed framework (5) is verified not to be equivalent to (4) but could still be treated as a good approximation. One can also observe that when the relaxation (5) coincides with the original non-convex problem (4), the performance of L_1-MDM is much better (even compared with the latest work BLDA). On the other hand, the performance of L_1-MDM might suffer from the non-equivalence since we need to truncate the eigenvectors of M in this case.

We then evaluate the effect of the parameter C in L_1-MDM. We run L_1-MDM 21 times and each time with a different C. Since this dataset has seven classes, C should lie within the interval $[\frac{1}{21}, 1]$, thus we start with $C = \frac{1}{21}$ and incrementally increase C by $\frac{1}{21}$. Results are tabulated in Table 3. One can observe that the performance of L_1-MDM is moderately stable in a large range of C. The second ($C = \frac{1}{21}$) and last ($C=1$) row in Table 3 deserve a detailed explanation. When $C = \frac{1}{21}$, L_1-MDM is expected to be equivalent to LDA, and their actual performances are indeed very close (the minor difference is possibly caused by numerical precisions). While $C=1$, L_1-MDM degenerates to MDM, however, the performance is not as good as those of other C's. After checking the pairwise distance of centroids in the whitened space, we find that several centroids are very near to each other, which explains the degradation of MDM's performance and in turn emphasizes the importance of introducing slack variables in MDM.

Table 2
Classification accuracies (%) of parametric methods on the image segmentation dataset (best accuracy in bold).

	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$
LDA	50.90	66.19	81.90	87.90	89.86	90.47
aPAC	64.81	66.10	83.29	87.48	89.71	***90.62***
Bayes [20]	***71.24***	80.86	85.24	88.05	89.95	90.47
L_1-MDM ($C = \frac{4}{21}$)	63.29	***81.81***	***87.43***	***90.71***	***90.29***	90.29

Table 3
Classification accuracies (%) of L_1-MDM vs. the value of C on the image segmentation dataset.

	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$
$C = \frac{1}{21}$	50.71	66.29	81.90	87.76	89.57	90.62
$C = \frac{2}{21}$	57.81	71.57	87.14	88.86	90.62	90.62
$C = \frac{3}{21}$	61.90	81.62	86.76	90.67	90.71	90.29
$C = \frac{4}{21}$	63.29	81.81	87.43	90.71	90.29	90.29
$C = \frac{5}{21}$	51.95	78.95	87.76	90.71	90.29	90.29
$C = \frac{6}{21}$	46.00	79.62	87.10	90.62	90.29	90.29
$C = \frac{7}{21}$	46.33	78.90	87.24	90.43	90.19	90.19
$C = \frac{8}{21} - \frac{20}{21}$	41.81	73.81	85.57	90.38	90.00	89.95
$C=1$	41.81	73.81	74.38	81.00	80.48	80.24

Table 4
Classification accuracies (%) of non-parametric methods on the image segmentation dataset (best accuracy in bold).

	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=8$	$r=10$	$r=12$	$r=14$	$r=18$
PCA	41.57	61.43	61.67	81.62	83.62	83.71	87.33	87.67	87.67	87.67	87.67
MFA	51.05	75.19	***85.00***	87.10	89.57	89.76	89.71	89.05	89.05	87.62	84.33
LMNN	50.43	72.76	74.86	75.95	75.95	75.95	75.95	76.19	84.86	90.81	***92.76***
MCML	55.90	69.00	79.10	84.14	84.38	84.38	84.38	84.43	84.62	84.62	84.67
aSMM	***67.71***	***78.62***	83.95	***89.24***	***89.90***	***91.67***	***91.86***	***91.90***	***92.00***	***92.00***	87.67

Table 5
Classification accuracies (%) of parametric methods on the satellite imagery dataset (best accuracy in bold).

	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$
LDA	53.60	72.35	82.65	83.10	***83.95***
aPAC	67.30	77.70	82.35	83.20	***83.95***
Bayes [20]	***69.65***	80.65	82.80	82.55	83.15
L_1-MDM ($C=\frac{2}{15}$)	57.10	***82.20***	***83.10***	***83.60***	***83.95***
L_1-MDM ($C=\frac{1}{15}$)	53.40	72.65	82.25	83.20	***83.95***
L_1-MDM ($C=1$)	48.80	70.75	73.60	80.40	71.90

The last experiment on this dataset is to assess the performance of non-parametric methods. Comparing Tables 2 and 4, it is interesting to note that the performances of non-parametric methods are even worse than those of parametric methods when the reduced dimension is fairly small. However, the performances of non-parametric methods continue to improve when dimensionality exceeds 6 and finally are comparable to or better than those of parametric methods. Our proposed method, aSMM, shows an encouraging behavior in all dimensions. We notice that aSMM's performance dropped to 87.67% in the full dimensional space ($r=m$), which is predictable since M is tightened to be a trivial identity matrix by the constraints $0\preceq M\preceq \mathbb{I}$, $\mathrm{Tr}(M)=m$ (recall that $M\in R^{m\times m}$) and the learned Mahalanobis metric M reduces to the ordinary Euclidean metric.

5.2. Satellite imagery database

Our second experiment is implemented in the satellite imagery database of the UCI repository [32]. This database consists of the multi-spectral values of pixels in 3×3 neighborhoods in a satellite image, and the classification is associated with the central pixel in each neighborhood. There are 6435 samples in total, with 4435 instances for training and the rest 2000 samples for testing. Each sample has 36 attributes and belongs to one of the six classes. The first class and the second class in this dataset can be regarded as "outliers" since their average distance (in the whitened space) to other classes are 4–5 times larger.

Table 5 lists the results of parametric methods. L_1-MDM ($C=\frac{2}{15}$) performs best in the last 4 dimensions while BLDA is again the best in the first dimension. L_1-MDM shows a consistent improvement upon LDA, however, as dimensionality increases, the discrepancy between different methods decreases. We have also evaluated the effect of the parameter C. Again, we find that when $C=\frac{1}{15}$ (recall that there are six classes in this database), L_1-MDM's performance is very close to LDA; and when $C=1$, the performance of MDM is not so good.

The performances of non-parametric methods are reported in Table 6. We only list the results of the first 12 dimensions since the performances saturate from the 12th dimension. Our method, aSMM, outperforms other methods in most dimensions especially when the reduced dimensionality is fairly small. An interesting phenomenon in this database is that PCA's performance is remarkably good when the reduced dimensionality exceeds 4.

5.3. Face recognition

To further test our approximation algorithm aSMM, we have applied it to the face recognition task. Two face databases, i.e., the ORL face dataset[6] and the YALE face dataset[7] are chosen in our next experiment.

The ORL face dataset contains 40 different people while each person has 10 images with size 112×92. The Yale face dataset is composed of 15 people and each person has 11 images with size 256×256. We have performed some pre-processing on the raw images to align the eyes and eliminate the non-relevant background. Each image is finally normalized to 32×32 pixels, with 256 gray levels per pixel.

In the experiment, five images per person in each database are randomly selected to constitute the training set and the rest are used as the testing set. Experiments are repeated 50 times for each database and the average result is reported. Note that since the original dimensionality of face images (1024 after pre-processing) are too high, PCA (with 100% energy) is employed to reduce the dimensionality first (to avoid the small sample size problem).

Figs. 4 and 5 show the error rates of all methods with respect to the reduced dimensionality on ORL and YALE face database, respectively. As can be seen, aSMM exhibits a consistent advantage over other methods on most dimensions. However, the computational cost of aSMM is the heaviest if one needs to search for the best reduced dimensionality r since we need to re-run aSMM with a different r. Error rates of all methods in the YALE database are observed to be much larger than those in the ORL database, possibly because images in the YALE database are acquired with much heavier illumination variations.

6. Conclusions

In this paper, under homoscedastic Gaussian assumption, we first provide a unified framework to analyze the multi-class LDR problem. After exposing the deficiency of existing methods, a new ML algorithm, called minimal distance maximization (MDM), is proposed to address the non-robustness issue in LDA. The principle behind MDM is to maximize the *minimal* between-class distance in the *output* space and is formulated as an SDP. To drop the homoscedastic Gaussian assumption, MDM is extended in a non-parametric way. Interestingly, the dual problems of MDM and SMM are shown to be related to, but fundamentally different from, the "weighted" LDR methods: our dual problems dynamically adjust the weights in the *output* space rather than

[6] http://www.cam-orl.co.uk/facedatabase.html
[7] http://cvc.yale.edu/projects/yalefaces/yalefaces.html

Table 6
Classification accuracies (%) of non-parametric methods on the satellite imagery dataset (best accuracy in bold).

	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$	$r=8$	$r=9$	$r=10$	$r=11$	$r=12$
PCA	40.80	78.35	84.35	85.55	87.40	89.00	89.35	*89.90*	89.50	89.80	89.70	89.95
MFA	47.05	66.00	71.30	71.80	73.85	73.85	75.90	77.90	77.40	78.05	80.50	81.95
LMNN	50.70	68.55	72.50	76.65	81.75	84.30	85.95	86.50	88.65	88.75	***89.80***	***90.35***
MCML	52.50	78.40	83.70	85.20	86.90	86.55	86.65	86.90	87.45	87.40	87.60	87.70
aSMM	***54.95***	***78.75***	***84.95***	***86.00***	***87.65***	***89.25***	***89.75***	89.55	***89.90***	***90.05***	***89.80***	89.75

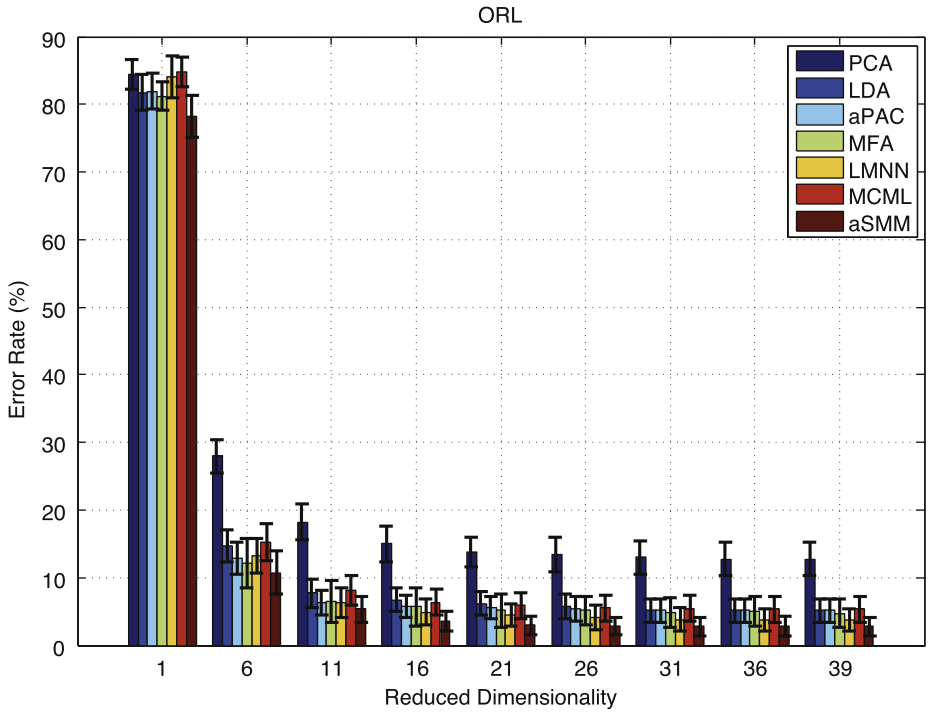

Fig. 4. Error rates on the ORL face database.

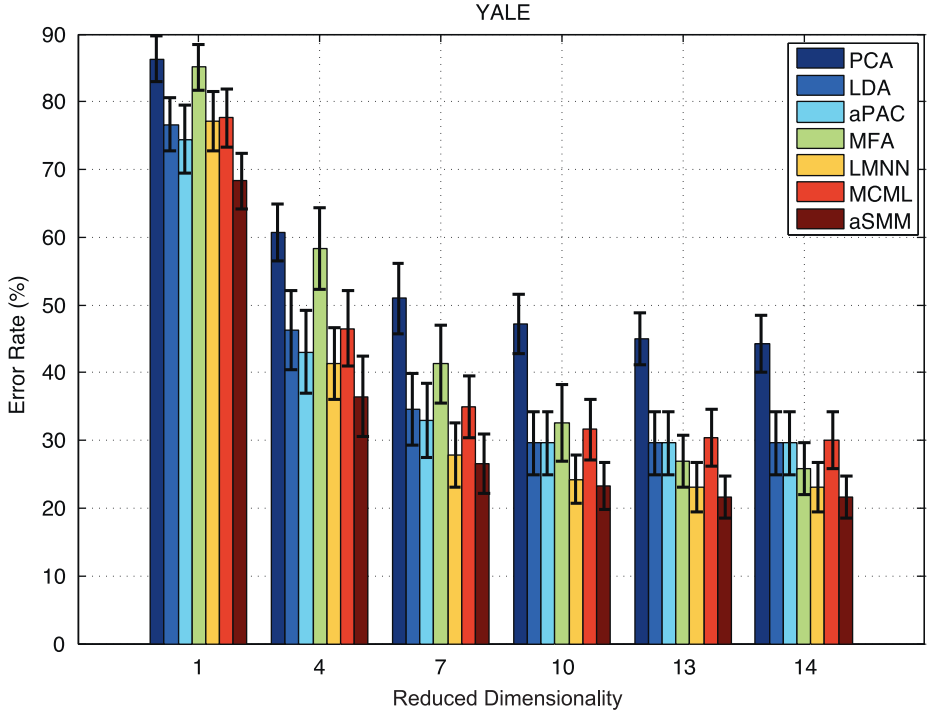

Fig. 5. Error rates on the YALE face database.

empirically or statically determines them in the *input* space. A soft version of MDM, in which LDA is subsumed as a special case, is developed to overcome the triviality when two (or more) centroids overlap. Finally, to significantly reduce the computational burden, we give a gradient-based smooth convex approximation algorithm to the original SDP formulation by employing the "softmax" inequality. Experimental results on various databases demonstrate the effectiveness of our proposed methods.

Evaluating the compactness of a class by the maximal distance within it, as we did in this paper, might be problematic for multimodal distributions. In the future work, we are planning to incorporate ideas like "sub-class" in [33] into our non-parametric method. On the other hand, since the proposed methods in this paper are linear in nature, we are expecting a further improvement by investigating their kernel counterparts.

Acknowledgments

The authors are supported by National Natural Science Foundation of China (60571052) and Shanghai Leading Academic Discipline Project (B112).

References

[1] Q. Ke, T. Kanade, Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming, in: IEEE conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Los Alamitos, CA, USA, 2005, pp. 739–746.
[2] C. Ding, D. Zhou, X. He, H. Zha, R_1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization, in: The 23rd International Conference on Machine Learning, ACM Press, New York, NY, USA, 2006, pp. 281–288.
[3] N. Kwak, Principal component analysis based on L_1-norm maximization, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (9) (2008) 1672–1680.
[4] J. Gao, Robust L_1 principal component analysis and its bayesian variational inference, Neural Computation 20 (2) (2008) 555–572.
[5] M. Loog, R. Duin, R. Haeb-Umbach, Multiclass linear dimension reduction by weighted pairwise fisher criteria, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (7) (2001) 762–766.
[6] R. Lotlikar, R. Kothari, Fractional-step dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (6) (2000) 623–627.
[7] R. Lotlikar, R. Kothari, Adaptive linear dimensionality reduction for classification, Pattern Recognition 33 (2) (2000) 185–194.
[8] O.C. Hamsici, A.M. Martinez, Bayes optimality in linear discriminant analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (4) (2008) 647–657.
[9] Y. Koren, L. Carmel, Robust linear dimensionality reduction, IEEE Transactions on Visualization and Computer Graphics 10 (4) (2004) 459–470.
[10] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using laplacianfaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (3) (2005) 328–340.
[11] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1) (2007) 40–51.
[12] M. Sugiyama, Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis, Journal of Machine Learning Research 8 (2007) 1027–1061.
[13] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2006.
[14] E.P. Xing, A.Y. Ng, M.I. Jordan, S. Russell, Distance metric learning with application to clustering with side information, in: S. Becker, S. Thrun, K. Obermayer (Eds.), Advances in Neural Information Processing Systems, vol. 15, MIT Press, Cambridge, MA2003, pp. 505–512.
[15] K.Q. Weinberger, J. Blitzer, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems, vol. 18, MIT Press, Cambridge, MA2006, pp. 1473–1480.
[16] A. Globerson, S.T. Roweis, Metric learning by collapsing classes, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems, vol. 18, MIT Press, Cambridge, MA2006, pp. 451–458.
[17] G.E. Hinton, S.T. Roweis, Stochastic neighbor embedding, in: S.T.S. Becker, K. Obermayer (Eds.), Advances in Neural Information Processing Systems, vol. 15, MIT Press, Cambridge, MA2003, pp. 833–840.
[18] J.T. Kwok, I.W. Tsang, Learning with idealized kernels, in: T. Fawcett, N. Mishra (Eds.), The 20th International Conference on Machine Learning, AAAI Press, Washington, DC2003, pp. 400–407.
[19] B.K. Sriperumbudur, O.A. Lang, G.R.G. Lanckriet, Metric embedding for kernel classification rules, in: the 25th International Conference on Machine Learning, 2008.
[20] S. Shalev-Shwartz, Y. Singer, A.Y. Ng, Online and batch learning of pseudo-metrics, in: C.E. Brodley (Ed.), The 21th International Conference on Machine Learning, ACM Press, New York, NY, USA2004, pp. 94–101.
[21] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: Z. Ghahramani (Ed.), The 24th International Conference on Machine Learning, Omnipress2007, pp. 209–216.
[22] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.
[23] I.T. Joliffe, Principal Component Analysis, second ed., Springer-Verlag, 2002.
[24] D.L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (8) (1996) 831–836.
[25] P.N. Belhumeur, J.ao P. Hespanda, D.J. Kriegeman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.
[26] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, Boston, 1990.
[27] F. Alizadeh, Interior point methods in semidefinite programming with applications to combinatorial optimization, SIAM Journal of Optimization 5 (1) (1995) 13–51.
[28] J.F. Sturm, Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones, Optimization Methods and Software 11–12 (1999) 625–653.
[29] B. Borchers, CSDP, a C library for semidefinite programming, Optimization Methods and Software 11 (1999) 613–623.
[30] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, An introduction to kernel-based learning algorithms, IEEE Transactions on Neural Networks 12 (2) (2001) 181–201.
[31] F. Sha, L.K. Saul, Large margin hidden Markov models for automatic speech recognition, in: B. Schölkopf, J. Platt, T. Hoffman (Eds.), Advances in Neural Information Processing Systems, vol. 19, MIT Press, Cambridge, MA2007, pp. 1249–1256.
[32] A. Asuncion, D. Newman, UCI machine learning repository. 〈http://www.ics.uci.edu/∼mlearn/MLRepository.html〉, 2007.
[33] M. Zhu, A.M. Martinez, Subclass discriminant analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (8) (2006) 1274–1286.

Yaoliang Yu received his B.S. and M.S. degrees from Electronic Engineering Department of Fudan University, Shanghai, China, in 2005 and 2008, respectively. He is currently a Ph.D. student at the Department of Computing Science, University of Alberta, Canada. His research interests include machine learning and pattern recognition.

Jiayan Jiang received his B.S. degree in 2004 from Electronic Engineering Department of Fudan University, Shanghai, China. Since 2004 he has been a doctoral student in the Image and Intelligence Laboratory. His research interests include face recognition, computer vision, and machine learning. He is now at University of California, Los Angeles, USA.

Liming Zhang received her B.S. degree in Physics from Fudan University, Shanghai, China, in 1965. From 1986 to 1988, she was a visiting scholar at the Electrical Engineering Department, University of Notre Dame, South Bend, IN, USA. In 1996, she was a senior visiting scholar at Munich Technology University, Munich, Germany. Now she is a full professor and Doctoral Advisor in the Department of Electronic Engineering, Fudan University and a leader of Image and Intelligence Laboratory. Since 1986, she has been engaged in artificial neural network, machine learning, feature selection and pattern recognition of images and objects, including face recognition, brain-like robot, etc. Her group has accomplished more than 10 projects supported by climbing programs, notional key projects, natural sciences foundation, Shanghai science and technology committee, etc. She has published more than 120 papers on important national and international journals and conference proceedings concentrated on pattern recognition, machine learning, and neural networks.