Robust Metric Learning on Grassmann Manifolds with Generalization Guarantees

Lei Luo[1], **Jie Xu**[1,2], **Cheng Deng**[2], **Heng Huang**[1,3*]
[1]Electrical and Computer Engineering, University of Pittsburgh, USA
[2]School of Electronic Engineering, Xidian University, Xian, Shaanxi, China, [3]JDDGlobal.com
lel94@pitt.edu, jie.xu@pitt.edu, chdeng.xd@gmail.com, heng.huang@pitt.edu

Abstract

In recent research, metric learning methods have attracted increasing interests in machine learning community and have been applied to many applications. However, the existing metric learning methods usually use a fixed L_2-norm to measure the distance between pairwise data samples in the projection space, which cannot provide an effective mechanism to automatically remove the noise that exist in data samples. To address this issue, we propose a new robust formulation of metric learning. Our new model constructs a projection from higher dimensional Grassmann manifold into the one in a relative low-dimensional with more discriminative capability, where the errors between sample points are considered as an MLE (maximum likelihood estimation)-like estimator. An efficient iteratively reweighted algorithm is derived to solve the proposed metric learning model. More importantly, we establish the generalization bounds for the proposed algorithm by utilizing the techniques of U-statistics. Experiments on six benchmark datasets clearly show that the proposed method achieves consistent improvements in discrimination accuracy, in comparison to state-of-the-art methods.

Introduction

A large number of machine learning algorithms involve the use of a distance metric over the original input space. An effective distance, which successfully captures the important interrelations among data, can significantly improve the performance of algorithms. The Mahalanobis distance generalizes the standard Euclidean distance by admitting arbitrary linear scalings and rotations of the feature space. It has shown promising results in many machine learning tasks. Recent studies on metric learning mainly focus on learning an optimal Mahalanobis distance in supervised settings.

The goal of Mahalanobis metric learning is to seek a square matrix $\mathbf{M} \in \mathcal{R}^{d\times d}$ from the training set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} (\in \mathcal{R}^{d\times n})$, under which the distance between any two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ in $\mathbf{X}$ can be computed as: $d_{\mathbf{M}}(\mathbf{x}_i,\ \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)}$. Most of Mahalanobis metric learning methods utilize weakly-supervised constraints to induce a powerful distance metric. Representative methods include Information Theoretic Metric Learning (ITML) (Davis et al. 2007), Large Margin Nearest Neighbor (LMNN) (Weinberger and Saul 2009), Logistic Discriminant Metric Learning (LDML) (Guillaumin, Verbeek, and Schmid 2009).

The most popular way to enhance robustness of metric learning is to incorporate the structural information of metric matrix $\mathbf{M}$ or each sample $\mathbf{x}_i$ into models. To this end, some regularized metric learning methods (Law, Thome, and Cord 2014; Huo, Nie, and Huang 2016) use the sparse or low-rank regularization to characterize the structure of $\mathbf{M}$.Robust Structural Metric Learning (RSML) (Lim, McFee, and Lanckriet 2013) enforces group sparsity structure on the learned transformation. The maximum correntropy criterion was introduced to improve the robustness (Xu et al. 2018c). In (Xu et al. 2018b), the Wasserstein distance was utilized to characterize the errors between pairwise samples. The matrix variate Gaussian mixture distribution was adopted to model the structure of the metric matrix (Luo and Huang 2018).

More recently, a trend of integrating different traditional techniques into metric learning has emerged. For instance, RDML (Yi et al. 2012) uses the matrix completion technique to rectify the noisy pairwise similarities in metric learning. Hamming Distance Metric Learning (Norouzi, Fleet, and Salakhutdinov 2012) formulates the hashing problem by preserving the relative similarity defined over triplets of items. Hierarchical multimodal metric learning (Zhang, Patel, and Chellappa 2017) efficiently learns multiple metrics for multimodal data while fully exploiting the relationships among these metrics. Joint intensity metric learning method (Makihara et al. 2017) induces the dissimilarity using a bilinear form of joint intensity and spatial metrics, and alternately optimizes it by linear SVM or ranking SVM. Metric Learning with multiple Kernels (Wang et al. 2011) combines metric learning and multiple kernel learning, while deep metric learning method in (Oh Song et al. 2016) defines a novel structured prediction objective on the lifted pairwise distance matrix during the neural network training. The bilevel model was introduced to unify the metric learning and dictionary learning (Xu et al. 2018a).

It is worth noting that the metric matrix $\mathbf{M}$ can be decomposed as: $\mathbf{M} = \mathbf{P}^T\mathbf{P}$, where $\mathbf{P} \in \mathcal{R}^{p\times d}$ and $p \leq d$,

*Corresponding Author. L. Luo, H. Huang were partially supported by U.S. NSF IIS 1836945, IIS 1836938, DBI 1836866, IIS 1845666, IIS 1852606, IIS 1838627, IIS 1837956.

since it is symmetric and positive semi-definite. As a result, $d_{\mathbf{M}}(\mathbf{x}_i,\ \mathbf{x}_j) = \|\mathbf{P}(\mathbf{x}_i - \mathbf{x}_j)\|_2$. This means that learning a Mahalanobis distance metric $\mathbf{M}$ is equivalent to seeking a linear transformation $\mathbf{P}$ which projects each sample $\mathbf{x}_i$ into a low dimensional subspace. In the above mentioned methods, the distance between $\mathbf{P}\mathbf{x}_i$ and $\mathbf{P}\mathbf{x}_j$ of the low dimensional subspace is measured via the traditional L_2-norm, which is not robust to data noise. On the other hand, these methods are all dependent on the hypothesis that the data are obtained from an Euclidean vector space. This is often not held in many practical applications, where the linear subspaces with the same dimensionality reside on a special type of Riemannian manifold, *i.e.*, Grassmann manifold, which has a nonlinear structure (Yu and Zhang 2010). A few existing works overcome the above drawbacks by exploiting typical Riemannian geometries for kernel embedding (Huang et al. 2015). However, they are only confined to pairwise constraints and use a fixed matrix norm to measure the differences between any two samples projected onto the low dimensional manifold. More importantly, the theoretical guarantee of metric learning with triplet constraints on manifolds is missing.

Our Contributions. In this paper, we address the Metric learning problem via constructing a projection from higher dimensional Grassmann manifold into the one in a relative low-dimensional with more discriminative capability. Different from other existing approaches which use a fixed norm to characterize errors, our method provides an effective mechanism to automatically remove the noise that exist in samples since our model actually performs a robust regression-like process on the Grassmann manifold. The contributions of this paper can be summarized as follows:

• We provide a robust metric learning method with triplet constraints based on Grassmann manifold. We measure the error between data points in the low-dimensional manifold using an MLE-like estimator, which minimizes some function associated with the distribution of errors. We apply this strategy to LMNN and transform the minimization problem, named Robust LMNN (RLMNN), into an iteratively reweighted regression-like problem.

• The complete theoretical guarantee for our model is established. This is the first time to establish the generalization analysis of metric learning with triplet constraints on Grassmann manifold. Leveraging the techniques of U-statistics analysis, we derive the generalization bound of our new model. Our results can be extended to other metric learning models on Grassmann manifolds.

• We conduct extensive experiments, including face verification (associated with images, videos, and kinship), video and image classification, on six benchmark datasets. The experimental results demonstrate that our method can achieve superior performance as compared to some newest methods.

Notations

Throughout this paper, let $\mathbf{y} = \{y_1, y_2, \cdots, y_n\}$ be the label set of input (or training) samples $\mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_n\}$, where each $\mathbf{A}_i \in \mathcal{R}^{d\times p}$ $(i = 1, 2, \cdots, n)$. For example, the label of sample $\mathbf{A}_i$ is y_i. $r(y_i, y_l) = 1$ if $y_i \neq y_l$ otherwise $r(y_i, y_l) = 0$. Suppose input samples $\mathbf{A}$ and labels $\mathbf{y}$ are contained in an input space $\mathcal{A}$ and a label space $\mathcal{Y}$, respectively. Denote $\mathcal{B} = \mathcal{A} \times \mathcal{Y}$ and assume $\mathcal{B} := \{\mathbf{B}_i = (\mathbf{A}_i, y_i) \in \mathcal{B} : i \in \mathcal{N}_n\}$, where $\mathcal{N}_n = \{1, 2, \cdots, n\}$. For any $x \in \mathcal{R}$, the function $f(x) = [x]_+$ is equal to x if $x > 0$ and zero otherwise. For any $\mathbf{X}, \mathbf{Z} \in \mathcal{R}^{d\times p}$, let $\langle \mathbf{X}, \mathbf{Z}\rangle = \mathbf{trace}(\mathbf{X}^T\mathbf{Z})$, where $\mathbf{trace}(\cdot)$ denotes the trace of a matrix. $\circ$ is the Hadamard product and $\mathcal{E}(\cdot)$ is the Expectation of a random variable. For any matrix-norm $\|\cdot\|$, its dual norm $\|\cdot\|_*$ is defined, for any $\mathbf{X}$, by $\|\mathbf{X}\|_* = sup_{\|\mathbf{Z}\|\leq 1}\, \mathbf{trace}(\mathbf{X}^T\mathbf{Z})$.

Backgrounds

Regularized LMNN

We revisit the well-known LMNN (Large Margin Nearest Neighbor) metric learning method (Weinberger and Saul 2009). The LMNN objective includes two terms (one for each neighborhood objective): First, it reduces the distance between an instance and its target neighbors (*i.e.*, the same labeled inputs), thus pulling them closer and making the inputs local neighborhood smaller. Second, it moves impostor neighbors (*i.e.*, differently labeled inputs) farther away so that the distances to impostors should exceed the distances to target neighbors by a large margin. Let

$$\begin{aligned}\varepsilon_{\mathbf{z}}(\mathbf{M}) = \sum_{i,j\rightsquigarrow i} \mathcal{D}_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) + \mu \sum_{i,j\rightsquigarrow i,l} r(y_i, y_l) \\ \cdot [1 + \mathcal{D}_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) - \mathcal{D}_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_l)]_+,\end{aligned} \tag{1}$$

where $\mathcal{D}_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)$ and the parameter $\mu > 0$. The notation $j \rightsquigarrow i$ indicates that input $\mathbf{x}_j$ is a target neighbor of input. When $r(y_i, y_l) = 1$, $\mathbf{x}_l$ is an impostor neighbor of $\mathbf{x}_i$.

The optimization problem for LMNN is as follows:

$$\min_{\mathbf{M}}\ \varepsilon_{\mathbf{z}}(\mathbf{M})\,. \tag{2}$$

To avoid the problem of overfitting, many regularizations over matrix $\mathbf{M}$ can be added into (2). Here, we focus on four matrix-norm regularizations: (I) Squared Frobenius-norm: $\|\mathbf{M}\|_F^2 = \sum_{i=1}^d \sum_{j=1}^d M_{ij}^2$, where M_{ij} denotes the matrix element at the i-th row and j-th column of $\mathbf{M}$; (II) L_1-norm: $\|\mathbf{M}\|_1 = \sum_{i=1}^d \sum_{j=1}^d |M_{ij}|$; (III) mixed (2,1)-norm: $\|\mathbf{M}\|_{(2,1)} = \sum_{i=1}^d (\sum_{j=1}^d M_{ij}^2)^{\frac{1}{2}}$; (IV) trace norm: $\|\mathbf{M}\|_{tr} = \sum_{i=1}^d \delta_i$, where $\delta_1, \delta_2, \cdots, \delta_d$ are the singular values of $\mathbf{M}$.

Denoting the objective function in (2) by $\varepsilon_{\mathbf{z}}(\mathbf{M})$, the optimization problem (2) with matrix-norm regularization can be simplified as:

$$\mathbf{M}_{\mathbf{z}} = \underset{\mathbf{M}\in\mathcal{M}}{argmin}\ (\varepsilon_{\mathbf{z}}(\mathbf{M}) + \lambda\Lambda(\mathbf{M})), \tag{3}$$

where $\Lambda(\cdot)$ is Squared Frobenius-norm, sparse L_1-norm, mixed $(2,1)$-norm or trace norm, $\lambda > 0$ is a balance parameter and $\mathcal{M}$ denotes the feasible region of $\mathbf{M}$.

Grassmann Manifold

A manifold is locally similar to Euclidean space around each point of the manifold. Especially, Grassmann manifold has been successfully applied to many research communities

such as subspace clustering (Shen, Krim, and Gu 2016) and dictionary learning (Harandi et al. 2015). The definition of Grassmann Manifold can be written as:

Definition 1 (***Grassmann Manifold***) (Absil, Mahony, and Sepulchre 2009) *The Grassmann manifold, denoted by* $G(p;d)$, *consists of all the* p*-dimensional subspaces embedded in* d*-dimensional Euclidean space* $\mathcal{R}^d (0 \leq p \leq d)$.

It easy to see that an element of $G(p;d)$ is a linear subspace $span(\mathbf{U})$. Here, we assume that $span(\mathbf{U})$ is spanned by its orthonormal basis matrix $\mathbf{U}$ of size $d \times p$ such that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$, where $\mathbf{I}_p$ is the identity matrix of size $p \times p$. Therefore, Grassmann manifold can be embedded into symmetric matrices space as, $\Pi : G(p;d) \rightarrow Sym(d)$, $\Pi(\mathbf{U}) = \mathbf{U}\mathbf{U}^T$. Under the projection mapping $\Pi(\cdot)$, we can represent the elements on the Grassmann manifold with projection matrices $\mathbf{U}\mathbf{U}^T$, which is helpful for designing algorithms on the Grassmann manifold.

There are several possible choices to define a distance on the Grassmann manifold. The most popular way is to embed the Grassmann manifold into symmetric matrices space where the Euclidean metric is available. Due to its effectiveness, we use Embedding Distance in this paper.

Definition 2 (*Embedding Distance*) (Harandi et al. 2015) *Given Grassmann points* $\mathbf{U}_1$ *and* $\mathbf{U}_2$, *the corresponding distance on Grassmann manifold can be defined as:*

$$dist_g^2(\mathbf{U}_1, \mathbf{U}_2) = \frac{1}{2} \parallel \Pi(\mathbf{U}_1) - \Pi(\mathbf{U}_2) \parallel_F^2 . \qquad (4)$$

Robust LMNN on Grassmann Manifold

As discussed in Section 1, conventional metric learning methods use a fixed L_2-norm to measure the distance between any two samples projected onto the low dimensional subspace, which is sensitive to both sample and feature noise. Especially the distributions of data in practical applications are extremely complicated and far from Gaussian or Laplace distribution (Luo et al. 2015). Thus, L_2-norm (even L_1-norm) cannot characterize these noise. To tackle this problem, in this section, we propose a new robust metric learning model based on Grassmann manifold, which can effectively deal with noisy data. We first formulate the problem of our method for the video classification task. After that, we describe the optimization of our problem.

The projection on Grassmann manifold. Assume n video sequences of face frames are given as $\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$, where each $\mathbf{X}_i \in \mathcal{R}^{D \times n_i}$ describes a data matrix of the i-th video containing n_i frames, each frame being expressed as a D-dimensional feature vector. In these data, each video belongs to one of face classes denoted by C_i. The i-th video $\mathbf{X}_i$ is represented by a q-dimensional linear subspace spanned by an orthonormal basis matrix $\mathbf{H}_i \in \mathcal{R}^{D \times P}$, *s.t.* $\mathbf{X}_i\mathbf{X}_i^T \simeq \mathbf{H}_i\mathbf{\Sigma}_i\mathbf{H}_i^T$, where $\mathbf{\Sigma}_i$ and $\mathbf{H}_i$ correspond to matrices of the p largest eigenvalues and eigenvectors respectively.

Remark 1. For images, we can use the similar strategy as in (Xu et al. 2014) to construct $\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$.

Next, we seek to learn a projection $\mathbf{P}$ that maps high-dimensional Grassmann point $\mathbf{H}_i \in G(p;D)$ to a point $\mathbf{L}_i$ in a relative low-dimensional Grassmann manifold $G(p;d)$, where $D > d$, that is, $\mathbf{L}_i = \mathbf{P}^T\mathbf{H}_i$, where $\mathbf{P} \in \mathcal{R}^{D \times d}$ is a transformation matrix of column full rank.

As discussed before, only the linear subspaces spanned by orthonormal basis matrix can form a valid Grassmann manifold. In (4), however, except the case that $\mathbf{P}$ is an orthogonal matrix, $\mathbf{P}^T\mathbf{H}_i$ is not generally an orthonormal basis matrix. To tackle this issue, we perform QR decomposition on matrix $\mathbf{L}_i$ as follows,

$$\mathbf{L}_i = \mathbf{P}^T\mathbf{H}_i = \mathbf{Q}_i\mathbf{R}_i \Rightarrow \mathbf{Q}_i = \mathbf{P}^T\widetilde{\mathbf{Z}}_i, \qquad (5)$$

where $\widetilde{\mathbf{Z}}_i = \mathbf{H}_i\mathbf{R}_i^{-1} \in \mathcal{R}^{D \times p}$ denotes the normalized $\mathbf{H}_i$, $\mathbf{Q}_i \in \mathcal{R}^{d \times p}$ is an orthogonal matrix and $\mathbf{R}_i \in \mathcal{R}^{p \times p}$ is an invertible upper-triangular matrix. Since both $\mathbf{L}_i$ and $\mathbf{Q}_i$ generate the same (columns) subspace, the orthogonal matrix $\mathbf{Q}_i$ (or $\mathbf{P}^T\widetilde{\mathbf{Z}}_i$) can be used as the representative of the low-dimensional Grassmann point mapped from $\mathbf{H}_i$. Therefore, the distance between $\mathbf{Q}_i$ and $\mathbf{Q}_j$ on Grassmann manifold can be computed as:

$$\begin{aligned} \mathrm{dist}_g^2(\mathbf{Q}_i, \mathbf{Q}_j) &= \mathrm{dist}_g^2(\mathbf{P}^T\widetilde{\mathbf{Z}}_i, \mathbf{P}^T\widetilde{\mathbf{Z}_j}) \\ &= \frac{1}{2}\|\mathbf{P}^T\widetilde{\mathbf{Z}}_i\widetilde{\mathbf{Z}}_i{}^T\mathbf{P} - \mathbf{P}^T\widetilde{\mathbf{Z}_j}\widetilde{\mathbf{Z}_j}{}^T\mathbf{P}\|_F^2 \\ &= \frac{1}{2}\|\mathbf{P}^T\mathbf{G}_{ij}\mathbf{P}\|_F^2, \end{aligned} \qquad (6)$$

where $\mathbf{G}_{ij} = \widetilde{\mathbf{Z}}_i\widetilde{\mathbf{Z}}_i{}^T - \widetilde{\mathbf{Z}_j}\widetilde{\mathbf{Z}_j}{}^T$, which is a symmetric matrix of size $D \times D$.

Robust LMNN model on Grassmann manifold. We propose to find an MLE-like solution of matrix $\mathbf{P}$. Denote $\mathbf{E}_{ij} = (e_{ij}^{lm})_{d \times d} = \mathbf{P}^T\widetilde{\mathbf{Z}}_i\widetilde{\mathbf{Z}}_i{}^T\mathbf{P} - \mathbf{P}^T\widetilde{\mathbf{Z}_j}\widetilde{\mathbf{Z}_j}{}^T\mathbf{P}$, where e_{ij}^{lm} is an element in the l-th row and m-th column of $\mathbf{E}_{ij}$. Assume that all e_{ij}^{lm} $(l = 1, \cdots, d, m = 1, \cdots, d)$ are independently and identically distributed according to some unknown probability density function (PDF) $\mathcal{P}_\Omega(e_{ij}^{lm})$, where Ω denotes the parameter set that characterizes the distribution. Denote $\mathbf{E} = \{\mathbf{E}_{ij} : i, j = 1, 2, \cdots, n, j \neq i\}$, then the likelihood of the estimator can be written as $\mathcal{L}_\Omega(\mathbf{E}) = \prod_{i,j \in \mathcal{N}_n j \neq i}^{n} \prod_{l,m=1}^{d} \mathcal{P}_\Omega(e_{ij}^{lm})$. The MLE aims to maximize this likelihood function or, equivalently, minimize the objective function: $-\ln\mathcal{L}_\Omega = \sum_{i,j \in \mathcal{N}_n, j \neq i} \sum_{l,m=1}^{d} \rho_\Omega(e_{ij}^{lm})$, where $\rho_\Omega(e_{ij}^{lm}) = -\ln\mathcal{P}_\Omega(e_{ij}^{lm})$.

With consideration of the constraint of $\mathbf{P}$, the MLE of $\mathbf{P}$ can be formulated as the following minimization:

$$\min_{\mathbf{P}} \sum_{i,j \in \mathcal{N}_n j \neq i} \sum_{l,m=1}^{d} \rho_\Omega((\mathbf{P}^T\mathbf{G}_{ij}\mathbf{P})^{lm}), \text{ s.t. } \Lambda(\mathbf{P}) \leq \nu, \qquad (7)$$

where $\nu > 0$ and $\mathbf{E}_{ij}^{lm}$ denotes an element in the l-th row and m-th column of $\mathbf{E}_{ij}$. In fact, the above problem is similar to an M-estimator, which is a popular robust technique. M-estimators try to reduce the effect of outlier by replacing the traditional squared residual by novel function $\rho_\Omega(\cdot)$. This technique requires $\rho_\Omega(\cdot)$ to be symmetrical, positive definite and monotonically increasing.

The model (7) can be ultimately approximated by

$$\min_{\mathbf{P}} \sum_{i,j \in \mathcal{N}_n j \neq i} \|\mathbf{W}_{ij} \circ (\mathbf{P}^T\mathbf{G}_{ij}\mathbf{P})\|_F^2, \text{ s.t. } \Lambda(\mathbf{P}) \leq \nu, \quad (8)$$

where the (l,m)-th element of $\mathbf{W}_{ij}$ is

$$w_\Omega(e_{ij}^{lm}) = \frac{1}{e_{ij}^{lm}} \frac{d\rho_\Omega(e_{ij}^{lm})}{de_{ij}^{lm}}, \tag{9}$$

and $\frac{d\rho_\Omega(e_{ij}^{lm})}{de_{ij}^{lm}}$ denotes the the derivative of $\rho_\Omega(e_{ij}^{lm})$ with regard to e_{ij}^{lm}. The detailed derivation is given in Supplemental Materials.

Let $\mathbf{B}_i = (\mathbf{A}_i, y_i)$, where $\mathbf{A}_i = \widetilde{\mathbf{Z}}_i \widetilde{\mathbf{Z}}_i^T$, and

$$\varepsilon_{\mathbf{B},\mathbf{W}}(\mathbf{P}) = \frac{1}{n(n-1)(n-2)} \sum_{i,j \rightsquigarrow i,l} \Phi_{\mathbf{P}}(\mathbf{B}_i, \mathbf{B}_j, \mathbf{B}_l), \tag{10}$$

where

$$\Phi_{\mathbf{P}}(\mathbf{B}_i, \mathbf{B}_j, \mathbf{B}_l) = \frac{1}{v}\|\mathbf{W}_{ij} \circ (\mathbf{P}^T \mathbf{G}_{ij} \mathbf{P})\|_F^2 + \mu r(y_i, y_l) \cdot [1 + \|\mathbf{W}_{ij} \circ (\mathbf{P}^T \mathbf{G}_{ij} \mathbf{P})\|_F^2 - \|\mathbf{W}_{il} \circ (\mathbf{P}^T \mathbf{G}_{il} \mathbf{P})\|_F^2]_+. \tag{11}$$

The above pairwise model (8) can be further extended to triplet model. Thereby we can propose a robust LMNN (RLMNN) model with regularization terms based on Grassmann manifold:

$$\mathbf{P}_{\mathbf{B},\mathbf{W}} = \underset{\mathbf{P} \in \mathcal{P}}{argmin} \, (\varepsilon_{\mathbf{B},\mathbf{W}}(\mathbf{P}) + \lambda \Lambda(\mathbf{P})), \tag{12}$$

where $\lambda > 0$ and $\mathcal{P}$ denotes the feasible region of $\mathbf{P}$. If we obtain the optimal objection matrix $\mathbf{P}$, we can use Eq. 4 to measure the distance between any two samples on Grassmann manifold.

Proposed algorithm. The performance of the model is dependent on the choice of weight function. Considering that the logistic function has properties similar to the hinge loss in SVM and has been widely applied to robust regression problem, in this paper, we choose it as the weight function, *i.e.*,

$$w_\Omega(e_{ij}^{lm}) = (exp(\eta\varrho - \eta(e_{ij}^{lm})^2)/(1+exp(\eta\varrho - \eta(e_{ij}^{lm})^2)))^{1/2}, \tag{14}$$

where η and ϱ are positive scalars. Parameter ϱ controls the decreasing rate from 1 to 0, and η controls the location of demarcation point. Especially, by $\rho(\cdot)$, we can find the PDF $\mathcal{P}(\cdot)$ of e_{ij}^{lm}.

We propose to use an iteratively reweighted algorithm to optimize model (12). The detailed iteration process for solving model (12) is listed in Algorithm 1. Each iteration of Algorithm 1 involves a key subproblem (13), which can be optimized via subgradient method, *i.e.*,

$$\mathbf{P} \Leftarrow \mathbf{P} - \mathfrak{a} \nabla_{\mathbf{P}}, \tag{15}$$

where $\nabla_{\mathbf{P}}$ is a subgradient of the objection (13) with regard to $\mathbf{P}$ and $\mathfrak{a} > 0$ is a step size.

Convergence. By the constraint conditions:

$$\varepsilon_{\mathbf{B},\mathbf{W}^{(t+1)}}(\mathbf{P}^{t+1}) + \lambda\Lambda(\mathbf{P}^{t+1}) \leq \varepsilon_{\mathbf{B},\mathbf{W}^{(t)}}(\mathbf{P}^t) + \lambda\Lambda(\mathbf{P}^t), \tag{16}$$

it is easy to see that the objective (12) is monotonically decreasing with the iteration number. Considering that the objective (12) is nonnegative, the convergence of Algorithm 1 can be guaranteed.

Algorithm 1 RLMNN via Iteratively Reweighted Method

Input: training samples $\mathbf{X}$ and parameters λ, μ, η, ϱ.
Initialization: $\mathbf{P}^0 = \mathbf{I}_{D\times d}$, where $\mathbf{I}_{D\times d}$ denotes the $D \times d$ identity matrix.
While *Stopping criteria is not satisfied* **do**

1. Calculate the error $\mathbf{E}_{ij}^{(t+1)} = {\mathbf{P}^{(t)}}^T \mathbf{G}_{ij} \mathbf{P}^{(t)}$;

2. Estimate weight matrix $\mathbf{W}_{ij}^{(t+1)}$ by (12). Denoting $e_{ij,t+1} = e_{ij}^{(t+1)}$, then (l,m)-th diagonal element of $\mathbf{W}_{ij}^{(t+1)}$ is

$$w_\Omega(e_{ij,t+1}^{lm}) = exp(\eta\varrho - \eta((e_{ij,t+1}^{lm})^2)/(1 + exp(\eta\varrho - \eta((e_{ij,t+1}^{lm})^2));$$

3. Calculate the optimal $\mathbf{P}^*$ of the following problem:

$$\mathbf{P}^* = \underset{\mathbf{P} \in \mathcal{P}}{argmin} \, (\varepsilon_{\mathbf{z},\mathbf{W}^{(t)}}(\mathbf{P}) + \lambda\Lambda(\mathbf{P})), \tag{13}$$

4. Update $\mathbf{P}^{(t+1)}$: If $t = 1$, $\mathbf{P}^{(t+1)} = \mathbf{P}^*$; if $t > 1$, $\mathbf{P}^{(t+1)} = \mathbf{P}^{(t)} + \varsigma^{(t+1)}(\mathbf{P}^* - \mathbf{P}^{(t)})$, where $0 < \varsigma^{(t+1)} < 1$ is the step size, and a suitable $\varsigma^{(t+1)}$ should make $\varepsilon_{\mathbf{z},\mathbf{W}^{(t+1)}}(\mathbf{P}^{t+1}) + \lambda\Lambda(\mathbf{P}^{t+1}) \leq \varepsilon_{\mathbf{z},\mathbf{W}^{(t)}}(\mathbf{P}^t) + \lambda\Lambda(\mathbf{P}^t)$. ($\varsigma^{(t+1)}$ can be searched from 1 to 0 by the standard line search process.)

Output: Optimal $\mathbf{P}^{(t+1)}$.

Generalization Bounds of Our Method

We will introduce some important definitions and Lemmas which are helpful for establishing the generalization bounds of our model.

Definition 3 (McDiarmid's inequality (McDiarmid 1989)). *We say the function $f : \prod_{k=1}^n \Delta_k \to \mathcal{R}$ (each Δ_k is a linear space) with bounded differences $\{c_k\}_{k=1}^n$ if, for all $1 \leq k \leq n$,*

$$\underset{\substack{z_1,\cdots,z_k',\\ z_k',\cdots,z_n}}{max} | f(z_1,\cdots,z_{k-1},z_k,z_{k+1},\cdots,\cdots,z_n) - f(z_1,\cdots,z_{k-1},z_k',z_{k+1},\cdots,\cdots,z_n) | \leq c_k. \tag{17}$$

Lemma 1 (McDiarmid's inequality (McDiarmid 1989)). Suppose $f : \prod_{k=1}^n \Delta_k \to \mathcal{R}$ with bounded differences $\{c_k\}_{k=1}^n$, then, for all $\iota > 0$, there holds

$$\mathcal{P}_{\mathbf{z}}\{f(\mathbf{z}) - \mathcal{E}_{\mathbf{z}} f(\mathbf{z}) \geq \iota\} \leq e^{-\frac{2\iota^2}{\sum_{k=1}^n c_k^2}}. \tag{18}$$

We need the following contraction property of the Rademacher averages which is essentially implied by Theorem 4.12 in (Ledoux and Talagrand 2013).

Lemma 2. Let $\mathcal{F}$ be a class of uniformly bounded real-valued function on (Δ, ϖ) and $h \in \mathcal{N}$. If for each $i \in \{1, 2, \cdots, h\}$, $\phi_i : \mathcal{R} \to \mathcal{R}$ is a function having a Lipschitz constant c_i, then for any $\{x_i\}_{i \in \mathcal{N}_h}$ and i.i.d. Rademacher

variable $\{\iota_i \in \{\pm 1\} : i \in \mathcal{N}_h\}$

$$\mathcal{E}_\iota(sup_{f\in\mathcal{F}} \sum_{i\in\mathcal{N}_h} \iota_i\phi_i(f(x_i))) \leq 2\mathcal{E}_\iota(\sup_{f\in\mathcal{F}} \sum_{i\in\mathcal{N}_h} c_i\iota_i f(x_i)). \tag{19}$$

Lemma 3. Given the i.i.d. random variables $\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_n \in \mathcal{B}$, then U-statistic of degree three with kernel $q : \mathcal{B}\times\mathcal{B}\times\mathcal{B} \to \mathcal{R}$ is defined as

$$U_n = \frac{1}{n!}\sum_{\pi} \frac{1}{\lfloor \frac{n}{3} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{3} \rfloor} q(\mathbf{B}_{\pi(i)}, \mathbf{B}_{\pi(\lfloor \frac{n}{3} \rfloor + i)}, \mathbf{B}_{\pi(2\lfloor \frac{n}{3} \rfloor + i)}). \tag{20}$$

where the sum is taken over all permutations π of $\{1, 2, \cdots n\}$.

Lemma 4. Let $q_\tau : \mathcal{B}\times\mathcal{B}\times\mathcal{B} \to \mathcal{R}$ be real-valued functions indexed by $\tau \in \mathcal{T}$, where $\mathcal{T}$ is some index set. If $\mathbf{B}_1, \cdots, \mathbf{B}_n$ are i.i.d., then we have that

$$\begin{aligned}
&\mathcal{E}[sup_{\tau\in\mathcal{T}} \frac{1}{n(n-1)(n-2)} \sum q_\tau(\mathbf{B}_{\pi(i)}, \mathbf{B}_{\lfloor \frac{n}{3} \rfloor + i}, \mathbf{B}_{2\lfloor \frac{n}{3} \rfloor + i})] \\
&\leq \mathcal{E}[sup_{\tau\in\mathcal{T}} \frac{1}{\lfloor \frac{n}{3} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{3} \rfloor} q_\tau(\mathbf{B}_{\pi(i)}, \mathbf{B}_{\lfloor \frac{n}{3} \rfloor + i}, \mathbf{B}_{2\lfloor \frac{n}{3} \rfloor + i})].
\end{aligned} \tag{21}$$

Proof. From the representation of U-statistics (20), we observe that

$$\begin{aligned}
&\mathcal{E}[sup_{\tau\in\mathcal{T}} \frac{1}{n(n-1)(n-2)} \sum q_\tau(\mathbf{B}_{\pi(i)}, \mathbf{B}_{\lfloor \frac{n}{3} \rfloor + i}, \mathbf{B}_{2\lfloor \frac{n}{3} \rfloor + i})] \\
&= \mathcal{E} sup_{\tau\in\mathcal{T}} \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor \frac{n}{3} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{3} \rfloor} q_\tau(\mathbf{B}_{\pi(i)}, \mathbf{B}_{\lfloor \frac{n}{3} \rfloor + i}, \mathbf{B}_{2\lfloor \frac{n}{3} \rfloor + i}) \\
&\leq \frac{1}{n!} \mathcal{E} sup_{\tau\in\mathcal{T}} \sum_{\pi} \frac{1}{\lfloor \frac{n}{3} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{3} \rfloor} q_\tau(\mathbf{B}_{\pi(i)}, \mathbf{B}_{\lfloor \frac{n}{3} \rfloor + i}, \mathbf{B}_{2\lfloor \frac{n}{3} \rfloor + i}) \\
&= \frac{1}{n!} \Sigma_\pi \mathcal{E} sup_{\tau\in\mathcal{T}} \frac{1}{\lfloor \frac{n}{3} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{3} \rfloor} q_\tau(\mathbf{B}_{\pi(i)}, \mathbf{B}_{\lfloor \frac{n}{3} \rfloor + i}, \mathbf{B}_{2\lfloor \frac{n}{3} \rfloor + i}) \\
&= \mathcal{E} sup_{\tau\in\mathcal{T}} \frac{1}{\lfloor \frac{n}{3} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{3} \rfloor} q_\tau(\mathbf{B}_{\pi(i)}, \mathbf{B}_{\lfloor \frac{n}{3} \rfloor + i}, \mathbf{B}_{2\lfloor \frac{n}{3} \rfloor + i}).
\end{aligned} \tag{22}$$

For clarity, $\Omega(\cdot)$ is written as $\|\cdot\|^2$. We can see $\| \mathbf{W}_{ij} \|$ is bounded since its each element $w_\Omega(e_{ij}^{lm}) \in [0\ 1]$. As a result, it is reasonable that we assume $\| \mathbf{W}_{ij} \|_F^2 \leq \gamma$, where $\gamma > 0$. Additionally, we define the expectation of $\varepsilon_{\mathbf{B},\mathbf{W}}(\mathbf{P})$ as

$$\varepsilon_{\mathbf{W}}(\mathbf{P}) = \int\int \Phi_{\mathbf{P}}(\mathbf{B}_i, \mathbf{B}_j, \mathbf{B}_l) d\rho(\mathbf{B}_i) d\rho(\mathbf{B}_j) d\rho(\mathbf{B}_l). \tag{23}$$

Let $\mathbf{P}_{\mathbf{B},\mathbf{W}}$ be the solution of formulation (12). With the help of the above results and techniques of U-statistic, we can estimate the bound of $\varepsilon_{\mathbf{W}}(\mathbf{P}_{\mathbf{B},\mathbf{W}}) - \varepsilon_{\mathbf{B},\mathbf{W}}(\mathbf{P}_{\mathbf{B},\mathbf{W}})$ in (12) as follows:

Theorem 1. For any $0 < \delta < 1$, with probability $1-\delta$ we have that

$$\begin{aligned}
&\varepsilon_{\mathbf{W}}(\mathbf{P}_{\mathbf{B},\mathbf{W}}) - \varepsilon_{\mathbf{B},\mathbf{W}}(\mathbf{P}_{\mathbf{B},\mathbf{W}}) \\
&\leq 2\left(\frac{\frac{1}{v}\gamma + 2\mu\gamma}{\sqrt{\lfloor \frac{n}{3} \rfloor}}\right) \frac{G_*}{\lambda^2} + \frac{\mu}{\sqrt{\lfloor \frac{n}{3} \rfloor}} + \sqrt{\frac{1}{2} n\varphi^2 ln\frac{1}{\delta}},
\end{aligned} \tag{24}$$

where $\varphi = \frac{2(nv+nu+uv)}{n(n-1)(n-2)} \left(\frac{\gamma}{v\lambda^2} G_* + \mu(1 + \frac{4\gamma}{\lambda^2} G_*)\right)$ and $G_* = max_{\mathbf{G}_{ij}\in\mathcal{G}} \| \mathbf{G}_{ij} \|_*^2$.

Theorem 1 provides a general framework for error bound of LMNN with any matrix norm regularization on Grassmann manifold and can be applied to other triplet models.

Experiments and Discussions

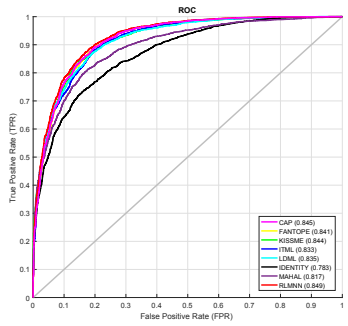
In this section, we evaluate the effectiveness of our method on six standard databases, including LFW database (Huang et al. 2007), PubFig database (Kumar et al. 2009), YouTube Face Database (Wolf, Hassner, and Maoz 2011), OSR database (Parikh and Grauman 2011) and Highway Traffic Database (Chan and Vasconcelos 2008). We design face verification experiments on PubFig, LFW and Youtube video datasets, video classification experiments on the Highway Traffic Database, image classification experiments on OSR, LFW and PubFig databases. Some existing methods, including R-MLR (Lim, McFee, and Lanckriet 2013), LMNN (Weinberger and Saul 2009), LMNN_Trace (Huo, Nie, and Huang 2016), LMNN_Fantope (Law, Thome, and Cord 2014), LMNN_Cap (Huo, Nie, and Huang 2016), KISSME (Koestinger et al. 2012), ITML (Davis et al. 2007), LDML (Guillaumin, Verbeek, and Schmid 2009), IDENTITY (Huo, Nie, and Huang 2016) and MAHAL (Huo, Nie, and Huang 2016), are compared with our method. We set $\lambda = 0.1, \mu = 0.5, \eta = 0.8$ and $\varrho = 0.1$ in our method.

Experiments on LFW Database

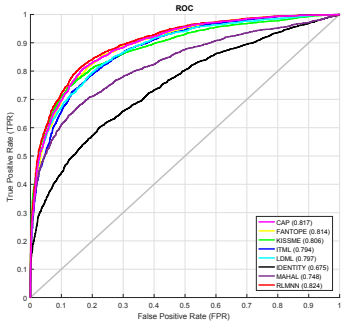
In this subsection, we implement face verification experiments on the Labeled Faces in the Wild (LFW) dataset. We utilize two different feature representations, *i.e.*, LFW Attribute feature and LFW SIFT feature datasets. The same experiment setting in (Huo, Nie, and Huang 2016) is adopted. ROC (Receiver Operator Characteristic) curves for all methods, including LMNN_Cap, Fantope, KISSME , ITML, LDML, Identity, MAHAL and RLMNN are plotted in Fig. 1(a) and 1(b). Meanwhile, we compute Equal Error Rate (EER) and use 1-EER values to evaluate the performance of these methods. Fig. 1(a) shows the results on LFW Attribute feature dataset. It is observed that Mahalanobis distance between two similar pairs performs better than Euclidean distance. It increases the performance from 78.3% to 81.7%. Comparing with Identity and Mahalanobis methods, KISSME achieves a significant improvement. The results of other methods with learning Mahalanobis distance are also competitive. For example, LMNN_Cap and Fantope reach 84.5% and 84.1%, respectively. For SIFT Feature dataset, we can find the similar phenomenon. However, our method consistently outperforms other methods.

Experiments on PubFig Face Database

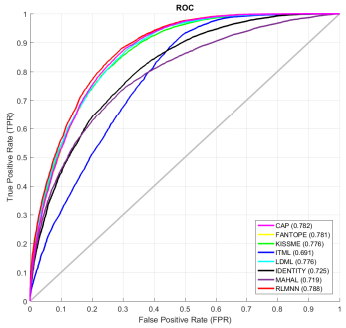
Two experiments are carried out on the PubFig Face Dataset (Kumar et al. 2009). In first experiment, we focus on face verification task using face verification benchmark dataset. Fig. 1(c) exhibits the performance of each compared method. It is obvious that our methods have remarkable priorities compared to the other algorithms. For example, the 1-EER values of LMNN_Cap, Fantope, KISSME,



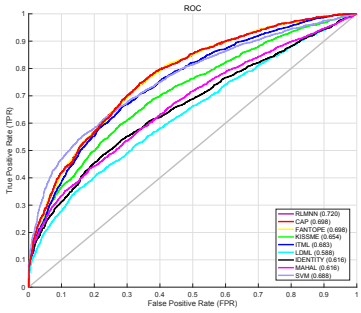
(a) Attribute Feature



(b) SIFT Feature



(c) PubFig datasets



(d) Youtube datasets

Figure 1: ROC curves of face verification on LFW (Attribute and SIFT features), Pubfig, and Youtube datasets

ITML, LDML, IDENTITY, MAHAL and RLMNN are 0.782, 0.781, 0.766, 0.725, 0.719 and 0.788, respectively. In the second experiment, we use a subset face images of Pubfig database, which include 771 images from 8 face categories. The experimental setting is the same as (Law, Thome, and Cord 2014). We run this experiment 5 times, where 30 images per person in training data are selected randomly each time, and the average performance is used as evaluation criterion. The average classification accuracies of all methods are shown in Table 1. We can find that RLMNN achieves an improvement of about 2.2% as compared to the second best method: LMNN_Cap. Meanwhile, it is clear that some methods based on LMNN, including LMNN_Trace, LMNN_fantope and LMNN_Cap, have the similar results.

Experiments on YouTube Face Database

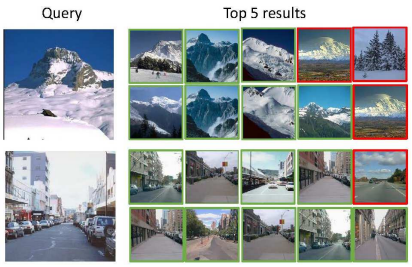
The YouTube Face (YTF) (Wolf, Hassner, and Maoz 2011) contains 3,425 videos of 1,595 different persons collected from the YouTube website. In this database, there exist large variations in pose, illumination, and expression in each video sequence. We follow the standard evaluation protocol (Wolf, Hassner, and Maoz 2011) to perform standard, ten-fold, cross validation, pairmatching tests. Specifically, we use the officially provided 5,000 video pairs, which are equally divided into 10 folds. Each fold contains 250 intrapersonal pairs and 250 interpersonal pairs. We directly crop the face images according to the provided data and then resize them into 24×40 pixels for YTF. The raw intensity feature of resized video frames is extracted. Fig. 1(d) shows the ROC for the video based face verification on YTF. It can observed that the performances of the low-rank metric learning methods such as LMNN_Cap and Fantope are similar. However, compared with state-of-the art methods, the advantages of our method (0.72) are still obvious.

Experiments on Outdoor Scene Recognition

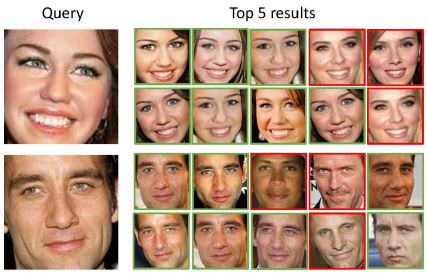
The Outdoor Scene Recognition (OSR) dataset from (Parikh and Grauman 2011) is used in this experiment. It includes 2688 images from 8 scene categories, which are described by high level attribute features. 30 images for each category are chosen as training data, and other images are used as testing data. The training data is randomly selected and this procedure is repeated 5 times. The average accuracy is used to evaluate the performance of each method. The detailed results of all methods are shown in the third line of Table 1. From Table 1, it is seen that the performance of some methods related to LMNN (including LMNN_cap, LMNN_trace and LMNN_cap) is comparative. As the previous experiment, the result of LMNN still lags behind some regularized methods, which implies the regularization is helpful for improving the performance of models. It is clear that our method achieves a leading performance: 78.1523%.

Databases	R-MLR	LMNN	LMNN_Trace	LMNN_Fantope	LMNN_Cap	RLMNN
PubFig	77.36 ±1.28	77.33 ±1.36	77.67 ±1.49	78.65±1.19	78.87 ±1.93	**82.75** ±1.32
OSR	76.26 ±1.63	75.32 ±1.61	76.02 ±1.84	76.47 ±1.97	76.51 ±1.45	**78.83** ±1.51
HT	87.37 ±1.92	83.87 ±1.86	83.87 ±2.37	91.34 ±1.32	83.87 ±1.29	**94.35** ±1.68

Table 1: The classification accuracy(%) and standard deviation of each method on three databases



(a) Results of 5 neareset neighbors when we query an image on OSR dataset. The first second shows the results of LMNN_Cap, and the second row is the results of our method.



(b) Results of 5 neareset neighbors when we query an image on Pubfig dataset. The first second shows the results of LMNN_Cap, and the second row is the results of our method.

Figure 2: Results of 5 neareset neighbors when we query an image. Green line means this neighbor is in the same class with query image, and red line denotes they are different.

Experiments on Highway Traffic Database

Highway Traffic (HT) dataset (Parikh and Grauman 2011) contains 253 video sequences of highway traffic. These sequences are labeled with three levels: 44 clips at heavy level, 45 clips at medium level and 164 clips at light level. Each video sequence has 42 to 52 frames. The video sequences are converted to gray images and each image is normalized to size 24×24. Experimental results for classification tasks are shown in the fourth line of Table 1. In addition to our method and LMNN_Fantope, the experimental accuracies of the other methods are very similar. Our method is at least 2 percent higher than the corresponding compared methods. Although both LMNN_Fantope and LMNN_Cap belong to the low-rank metric learning methods, LMNN_Fantope performs better than LMNN_Cap. Experimental results demonstrate that the low-dimensional Grassmann points generated by our proposed method include more discrimination than the traditional Euclidean vector space.

Conclusions

In this paper, we proposed a robust LMNN algorithm based on Grassmann manifold. Traditional metric learning methods are sensitive to the practical noise because they use L_2-norm to characterize the error between two samples in the projection space. Meanwhile, they assume that the data are from Euclidean vector space. Our method constructs a projection from higher dimensional Grassmann manifold into the one in a relative low-dimensional with more discriminative capability, where the errors between sample points are considered as an MLE (maximum likelihood estimation)-like estimator. We provided the complete theoretical guarantee for our method. Experimental results demonstrate the effectiveness of the proposed method.

References

[Absil, Mahony, and Sepulchre 2009] Absil, P.-A.; Mahony, R.; and Sepulchre, R. 2009. *Optimization algorithms on matrix manifolds*. Princeton University Press.

[Chan and Vasconcelos 2008] Chan, A. B., and Vasconcelos, N. 2008. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE transactions on pattern analysis and machine intelligence* 30(5):909–926.

[Davis et al. 2007] Davis, J. V.; Kulis, B.; Jain, P.; Sra, S.; and Dhillon, I. S. 2007. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, 209–216. ACM.

[Guillaumin, Verbeek, and Schmid 2009] Guillaumin, M.; Verbeek, J.; and Schmid, C. 2009. Is that you? metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th international conference on*, 498–505. IEEE.

[Harandi et al. 2015] Harandi, M.; Hartley, R.; Shen, C.; Lovell, B.; and Sanderson, C. 2015. Extrinsic methods for coding and dictionary learning on grassmann manifolds. *International Journal of Computer Vision* 114(2-3):113–136.

[Huang et al. 2007] Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst.

[Huang et al. 2015] Huang, Z.; Wang, R.; Shan, S.; and Chen, X. 2015. Projection metric learning on grassmann manifold with application to video based face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 140–149.

[Huo, Nie, and Huang 2016] Huo, Z.; Nie, F.; and Huang, H. 2016. Robust and effective metric learning using capped trace norm: Metric learning via capped trace norm. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1605–1614. ACM.

[Koestinger et al. 2012] Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2288–2295. IEEE.

[Kumar et al. 2009] Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, 365–372.

[Law, Thome, and Cord 2014] Law, M. T.; Thome, N.; and Cord, M. 2014. Fantope regularization in metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1051–1058.

[Ledoux and Talagrand 2013] Ledoux, M., and Talagrand, M. 2013. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

[Lim, McFee, and Lanckriet 2013] Lim, D.; McFee, B.; and Lanckriet, G. R. 2013. Robust structural metric learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 615–623.

[Luo and Huang 2018] Luo, L., and Huang, H. 2018. Matrix variate gaussian mixture distribution steered robust metric learning. In *Thirty-Second AAAI Conference on Arti cial Intelligence*, 3722–3729.

[Luo et al. 2015] Luo, L.; Yang, J.; Qian, J.; and Tai, Y. 2015. Nuclear-l 1 norm joint regression for face reconstruction and recognition with mixed noise. *Pattern Recognition* 48(12):3811–3824.

[Makihara et al. 2017] Makihara, Y.; Suzuki, A.; Muramatsu, D.; Li, X.; and Yagi, Y. 2017. Joint intensity and spatial metric learning for robust gait recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5705–5715.

[McDiarmid 1989] McDiarmid, C. 1989. Surveys in combinatorics, chapter on the methods of bounded differences.

[Norouzi, Fleet, and Salakhutdinov 2012] Norouzi, M.; Fleet, D. J.; and Salakhutdinov, R. R. 2012. Hamming distance metric learning. In *Advances in neural information processing systems*, 1061–1069.

[Oh Song et al. 2016] Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4004–4012.

[Parikh and Grauman 2011] Parikh, D., and Grauman, K. 2011. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 503–510. IEEE.

[Shen, Krim, and Gu 2016] Shen, X.; Krim, H.; and Gu, Y. 2016. Beyond union of subspaces: Subspace pursuit on grassmann manifold for data representation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 4079–4083. IEEE.

[Wang et al. 2011] Wang, J.; Do, H. T.; Woznica, A.; and Kalousis, A. 2011. Metric learning with multiple kernels. In *Advances in neural information processing systems*, 1170–1178.

[Weinberger and Saul 2009] Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10(Feb):207–244.

[Wolf, Hassner, and Maoz 2011] Wolf, L.; Hassner, T.; and Maoz, I. 2011. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 529–534. IEEE.

[Xu et al. 2014] Xu, C.; Wang, T.; Gao, J.; Cao, S.; Tao, W.; and Liu, F. 2014. An ordered-patch-based image classification approach on the image grassmannian manifold. *IEEE transactions on neural networks and learning systems* 25(4):728–737.

[Xu et al. 2018a] Xu, J.; Luo, L.; Deng, C.; and Huang, H. 2018a. Bilevel distance metric learning for robust image recognition. In *Neural Information Processing Systems*.

[Xu et al. 2018b] Xu, J.; Luo, L.; Deng, C.; and Huang, H. 2018b. Multi-level metric learning via smoothed wasserstein distance. In *27th International Joint Conference on Artificial Intelligence*, 2919–2925.

[Xu et al. 2018c] Xu, J.; Luo, L.; Deng, C.; and Huang, H. 2018c. New robust metric learning model using maximum correntropy criterion. In *24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2555–2564.

[Yi et al. 2012] Yi, J.; Jin, R.; Jain, S.; Yang, T.; and Jain, A. K. 2012. Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning. In *Advances in neural information processing systems*, 1772–1780.

[Yu and Zhang 2010] Yu, K., and Zhang, T. 2010. Improved local coordinate coding using local tangents. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 1215–1222.

[Zhang, Patel, and Chellappa 2017] Zhang, H.; Patel, V. M.; and Chellappa, R. 2017. Hierarchical multimodal metric learning for multimodal classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3057–3065.