Learning Semi-Riemannian Metrics for Semisupervised Feature Extraction

Wei Zhang, Zhouchen Lin, *Senior Member*, *IEEE*, and Xiaoou Tang, *Fellow*, *IEEE*

Abstract—Discriminant feature extraction plays a central role in pattern recognition and classification. Linear Discriminant Analysis (LDA) is a traditional algorithm for supervised feature extraction. Recently, unlabeled data have been utilized to improve LDA. However, the intrinsic problems of LDA still exist and only the similarity among the unlabeled data is utilized. In this paper, we propose a novel algorithm, called Semisupervised Semi-Riemannian Metric Map (S^3RMM), following the geometric framework of semi-Riemannian manifolds. S^3RMM maximizes the discrepancy of the separability and similarity measures of scatters formulated by using semi-Riemannian metric tensors. The metric tensor of each sample is learned via semisupervised regression. Our method can also be a general framework for proposing new semisupervised algorithms, utilizing the existing discrepancy-criterion-based algorithms. The experiments demonstrated on faces and handwritten digits show that S^3RMM is promising for semisupervised feature extraction.

Index Terms—Linear discriminant analysis, semisupervised learning, semi-Riemannian manifolds, feature extraction.

1 Introduction

Discriminant feature extraction is a central topic in pattern recognition and classification. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two traditional algorithms for linear feature extraction [1]. As the underlying structure of data may not be linear, some nonlinear feature extraction algorithms, e.g., Locality Preserving Projections (LPP) [11] and Linear Laplacian Discrimination (LLD) [51], have been developed. In addition, the kernel trick [19] is also widely used to extend linear feature extraction algorithms to nonlinear ones by performing linear operations in a higher or even infinite-dimensional space transformed by a kernel mapping function. Despite the success of LDA and its variants [13], [42], [51], it has been found to have some intrinsic problems [40]: singularity of within-class scatter matrices and limited available projection directions. Much work has been done to deal with these problems [7], [10], [35], [36], [37], [38], [39], [40], [41], [44]. Most of such work can be traced back to LDA and Fisher criterion, i.e., the structural analysis of classes by simultaneously maximizing the between-class scatter and minimizing the within-class scatter via the ratio of them.

The discrepancy criterion has been developed recently as an alternative way to avoid the intrinsic problems of LDA. Such kind of methods include Maximum Margin Criterion (MMC) [16], Kernel Scatter-Difference Analysis (KSDA) [18], Stepwise Nonparametric Maximum Margin Criterion (SNMMC) [25], Local and Weighted Maximum Margin Discriminant Analysis (LWMMDA) [33], Average Neighborhood Margin Maximization (ANMM) [32], and Discriminative Locality Alignment (DLA) [46]. It has also been found that the Fisher criterion can be well solved by iterative discrepancy criterions [34]. Zhao et al. have found that the discrepancy criterion can be adapted into the framework of semi-Riemannian manifolds [50]. They developed Semi-Riemannian Discriminant Analysis (SRDA) using this framework [50]. All these discrepancy-criterion-based methods are supervised methods.

In many real-world applications, labeled data are hard or expensive to obtain. This makes it necessary to utilize unlabeled data. Both labeled and unlabeled data can contribute to the learning process [3], [53]. Consequently, semisupervised learning, which aims at learning from both labeled and unlabeled data, has been a hot topic within the machine learning community [53]. Many semisupervised learning methods have been proposed, e.g., Transductive SVM (TSVM) [31], Cotraining [5], and graph-based semisupervised learning algorithms [3], [28], [52]. Semisupervised dimensionality reduction has been considered recently, e.g., semisupervised discriminant analysis (SDA [6] and SSDA [48]). However, SDA and SSDA also suffer from the problems of the Fisher criterion, as a result of which both of them use Tikhonov regularization to deal with the singularity problem as in regularized discriminant analysis [7]. In [43] a graph-based subspace semisupervised learning framework (SSLF) has been developed as a semisupervised extension of graph embedding [41] and several semisupervised algorithms, including SSLDA, SSLPP, and SSMFA, are provided. Supervised methods based on the discrepancy criterion have also been extended to the semisupervised case, e.g., Semisupervised Discriminative Locality Alignment (SDLA) is the semisupervised counterpart of DLA [46]. SDA, SSLF, and SDLA only utilize the smooth regularization on unlabeled or all data, while SSDA adds a term to capture the similarity between unlabeled data points and class centers of labeled data.

- *W. Zhang is with the Department of Information Engineering, The Chinese University of Hong Kong, P.R. China. E-mail: zw007@ie.cuhk.edu.hk.*
- *Z. Lin is with Visual Computing Group, Microsoft Research Asia, 5th Floor, Sigma Building, Zhichun Road 49#, Haidian District, Beijing 100190, P.R. China. E-mail: zhoulin@microsoft.com.*
- *X. Tang is with the Department of Information Engineering, The Chinese University of Hong Kong, P.R. China, and Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, P.R. China. E-mail: xtang@ie.cuhk.edu.hk.*

Manuscript received 29 Mar. 2009; revised 25 Sept. 2009; accepted 29 Dec. 2009; published online 24 Aug. 2010.
Recommended for acceptance by S. Zhang.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2009-03-0219.
Digital Object Identifier no. 10.1109/TKDE.2010.143.

However, the smooth regularization may not be the optimal constraints on samples. First, not all the neighbors of a sample have the same label. Second, they set the size of neighborhoods in advance, and then, there are no constraints between two samples if they are not neighbors. Thus, the discriminant information among unlabeled data is not well used.

In this paper, we propose a novel algorithm, Semisupervised Semi-Riemannian Metric Map (S^3RMM), for semisupervised dimensionality reduction. Our algorithm consists of two steps: learning semi-Riemannian metrics and pursuing the optimal low-dimensional projection. We formulate the problem of learning semi-Riemannian metric tensors as semisupervised regression. Labeled data are used to initialize the regression. Then, a fast and efficient graph-based semisupervised learning scheme is adopted and closed-form solutions are given. The optimal low-dimensional projection is obtained via maximizing the total margin of all samples encoded in semi-Riemannian metric tensors. Unlike previous manifold-based algorithms [2], [3], [26], [49] in which learning the manifold structure does not use any class labels, we construct the manifold structure using the partial labels. Labeled samples can help discover the structure, so our semi-Riemannian manifolds can be more discriminative. We utilize unlabeled data in two aspects: First, the unlabeled data help to estimate the geodesic distances between samples, so that the structure of all data is captured; second, the separability and similarity criteria between all sample points, including labeled and unlabeled data, are considered. In addition, our method provides a new general framework for semisupervised dimensionality reduction.

The rest of this paper is organized as follows: Section 2 recalls basic concepts of semi-Riemannian spaces. In Section 3, we begin with the discrepancy criterion and the semi-Riemannian geometry framework, then present our method of learning semi-Riemannian metrics, and finally summarize the S^3RMM algorithm. Section 4 discusses its extensions and relationships to the previous research. Section 5 shows the experimental results on face and handwritten digit recognition. Finally, we conclude this paper in Section 6.

2 Semi-Riemannian Spaces

Semi-Riemannian manifolds were first applied to supervised discriminant analysis by Zhao et al. [50].

A semi-Riemannian space is a generalization of a Riemannian space. The key difference between Riemannian and semi-Riemannian spaces is that in a semi-Riemannian space the metric tensor need not be positive definite. Semi-Riemannian manifolds (also called pseudo-Riemannian manifolds) are smooth manifolds furnished with semi-Riemannian metric tensors. The geometry of semi-Riemannian manifolds is called semi-Riemannian geometry. Semi-Riemannian geometry has been applied to Einstein's general relativity, as a basic geometric tool of modeling space-time in physics. One may refer to [21] for more details.

The metric of a semi-Riemannian manifold $\mathbb{N}_\nu^n$ is of the form

$$\mathbf{\Lambda} = \begin{bmatrix} \check{\mathbf{\Lambda}}_{p\times p}, & 0 \\ 0, & -\hat{\mathbf{\Lambda}}_{\nu\times\nu} \end{bmatrix},$$

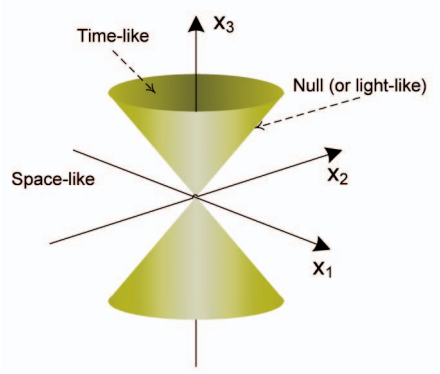


Fig. 1. An illustration of the space-time interval $\mathrm{d}s^2 = \mathrm{d}x_1^2 + \mathrm{d}x_2^2 - \mathrm{d}x_3^2$. The space-time interval is space-like outside the cone, null (or light-like) on the cone and time-like inside the cone.

where $\check{\mathbf{\Lambda}}_{p\times p}$ and $\hat{\mathbf{\Lambda}}_{\nu\times\nu}$ are diagonal and their diagonal entries are positive, and $p+\nu = n$. ν is called the index of $\mathbb{N}_\nu^n$. With $\mathbf{\Lambda}$, the space-time interval $\mathrm{d}s^2$ in $\mathbb{N}_\nu^n$ can be written as

$$\mathrm{d}s^2 = (\mathrm{d}\mathbf{x})^T \mathbf{\Lambda} \mathrm{d}\mathbf{x} = \sum_{i=1}^{p} \check{\mathbf{\Lambda}}(i,i)\mathrm{d}x_i^2 - \sum_{i=1}^{\nu} \hat{\mathbf{\Lambda}}(i,i)\mathrm{d}x_i^2.$$

The interval is called space-like if it is positive, time-like if it is negative, and null (or light-like) if it is zero. One may refer to Fig. 1 for an illustration of the space-time interval.

3 Semisupervised Semi-Riemannian Metric Map

In this paper, we focus on the problem of pursuing the optimal projection matrix $\mathbf{U}$ under the semisupervised setting, i.e., given l labeled samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$, each of which has a class label $c_i \in \{1, \ldots, c\}$, and m unlabeled samples $\{\mathbf{x}_{l+1}, \ldots, \mathbf{x}_n\}$ with unknown class memberships. In addition, we have $l+m=n$ and $\mathbf{x}_i \in \mathbb{R}^D$. With the optimal projection matrix, we project the samples into a low-dimensional space: $\mathbf{y}_i = \mathbf{U}^T\mathbf{x}_i$, $i = 1, \ldots, n$. Denote $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$.

3.1 The Discrepancy Criterion

Given only the labeled training set $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$ and the labels of each sample $c_i \in \{1, \ldots, c\}$, many dimensionality reduction methods aim at learning a projection $\mathbf{U}$ to maximize the within-class similarity and between-class separability. Traditional LDA maximizes the following ratio:

$$J = \frac{\mathrm{tr}(\mathbf{U}^T\mathbf{S}_b\mathbf{U})}{\mathrm{tr}(\mathbf{U}^T\mathbf{S}_w\mathbf{U})},$$

where $\mathbf{S}_b = \sum_{k=1}^{c} l_k(\overline{\mathbf{x}}_k - \overline{\mathbf{x}})(\overline{\mathbf{x}}_k - \overline{\mathbf{x}})^T$ is the between-class scatter matrix, $\mathbf{S}_w = \sum_{k=1}^{c}\sum_{c_i=k}(\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T$ is the within-class scatter matrix, and $\mathrm{tr}(\cdot)$ is the trace operator. $\overline{\mathbf{x}}_k = \frac{1}{l_k}\sum_{c_i=k}\mathbf{x}_i$ is the mean of the kth class, $\overline{\mathbf{x}} = \frac{1}{l}\sum_{i=1}^{n}\mathbf{x}_i$ is the mean of all data samples, and l_k is the number of samples in the kth class. This ratio is known as the *Fisher criterion*.

The *discrepancy criterion* [16], [32] defines two types of neighborhoods:

- *Homogeneous Neighborhoods* $\hat{\mathcal{N}}_i^{\hat{K}}$: the set of $\hat{K}$ most similar data in the same class of $\mathbf{x}_i$.

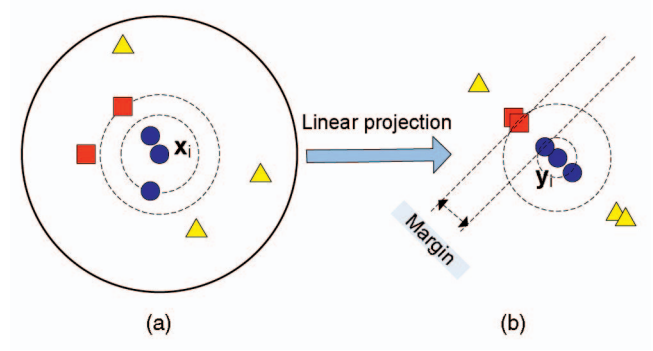


Fig. 2. An illustration of the margin maximization in the discrepancy criterion. The elements with the same shape belong to the same class. (a) $\mathbf{x}_i$ and its neighbors in the original 2D plane, among which circles except $\mathbf{x}_i$ are homogeneous neighbors, while squares and triangles belong to heterogeneous neighbors. (b) $\mathbf{y}_i$ and the projected neighbors.

- *Heterogeneous Neighborhoods* $\check{\mathcal{N}}_i^{\check{K}}$: the set of $\check{K}$ most similar data not in the same class of $\mathbf{x}_i$.

Taking ANMM [32] as an example, the average neighborhood margin γ_i for $\mathbf{x}_i$ in the projected space can be measured as

$$\gamma_i = \sum_{j \in \check{\mathcal{N}}_i^{\check{K}}} \frac{1}{\check{K}} \|\mathbf{y}_j - \mathbf{y}_i\|^2 - \sum_{j \in \hat{\mathcal{N}}_i^{\hat{K}}} \frac{1}{\hat{K}} \|\mathbf{y}_j - \mathbf{y}_i\|^2, \tag{1}$$

where $\|\cdot\|$ is the L2-norm. The maximization of such a margin can project high-dimensional data into a low-dimensional feature space with high within-class similarity and between-class separability. Fig. 2 gives an intuitive illustration of the discrepancy criterion.

3.2 Semi-Riemannian-Geometry-Based Feature Extraction Framework

The average neighborhood margin can be generalized in the framework of semi-Riemannian geometry. In contrast to the local semi-Riemannian metric tensors and the global alignment of local semi-Riemannian geometry in [50], we define global semi-Riemannian metric tensors to unify the discrepancy criterion. A global metric tensor encodes the structural relationship of all data samples to a sample, while in a local metric tensor only samples in neighborhoods are chosen. For a sample $\mathbf{x}_i$, its metric tensor $\mathbf{\Lambda}_i$ is a diagonal matrix with positive, negative, or zero diagonal elements:

$$\mathbf{\Lambda}_i(j,j) \begin{cases} > 0, & \text{if } \mathbf{x}_j \in \check{\mathcal{N}}_i^{\check{K}}, \\ < 0, & \text{if } \mathbf{x}_j \in \hat{\mathcal{N}}_i^{\hat{K}}, \\ = 0, & \text{if } \mathbf{x}_j \notin \check{\mathcal{N}}_i^{\check{K}} \quad \text{and} \quad \mathbf{x}_j \notin \hat{\mathcal{N}}_i^{\hat{K}}. \end{cases}$$

Then, the construction of the homogeneous and heterogeneous neighborhoods as well as the metric tensor do not need to follow those in Section 3.1.

The margin γ_i can be written as

$$\gamma_i = \sum_j \mathbf{\Lambda}_i(j,j) \|\mathbf{y}_j - \mathbf{y}_i\|^2, \tag{2}$$

which is in the same form of the space-time interval. So, we consider the sample space with class structures as a semi-Riemannian manifold. Unlike Riemannian metric tensors, which are positive-definite, semi-Riemannian metric tensors can naturally encode the class structures. Thus, a semi-Riemannian manifold is more discriminative.

We define a metric matrix $\mathbf{G}$, where the ith column of $\mathbf{G}$ (denoted as $\mathbf{g}_i$) is the diagonal of $\mathbf{\Lambda}_i$, i.e., $\mathbf{g}_i = [g_{1i}, \ldots, g_{ni}]^T$ and $g_{ji} = \mathbf{\Lambda}_i(j,j)$ $(j = 1, \ldots, n)$. An entry g_{ji} in $\mathbf{G}$ is called a metric component of a metric tensor $\mathbf{g}_i$. The projections can be learned via maximizing the total margin

$$\begin{aligned} \Phi &= \frac{1}{2}\sum_{i=1}^{n} \gamma_i = \frac{1}{2}\sum_{i,j=1}^{n} g_{ji}(\mathbf{y}_j - \mathbf{y}_i)^T(\mathbf{y}_j - \mathbf{y}_i), \\ &= \mathrm{tr}(\mathbf{Y}\mathbf{L}_\mathbf{G}\mathbf{Y}^T) = \mathrm{tr}(\mathbf{U}^T\mathbf{X}\mathbf{L}_\mathbf{G}\mathbf{X}^T\mathbf{U}), \end{aligned} \tag{3}$$

i.e., pulling the structures of samples in the embedded low-dimensional space toward the space-likeness, where $\mathbf{L}_\mathbf{G}$ is the Laplacian matrix of $\frac{1}{2}(\mathbf{G} + \mathbf{G}^T)$. If $\mathbf{G}$ is already learned (detailed in Section 3.3), the optimal linear projection matrix $\mathbf{U}$, which projects the samples into a d-dimensional euclidean space and satisfies $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{d\times d}$ and $\mathbf{Y} = \mathbf{U}^T\mathbf{X}$, can be found to be composed of the eigenvectors of $\mathbf{X}\mathbf{L}_\mathbf{G}\mathbf{X}^T$ corresponding to its first d largest eigenvalues.

The cases of nonlinear and multilinear embedding can be easily extended via the kernel method and tensorization, respectively, as in [29], [32], [47].

3.3 Semisupervised Learning of Semi-Riemannian Metrics

The key problem in the semi-Riemannian geometry framework is to determine the metric matrix $\mathbf{G}$. Under the semisupervised setting, the metric matrix $\mathbf{G}$ can be divided into four blocks:

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{LL}, & \mathbf{G}_{LU} \\ \mathbf{G}_{UL}, & \mathbf{G}_{UU} \end{bmatrix}, \tag{4}$$

where $\mathbf{G}_{LL}$ are the metric components between labeled samples, $\mathbf{G}_{LU}$ and $\mathbf{G}_{UL}$ between labeled and unlabeled samples, and $\mathbf{G}_{UU}$ between unlabeled samples. $\mathbf{G}_{UL}$, $\mathbf{G}_{LU}$, and $\mathbf{G}_{UU}$ are estimated via information propagation from labeled data to unlabeled data, which is a common technique in semisupervised learning [53]. Label propagation, as a kind of information propagation, also appeared in some recent papers on semisupervised feature extraction, e.g., [20].

In brief, the metric matrix is learned in three steps. First of all, the metric tensors at labeled sample points, i.e., the blocks $\mathbf{G}_{LL}$ and $\mathbf{G}_{UL}$, are learned. Then, the neighborhood relationships are propagated from metric tensors at labeled sample points to unlabeled sample points, i.e., from $\mathbf{G}_{UL}$ to $\mathbf{G}_{LU}$. Finally, the metric tensors at unlabeled sample points, i.e., $\mathbf{G}_{LU}$ and $\mathbf{G}_{UU}$, are learned. Then, the metric tensor at a point $\mathbf{x}_i$ is a column vector $\mathbf{g}_i$ of $\mathbf{G}$. Similar to (4), $\mathbf{g}_i$ can be divided into two parts $\mathbf{g}_i^L$ and $\mathbf{g}_i^U$, where $\mathbf{g}_i^L = [g_{1i}, \ldots, g_{li}]^T$ and $\mathbf{g}_i^U = [g_{l+1,i}, \ldots, g_{ni}]^T$.

3.3.1 Local Nullity of Semi-Riemannian Manifolds

Null manifolds are a typical class of semi-Riemannian manifolds, on which each point has a zero space-time interval, being neither space-like nor time-like (see Fig. 1) [21]. Inspired by the neutrality of null manifolds, we assume that the samples in the original high-dimensional space lie on a null manifold, so that the contributions of the homogeneous

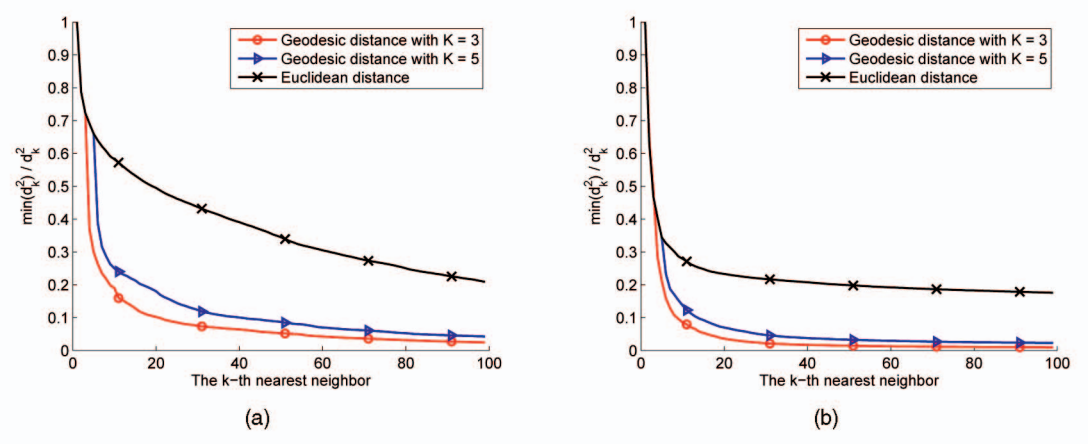


Fig. 3. Comparison of geodesic distances and euclidean distances. (a) only labeled data are available; (b) both labeled and unlabeled data are available. d_k^2 is the squared distance from a sample to its kth nearest neighbor. X-axis sorts the neighbors with increasing distances and the first 100 neighbors are presented. Y-axis offers the ratio between minimum squared distances and d_k^2. The results are averaged over 50 randomly selected samples.

and heterogeneous neighborhoods are balanced. This leads to a *local nullity condition* to each metric tensor

$$\mathbf{g}_i^T \mathbf{d}_i = \sum_{j=1}^{n} g_{ji} d_{ji}^2 = 0, \quad \forall i = 1, \ldots, n, \tag{5}$$

where d_{ji} is the pairwise distance from $\mathbf{x}_j$ to $\mathbf{x}_i$ on the data manifold and $\mathbf{d}_i = [d_{1i}^2, \ldots, d_{ni}^2]^T$.

In [50] the distance d_{ji} is chosen as the known metric of the high-dimensional feature space, e.g., the euclidean distance and χ^2 distance are used for raw image features and local binary pattern features, respectively. However, we often do not know the appropriate metrics a priori. Besides, the local structure of samples has shown its power in unsupervised manifold learning [2], [26] and supervised dimensionality reduction [11], [41], [46]. Inspired by the ISOMAP algorithm [30], we use geodesic distances, approximated by graph distances.[1] It is a great advantage of the semisupervised setting that a number of unlabeled data exist and can be utilized in the graph approximation of geodesic distances. So, the geodesic distances capture the manifold structure of all data. As a result, our global semi-Riemannian metric tensors can achieve good performance even without careful tuning of the sizes of the homogeneous and heterogeneous neighborhoods. For example in Fig. 6, it is shown that the performance is affected very slightly when the choice of $\hat{K}$ varies in a large range. To testify, we use two labeled and 28 unlabeled images per person of 68 persons from the CMU PIE facial database (with detailed descriptions in Section 5). We compare geodesic distances and euclidean distances in two cases: with only labeled data and with both labeled and unlabeled data. The observations from Fig. 3 are as follows:

- According to (5), when d_{ji}^2 is large, the weight of $\mathbf{x}_j$ in the margin of $\mathbf{x}_i$ is suppressed. The geodesic distance of the kth nearest neighbor increases much faster than the euclidean distance when k increases, so the homogeneous and heterogeneous neighborhoods can be selected automatically, i.e., setting $\hat{K}$ and $\check{K}$ to large values has almost no influence on the performance of our algorithm.
- With a number of unlabeled data the geodesic distances perform better than when only labeled data are available.
- The performance of geodesic distances is robust to the varying parameter K, the size of neighborhoods for computing geodesic distances. So, we simply choose $K = 5$ in our implementation.

1. The geodesic distances are computed as follows: First, the K-nearest-neighbor graph is constructed for all samples and the weight of an edge connecting two samples is their euclidean distance. Then, the geodesic distance between two samples is the length of the shortest path connecting them.

3.3.2 Metric Tensors of Labeled Samples

To determine $\mathbf{g}_i^L$, we consider margins of labeled data first. In ANMM [32] and DLA [46], the samples in the same kind of neighborhood have equal weights in a margin. Such a definition of margins only weakly models the intrinsic structure of the training data. To overcome this drawback, we define the average neighborhood margin normalized by geodesic distances at a labeled point $\mathbf{x}_i (i = 1, \ldots, l)$ as

$$\gamma_i = \sum_{j \in \check{\mathcal{N}}_i^{\check{K}}} \frac{1}{\check{K}} \left(\frac{\|\mathbf{y}_j - \mathbf{y}_i\|}{d_{ji}} \right)^2 - \sum_{j \in \hat{\mathcal{N}}_i^{\hat{K}}} \frac{1}{\hat{K}} \left(\frac{\|\mathbf{y}_j - \mathbf{y}_i\|}{d_{ji}} \right)^2, \tag{6}$$

where the homogeneous and heterogeneous neighborhoods are chosen as in Section 3.1. Here, the importance of a marginal sample $\mathbf{x}_j$ is quantified by the distance d_{ji} to $\mathbf{x}_i$. Then for $\mathbf{x}_i$, we have the metric components

$$g_{ji} = \begin{cases} \frac{1}{|\check{\mathcal{N}}_i^{\check{K}}| d_{ji}^2}, & \text{if } x_j \in \check{\mathcal{N}}_i^{\check{K}}, \\ -\frac{1}{|\hat{\mathcal{N}}_i^{\hat{K}}| d_{ji}^2}, & \text{if } x_j \in \hat{\mathcal{N}}_i^{\hat{K}}, \\ 0, & \text{if } x_j \notin \check{\mathcal{N}}_i^{\check{K}} \text{ and } x_j \notin \hat{\mathcal{N}}_i^{\hat{K}}, \end{cases} \tag{7}$$

for all $j = 1, \ldots, l$, where $|\cdot|$ is the cardinality of a set. Equation (7) can also be obtained by the smoothness and local nullity conditions as in [50] (please refer to the Appendix A).

Now, we come to the metric components in $\mathbf{g}_i^U$. The metric component g_{ji} of $\mathbf{g}_i$ can be regarded as a function of

$\mathbf{x}_j$ in the sample space. So, $\mathbf{g}_i^U$ can be inferred from $\mathbf{g}_i^L$ by semisupervised regression as follows.

We assume that nearby points are likely to have close function values, which is known as the smoothness assumption. So, g_{ji} should be close to the metric components of $\mathbf{g}_i$ corresponding to $\mathbf{x}_j$'s neighbors. For example, if $\mathbf{x}_j$ is surrounded by heterogenous neighbors of $\mathbf{x}_i$, g_{ji} should be nonnegative. We choose the similarity measure a_{jk} between samples $\mathbf{x}_j$ and $\mathbf{x}_k$ as

$$a_{jk} = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_j-\mathbf{x}_k\|^2}{2\sigma^2}\right), & \text{if } \mathbf{x}_j \in \mathcal{N}_k^K \text{or } \mathbf{x}_k \in \mathcal{N}_j^K, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathcal{N}_j^K$ and $\mathcal{N}_k^K$ are the K-nearest neighborhoods of $\mathbf{x}_j$ and $\mathbf{x}_k$, respectively. In our experiments, $K = 5$ and σ is the average distance of all sample points to their 6th nearest neighbors.

Then, we estimate the metric tensor $\mathbf{g}_i$ by minimizing

$$\begin{aligned} \phi_i(\mathbf{g}_i) &= \frac{1}{2}\sum_{j,k=1}^{n} a_{jk}(g_{ji}-g_{ki})^2 + \lambda_l \sum_{j=1}^{n} g_{ji}^2, \\ &= \mathbf{g}_i^T(\mathbf{L_A} + \lambda_l \mathbf{I}_{n\times n})\mathbf{g}_i, \\ s.t. \quad & \mathbf{g}_i^T \mathbf{d}_i = 0 \text{ and } \mathbf{g}_i^L \text{ is fixed as in (7)}, \end{aligned} \tag{8}$$

where λ_l is a regularization parameter to avoid singularity of $\mathbf{L_A}$, which is empirically chosen as 0.01, $\mathbf{A} = [a_{jk}]_{n\times n}$, and $\mathbf{L_A}$ is the Laplacian matrix of $\frac{1}{2}(\mathbf{A} + \mathbf{A}^T)$.

By the Lagrangian multiplier, we get the solution of (8):

$$\begin{aligned} \mathbf{g}_i &= \begin{bmatrix} \mathbf{g}_i^L \\ \mathbf{g}_i^U \end{bmatrix}, \\ \mathbf{g}_i^U &= \mathbf{L}_{UU}^{-1}\left(\frac{(\mathbf{L}_{UL}\mathbf{g}_i^L)^T\mathbf{L}_{UU}^{-1}\mathbf{d}_i^U}{(\mathbf{d}_i^U)^T\mathbf{L}_{UU}^{-1}\mathbf{d}_i^U}\mathbf{d}_i^U - \mathbf{L}_{UL}\mathbf{g}_i^L\right), \end{aligned} \tag{9}$$

where the symmetric matrix $(\mathbf{L_A} + \lambda_l \mathbf{I}_{n\times n})$ is divided into

$$\begin{bmatrix} \mathbf{L}_{LL} & \mathbf{L}_{LU} \\ \mathbf{L}_{UL} & \mathbf{L}_{UU} \end{bmatrix}$$

similar to (4) and $\mathbf{d}_i^U = [d_{l+1,i}^2, \ldots, d_{ni}^2]^T$.

3.3.3 Neighborhood Relationship Propagation

Metric tensors encode the structure of the sample space. The metric components g_{ji} and g_{ij} are not independent because if $\mathbf{x}_j$ is in the homogenous or heterogenous neighborhood of $\mathbf{x}_i$, $\mathbf{x}_i$ is probably in the same type of neighborhood of $\mathbf{x}_j$. So, metric tensors of labeled samples provide a priori information for those of unlabeled samples. However, we do not propagate all information in $\mathbf{G}_{UL}$ as components with small values are disturbed more easily by noise. So, we initialize the neighborhoods of unlabeled samples as follows: In the metric tensor of each labeled sample $\mathbf{x}_i (i = 1, \ldots, l)$, we choose $\frac{m\check{K}}{l}$ negative and $\frac{m\check{K}}{l}$ positive entries from $\mathbf{g}_i^U$ with the largest absolute values, and then, put $\mathbf{x}_i$ in the homogeneous or heterogeneous neighborhoods to the corresponding unlabeled samples according to these entries' signs. We also put the $\hat{K}$-nearest and $\check{K}$-farthest samples of an unlabeled sample in its homogeneous and heterogeneous neighborhoods, respectively, as it is more likely for the $\hat{K}$-nearest samples to be in the same class as the sample and the $\check{K}$-farthest samples to be in different classes from the sample.

3.3.4 Metric Tensors of Unlabeled Samples

We initialize the metric tensor of an unlabeled sample $\mathbf{x}_i$ $(i = l+1, \ldots, n)$ as (7) for $j = 1, \ldots, n$, where $\check{\mathcal{N}}_i^{\check{K}}$ and $\hat{\mathcal{N}}_i^{\hat{K}}$ have been constructed in Section 3.3.3, and denote this initial value as $\widetilde{\mathbf{g}}_i$, where $\widetilde{\mathbf{g}}_i = [\widetilde{g}_{1i}, \ldots, \widetilde{g}_{ni}]^T$.

Also by the smoothness of metric components, the metric tensor $\mathbf{g}_i$ can be estimated by minimizing

$$\begin{aligned} \psi_i(\mathbf{g}_i) &= \frac{1}{2}\sum_{j,k=1}^{n} a_{jk}(g_{ji}-g_{ki})^2 + \lambda_u \sum_{j=1}^{n}(g_{ji}-\widetilde{g}_{ji})^2, \\ &= \mathbf{g}_i^T(\mathbf{L_A} + \lambda_u \mathbf{I}_{n\times n})\mathbf{g}_i - 2\lambda_u \widetilde{\mathbf{g}}_i^T\mathbf{g}_i + \lambda_u \widetilde{\mathbf{g}}_i^T\widetilde{\mathbf{g}}_i, \\ s.t. \quad & \mathbf{g}_i^T\mathbf{d}_i = 0, \end{aligned} \tag{10}$$

where λ_u is a control parameter ($\lambda_u > 0$), which is chosen as $\lambda_u = 10$ in our experiments. The regularization term with the weight λ_u requires that the estimated metric tensor is not far from its initial value.

By the Lagrangian multiplier, $\mathbf{g}_i$ can be found as

$$\mathbf{g}_i = \widetilde{\mathbf{L}}^{-1}\left(\widetilde{\mathbf{g}}_i - \frac{\widetilde{\mathbf{g}}_i^T\widetilde{\mathbf{L}}^{-1}\mathbf{d}_i}{\mathbf{d}_i^T\widetilde{\mathbf{L}}^{-1}\mathbf{d}_i}\mathbf{d}_i\right), \tag{11}$$

where $\widetilde{\mathbf{L}} = \frac{1}{\lambda_u}\mathbf{L_A} + \mathbf{I}_{n\times n}$.

3.3.5 S^3RMM Algorithm

The learned matrix $\mathbf{G}$ in the above sections is not the final form. We shall adjust it in two steps.

Noise reduction. Metric components in $\mathbf{G}_{UL}$, $\mathbf{G}_{LU}$, and $\mathbf{G}_{UU}$ are only estimations, so we need to reduce the effect of incorrect components. Metric components close to zero are regarded as unreliable and of little importance in a margin. Thus, for each metric tensor $\mathbf{g}_i$, we set an entry g_{ji} to be zero if $\mathbf{x}_i$ or $\mathbf{x}_j$ is unlabeled and $|g_{ji}| < \frac{1}{10}\max_{g_{ji}g_{ki}>0}|g_{ki}|$. Besides, g_{ji} and g_{ij} should reach an agreement on whether $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same class. So, we split the metric matrix $\mathbf{G}$ to $\mathbf{G}^+ + \mathbf{G}^-$, where $\mathbf{G}^+$ and $\mathbf{G}^-$ keep the positive and negative entries of $\mathbf{G}$, respectively, while leaving the remaining entries zero. Then, update $\widetilde{\mathbf{G}}^+ = \min\{\mathbf{G}^+, (\mathbf{G}^+)^T\}$ and $\widetilde{\mathbf{G}}^- = \max\{\mathbf{G}^-, (\mathbf{G}^-)^T\}$. Finally, we combine them with a factor $\gamma \in [0.5, 1]$: $\widetilde{\mathbf{G}} = (1-\gamma)\widetilde{\mathbf{G}}^+ + \gamma\widetilde{\mathbf{G}}^-$, to make the metric tensors tend to be time-like [50].[2] γ can be estimated by cross validation.

Balancing contributions of labeled and unlabeled samples. Because the target samples of classification are only labeled samples, we suppress the contribution of unlabeled samples as

$$\widetilde{\mathbf{G}}' = \begin{bmatrix} \widetilde{\mathbf{G}}_{LL}, & \alpha_1\widetilde{\mathbf{G}}_{LU} \\ \alpha_1\mathbf{G}_{UL}, & \alpha_2\widetilde{\mathbf{G}}_{UU} \end{bmatrix},$$

where

$$\widetilde{\mathbf{G}} = \begin{bmatrix} \widetilde{\mathbf{G}}_{LL} & \widetilde{\mathbf{G}}_{LU} \\ \widetilde{\mathbf{G}}_{UL} & \widetilde{\mathbf{G}}_{UU} \end{bmatrix}$$

2. Note that the feature extraction process pulls the initially time-like semi-Riemannian manifold toward the space likeness.

TABLE 1
S^3RMM Algorithm

Semi-Supervised Semi-Riemannian Metric Map (S^3RMM) Algorithm

Given l labeled samples $\{\mathbf{x}_1, ..., \mathbf{x}_l\}$, each of which has a class label $c_i \in \{1, ..., c\}$, and m unlabeled samples $\{\mathbf{x}_{l+1}, ..., \mathbf{x}_n\}$ with unknown class memberships ($l + m = n$), where $\mathbf{x}_i \in \mathbb{R}^D$. We want to compute a matrix $\mathbf{U}$ to project samples linearly as $\mathbf{Y} = \mathbf{U}^T\mathbf{X}$, where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$, $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]$ and $\mathbf{y}_i \in \mathbb{R}^d$ is the low-dimensional embedding of $\mathbf{x}_i$.

1) Compute the affinity matrix $\mathbf{A}$ and pairwise geodesic distances $d_{ij}(i, j = 1, ..., n)$;
2) Construct the homogeneous and heterogeneous neighborhoods of labeled samples and then learn metric tensors $\mathbf{g}_i$ for $i = 1, ..., l$ as described in Section 3.3.2;
3) Construct the homogeneous and heterogeneous neighborhoods of unlabeled samples as in Section 3.3.3 and learn metric tensors $\mathbf{g}_i$ for $i = l + 1, ..., n$ as described in Section 3.3.4;
4) Modify $\mathbf{G}$ as described at the beginning of Section 3.3.5 and replace $\mathbf{G}$ with $\widetilde{\mathbf{G}}'$;
5) Obtain the optimal projection $\mathbf{U}$ by computing the eigenvectors of $\mathbf{X}\mathbf{L}_{\mathbf{G}}\mathbf{X}^T$ associated with the first d largest eigenvalues.

is the metric matrix obtained after noise reduction and the values of α_1 and α_2 are chosen to be close to $\min\{1, \frac{l}{m}\}$ and $\min\{1, (\frac{l}{m})^2\}$, respectively. α_1 and α_2 do not exceed 1 because the contribution of unlabeled samples should not be more than that of labeled ones in the total margin (see (3)).

The whole procedure of S^3RMM is summarized in Table 1.

4 Discussion

In this section, we would like to discuss and highlight some aspects of our S^3RMM algorithm.

4.1 A General Framework for Semisupervised Dimensionality Reduction

S^3RMM can be viewed as a general framework for semisupervised dimensionality reduction. First, our margin maximization reformulation of SRDA [50] provides the connection between the semi-Riemannian geometry framework and the discrepancy criterion. So, S^3RMM can be integrated with any dimensionality reduction algorithm based on the discrepancy criterion, e.g., MMC [16], ANMM [32], and DLA [46], to obtain semisupervised extensions of them. To create new algorithms, we only need to change the structural properties of semi-Riemannian metric tensors, i.e., the constraints in (5) and (7). Second, in this framework we utilize the separability and similarity between samples including labeled and unlabeled ones, instead of the regularization term on the graph of unlabeled or all samples used in SDA [6], SSLF [43], and SDLA [46]. The traditional regularization term is considered as a special case under our framework (please refer to Appendix B). Finally, we only use a simple yet efficient way to learn semi-Riemannian metrics in this paper, and our method may be incorporated with a number of semisupervised regression methods [53].

4.2 Comparison to SRDA

The major differences between our method and SRDA [50] are threefold: First, we define *global* semi-Riemannian metric tensors rather than *local* metric tensors as in SRDA. Second, in SRDA asymmetric semi-Riemannian metrics are learned locally at each sample $\mathbf{x}_i$ independently, supervised by the label information. The relationship among the metrics at different data samples is not considered. In contrast, in our method we learn asymmetric metrics from labeled examples, local consistency in metric tensors and weak propagation between metric tensors globally. Third, different from the euclidean/χ^2 distances assumed known in SRDA, we use geodesic distances from *unsupervised* manifold learning, which do not require any a priori knowledge of the sample space, to capture the manifold structure of data.

4.3 Advantages over Semisupervised Discriminant Analysis

S^3RMM has several advantages over semisupervised discriminant analysis (SDA [6] and SSDA [48]). First of all, our algorithm can be applied to semisupervised dimensionality reduction with pairwise constraints directly, i.e., we only need to know pairwise constraints on partial samples, for learning semi-Riemannian metrics. A pairwise constraint between two samples, another kind of supervision information usually used in semisupervised dimensionality reduction [12], [45], describes whether they belong to the same class or not, rather than provides the labels. It might be too expensive to obtain explicit class memberships in many real-world applications. For example, in image retrieval it is much easier to know the relevance relationship between images, with the logs of user relevance feedback, while obtaining the exact class label of images requires quite expensive efforts of image annotation. Second, it is easy to see that S^3RMM avoids the intrinsic problems of LDA [40]: the singularity problem and limited available projection dimensions. SDA and SSDA alleviate, but not resolve, these problems, as their optimization models are in the form of

$$J = \frac{\mathrm{tr}(\mathbf{U}^T\mathbf{S}_b\mathbf{U})}{\mathrm{tr}(\mathbf{U}^T\mathbf{S}_w\mathbf{U}) + R(\mathbf{U})},$$

where $R(\mathbf{U})$ is some regularization term on the unlabeled data [6], [48]. In contrast, S^3RMM avoids the singularity problem, as there is no matrix inversion involved. The number of possible projection dimensions in S^3RMM is not limited to $(c-1)$ either because this limitation of LDA results from the limited ranks of the scatter matrices.

4.4 Connection to Indefinite Kernel PCA

The maximization of the total margin in (3), after learning the metric matrix $\mathbf{G}$, aims at finding the optimal linear

projection matrix $\mathbf{U}$ satisfying that $\mathbf{Y} = \mathbf{U}^T\mathbf{X}$. The manifold learning counterpart of (3) is

$$\max_{\mathbf{Y}} \operatorname{tr}\left(\mathbf{Y}\mathbf{L}_{\mathbf{G}}\mathbf{Y}^T\right),$$

which is equivalent to kernel PCA using $(-\mathbf{L}_{\mathbf{G}})^{\dagger}$ (i.e.,, the pseudo inverse of the matrix $-\mathbf{L}_{\mathbf{G}}$) as the kernel matrix [8]. From this view, learning the semi-Riemannian metric matrix can be interpreted as indefinite kernel learning [22]. However, there are significant differences between S^3RMM and indefinite kernel learning. First, learning the metric matrix is different from the existing indefinite kernel learning approaches, such as [15]. Our algorithm learns a *sparse metric* matrix, while does not learn the corresponding *nonsparse kernel* matrix. Second, we aim at proposing a linear feature extraction method, which is not so computationally expensive as kernel feature extraction and does not have the out-of-sample problem [4]. Third, our semi-Riemannian metric matrix encodes the structural relationship of all data samples, but the kernel matrix is derived from the mapping from the original sample space to some infinite-dimensional Krein space [22]. Interested readers may refer to [9], [14], [22] for the theory of Krein spaces. Finally, the existing feature extraction methods based on indefinite kernels define the kernel a priori [23]. To the best of our knowledge, there are no feature extraction methods utilizing indefinite kernel learning.

5 Experiments

We compare our method to several recently proposed semisupervised dimensionality reduction methods: SDA [6], SSDA [48], SSLDA [43], and SDLA [46]. The first three are different semisupervised extensions of LDA and the last one is a discrepancy-criterion-based method. We also list the results of traditional unsupervised and supervised algorithms, including PCA, LDA, LPP [11], MFA [41], and MMC [16], for reference.[3] Results of DLA [46] and SRDA [50] are presented for comparisons of supervised and semisupervised methods. Note that like S^3RMM, DLA/SDLA [46] also utilizes the nonlinear structure of the sample space and extracts linear projections. We test the performance of S^3RMM on two benchmark facial databases (CMU PIE and FRGC 2.0) and the USPS handwritten digit data set.

Before the experiments on real imagery data, we conduct simulations on synthetic data to show how well S^3RMM works.

5.1 Toy Experiments

To test our algorithm, we generate two kinds of 2D toy data sets: two-line data shown in Fig. 4a and two-half-circle data shown in Fig. 4b. The sample points are uniformly sampled and perturbed by random Gaussian noise. The two classes of labeled samples are shown with circles and crosses, respectively, and the unlabeled samples are shown with points. We present the results of SDLA and S^3RMM, as SDLA and S^3RMM are both discrepancy-criterion-based methods with different regularization terms on the unlabeled data. SDLA adopts the graph-based regularization, which only utilizes similarities between neighboring samples, while S^3RMM takes advantage of both similarities and dissimilarities between all samples, which are encoded in metric tensors. To show how the unlabeled data affect the projection direction, we set the weight of the term related to the unlabeled data to be sufficiently large, and we also give the result of DLA (i.e., the weight of the unlabeled data in SDLA is set to be zero). The classification boundary, perpendicular to the projection direction, is shown instead for each method, to give a more clear illustration. From Fig. 4, we can see that the graph-based regularization on the unlabeled data in SDLA works well for two-line data, and does not change the projection direction much for two-half-circle data. It is because for two-line data the similarities of neighboring samples vary a lot, when we choose different projection directions, and for two-half-circle data they vary little. However, S^3RMM gives a nearly ideal boundary between the two manifolds in both data.

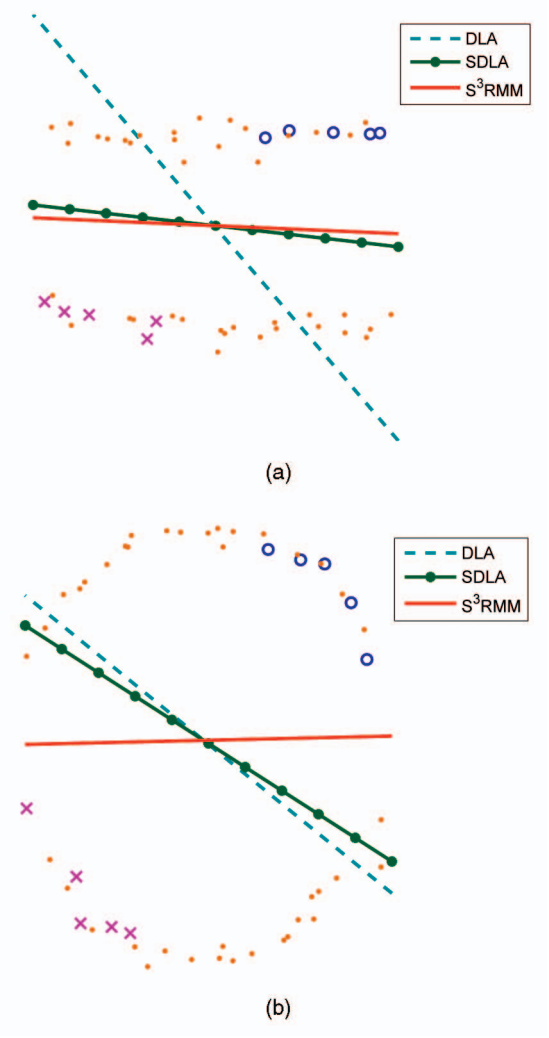


Fig. 4. Results of 2D toy experiments on two kinds of toy data. (a) Simulation results on two-line data; (b) simulation results on two-half-circle data.

We also test the robustness of our algorithm to noise on the toy data. We repeat the above experiments 20 times and find that the correlation between projection directions in any two experiments is larger than 0.99. This result verifies that our algorithm is robust to noisy data.

3. We use implementations of PCA, LDA, LPP, MFA, and SDA from http://www.zjucadcg.cn/dengcai/Data/data.html.

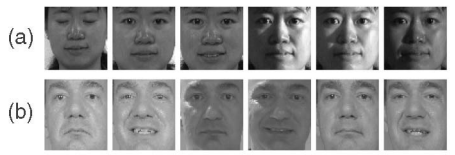


Fig. 5. Sample images from (a) CMU PIE database and (b) FRGC 2.0 database.

5.2 Setup of Experiments on Real Data

In each experiment, we randomly select $l+m$ images of each person for the training set. Then, among the $l+m$ images, l images are randomly selected and labeled, forming the labeled set, and the other m images form the unlabeled set. The remaining images in the database are used for testing. We test 50 trials of random splits and report the averaged results.

For unsupervised and supervised methods, the labeled set is used for training. For semisupervised methods, the unlabeled set is added to the training set. In all result tables, we use *US*, *S*, *SS* as short for *Unsupervised*, *Supervised*, and *Semisupervised*, respectively. A simple nearest-neighbor classifier is employed on the extracted low-dimensional features to classify samples in the unlabeled and test set. The result of the nearest-neighbor classifier on raw features without dimensionality reduction is used as the baseline.

All parameters of the involved methods are tuned on the training set, by the full search over a relatively wide range which is discretized by some step-size, e.g., for PCA preprocessing, we test with the preserved energy being between 90 percent and 100 percent.

5.3 Face Recognition

In our experiments, two benchmark face databases, CMU PIE [27], and experiment 4 in FRGC 2.0 [24], are used. The CMU PIE database contains more than 40,000 facial images of 68 people. The images were acquired in different poses, under various illumination conditions and with different facial expressions. In our experiments, a subset, the frontal pose (C 27) with varying lighting and illumination, is used. So, each person has about 49 images and in total 3,329 images are collected. All the images are aligned by fixing the locations of eyes, and then, normalized to 32×32 pixels (Fig. 5a). The training set of experiment 4 in FRGC 2.0 consists of controlled and uncontrolled still images. We search all images of each person in this set and take the first 60 images of the first 50 individuals the number of whose facial images is more than 60. Thus, we collect 3,000 facial images for our experiments. All the images are aligned according to the positions of eyes and mouths, and then, cropped to a size of 36×32 (Fig. 5b).

In the first experiment, $l=2$, $m=28$, the number of test images per individual is about 19, and the number of individuals is 68. In the second experiment, $l=5$, $m=35$, the number of test images per person is 20, and the number of persons is 50. Table 2 provides results of each method. The unsupervised method, PCA, only performs a little better than the baseline without any feature extraction. LDA, LPP, and MFA have good performance on PCA features and SRDA, as reported in [50], outperforms the supervised Fisher-criterion-based methods even if it is applied to the raw data directly. The recognition results of semisupervised methods are generally better than their corresponding supervised methods as they utilize the unlabeled data. S^3RMM is the best in the semisupervised methods and improves the results of SRDA. Besides, the improvement of S^3RMM is more than the differences between other methods and their supervised counterparts.[4]

It is interesting to know the sensitivity to the sizes of homogenous and heterogenous neighborhoods as our method is based on maximizing the margins of such neighborhoods. The size of homogenous neighborhoods is limited by the number of samples per class, while the size of heterogenous neighborhoods can be much larger, as the number of samples not in the same class of a sample is very large. Because there are many possible choices of the size of heterogenous neighborhoods, we test the robustness of our method when this size changes, although the robustness may be implied by the properties of geodesic distances discussed in Section 3.3.1. In the test, all parameters except $\check{K}$ are fixed. Fig. 6 shows the error rates on unlabeled and test data of both databases with a varying number of initial heterogenous neighbors. We see here that S^3RMM is surprisingly robust.

We also test the sensitivity to λ_l in (8) and λ_u in (10). All parameters except the tested parameter (λ_l or λ_u) are fixed. Fig. 7 shows the error rates on unlabeled and test data of both databases with a varying value of the tested parameter. We see that S^3RMM is also robust against the variance of these parameters in a large range.

5.4 Handwritten Digit Classification

The USPS data set contains grayscale handwritten digit images scanned from envelopes by the US Postal Service (Fig. 8). The images are of size 16×16. The original training set contains 7,291 images, and the test set contains 2,007 images.[5] We used digits 1, 2, 3, 4, and 5 in our experiments as the five classes.

On the USPS data set we choose $l=5$, $m=95$, the number of test samples per class as 1,000 and the number of classes as 5, respectively. The classification results are listed in Table 3. PCA is only better than the baseline. The Fisher-criterion-based methods, LDA, LPP, MFA, and SSLDA, do not improve PCA features much. The discrepancy-criterion-based methods, MMC, DLA, and SRDA, are better than other supervised methods. Unlabeled data can improve the classification accuracy and S^3RMM is the best again.

6 Conclusion

In this paper, we address the problem of semisupervised feature extraction via the semi-Riemannian geometry framework. Under this framework, the margins of samples in the high-dimensional space are encoded in the metric tensors. We explicitly model the learning of semi-Riemannian metric tensors as a semisupervised regression. Then, the optimal projection is pursued by maximizing the

4. PCA+LDA is compared with SSLDA because SSDLA is applied on PCA transformed subspace while SDA and SSDA use the Tikhonov regularization and are directly applied to the raw data.

5. We downloaded the set of 1,100 samples per class from http://www.cs.toronto.edu/~roweis/data.html.

TABLE 2
Recognition Error Rates (Percent, in $\mathrm{Mean}\pm\mathrm{Std\text{-}Dev}$) on the CMU PIE and FRGC 2.0 Databases

Method		CMU PIE		FRGC 2.0	
		unlabeled	test	unlabeled	test
US	Baseline	67.23±1.51	67.25±1.78	44.82±1.34	44.66±1.76
	PCA [1]	61.45±1.64	61.71±1.83	31.74±1.67	31.38±1.85
S	PCA+LDA [1]	24.01±2.22	24.05±2.35	12.64±1.36	12.65±1.66
	PCA+LPP [11]	23.83±1.76	23.79±2.08	11.39±1.13	11.35±1.51
	PCA+MFA [41]	22.03±1.80	22.04±1.80	11.58±1.28	11.23±1.21
	MMC [16]	30.60±4.54	30.67±4.87	12.08±1.48	12.07±1.73
	DLA [46]	19.78±1.44	19.74±1.73	11.08±1.49	10.97±1.59
	SRDA [50]	19.43±1.39	19.44±1.70	11.26±1.37	11.13±1.65
SS	SDA [6]	20.30±1.61	19.97±1.95	10.55±1.52	10.45±1.58
	SSDA [48]	18.80±1.55	18.65±1.97	10.18±1.47	10.21±1.71
	SSLDA [43]	22.32±1.79 (−1.69)	22.40±1.97 (−1.65)	10.73±1.42 (−1.91)	10.84±1.71 (−1.81)
	SDLA [46]	18.58±1.49 (−1.20)	18.52±1.72 (−1.22)	9.45±1.34 (−1.63)	9.36±1.40 (−1.61)
	S^3RMM	**17.08±1.25** (−2.35) ($\hat{K}=1, \check{K}=50$)	**17.15±1.53** (−2.29)	**7.70±1.46** (−3.56) ($\hat{K}=3, \check{K}=60$)	**7.80±1.76** (−3.33)

The reduced error rates of semisupervised methods over their supervised counterparts are given in brackets.

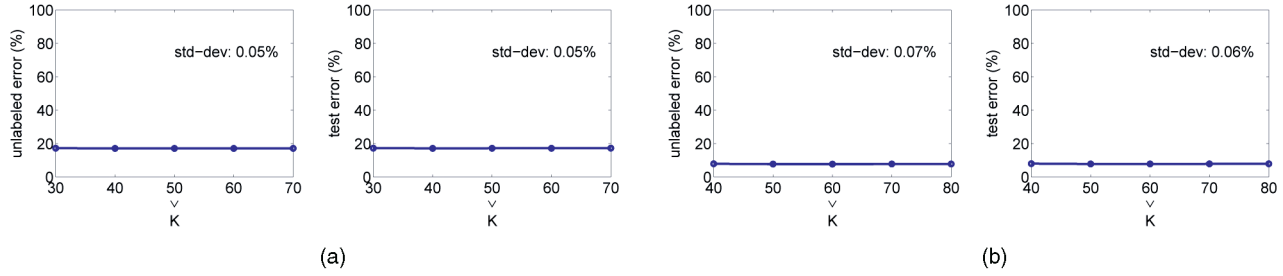


Fig. 6. Recognition error rates of S^3RMM against the variations of $\check{K}$ on unlabeled and test data of the CMU PIE and FRGC 2.0 databases. The standard deviations of error rates against the variations of $\check{K}$ are all less than 0.1 percent. (a) CMU PIE and (b) FRGC 2.0.

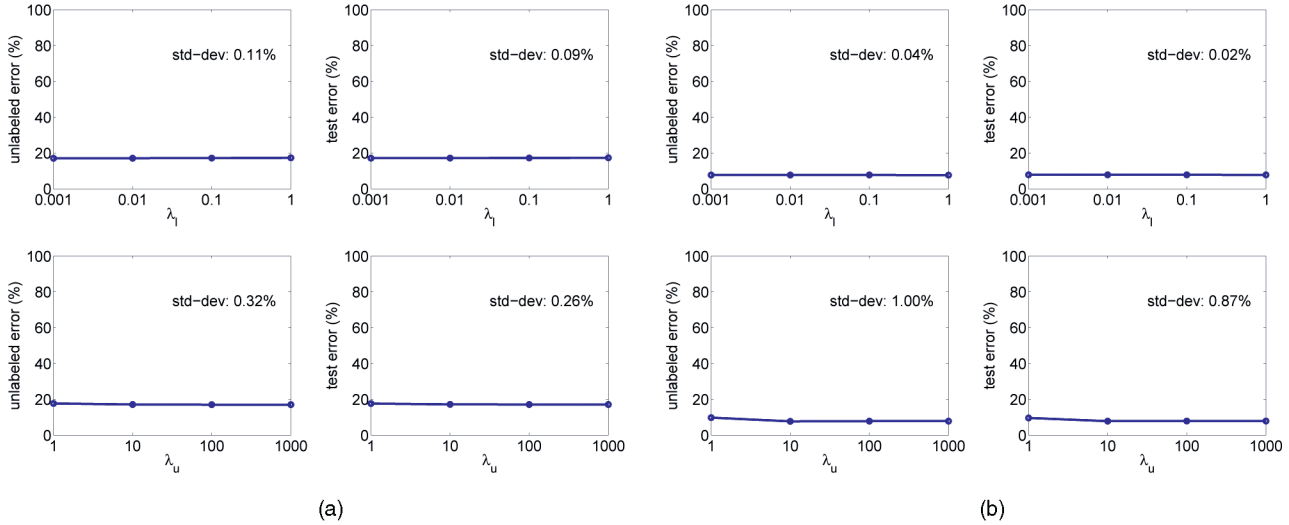


Fig. 7. Recognition error rates of S^3RMM against the variations of λ_l and λ_u on unlabeled and test data of the CMU PIE, and FRGC 2.0 databases. The standard deviations of error rates against the variations of the parameters are also given (a) CMU PIE and (b) FRGC 2.0.

margins of samples in the embedded low-dimensional space. Our algorithm can be a general framework for semisupervised dimensionality reduction. Compared to previous semisupervised methods, we utilize both the separability and similarity criteria of labeled and unlabeled samples. The links between our method and previous research are discussed. The effectiveness is tested on face recognition and handwritten digit classification.

For future work, it would be interesting to see whether our algorithm can be integrated into an active learning

Fig. 8. Samples from the USPS data set.

framework. Those zero entries corresponding to unlabeled data in the metric matrix might indicate the marginal samples of a sample. Therefore, it is possible to design a strategy on how to select the most informative samples to label. It is attractive to explore in this direction.

Appendix A

Connection between SRDA and ANMM

In SRDA [50], by the smoothness and local nullity condition they learn semi-Riemannian metrics as

$$\check{\mathbf{g}}_i = \frac{\check{\mathbf{D}}_i^{-1}\mathbf{e}_{\check{K}}}{\mathbf{e}_{\check{K}}^T\check{\mathbf{D}}_i^{-1}\mathbf{e}_{\check{K}}}, \quad \hat{\mathbf{g}}_i = \frac{\mathbf{e}_{\check{K}}^T\check{\mathbf{D}}_i\check{\mathbf{g}}_i}{\check{K}}\hat{\mathbf{D}}_i^{-1}\mathbf{e}_{\hat{K}},$$

where $\check{\mathbf{D}}_i = \mathrm{diag}([d_{ji}^2, j \in \check{N}_i^{\check{K}}])$, $\hat{\mathbf{D}}_i = \mathrm{diag}([d_{ji}^2, j \in \hat{N}_i^{\hat{K}}])$ and $\mathbf{e}_{\check{K}}$, $\mathbf{e}_{\hat{K}}$ are all-one column vectors. Then, the margin in the projected space for a sample $\mathbf{x}_i$ can be written as $\gamma_i = \sum_{j\in\check{\mathcal{N}}_i^{\check{K}}} \frac{\check{K}}{d_{ji}^{-2}}\gamma_i'$, where

$$\gamma_i' = \sum_{j\in\check{\mathcal{N}}_i^{\check{K}}} \frac{1}{\check{K}}\left(\frac{\|\mathbf{y}_j - \mathbf{y}_i\|}{d_{ji}}\right)^2 - \sum_{j\in\hat{\mathcal{N}}_i^{\hat{K}}} \frac{1}{\hat{K}}\left(\frac{\|\mathbf{y}_j - \mathbf{y}_i\|}{d_{ji}}\right)^2. \quad (12)$$

The only difference between (12) and (1) is the distance normalization, which can capture the structure of data better.

Appendix B

A Special Case of Semisupervised Semi-Riemannian Framework

In this appendix, we would like to show that the intrinsic relationship between the conventional graph-based semi-supervised dimensionality reduction methods, e.g., [43], and our semisupervised semi-Riemannian framework.

Let $\mathbf{A} = \mathbf{0}$ (which can be achieved by choosing a very small σ), i.e., remove the consistency constraints inside the metric tensors, and we have $\mathbf{G}_{UL} = \mathbf{0}$ from (8). Following the neighborhood propagation, we only add $\hat{K}$-nearest and $\check{K}$-farthest neighbors of an unlabeled sample in its homogenous and heterogeneous neighborhoods, respectively. Thus, we have

$$\widetilde{g}_{ji} = \begin{cases} \frac{1}{\check{K}d_{ji}^2}, & \text{if } x_j \in \check{\mathcal{N}}_i^{\check{K}}, \\ -\frac{1}{\hat{K}d_{ji}^2}, & \text{if } x_j \in \hat{\mathcal{N}}_i^{\hat{K}}, \\ 0, & \text{if } x_j \notin \check{\mathcal{N}}_i^{\check{K}} \text{ and } x_j \notin \hat{\mathcal{N}}_i^{\hat{K}}, \end{cases} \quad (13)$$

From Fig. 3 it is easy to see that

$$\frac{1}{d_{ji}^2}\bigg|_{x_j\in\check{\mathcal{N}}_i^{\check{K}}} \ll \frac{1}{d_{ji}^2}\bigg|_{x_j\in\check{\mathcal{N}}_i^{\hat{K}}}$$

TABLE 3
Recognition Error Rates (Percent, in $\mathrm{Mean}\pm\mathrm{Std\text{-}Dev}$) on the USPS Handwritten Digit Database

	Method	unlabeled	test
US	Baseline	23.67±2.82	24.40±2.48
	PCA [1]	21.90±2.75	22.98±2.71
S	PCA+LDA [1]	21.89±2.75	22.56±2.78
	PCA+LPP [11]	21.57±2.61	22.48± 2.58
	PCA+MFA [41]	21.45±2.76	22.22±2.50
	MMC [16]	20.93±3.20	22.05±2.77
	DLA [46]	20.00±2.46	20.92±2.09
	SRDA [50]	18.87±1.46	20.17±1.83
SS	SDA [6]	19.97±2.26	20.82±2.03
	SSDA [48]	19.13±2.07	20.29±1.88
	SSLDA [43]	20.46±2.57 (−1.43)	21.16±2.52 (−1.40)
	SDLA [46]	18.17±1.68 (−1.83)	19.38±1.73 (−1.54)
	S^3RMM	**15.73±2.07** (−3.14) ($\check{K}=3, \hat{K}=20$)	**17.05±2.13** (−3.12)

The reduced error rates of semisupervised methods over their supervised counterparts are given in brackets.

as $\check{\mathcal{N}}_i^{\check{K}}$ and $\check{\mathcal{N}}_i^{\hat{K}}$ include $\check{K}$-farthest and $\hat{K}$-nearest neighbors, respectively. Without loss of generality, let $\widetilde{\mathbf{g}}_j = \hat{K}\widetilde{\mathbf{g}}_j$ and we rewrite $\widetilde{g}$ as

$$\widetilde{g}_{ji} = \begin{cases} -\frac{1}{d_{ji}^2}, & \text{if } x_j \in \mathcal{N}_i^{\hat{K}}, \\ 0, & \text{if } x_j \notin \mathcal{N}_i^{\hat{K}}, \end{cases} \quad (14)$$

Still by $\mathbf{A} = 0$, we have $\mathbf{g}_j = \widetilde{\mathbf{g}}_j$. If $\hat{K} = K$, then $x_j \in \mathcal{N}_i^K$, and thus,

$$g_{ji} = \begin{cases} -\frac{1}{\|\mathbf{x}_j - \mathbf{x}_i\|^2}, & \text{if } x_j \in \mathcal{N}_i^K, \\ 0, & \text{if } x_j \notin \mathcal{N}_i^K. \end{cases} \quad (15)$$

This leads to the widely used regularization term

$$g_{ji} = \begin{cases} -f(\|\mathbf{x}_j - \mathbf{x}_i\|), & \text{if } x_j \in \mathcal{N}_i^K, \\ 0, & \text{if } x_j \notin \mathcal{N}_i^K. \end{cases} \quad (16)$$

The function $f(\cdot)$ is chosen as $f(\cdot) = 1$ in SDA [6], SSLF [43], and SDLA [46]. Another popular choice is $f(\cdot) = e^{-\frac{(\cdot)^2}{\sigma^2}}$.

Acknowledgments

This work was partially supported by the Research Grants Council of the Hong Kong SAR (Project No. CUHK 416510). The authors would like to thank the anonymous reviewers for their comments, which substantially improved the manuscript. The first author would also like to thank Guangcan Liu et al. for sharing their manuscript on unsupervised semi-Riemannian metric map [17] and Deli Zhao for his valuable comments.

References

[1] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces versus Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, July 1997.

[2] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation*, vol. 15, pp. 1373-1396, 2003.

[3] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *J. Machine Learning Research,* vol. 7, pp. 2399-2434, 2006.

[4] Y. Bengio, J. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering," *Advances in Neural Information Processing Systems,* MIT Press, 2003.

[5] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," *Proc. Workshop Computational Learning Theory (COLT),* 1998.

[6] D. Cai, X. He, and J. Han, "Semi-Supervised Discriminant Analysis," *Proc. Int'l Conf. Computer Vision,* 2007.

[7] J. Friedman, "Regularized Discriminant Analysis," *J. Am. Statistical Assoc.,* vol. 84, no. 405, pp. 165-175, 1989.

[8] J. Ham, D. Lee, S. Mika, and B. Schölkopf, "A Kernel View of the Dimensionality Reduction of Manifolds," *Proc. Int'l Conf. Machine Learning,* 2004.

[9] B. Hassibi, A. Sayed, and T. Kailath, "Linear Estimation in Krein Spaces-Part I: Theory," *IEEE Trans. Automatic Control,* vol. 41, no. 1, pp. 18-33, Jan. 1996.

[10] T. Hastie, A. Buja, and R. Tibshirani, "Penalized Discriminant Analysis," *The Annals of Statistics,* vol. 23, no. 1, pp. 73-102, 1995.

[11] X. He and P. Niyogi, "Locality Preserving Projections," *Advances in Neural Information Processing Systems,* MIT Press, 2003.

[12] S.C.H. Hoi, W. Liu, M.R. Lyu, and W.-Y. Ma, "Learning Distance Metrics with Contextual Constraints for Image Retrieval," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '06),* 2006.

[13] P. Howland and H. Park, "Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 8, pp. 995-1006, Aug. 2004.

[14] I. Iokhvidov, M. Krein, and H. Langer, *Introduction to the Spectral Theory of Operators in Spaces with an Indefinite Metric.* Akademie-Verlag, 1982.

[15] M. Kowalski, M. Szafranski, and L. Ralaivola, "Multiple Indefinite Kernel Learning with Mixed Norm Regularization," *Proc. Int'l Conf. Machine Learning,* 2009.

[16] H. Li, T. Jiang, and K. Zhang, "Efficient and Robust Feature Extraction by Maximum Margin Criterion," *IEEE Trans. Neural Networks,* vol. 17, no. 1, pp. 157-165, Jan. 2006.

[17] G. Liu, Z. Lin, and Y. Yu, "Learning Semi-Riemannian Manifolds for Unsupervised Dimensionality Reduction," submitted to *Pattern Recognition.*

[18] Q. Liu, X. Tang, H. Lu, and S. Ma, "Face Recognition Using Kernel Scatter-Difference-Based Discriminant Analysis," *IEEE Trans. Neural Networks,* vol. 17, no. 4, pp. 1081-1085, July 2006.

[19] K. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An Introduction to Kernel-Based Learning Algorithms," *IEEE Trans. Neural Networks,* vol. 12, no. 2, pp. 181-201, Mar. 2001.

[20] F. Nie, S. Xiang, Y. Jia, and C. Zhang, "Semi-Supervised Orthogonal Discriminant Analysis via Label Propagation," *Pattern Recognition,* vol. 42, no. 11, pp. 2615-2627, 2009.

[21] B. O'Neill, *Semi-Riemannian Geometry with Application to Relativity.* Academic Press, 1983.

[22] C. Ong, X. Mary, S. Canu, and A. Smola, "Learning with Non-Positive Kernels," *Proc. Int'l Conf. Machine Learning,* 2004.

[23] E. Pekalska and B. Haasdonk, "Kernel Discriminant Analysis for Positive Definite and Indefinite Kernels," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 6, pp. 1017-1032, June 2009.

[24] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '05),* 2005.

[25] X. Qiu and L. Wu, "Face Recognition by Stepwise Nonparametric Margin Maximum Criterion," *Proc. Int'l Conf. Computer Vision,* 2005.

[26] L.K. Saul, S.T. Roweis, and Y. Singer, "Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds," *J. Machine Learning Research,* vol. 4, pp. 119-155, 2003.

[27] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression Database," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 12, pp. 1615-1618, Dec. 2003.

[28] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the Point Cloud: From Transductive to Semi-Supervised Learning," *Proc. Int'l Conf. Machine Learning,* 2005.

[29] Y. Song, F. Nie, C. Zhang, and S. Xiang, "A Unified Framework for Semi-Supervised Dimensionality Reduction," *Pattern Recognition,* vol. 41, no. 9, pp. 2789-2799, 2008.

[30] J.B. Tenenbaum, V. de Silva, and J.C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science,* vol. 290, pp. 2319-2323, 2000.

[31] V. Vapnik, *Statistical Learning Theory.* John Wiley, 1998.

[32] F. Wang and C. Zhang, "Feature Extraction by Maximizing the Average Neighborhood Margin," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '07),* 2007.

[33] H. Wang, W. Zheng, Z. Hu, and S. Chen, "Local and Weighted Maximum Margin Discriminant Analysis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '07),* 2007.

[34] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace Ratio versus Ratio Trace for Dimensionality Reduction," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '07),* 2007.

[35] X. Wang and X. Tang, "Unified Subspace Analysis for Face Recognition," *Proc. Int'l Conf. Computer Vision,* 2003.

[36] X. Wang and X. Tang, "Dual-Space Linear Discriminant Analysis for Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '04),* 2004.

[37] X. Wang and X. Tang, "Random Sampling LDA for Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '04),* 2004.

[38] X. Wang and X. Tang, "A Unified Framework for Subspace Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 9, pp. 1222-1228, Sept. 2004.

[39] X. Wang and X. Tang, "Random Sampling for Subspace Face Recognition," *Int'l J. Computer Vision,* vol. 70, no. 1, pp. 91-104, 2006.

[40] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, "Multilinear Discriminant Analysis for Face Recognition," *IEEE Trans. Image Processing,* vol. 16, no. 1, pp. 212-220, Jan. 2007.

[41] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 1, pp. 40-51, Jan. 2007.

[42] J. Yang, A. Frangi, J. Yang, D. Zhang, and Z. Jin, "KPCA Plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 2, pp. 230-244, Feb. 2005.

[43] W. Yang, S. Zhang, and W. Liang, "A Graph Based Subspace Semi-Supervised Learning Framework for Dimensionality Reduction," *Proc. European Conf. Computer Vision,* 2008.

[44] J. Ye, R. Janardan, and Q. Li, "Two-Dimensional Linear Discriminant Analysis," *Advances in Neural Information Processing Systems,* MIT Press, 2004.

[45] D. Zhang, Z.-H. Zhou, and S. Chen, "Semi-Supervised Dimensionality Reduction," *Proc. SIAM Int'l Conf. Data Mining,* 2007.

[46] T. Zhang, D. Tao, and J. Yang, "Discriminative Locality Alignment," *Proc. European Conf. Computer Vision,* 2008.

[47] W. Zhang, Z. Lin, and X. Tang, "Tensor Linear Laplacian Discrimination (TLLD) for Feature Extraction," *Pattern Recognition,* vol. 42, no. 9, pp. 1941-1948, 2009.

[48] Y. Zhang and D.-Y. Yeung, "Semi-Supervised Discriminant Analysis Using Robust Path-Based Similarity," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[49] Z. Zhang and H. Zha, "Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment," *SIAM J. Scientific Computing,* vol. 26, no. 1, pp. 313-338, 2005.

[50] D. Zhao, Z. Lin, and X. Tang, "Classification via Semi-Riemannian Spaces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '08),* 2008.

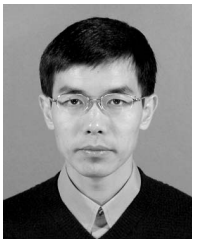
[51] D. Zhao, Z. Lin, R. Xiao, and X. Tang, "Linear Laplacian Discrimination for Feature Extraction," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '07),* 2007.

[52] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf, "Learning with Local and Global Consistency," *Advances in Neural Information Processing Systems,* pp. 321-328, MIT Press, 2004.

[53] X. Zhu, "Semi-Supervised Learning Literature Survey," technical report, Computer Science Dept., Univ. of Wisconsin-Madison, 2005.

Wei Zhang received the BEng degree in electronic engineering from the Tsinghua University, Beijing, in 2007, and the MPhil degree in information engineering in 2009 from The Chinese University of Hong Kong, where he is currently working toward the PhD degree at the Department of Information Engineering. His research interests include machine learning, computer vision, and image processing.

Zhouchen Lin received the PhD degree in applied mathematics from Peking University, in 2000. He is currently a lead researcher in Visual Computing Group, Microsoft Research, Asia. His research interests include computer vision, image processing, machine learning, pattern recognition, numerical computation and optimization, and computer graphics. He is a senior member of the IEEE.



Xiaoou Tang received the BS degree from the University of Science and Technology of China, Hefei, in 1990, the MS degree from the University of Rochester, Rochester, New York, in 1991, and the PhD degree from the Massachusetts Institute of Technology, Cambridge, in 1996. He is a professor in the Department of Information Engineering and associate dean (research) of the Faculty of Engineering at the Chinese University of Hong Kong. He worked as the group manager of the Visual Computing Group at Microsoft Research Asia from 2005 to 2008. His research interests include computer vision, pattern recognition, and video processing. He received the Best Paper Award from the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2009. He is a program chair of the IEEE International Conference on Computer Vision (ICCV) 2009 and an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),* and the *International Journal of Computer Vision (IJCV)*. He is a fellow of the IEEE.