Classification via Semi-Riemannian Spaces*

Deli Zhao[1] Zhouchen Lin[2] Xiaoou Tang[1,2]

[1]Dept. of Information Engineering
The Chinese University of Hong Kong
{dlzhao,xtang}@ie.cuhk.edu.hk

[2]Microsoft Research Asia
Beijing, China
{zhoulin,xitang}@microsoft.com

Abstract

In this paper, we develop a geometric framework for linear or nonlinear discriminant subspace learning and classification. In our framework, the structures of classes are conceptualized as a semi-Riemannian manifold which is considered as a submanifold embedded in an ambient semi-Riemannian space. The class structures of original samples can be characterized and deformed by local metrics of the semi-Riemannian space. Semi-Riemannian metrics are uniquely determined by the smoothing of discrete functions and the nullity of the semi-Riemannian space. Based on the geometrization of class structures, optimizing class structures in the feature space is equivalent to maximizing the quadratic quantities of metric tensors in the semi-Riemannian space. Thus supervised discriminant subspace learning reduces to unsupervised semi-Riemannian manifold learning. Based on the proposed framework, a novel algorithm, dubbed as Semi-Riemannian Discriminant Analysis (SRDA), is presented for subspace-based classification. The performance of SRDA is tested on face recognition (singular case) and handwritten capital letter classification (nonsingular case) against existing algorithms. The experimental results show that SRDA works well on recognition and classification, implying that semi-Riemannian geometry is a promising new tool for pattern recognition and machine learning.

1. Introduction

Classification is a fundamental task in pattern recognition. Linear discriminant analysis is a popular fashion of performing classification, of which researchers are fond due to its simplicity, principled treatment, and comparable performance. We devote this paper to addressing the linear classification issue from the perspective of semi-Riemannian geometry [18].

1.1. Fisher Criterion and Discrepancy Criterion

Fisher's Linear Discriminant Analysis (LDA) [7] is well known as the classic work on discriminant analysis. Fisher performed the structural analysis of classes by maximizing the between-class scatter and simultaneously minimizing the within-class scatter via the ratio of them — known as *Fisher criterion*. Fisher criterion now works as a fundamental way of integrating dual quantities between classes and within classes in classification. However, the singularity of the within-class scatter matrix (or its analogues) usually leads to the computational issue when performing the generalized eigen-analysis.

In recent decades, a great deal of effort on quadratic or linear discrimination has been devoted towards tackling the singularity problem. Overall, there are mainly three types of approaches: 1) the regularization of the within-class covariance matrix such as the work in [8, 9], 2) Principal Component Analysis (PCA) based dimensionality reduction such as Fisherfaces [2], and 3) subspace-based variants of LDA such as [2, 5, 34, 29, 32, 28, 30]. There are also matrix-decomposition-based approaches like [13, 33] and the correlation-based methods such as [17]. However, less attention has been paid to investigating class structures since Fisher's LDA. Most works for subspace-based classification can be traced back to LDA and Fisher criterion.

Recently, the development of manifold learning [23, 22] leads researchers' attention to the investigation of local structures of data in the pattern recognition community. Such kind of analysis is necessary in cases where data structures are complex. Linear methods related to manifold learning have been proposed for subspace-based recognition [11, 31].

Another recent development on linear discrimination is that *discrepancy criterions* took the role of integrating (global or local) between-class scatters and (global or local) within-class scatters instead of ratios like the traditional Fisher criterion. Global methods include Maximum Margin Criterion (MMC) [14] and Kernel Scatter-Difference Analysis (KSDA) [15, 16], and local ones include Stepwise Non-

[*] The work was performed when Deli Zhao worked in Microsoft Research Asia.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

parametric Maximum Margin Criterion (SNMMC) [21], Local and Weighted Maximum Margin Discriminant Analysis (LWMMDA) [26], and Average Neighborhood Margin Maximization (ANMM) [24]. A discrepancy criterion is also implicitly contained in [25]. Such kinds of methods successfully avoid the generalized eigen-decomposition problem, thereby are free from the computational dilemma of singularity.

Table 1. Notations.

tr($\mathbf{A}$)	The trace of the matrix $\mathbf{A}$.
$\mathbf{A}^T$	The transpose of $\mathbf{A}$.
$\mathbf{I}_{p\times p}$	The identity matrix of size $p \times p$.
$\mathbf{e}_p$	The all-one column vector of length p.
$\mathbb{R}^n$	The n-dimensional Euclidean space.
$\mathbb{S}^n_{\mathbf{x}}$	The n-dimensional original sample space.
$\mathbf{x}_i$	The i-th sample, $\mathbf{x}_i \in \mathbb{S}^n_{\mathbf{x}}$, $i = 1, \ldots, m$.
$\mathcal{S}_{\mathbf{x}}$	$\mathcal{S}_{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$.
$\mathbf{X}$	$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$.
$\mathcal{C}_j$	The j-th class, $j = 1, \ldots, c$.
$\bar{\mathbf{x}}_j$	The centroid of class j.
$\omega(\mathbf{x}_i)$	The label of class that $\mathbf{x}_i$ belongs to.
$\hat{\mathbf{x}}_{i_{\hat{k}}}$	The $\hat{k}$-th neighbor of $\mathbf{x}_i$ in $\mathcal{C}_{\omega(\mathbf{x}_i)}$, $\hat{k} = 1, \ldots, \hat{K}$ and $\hat{K} \leq \lvert\mathcal{C}_{\omega(\mathbf{x}_i)}\rvert - 1$.
$\check{\mathbf{x}}^j_{i_{\check{k}}}$	The $\check{k}$-th neighbor of $\mathbf{x}_i$ in $\mathcal{C}_j$, $\check{k} = 1, \ldots, \check{K}$ and $\check{K} \leq \lvert\mathcal{C}_j\rvert$.
$\hat{\mathcal{S}}_{\mathbf{x}_i}$	$\hat{\mathcal{S}}_{\mathbf{x}_i} = \{\hat{\mathbf{x}}_{i_{\hat{1}}}, \ldots, \hat{\mathbf{x}}_{i_{\hat{K}}}, \mathbf{x}_i\}$.
$\check{\mathcal{S}}^j_{\mathbf{x}_i}$	$\check{\mathcal{S}}^j_{\mathbf{x}_i} = \{\check{\mathbf{x}}^j_{i_{\check{1}}}, \ldots, \check{\mathbf{x}}^j_{i_{\check{K}}}\}$.
$\check{\mathcal{S}}_{\mathbf{x}_i}$	$\check{\mathcal{S}}_{\mathbf{x}_i} = \{\check{\mathcal{S}}^1_{\mathbf{x}_i}, \ldots, \check{\mathcal{S}}^K_{\mathbf{x}_i}\}$.
$\mathcal{S}_{\mathbf{x}_i}$	$\mathcal{S}_{\mathbf{x}_i} = \{\check{\mathcal{S}}_{\mathbf{x}_i}, \hat{\mathcal{S}}_{\mathbf{x}_i}\}$.
I_i	The index set of elements in $\mathcal{S}_{\mathbf{x}_i}$.
$\hat{\mathbf{Y}}_i$	$\hat{\mathbf{Y}}_i = [\hat{\mathbf{y}}_{i_{\hat{1}}}, \ldots, \hat{\mathbf{y}}_{i_{\hat{K}}}, \mathbf{y}_i]$.
$\check{\mathbf{Y}}^j_i$	$\check{\mathbf{Y}}^j_i = [\check{\mathbf{y}}^j_{i_{\check{1}}}, \ldots, \check{\mathbf{y}}^j_{i_{\check{K}}}]$.
$\check{\mathbf{Y}}_i$	$\check{\mathbf{Y}}_i = [\check{\mathbf{Y}}^1_i, \ldots, \check{\mathbf{Y}}^K_i]$.
$\mathbf{Y}_i$	$\mathbf{Y}_i = [\check{\mathbf{Y}}_i, \hat{\mathbf{Y}}_i]$.
$\hat{d}_{\mathbf{x}_i, \mathbf{x}_{i_{\hat{k}}}}$	The distance between $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_{i_{\hat{k}}}$.
$\check{d}^j_{\mathbf{x}_i, \mathbf{x}_{i_{\check{k}}}}$	The distance between $\hat{\mathbf{x}}_i$ and $\check{\mathbf{x}}^j_{i_{\check{k}}}$.
$\hat{\mathbf{d}}_{\mathbf{x}_i}$	$\hat{\mathbf{d}}_{\mathbf{x}_i} = [\hat{d}_{\mathbf{x}_i, \mathbf{x}_{i_{\hat{1}}}}, \ldots, \hat{d}_{\mathbf{x}_i, \mathbf{x}_{i_{\hat{K}}}}]^T$,
$\check{\mathbf{d}}^j_{\mathbf{x}_i}$	$\check{\mathbf{d}}^j_{\mathbf{x}_i} = [\check{d}^j_{\mathbf{x}_i, \mathbf{x}_{i_{\check{1}}}}, \ldots, \check{d}^j_{\mathbf{x}_i, \mathbf{x}_{i_{\check{K}}}}]^T$.
$\check{\mathbf{d}}_{\mathbf{x}_i}$	$\check{\mathbf{d}}_{\mathbf{x}_i} = [(\check{\mathbf{d}}^1_{\mathbf{x}_i})^T, \ldots, (\check{\mathbf{d}}^K_{\mathbf{x}_i})^T]^T$.
$\mathbf{d}_{\mathbf{x}_i}$	$\mathbf{d}_{\mathbf{x}_i} = [\check{\mathbf{d}}^T_{\mathbf{x}_i}, \hat{\mathbf{d}}^T_{\mathbf{x}_i}]^T$.
$\hat{\mathbf{D}}_{\mathbf{x}_i}$	$\hat{\mathbf{D}}_{\mathbf{x}_i} = \mathrm{diag}((\hat{d}_{\mathbf{x}_i, \mathbf{x}_{i_{\hat{1}}}})^2, \ldots, (\hat{d}_{\mathbf{x}_i, \mathbf{x}_{i_{\hat{K}}}})^2)$.
$\check{\mathbf{D}}_{\mathbf{x}_i}$	$\check{\mathbf{D}}_{\mathbf{x}_i} = \mathrm{diag}((\check{d}^1_{\mathbf{x}_i, \mathbf{x}_{i_{\check{1}}}})^2, \ldots, (\check{d}^1_{\mathbf{x}_i, \mathbf{x}_{i_{\check{K}}}})^2,$ $\ldots, (\check{d}^K_{\mathbf{x}_i, \mathbf{x}_{i_{\check{K}}}})^2)$.

1.2. Our Work

1.2.1 From Data to Semi-Riemannian Manifold

Our motivations are two-fold: the viewpoint from manifold learning and the success of discrepancy criterions in classification. The theory and the algorithm in this paper are based on our perspective that *the intrinsic structure of a group of classes is, independent of ambient vector-valued representations, a low-dimensional curved manifold which is tightly related to structural associations between local classes and within classes*.

On one hand, the manifold-related manipulations are only allowed on local neighborhoods, which drives us to define the K nearest neighbor (KNN) classes of a sample (the beginning of Section 3). Treating each class as a unit and considering that discrimination relies on the relationship between a sample and its KNN classes at the same time, we introduce the concept of free degrees of discriminability of a sample and naturally form a discriminant manifold for class structures (Section 3.1.1). On the other hand, to optimize class structures, we usually need to perform the discrepancies of intra-class quantities and inter-class quantities. To do so, we introduce *semi-Riemannian metrics* [18] (Section 2) *which are the unique tools to locally integrate such kinds of dual quantities from the mathematical point of view*. Thus, the structure of classes is initially modeled as a semi-Riemannian manifold (Section 3.1.1).

Furthermore, the computation on the discriminant manifold is allowed when the coordinates of each point on it are available. To this end, we represent the coordinate of each dimension using the dissimilarities between the sample and several sampled points in each of its KNN classes (see Figure 2). Thus we obtain an ambient space with semi-Riemannian metrics where coordinates are characterized by dissimilarities between local sample pairs in intra classes and in inter classes. The discriminant manifold is considered as a semi-Riemannian submanifold of the ambient semi-Riemannian space and points on it are represented by the ambient coordinates. Thus, we complete the semi-Riemannian manifold model of class structures (Section 3.1.2).

1.2.2 Learning on Semi-Riemannian manifold

By virtue of the geometrization of class structures, learning a discriminant subspace reduces to learning the geometry of a semi-Riemannian manifold. Thus, classification is coupled with semi-Riemannian manifold learning. Moreover, we present an approach to optimize class structures in the feature space using metric tensors learnt from the ambient semi-Riemannian space (Section 3.2.1). Semi-Riemannian metric learning is developed via the discretized Laplacian smoothing of discrete functions and the nullity of the ambient space which is the special nature of semi-Riemannian spaces (Section 3.2.2). In fact, the role of semi-Riemannian metrics in semi-Riemannian manifold learning is equivalent to the media of transferring geometry from the sample space to the feature space (Section 3.2.3). Finally, a specific algorithm, dubbed as Semi-Riemannian Discriminant Analysis (SRDA), is presented for subspace-based classification (Section 3.2.4).

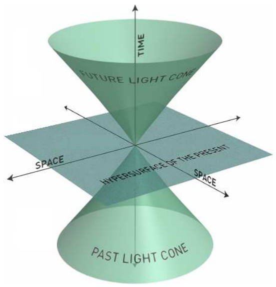


Figure 1. Illustration of a space-time. The plane is the space-time of the present. On the top is the future light cone and at the bottom the past light cone. Inside the light cone is the time-like space-time and outside the space-like space-time.

2. Fundamentals of Semi-Riemannian Spaces

Semi-Riemannian manifolds[1] are smooth manifolds furnished with semi-Riemannian metric tensors. The geometry of semi-Riemannian manifolds is called semi-Riemannian geometry. The semi-Riemannian geometry has been extensively applied, due to the success of Einstein's general relativity, as a basic geometric tool of modeling space-times in physics. To the best of our knowledge, however, it has not been explicitly applied in pattern recognition before. Here we give a concise introduction to semi-Riemannian spaces. One may refer to [18, 6] for more details.

Geometric spaces are specified by their metrics. The metric matrix in the semi-Riemannian space $\mathbb{N}^n_\nu$ is of form

$$\mathbf{G} = \begin{bmatrix} \check{\mathbf{\Lambda}}_{p\times p} & \mathbf{0} \\ \mathbf{0} & -\hat{\mathbf{\Lambda}}_{\nu\times\nu} \end{bmatrix}, \tag{1}$$

where $\check{\mathbf{\Lambda}}_{p\times p}$ and $\hat{\mathbf{\Lambda}}_{\nu\times\nu}$ are diagonal and their diagonal entries are positive, and $p+\nu=n$. With $\mathbf{G}$, the space-time interval ds^2 in $\mathbb{N}^n_\nu$ can be expressed as

$$\mathrm{d}s^2 = \sum\nolimits_{i=1}^{p} \check{\mathbf{\Lambda}}(i,i)\mathrm{d}x_i^2 - \sum\nolimits_{i=p+1}^{p+\nu} \hat{\mathbf{\Lambda}}(i-p,i-p)\mathrm{d}x_i^2, \tag{2}$$

where ν is called the index of $\mathbb{N}^n_\nu$. $\mathbb{N}^n_\nu$ is a semi-Euclidean space if $\check{\mathbf{\Lambda}}_{p\times p} = \mathbf{I}_{p\times p}$ and $\hat{\mathbf{\Lambda}}_{\nu\times\nu} = \mathbf{I}_{\nu\times\nu}$, and a Lorentz (Minkowski) space if $\check{\mathbf{\Lambda}}_{p\times p} = \mathbf{I}_{p\times p}$ and $\nu = 1$. The space-time in Einstein's relativity theory is the case of $n=4$ and $\nu=1$. $\mathbb{N}^n_\nu$ degenerates to the Euclidean space $\mathbb{R}^n$ if $\nu=0$. Semi-Riemannian spaces are more general curved spaces with many special properties in their own right than Riemannian spaces.

Suppose that $\mathbf{r} = [\check{\mathbf{r}}^T, \hat{\mathbf{r}}^T]^T$ is a vector in $\mathbb{N}^n_\nu$. Then a metric tensor $g(\mathbf{r},\mathbf{r})$ with respect to $\mathbf{G}$ is expressible as

$$g(\mathbf{r},\mathbf{r}) = \mathbf{r}^T\mathbf{G}\mathbf{r} = \check{\mathbf{r}}^T\check{\mathbf{\Lambda}}\check{\mathbf{r}} - \hat{\mathbf{r}}^T\hat{\mathbf{\Lambda}}\hat{\mathbf{r}}. \tag{3}$$

The vector $\mathbf{r}$ is called space-like if $g(\mathbf{r},\mathbf{r}) > 0$ or $\mathbf{r} = 0$, time-like if $g(\mathbf{r},\mathbf{r}) < 0$, and null (or light-like, isotropic) if $g(\mathbf{r},\mathbf{r}) = 0$ and $\mathbf{r} \neq 0$. Figure 1 illustrates a space-time.

3. Classification via Semi-Riemannian Spaces

What our framework differs from traditional ones on classification is that class structures are modeled as a semi-Riemannian submanifold embedded in an ambient semi-Riemannian space. As a result, learning a discriminant subspace for classification reduces to learning the geometry of the semi-Riemannian submanifold. Therefore, classification is coupled with manifold learning in semi-Riemannian spaces.

In our framework, the K nearest neighbor (KNN) classes of a sample $\mathbf{x}_i$ are involved.

Definition 1. ***KNN Classes****. For a sample $\mathbf{x}_i$, its KNN classes are defined as:*

$$\{i_1,\ldots,i_K\} = \arg\min_j \|\bar{\mathbf{x}}_j - \mathbf{x}_i\|_{\mathbb{S}^n_{\mathbf{x}}},\ j=1,\ldots,c. \tag{4}$$

The distance $\|\bar{\mathbf{x}}_j - \mathbf{x}_i\|_{\mathbb{S}^n_{\mathbf{x}}}$ depends on the attributes of the sample space $\mathbb{S}^n_{\mathbf{x}}$. It may be the Euclidean distance, one of statistical distances like the Chi-square [10], or the approximated geodesic distance [23].

It suffices to emphasize that the original motivation of the definition of KNN classes comes from the surprising effectiveness of discriminant subspaces learnt only from several nearest neighbor classes of a *query* sample in some resulting feature spaces [27][2]. Readers may refer to [27] for more details.

3.1. Modeling Class Structures as a Semi-Riemannian Submanifold

3.1.1 Associating Class Structures with a Semi-Riemannian Manifold

First, let us introduce the concept of "degrees of discriminability" of a sample. We contend that what is crucial to the discrimination of a sample is its KNN classes rather than all the involved classes. Namely, only KNN classes of the sample dominate the capability of discriminating it. Therefore, our concerns are only focused on mining the structural relationship between the sample and its related KNN classes. For a specific sample $\mathbf{x}_i$, one of its KNN classes accounts for one degree of discriminating it from other samples in different classes. So KNN classes account for K degrees of the discriminability. On the other hand, class $\omega(\mathbf{x}_i)$ in question accounts for one degree of associating $\mathbf{x}_i$ with its own class. Putting the inter-class degrees and the intra-class degree together, we say that the discriminability of the sample is of degree $K+1$. Furthermore, suppose that spanning axes are constructed from $\mathbf{x}_i$ to class $\omega(\mathbf{x}_i)$ and each of its KNN classes. Therefore, the discrimination admits a space that is supported by $K+1$ spanning axes. Denote this space by $\mathbb{M}^{K+1}_1$. From the above analysis, we know the discriminant space $\mathbb{M}^{K+1}_1$ has $K+1$ free degrees and thus is a manifold of dimension $K+1$. The left part in Figure 2 illustrates a toy example of $\mathbb{M}^{K+1}_1$. For each point on $\mathbb{M}^{K+1}_1$,

[1] Semi-Riemannian manifolds are also called pseudo-Riemannian manifolds.

[2] Note that the content related to KNN classes in [27] was not presented in the journal version [29].

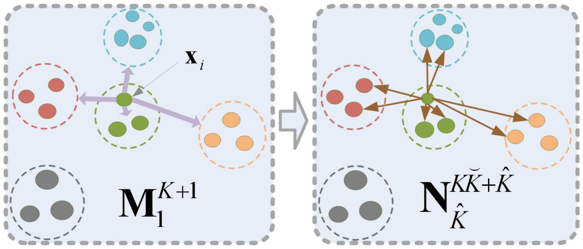


Figure 2. Schematic illustrations of the semi-Riemannian submanifold $\mathbb{M}_1^{K+1}$ and the ambient semi-Riemannian space $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$. Here $K = 3$, $\check{K} = \hat{K} = 2$. The dots with the same color belong to the same class. The left figure depicts the abstract discriminant manifold $\mathbb{M}_1^4$ and the right figure depicts the ambient $\mathbb{N}_2^8$.

we may endow the i-th axis a coordinate s_i. Thus, the vector $\mathbf{s} = [s_1, \ldots, s_K, s_{K+1}]^T$ is the coordinate representation of a point on $\mathbb{M}_1^{K+1}$. Furnishing $\mathbb{M}_1^{K+1}$ with a metric of form

$$\mathbf{G}^{\mathbb{M}} = \begin{bmatrix} \mathbf{\Lambda}_{K\times K} & \mathbf{0} \\ \mathbf{0} & -\phi \end{bmatrix}, \tag{5}$$

then $\mathbb{M}_1^{K+1}$ is a semi-Riemannian manifold with the index one, i.e., a Lorentz manifold. Thus, on the tangent space $\mathbb{T}\mathbb{M}_1^{K+1}$, the quadratic form of the vector, with respect to $\mathbf{G}^{\mathbb{M}}$, is measured by $g(\mathbf{s}, \mathbf{s}) = \sum_{i=1}^{K} \mathbf{\Lambda}(i,i)s_i^2 - \phi s_{K+1}^2$. Intuitively, the positive definite part of $\mathbf{G}^{\mathbb{M}}$ measures the inter-class quantity and the negative definite part of $\mathbf{G}^{\mathbb{M}}$ measures the intra-class quantity. To make the concept easily understandable, let us relate the semi-Riemannian manifold $\mathbb{M}_1^4$ with the space-time in the relativity theory. The role of the first three inter-class degrees of discriminability corresponds to that of dimensions of spatial location in the space-time and the role of the intra-class degree is equivalent to that of the dimension of time. Hence, we construct a semi-Riemannian manifold for the discrimination problem.

3.1.2 Embedding Discriminant Manifolds into Ambient Semi-Riemannian Spaces

Questions naturally arise from the conceptualization of class structures as a semi-Riemannian manifold: how to form the coordinates of $\mathbb{M}_1^{K+1}$ and how to parameterize it for computation?

The structural relationship in pattern analysis is in general characterized by distances or more general dissimilarities between samples. We may apply dissimilarities from each sample to its KNN classes as the coordinates of representation. Each of the degrees of discriminability is represented by the corresponding sample-to-class dissimilarity. The problem is, however, that it is unclear how to determine the sample-to-class dissimilarity. Hopefully, it can be handled by sampling points in KNN classes. More specifically, we sample $\check{K}$ points in each KNN class of $\mathbf{x}_i$. Thus, we exploit the distances between $\mathbf{x}_i$ and its $\check{K}$ points in each KNN class as an ambient representation of the general sample-to-interclass dissimilarity, denoting it by $(\check{d}^j_{\mathbf{x}_i,\mathbf{x}_{i_{\check{1}}}}, \ldots, \check{d}^j_{\mathbf{x}_i,\mathbf{x}_{i_{\check{K}}}})^T$. Similarly, we employ the distances between $\mathbf{x}_i$ and $\hat{K}$ points in class $\omega(\mathbf{x}_i)$ to represent the sample-to-intraclass dissimilarity, denoting it by $(\hat{d}_{\mathbf{x}_i,\mathbf{x}_{i_{\hat{1}}}}, \ldots, \hat{d}_{\mathbf{x}_i,\mathbf{x}_{i_{\hat{K}}}})^T$. Putting them together, we eventually get an explicit coordinate representation of a point on $\mathbb{M}_1^{K+1}$ induced at the sample $\mathbf{x}_i$, i.e., $\mathbf{d}_{\mathbf{x}_i} = [\check{\mathbf{d}}^T_{\mathbf{x}_i}, \hat{\mathbf{d}}^T_{\mathbf{x}_i}]^T$.

The above manipulations of up-sampling (enlarging dimension) are essentially to explicate one intrinsic coordinate with more extrinsic parameters (or ambient representation), which is the process of embedding a low-dimensional manifold into a high-dimensional ambient space whose metrics and coordinates are non-ambiguous.

With the up-sampling, each point on $\mathbb{M}_1^{K+1}$ is endowed with a $(K\check{K} + \hat{K})$-tuple coordinate representation $\mathbf{d}_{\mathbf{x}_i}$. Henceforth, we obtain a new semi-Riemannian space $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$ furnished with the metric

$$\mathbf{G}^{\mathbb{N}} = \begin{bmatrix} \check{\mathbf{\Lambda}}_{(K\check{K})\times(K\check{K})} & \mathbf{0} \\ \mathbf{0} & -\hat{\mathbf{\Lambda}}_{\hat{K}\times\hat{K}} \end{bmatrix}. \tag{6}$$

In the manifold language, $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$ is called the ambient space of $\mathbb{M}_1^{K+1}$, and $\mathbb{M}_1^{K+1}$ is a semi-Riemannian submanifold of $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$. In general, K, $\check{K}$, and $\hat{K}$ are small positive integers. So $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$ is a low-dimensional semi-Riemannian space, implying that, even in the ambient space, the global structure of classes is a low-dimensional semi-Riemannian manifold. It is necessary to point it out that the dimensions of $\mathbb{M}_1^{K+1}$ and $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$ are completely independent of the dimension of the sample space $\mathbb{S}^n_{\mathbf{x}}$. The right part in Figure 2 illustrates a toy example of $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$.

3.2. Learning Discriminant Subspaces via Semi-Riemannian Manifold Learning

With the newly built space, discriminant subspaces can be learnt from the semi-Riemannian geometry of $\mathbb{M}_1^{K+1}$. To this end, we need to handle two matters. The first is the optimization framework for learning discriminant subspaces from local semi-Riemannian geometry embodied by the ambient metric $\mathbf{G}_i^{\mathbb{N}}$ at $\mathbf{x}_i$. The second is the feasible solution of the metric $\mathbf{G}_i^{\mathbb{N}}$ that is favorable of discrimination.

In this paper, we assume that the feature space is Euclidean, meaning that the length of $\mathbf{y}$ is measured by the ℓ_2 norm [12]: $\|\mathbf{y}\|^2_{\ell_2} = \mathbf{y}^T\mathbf{y} = \mathrm{tr}(\mathbf{y}\mathbf{y}^T)$.

3.2.1 Alignment of Metric Tensors in Semi-Riemannian Space

Suppose that the metric matrix $\mathbf{G}_i^{\mathbb{N}}$ at $\mathbf{x}_i$ has already been determined. If we penalize the feature space $\mathbb{S}^d_{\mathbf{y}}$ using $\mathbf{G}_i^{\mathbb{N}}$, meaning that the metric keeps invariant in $\mathbb{S}^d_{\mathbf{y}}$, then the optimization of learning discriminant subspaces can be performed using the metric tensor $g(\mathbf{d}_{\mathbf{y}_i}, \mathbf{d}_{\mathbf{y}_i})$ in $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$. It is straightforward to see that $g(\mathbf{d}_{\mathbf{y}_i}, \mathbf{d}_{\mathbf{y}_i})$ can be written as

$$g(\mathbf{d}_{\mathbf{y}_i}, \mathbf{d}_{\mathbf{y}_i}) = \mathbf{d}_{\mathbf{y}_i}^T \mathbf{G}_i^{\mathbb{N}} \mathbf{d}_{\mathbf{y}_i} = \check{\mathbf{d}}_{\mathbf{y}_i}^T \check{\Lambda}_i \check{\mathbf{d}}_{\mathbf{y}_i} - \hat{\mathbf{d}}_{\mathbf{y}_i}^T \hat{\Lambda}_i \hat{\mathbf{d}}_{\mathbf{y}_i}. \tag{7}$$

Note here that $\mathbf{G}_i^{\mathbb{N}}$ is learnt from the structure of original samples[3] and applied to the feature space. What we desire is the larger inter-class margins and at the same time the smaller intra-class margins in $\mathbb{S}_{\mathbf{y}}^d$, which can be fulfilled by maximizing the metric tensor $g(\mathbf{d}_{\mathbf{y}_i}, \mathbf{d}_{\mathbf{y}_i})$. The maximization of $g(\mathbf{d}_{\mathbf{y}_i}, \mathbf{d}_{\mathbf{y}_i})$ is in effect the principal component analysis in $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$. We may handle the maximization by taking advantage of Zhao *et al.*'s theoretic framework [35] on the alignment of local geometry. More specifically, let the difference operator $\mathbf{D}$ be

$$\mathbf{D} = \begin{bmatrix} \mathbf{I}_{(K\check{K}+\hat{K})\times(K\check{K}+\hat{K})} \\ -\mathbf{e}_{K\check{K}+\hat{K}}^T \end{bmatrix}. \tag{8}$$

Then we have the following theorem pertaining to the metric alignment[4] $\sum_{i=1}^m g(\mathbf{d}_{\mathbf{y}_i}, \mathbf{d}_{\mathbf{y}_i})$.

Theorem 1. *$\sum_{i=1}^m g(\mathbf{d}_{\mathbf{y}_i}, \mathbf{d}_{\mathbf{y}_i}) = tr(\mathbf{YLY}^T)$, where $\mathbf{L} = \sum_{i=1}^m \mathbf{S}_i \mathbf{L}_i \mathbf{S}_i^T$, $\mathbf{L}_i = \mathbf{DG}_i^{\mathbb{N}}\mathbf{D}^T$, and $\mathbf{S}_i$ is the binary matrix of size $m\times(K\check{K}+\hat{K}+1)$ whose structure is that $(\mathbf{S}_i)_{pq} = 1$ if the q-th vector in $\mathbf{Y}_i$ is the p-th vector in $\mathbf{Y}$.*

With Theorem 1, it is easy to know that the optimal nonlinear embedding of class structures is the d-column eigenvectors of $\mathbf{L}$ corresponding to the first d largest eigenvalues. This type of nonlinear embedding can be exploited for class visualization and the efficient computation of linear subspace [4]. If there is a linear isometric transformation between the low-dimensional feature vector $\mathbf{y}$ and the original sample $\mathbf{x}$, i.e., $\mathbf{y} \mapsto \mathbf{Uy} = \mathbf{x}$, where $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{d\times d}$, then the linear discriminant subspace $\mathbf{U}$ can be derived as the principal subspace of $\mathbf{XLX}^T$. The principal subspace $\mathbf{U}$ learnt via the semi-Riemannian subspace $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$ is the optimal subspace for discrimination, in the sense that the local inter-class structures are enlarged while the local intra-class structures are contracted.

It suffices to note that one may form the difference relationship by various operators $\mathbf{D}$ in (8). The alignment framework is still applicable for such modifications.

3.2.2 Semi-Riemannian Metric Learning

The metric $\mathbf{G}_i^{\mathbb{N}}$ is one of the crucial factors that govern the geometry of $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$. We may apply $\mathbf{G}_i^{\mathbb{N}}$ to deform local spaces of $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$ towards the optimization of class structures. Therefore, we can determine appropriate metrics that are favorable of discrimination in $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$. The metric $\mathbf{G}_i^{\mathbb{N}}$ consists of two parts: the positive definite part $\check{\Lambda}_i$ and the negative definite part $-\hat{\Lambda}_i$. In this section, we introduce an alternative way to determine $\check{\Lambda}_i$ and $-\hat{\Lambda}_i$, e.g., by the smoothing of discrete functions and the nullity of $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$.

A. Smoothing. We may write $\check{\mathbf{d}}_{\mathbf{y}_i}^T \check{\Lambda}_i \check{\mathbf{d}}_{\mathbf{y}_i}$ in the form of components:

$$\check{\mathbf{d}}_{\mathbf{y}_i}^T \check{\Lambda}_i \check{\mathbf{d}}_{\mathbf{y}_i} = \sum_{\check{k}=1}^{K\check{K}} \check{\mathbf{d}}_{\mathbf{y}_i}(\check{k}) \check{\Lambda}_i(\check{k},\check{k}) \check{\mathbf{d}}_{\mathbf{y}_i}(\check{k}). \tag{9}$$

It is evident that the large component $\check{\mathbf{d}}_{\mathbf{y}_i}(\check{k})\check{\Lambda}_i(\check{k},\check{k})\check{\mathbf{d}}_{\mathbf{y}_i}(\check{k})$ will suppress the small ones when maximizing $g(\mathbf{d}_{\mathbf{y}_i}, \mathbf{d}_{\mathbf{y}_i})$. This functional non-uniformness is harmful for learning an optimal discriminant subspace. However, this weakness can be allievated by smoothing the elements in $\{\check{\mathbf{d}}_{\mathbf{x}_i}(\check{1})\check{\Lambda}_i(\check{1},\check{1})\check{\mathbf{d}}_{\mathbf{x}_i}(\check{1}),\ldots,\check{\mathbf{d}}_{\mathbf{x}_i}(K\check{K})\check{\Lambda}_i(K\check{K},K\check{K})\check{\mathbf{d}}_{\mathbf{x}_i}(K\check{K})\}$. Let $\check{\mathbf{g}}_i=[\check{\Lambda}_i(\check{1},\check{1}),\ldots,\check{\Lambda}_i(K\check{K},K\check{K})]^T$ and $\hat{\mathbf{g}}_i=[\hat{\Lambda}_i(\hat{1},\hat{1}),\ldots,\hat{\Lambda}_i(\hat{K},\hat{K})]^T$. Then the smoothing can be performed on $\check{\mathbf{D}}_{\mathbf{x}_i}\check{\mathbf{g}}_i$ due to that $\check{\mathbf{d}}_{\mathbf{x}_i}^T \check{\Lambda}_i \check{\mathbf{d}}_{\mathbf{x}_i} = \mathbf{e}^T \check{\mathbf{D}}_{\mathbf{x}_i} \check{\mathbf{g}}_i$. Here we employ the following discretized Laplacian smoothing

$$\begin{cases} \arg\min\limits_{\check{\mathbf{g}}_i} \|\check{\mathbf{F}}\check{\mathbf{D}}_{\mathbf{x}_i}\check{\mathbf{g}}_i\|^2, \\ \text{s.t. } \mathbf{e}^T\check{\mathbf{g}}_i = 1, \end{cases} \tag{10}$$

where $\check{\mathbf{F}}$ is the first-order difference operator

$$\check{\mathbf{F}} = [\mathbf{I}_{(K\check{K}-1)\times(K\check{K}-1)} \;\; \mathbf{0}_{(K\check{K}-1)\times 1}] + \tag{11}$$

$$[\mathbf{0}_{(K\check{K}-1)\times 1} \;\; -\mathbf{I}_{(K\check{K}-1)\times(K\check{K}-1)}]. \tag{12}$$

Note that $\check{\mathbf{F}}^T\check{\mathbf{F}}$ is the Neuman discretization of Laplacian [19, 3].

B. Setting $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$ Locally Null. Null (or light-like) manifolds are typical examples in semi-Riemannian spaces [6]. In classification, a null manifold has its physical nature in its own right. As introduced in the preceding section, a null vector $\mathbf{d}_{\mathbf{x}_i}$ in $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$ is the vector that vanishes the metric tensor: $g(\mathbf{d}_{\mathbf{x}_i}, \mathbf{d}_{\mathbf{x}_i}) = 0$, i.e., $\check{\mathbf{d}}_{\mathbf{x}_i}^T \check{\Lambda}_i \check{\mathbf{d}}_{\mathbf{x}_i} = \hat{\mathbf{d}}_{\mathbf{x}_i}^T \hat{\Lambda}_i \hat{\mathbf{d}}_{\mathbf{x}_i}$. Equivalently, we have $\mathbf{e}_{\hat{K}}^T \hat{\mathbf{D}}_{\mathbf{x}_i} \hat{\mathbf{g}}_i = \mathbf{e}_{K\check{K}}^T \check{\mathbf{D}}_{\mathbf{x}_i} \check{\mathbf{g}}_i$. Putting the smoothing and the nullity together, we get the optimization for the negative definite part of metric $\mathbf{G}_i^{\mathbb{N}}$.

$$\begin{cases} \arg\min\limits_{\hat{\mathbf{g}}_i} \|\hat{\mathbf{F}}\hat{\mathbf{D}}_{\mathbf{x}_i}\hat{\mathbf{g}}_i\|^2, \\ \text{s.t. } \mathbf{e}_{\hat{K}}^T \hat{\mathbf{D}}_{\mathbf{x}_i} \hat{\mathbf{g}}_i = \mathbf{e}_{K\check{K}}^T \check{\mathbf{D}}_{\mathbf{x}_i} \check{\mathbf{g}}_i, \end{cases} \tag{13}$$

where $\hat{\mathbf{F}}$ is the difference operator similar to $\check{\mathbf{F}}$. For optimizations (10) and (13), we have the following theorem.

Theorem 2. $\check{\mathbf{g}}_i = \frac{\check{\mathbf{D}}_{\mathbf{x}_i}^{-1}\mathbf{e}_{K\check{K}}}{\mathbf{e}_{K\check{K}}^T \check{\mathbf{D}}_{\mathbf{x}_i}^{-1}\mathbf{e}_{K\check{K}}}$ *and* $\hat{\mathbf{g}}_i = \frac{\mathbf{e}_{K\check{K}}^T \check{\mathbf{D}}_{\mathbf{x}_i}\check{\mathbf{g}}_i}{\hat{K}} \hat{\mathbf{D}}_{\mathbf{x}_i}^{-1}\mathbf{e}_{\hat{K}}$.

From Theorem 2, we see that $\check{\mathbf{g}}_i$ and $\hat{\mathbf{g}}_i$ are independent of the difference operators $\check{\mathbf{F}}$ and $\hat{\mathbf{F}}$, respectively.

3.2.3 Local Geometry Transfer via Metrics

Readers may notice that $\mathbf{G}_i^{\mathbb{N}}$ is learnt from $\mathcal{S}_{\mathbf{x}}$ (in Section 3.2.2) but employed for learning $\mathcal{S}_{\mathbf{y}}$ (in Section 3.2.1), the process of which is the geometry transfer. The ambient $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$ is governed by many local $\mathbf{G}_i^{\mathbb{N}}$s which reveal the geometric distribution of class structures of original samples. $\mathcal{S}_{\mathbf{x}}$ and $\mathcal{S}_{\mathbf{y}}$ are investigated in the same

[3] It will be presented in the next section.

[4] We omit the proofs of the theorems in this paper due to lack of space.

Table 2. Algorithm of SRDA

1. For each $\mathbf{x}_i$, search the NN point sets $\check{\mathcal{S}}_{\mathbf{x}_i}$ and $\hat{\mathcal{S}}_{\mathbf{x}_i}$, record the index set I_i of $\mathcal{S}_{\mathbf{x}_i}$, and form the dissimilarity vector $\check{\mathbf{d}}_{\mathbf{x}_i}$ and $\hat{\mathbf{d}}_{\mathbf{x}_i}$.
2. Compute the metric matrix $\mathbf{G}_i^{\mathbb{N}}$ using Theorem 2, and form $\mathbf{L}$ by $\mathbf{L}(I_i, I_i) \longleftarrow \mathbf{L}(I_i, I_i) + \mathbf{D}\mathbf{G}_i^{\mathbb{N}}\mathbf{D}^T$, where $\mathbf{L}$ is initialized by a zero matrix.
3. Obtain $\mathbf{U}$ by computing the eigenvectors of $\mathbf{X}\mathbf{L}\mathbf{X}^T$ associated with the first d largest eigenvalues, and project samples: $\mathbf{Y} = \mathbf{U}^T\mathbf{X}$.
4. Choose an optimal γ in $[0.5, 1]$ with the adaption $\hat{\Lambda}_i \leftarrow \gamma\hat{\Lambda}_i$ and $\check{\Lambda}_i \leftarrow (1-\gamma)\check{\Lambda}_i$ by cross validation.

semi-Riemannian space. Henceforth, $\mathcal{S}_{\mathbf{y}}$ admit the metric $\mathbf{G}_i^{\mathbb{N}}$ in $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$. The functionality of $\mathbf{G}_i^{\mathbb{N}}$ for $\mathcal{S}_{\mathbf{y}}$ is to locally penalize the corresponding Euclidean distances $\{\mathbf{d}_{\mathbf{y}_1}, \ldots, \mathbf{d}_{\mathbf{y}_m}\}$ according to the learnt geometric structures when maximizing the metric tensor $g(\mathbf{d}_{\mathbf{y}_i}, \mathbf{d}_{\mathbf{y}_i})$. The role of $\mathbf{G}_i^{\mathbb{N}}$ in semi-Riemannian manifold learning is similar to that of locally linear fittings in the LLE algorithm [22] in traditional manifold learning, transferring the local geometry from the sample space to the feature space.

The enforcement of nullity of $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$ is in effect to balance the inter-class scatter $\check{\mathbf{d}}_{\mathbf{x}_i}^T\check{\Lambda}_i\check{\mathbf{d}}_{\mathbf{x}_i}$ and the intra-class scatter $\hat{\mathbf{d}}_{\mathbf{x}_i}^T\hat{\Lambda}_i\hat{\mathbf{d}}_{\mathbf{x}_i}$, thus leading $\mathbf{G}_i^{\mathbb{N}}$ to be the baseline of determining the final attribute of $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$ for classification. Maximizing $g(\mathbf{d}_{\mathbf{y}_i}, \mathbf{d}_{\mathbf{y}_i})$ means pulling $\mathcal{S}_{\mathbf{y}}$ towards the space-likeness in $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$. We empirically find that the discriminability will be enhanced if $\mathcal{S}_{\mathbf{x}}$ is time-like in $\mathbb{N}_{\hat{K}}^{K\check{K}+\hat{K}}$. The time-likeness of $\mathcal{S}_{\mathbf{x}}$ is easily achievable by multiplying a positive factor γ to $\hat{\Lambda}_i$, i.e, $\hat{\Lambda}_i \leftarrow \gamma\hat{\Lambda}_i$, where $\gamma \in [0.5, 1]$, and performing $\check{\Lambda}_i \leftarrow (1-\gamma)\check{\Lambda}_i$ at the same time.

3.2.4 Semi-Riemannian Discriminant Analysis

By means of the formulated framework in a semi-Riemannian space, we now give a specific algorithm, Semi-Riemannian Discriminant Analysis (SRDA), for classification or discriminant subspace learning. The elements in $\check{\mathcal{S}}_{\mathbf{x}_i}^j$ are determined by the nearest neighbor points in the j-th KNN class of $\mathbf{x}_i$ and $\hat{\mathcal{S}}_{\mathbf{x}_i}$ by the nearest neighbor points of $\mathbf{x}_i$ in class $\omega(\mathbf{x}_i)$. The algorithm of SRDA is summarized in Table 2.

4. Experiments

Experiments are conducted on face recognition and handwritten capital letter classification to test the performance of SRDA against traditional and newly proposed algorithms on recognition and classification. The former is the singular case (small sample size) while the latter is not.

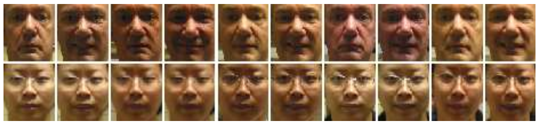


Figure 3. Facial images of two subjects in the FRGC version 2.

Table 3. Recognition results on experiment 4 of FRGC version 2.

–	On raw data	On PCA features	On LBP features
LBP	90.53 ± 0.74 (2891)	–	–
PCA	86.85 ± 1.17 (150)	–	93.48 ± 0.90 (300)
LDA	–	93.83 ± 0.83 (50)	–
LPP	–	91.32 ± 0.75 (65)	–
MFA	–	94.08 ± 0.96 (35)	–
MMC	87.48 ± 0.81 (570)	90.38 ± 0.82 (30)	94.72 ± 0.62 (540)
SNMMC	91.69 ± 0.66 (120)	91.82 ± 0.75 (100)	96.47 ± 0.61 (510)
ANMM	91.35 ± 0.97 (170)	91.69 ± 0.71 (105)	96.18 ± 0.60 (510)
SRDA	**94.19** ± 0.54 (80)	**94.24** ± 0.76 (140)	**98.09** ± 0.49 (850)

The methods for comparison include PCA, LDA, LPP [11], MFA [31], MMC [14], SNMMC [21], and ANMM [24].

For simplicity and generality, we directly use the ℓ_2 norm (for raw data and PCA features) and the Chi-square (for LBP features) to compute $\|\mathbf{x}_j - \mathbf{x}_i\|_{\mathbb{S}_{\mathbf{x}}^n}$. The nearest neighbor classifier is employed on extracted features for recognition and classification.

4.1. Singular Case: Face Recognition

We perform the experiments on a subset selected from the query set of experiment 4 in FRGC version 2 [20]. The facial data set was used in [36]. There are 200 subjects in the gallery and probe set and 116 subjects in the training set. There are ten facial images for each subject. The identities of subjects in the training set is different from those of subjects in the gallery and probe set. The facial images are aligned according to the positions of eyes and mouths, and cropped to the size of 51×57. Figure 3 shows facial images of two subjects. For each subject, five images are randomly selected as the gallery set and the remaining for the probe set. Such a trial is repeated 20 times.

We apply the Local Binary Pattern (LBP) algorithm to extract visual features. The usage of LBP here is consistent with that in [1]: pattern $(8, 2)$, 59 bins, and 7×7 image blocks. For PCA-combined methods, the number of principal components is optimally determined. Besides, for LPP, the number of nearest neighbors is chosen as 3, and for MFA the number of the inter-class and intra-class nearest neighbors are chosen as 40 and 3, respectively. These parameters are tuned optimally in the training phase. For ANMM, as suggested by authors [24], we take ten inter-class and intra-class nearest neighbors, respectively. For SRDA, we take $K = 5$, $\check{K} = 2$, and $\hat{K} = 9$. The results are shown in Table 3.

From Table 3, we see that SRDA performs better than the other methods on the raw data whereas LDA and MFA have the comparable performance with SRDA on PCA features. What's interesting is that the performance of SRDA

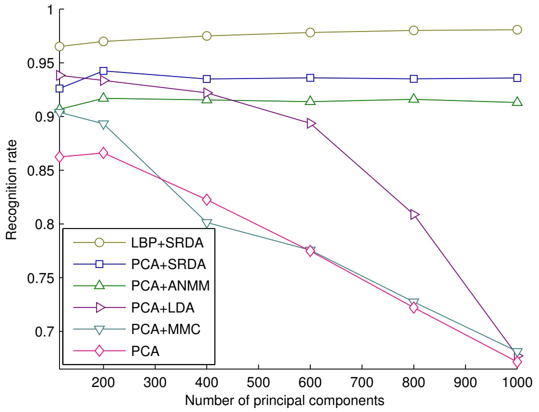
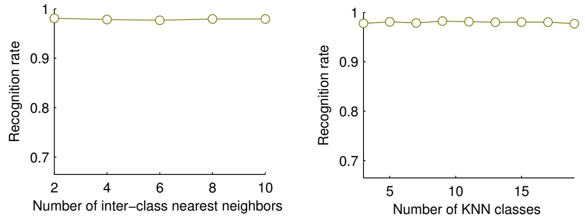


Figure 4. Recognition rates of involved algorithms over the variation of number of principal components. The related parameters in all algorithms keep invariant.



(a) (b)

Figure 5. Recognition rates of SRDA against the variations of numbers of KNN classes and inter-class nearest neighbors. (a) $K = 5$ and $\hat{K} = 9$. (b) $\check{K} = 2$ and $\hat{K} = 9$.

keeps almost invariant on the raw data and on PCA features, implying that PCA does not contribute much to enhance the discrimination. The major contribution of PCA in classification is on dimension reduction and reducing the computational complexity. Figure 4 illustrates the robustness of involved algorithms over the variation of the number of principal components in PCA. We can see that SRDA and ANMM behave robustly. This is because the distances between neighboring projected samples only slightly vary with the increment of the number of principal components when they are sufficiently large. Besides, the recognition performance of SRDA is improved on LBP visual features. Notice that the discriminability of SRDA depends on the accuracy of characterization of local class structures. And LBP features are superior to the raw data and PCA features on measuring the similarities among faces. Thus, it is not surprising that SRDA performs better on LBP features than on the raw data and on PCA features. Figure 5 shows the recognition rates of SRDA against the variations of numbers of KNN classes and inter-class nearest neighbors. Again, SRDA exhibits strong robustness.

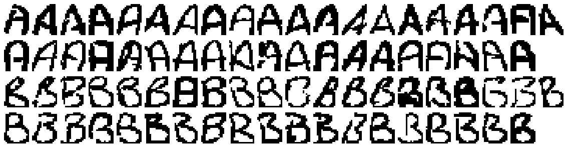


Figure 6. Handwritten capital letters.

Table 4. Classification results on handwritten capital letters.

Algorithm	Accuracy (%)	Algorithm	Accuracy (%)
raw data	72.6 ± 1.60		
PCA	71.62 ± 1.86 (20)	MMC	73.56 ± 1.28 (30)
LDA	38.76 ± 1.76 (30)	SNMMC	60.03 ± 1.86 (40)
LPP	53.63 ± 2.72 (30)	ANMM	77.79 ± 1.79 (30)
MFA	45.18 ± 2.11 (70)	SRDA	**82.71** ± 2.49 (20)

4.2. Nonsingular Case: Handwritten Capital Letter Classification

The capital letter data set (including the USPS handwritten digits) used in this experiment comes from Sam Roweis's homepage[5]. The capital letters are the cropped 20 × 16 images of 'A' through 'Z'. There are 39 examples for each class. Figure 6 shows the examples of 'A' and 'B'. We randomly select 19 samples from each class for training. So, there are all together 494 images in the training set. Therefore, the methods such as LDA, LPP, and MFA will not encounter the singularity problem of computation. So, we can directly employ them for discrimination. The trial is repeated 50 times.

The classification results are listed in Table 4. PCA performs comparably well with using the raw data directly. Surprisingly, the performance of LDA, LPP, and MFA is less effective than that directly on the raw data. This is because the performance of these methods may be affected by the numerical instability of generalized eigen-analysis on *complex* or *fairly noisy* data. One may resort to the methods in [9, 8, 25, 3] to improve the numerical stability. In contrast, the methods like MMC, ANMM, and our SRDA perform better. Particularly, the performance of classification is improved by 10% on SRDA discriminant features over using the raw data. SNMMC is a bit sensitive to the structural variation when the number of samples in each class is large, because the method exploits the distance between the point in question and its farthest point to represent the intra-class association.

The experiments are also performed on the classification of the USPS handwritten digits. The first 100 samples are selected from 1100 samples of each digit (ten digits all together) for training and the remaining for testing. The classification accuracies are 88.2% using the raw data, 88.35% on PCA features, 82.73% on LDA features, and 89.19% on MMC features, 92.05% on ANMM features, and 92.72% on SRDA features.

5. Conclusion

The classification problem is investigated via semi-Riemannian spaces in this paper. The structural relationship between classes is locally described as a low-dimensional

[5] http://www.cs.toronto.edu/˜roweis/data.html.

semi-Riemannian submanifold of index one, or equivalently a Lorentz manifold embedded in an ambient semi-Riemannian space. The dimension and structure of the discriminant sub-manifold are determined by the class and neighboring classes of a sample. The dissimilarities between the sample and its intra-class neighbors and inter-class neighbors are considered as the natural coordinate representation of a point in the ambient space. Therefore, the built semi-Riemannian space is not restricted by metrics of diverse original sample spaces. This property is similar to those of kernel-based methods. The structures of classes can be characterized and reshaped by metrics in the semi-Riemannian space. The linear and nonlinear discriminant subspaces can be obtained by virtue of the alignment of local metric tensors, which reduces to a simple eigen-decomposition like those in traditional manifold learning. Furthermore, we present a feasible determination of local metrics via the smoothing of discrete functions and the nullity of a semi-Riemannian space. Based on the proposed framework, a new method, Semi-Riemannian Discriminant Analysis (SRDA), is presented for supervised discriminant subspace learning. The effectiveness of SRDA is tested on face recognition and handwritten capital letter classification.

Our future work will be focused on developing algorithms for classification by means of the intrinsic geometry of semi-Riemannian submanifolds in semi-Riemannian spaces.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face decription with local binary patterns: application to face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006. 6

[2] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(1):711–720, 1997. 1

[3] D. Cai, X. He, Y. Hu, J. Han, and T. Huang. Learning a spatially smooth subspace for face recognition. In *CVPR*, 2007. 5, 7

[4] D. Cai, X. He, Y. Hu, J. Han, and T. Huang. Spectral regression for efficient regularized subspace learning. In *ICCV*, 2007. 5

[5] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33:1713–1726, 2000. 1

[6] K. Duggal and A. Bejancu. *Lightlike Submanifolds of Semi-Riemannian Manifolds and Applications*. Kluwer Academic Publisher, 1996. 3, 5

[7] R. Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376–386, 1938. 1

[8] J. Friedman. Regularized discriminant analysis. *American Statistical Association*, 84(405):165–175, 1989. 1, 7

[9] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995. 1, 7

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001. 3

[11] X. He, S. Yan, P. N. Y.X. Hu, and H. Zhang. Face recognition using Laplacianfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005. 1, 6

[12] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1985. 4

[13] P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(8):995–1006, 2004. 1

[14] H. Li, T. Jiang, and K. Zhang. Efficient robust feature extraction by maximum margin criterion. In *NIPS*, 2003. 1, 6

[15] Q. Liu, X. Tang, H. Lu, and S. Ma. Kernel scatter-difference based discriminant analysis for face recognition. In *ICPR*, 2004. 1

[16] Q. Liu, X. Tang, H. Lu, and S. Ma. Face recognition using kernel scatter-difference-based discriminant analysis. *IEEE Trans. on Neural Networks*, 17:1081–1085, 2006. 1

[17] A. Martinez and M. Zhu. Where are linear feature extraction methods applicable? *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(12):1934–1944, 2006. 1

[18] B. O'Neill. *Semi-Riemannian Geometry with Applications to Relativity*. Academic Press, New York, 1983. 1, 2, 3

[19] F. O'Sullivan. Discretized laplacian smoothing by Fourier methods. *JASA*, 86(415):634–642, 1991. 5

[20] P. Philips, P. Flynn, T. Scruggs, and K. Bowyer. Overview of the face recognition grand challenge. *CVPR*, 2005. 6

[21] X. Qiu and L. Wu. Face recognition by stepwise nonparameteric margin maximum criterion. In *ICCV*, 2005. 2, 6

[22] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000. 1, 6

[23] J. Tenenbaum, V. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000. 1, 3

[24] F. Wang and C. Zhang. Feature extraction by maximizing the average neighborhood margin. In *CVPR*, 2007. 2, 6

[25] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *CVPR*, 2007. 2, 7

[26] H. Wang, W. Zheng, Z. Hu, and S. Chen. Local and weighted maximum margin discriminant analysis. In *CVPR*, 2007. 2

[27] X. Wang and X. Tang. Unified subspace analysis for face recognition. *International Conf. on Computer Vision*, 2003. 3

[28] X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. *International Conf. on Computer Vision and Pattern Recognition*, pages 564–569, 2004. 1

[29] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(9):1222–1228, 2004. 1, 3

[30] X. Wang and X. Tang. Random sampling for subspace face recognition. *International Journal of Computer Vision*, 70(1):91–104, 2006. 1

[31] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007. 1, 6

[32] 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. on Pattern Analysis and Machine Intelligence*, 27(2):230–244, 2005. 1

[33] J. Ye and Q. Li. A two-stage linear discriminant analysis via QR-decomposition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(6):929–941, 2005. 1

[34] H. Yu and J. Yang. A direct LDA algorithm for high-dimensional datałwith application to face recognition. *Pattern Recognition*, 34(10):2067–2070, 2001. 1

[35] D. Zhao, Z. Lin, and X. Tang. Laplacian PCA and its applications. In *ICCV*, 2007. 5

[36] D. Zhao, Z. Lin, R. Xiao, and X. Tang. Linear Laplacian discrimination for feature extraction. In *CVPR*, 2007. 6