2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Recursive Soft Margin Subspace Learning

Q. L. Ye, and N. Ye
College of Information Science and Technology
Nanjing Forestry University
Nanjing, China
{yening@njfu.edu.cn; yqlcom@njfu.edu.cn}

C. X. Zhao
College of Computer Science and Technology
Nanjing University of Science and Technology
Nanjing, China
{zhaochx@mail.njust.edu.cn}

Abstract—**In this paper, we propose a recursive soft margin (RSM) subspace learning framework for dimension reduction of high-dimensional data, which has strong recognition ability. RSM is motivated by the soft margin criterion of support vector machines (SVMs), which allows some training samples to be misclassified for a certain cost to achieve higher recognition results. Instead of maximizing the sum of squares of Euclidean interclass (called intracluster in unsupervised learning) pairwise distances over all the similar points in previous work, RSM seeks to maximize every pairwise interclass distance between two similar points, and this distance is represented in absolute. Then, we introduce a symmetrical Hingle loss function into the RSM framework. Doing so is to allow some pairwise interclass distances to violate the maximization constraint, such that we can get satisfactory classification performance by losing some training performance. To find multiple projection vectors, a recursive procedure is designed. Our framework is illustrated with Graph Embedding (GE). For any dimension reduction method expressible by the GE, it can thus be generalized by the proposed framework to boost their recognition power by reformulating the original problems.**

Keywords—linear discriminant analysis; orthogonal linear discriminant analysis; orthogonal projection vectors; QR decomposition

I. INTRODUCTION

Dimension reduction (DR) is one of the fundamental topics in data mining, pattern recognition and computer vision, etc.. The primary goal of DR is to seek for such an optimal low-dimensional space that can help speed up the computation of any pattern classifier and gain the advantage of better analyzing the intrinsic data structures for large volumes of real-world applications. To be specific, the DR techniques are to construct a meaningful lower-dimensional representation in the reduced space of high-dimension data in the input space.

Over past decades, a family of DR techniques has been widely studied. Two most notable linear DR techniques are Linear Discriminant Analysis (FDA) [1] and Principal Component Analysis (PCA) [2]. Recently, many research efforts have shown that many forms of real data, such as faces [3] [4] and webpages [5], exhibit the essential nonlinear manifold structure. Numerous manifold learning based techniques have been proposed to discover the nonlinear manifold structure, e.g., Isometric Feature Mapping (ISOMAP) [6], Local Linear Embedding (LLE) [7], Laplacian Eigenmap (LE) [8]. Despite the exhibited promising results, these nonlinear methods cannot solve the so-called "out-of-sample" problem. Locality Preserving Projections (LPP) [4] is proposed to address this problem.

The above nonlinear techniques take into account only the local geometry of the data manifold. In subspace learning systems, the recognition performance of DR algorithms is known to be greatly improved with large margin training. Large margin DR (LMDR) techniques try to find the maximal margin projection by taking both interclass geometry (called intercluster geometry in unsupervised learning) and intraclass geometrical information (called intracluster geometry in unsupervised learning) into account, which is categorized into two classes: global and local. Global LMDA algorithms include FDA [1]. The past years can see many local LMDR algorithms, e.g., Marginal Isomap (M-Isomap) [9], Marginal Fisher analysis (MFA) [10], Local Fisher Discriminant Analysis (LFDA) [11], Locality Sensitive Discriminant Analysis (LSDA) [12], Maximum Margin Projection (MMP) [13], and Unsupervised Discriminant Projection (UDP) [14]. M-Isomap, MFA, and LFDA are supervised. M-Isomap, like ISOMAP, cannot address the "out-of-sample" problem. MFA, LFDA and LSDA are similar in sprit, which construct two graphs, i.e. interclass graph and intraclass graph, by using class information and neighborhood information. UDP, an unsupervised algorithm, characterizes locality and nonlocality to represent the intracluster and intercluster scatters by using neighborhood information. UDP can be viewed as an unsupervised version of MFA, FDA, MMP and LFDA. The success of UDP is largely based on the manifold assumption. Yan et al. [10] pointed out that each of the LMDA algorithms can be expressible by the Graph Embedding (GE) framework with a penalty graph.

Recall that the primary goals of SOMAP, LLE and LE, are focusing on best visualizing the given data. The LMDA algorithms discussed above devote to evaluating the maps on test data points without losing the primary goals of SOMAP, LLE and LE. Those studies are worthwhile in the endeavor of achieving large margin models or competitive visualized performance (or training performance) of given data; nevertheless, what makes us most interesting is that the map functions on test data can help obtain the optimal recognition results on the test data and simultaneously reflect and describe the rule of the given data in real applications. Furthermore, the given data are usually

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

disturbed by many negative factors, such as noise and not supported manifold assumption, etc., which usually leads to the unreliable visualization result and the undesired recognition performance.

Support Vector Machines (SVMs) [15], as well-known large margin classifiers, use a soft margin criterion, which allow some samples to appear on the wrong side of the hyperplane in the training phase to yield higher generalization performance. With the soft margin criterion, training patterns are allowed to be misclassified for a certain cost. Inspired by SVMs, we, in this paper, develop a novel dimension reduction framework, referred to as Recursive Soft Margin (RSM). Many existing DR algorithms can naturally be generalized by our framework to boost their generalization power. RSM maximizes every pairwise interclass distance between two patterns rather than their Euclidean squares sum. We represent this distance in absolute to simplify the reformulated optimization problem. Then, we introduce a symmetrical Hingle loss function into the RSM framework, which leads to allow some interclass pairwise distances to violate the maximization constraint. Doing so helps us achieve higher generalization by losing some visual performance on given training data. Note that for our framework in unsupervised setting, "interclass" is called "interclass", as in [14]. The constraint of the resultant optimization is nonconvex. We use the Concave-Convex Procedure (CCCP) [16] to solve our nonconvex optimization. In order to seek to find all the projections axes, a recursive algorithm is developed. In each time of iterations, only a projection axis is computed based on the updated dataset in which the "old" information represented by previously-computed projection axes has been discarded. We illustrate our framework with Graph Embedding (GE) with the penalty graph.

II. Graph Embedding

We consider the problem of representing all of the vectors in a set $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n]$, $\mathbf{x}_i \in \mathbb{R}^d$, where $\mathbf{X}$ denotes the matrix of all the training samples, where n is the sample size and d the dimensionality. For supervised learning, the class label of the sample $\mathbf{x}_i$ is from the set $\{1,2,...,c\}$, where c is the number of classes. Define by $\mathbf{z} \in \mathbb{R}^d (1 \le r \le n)$ a low-dimensional representation of a high-dimensional sample $\mathbf{x}$ in the original input space, where r is the dimensionality of the reduced space. The purpose of DR is to seek for a transformation matrix $\mathbf{W}$, such that a lower representation $\mathbf{z}$ of the sample $\mathbf{x}$ can be yielded as $\mathbf{z} = \mathbf{W}^T\mathbf{x}$, where "$T$" denotes the transpose.

A large family of dimension reduction algorithms has been designed for various motivations and application problems. GE, as a general formulation, can unify them within a common framework, which can be also used as a platform for developing new dimension reduction algorithms [10]. Let $G=\{\mathbf{X},\mathbf{V}\}$ denote a complete undirected graph with similarity matrix $\mathbf{V} \in \mathrm{R}^{n\times n}$. Each element of the matrix records the edge weight that measures the similarity between a pair of vertices. The matrix can be defined by various similarity criteria. In GE, the optimal projection can be yielded by solving the following graph-preserving criterion

$$\min_{\mathbf{w},\mathbf{w}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{w}=1} \sum_{i,j=1}^{n} \left\|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right\|^2 \mathbf{V}_{ij} = \min_{\mathbf{w},\mathbf{w}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{w}=1} \mathbf{w}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{w} \quad (1)$$

where $\mathbf{L}=\mathbf{D}-\mathbf{V}$ is the graph Laplacian of G, and $\mathbf{D}$ is a diagonal matrix whose elements on diagonal are the column sum of $\mathbf{V}$, i.e., $\mathbf{D}_{ii}=\sum_{j=1}^{n}\mathbf{V}_{ij}$. $\mathbf{M}$ from the constraint has two-fold choices. First, it is typically a diagonal matrix for scale normalization, that is, $\mathbf{M}=\mathbf{D}$. Second, it can also be selected as a graph Laplacian matrix of the penalty graph G', that is, $\mathbf{M}=\mathbf{D}'-\mathbf{V}'$, where $\mathbf{V}'$ is a similarity matrix of G' and $\mathbf{D}'$ is a diagonal matrix whose elements on diagonal are the column sum of $\mathbf{V}'$. As a result, the constraint can be formulated as $\sum_{i=1}\left\|\mathbf{w}^T\mathbf{x}_i\right\|^2\mathbf{V}_{ii}$ for scale normalization or $\sum_{i,j=1}\left\|\mathbf{w}^T\mathbf{x}_i-\mathbf{w}^T\mathbf{x}_j\right\|^2\mathbf{V}'_{ij}$ for the graph G'. (1) can be split into the following problems

$$\min_{\mathbf{w}} \sum_{i,j=1}^{n} \left\|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right\|^2 \mathbf{V}_{ij} \quad (2)$$

$$\text{s.t.} \sum_{i=1} \left\|\mathbf{w}^T\mathbf{x}_i\right\|^2 \mathbf{V}_{ii} = 1, \quad (3)$$

$$\text{or} \quad \sum_{i,j=1}^{n} \left\|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right\|^2 \mathbf{V}'_{ij} = 1. \quad (4)$$

The optimal solution of (2) can be found by solving an eigen-equation. Yan et al., [10] has shown that a large family of dimension reduction algorithms, such as PCA, LPP, FDA and UDP can be expressed by simply defining the similarity matrix $\mathbf{V}$ or $\mathbf{V}'$. We note that in FDA the similarity matrix is formed by using prior class information and each edge weight is $1/n_k$, where n_k denote the number of the samples in the kth class.

III. Recursive Soft Marge Subspace Learning

The large margin dimension reduction (LMDR) algorithms are expressible by the framework (2) with the constraint in (4). In this section, we introduce our RSM algorithm, which is based the problem (2) with the constraint in (4).

A. The Soft-margin Objective Function

As we have described, previously, the distribution of training samples diverges from the distribution of new occurring test samples. To obtain high recognition results, we maximize a soft margin between any pair of vertices of graph G'. Specifically, we define the objective function of RSM as follows:

$$\begin{aligned} &\min \ \frac{1}{2}\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} + \delta \sum_{i,j=1}^{n} \boldsymbol{\xi}_{ij} \\ &\text{s.t.} \ \left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right|\mathbf{V}'_{ij} + \boldsymbol{\xi}_{ij} \ge f, \boldsymbol{\xi}_{ij} \ge 0 \end{aligned} \quad (5)$$

where $\delta \geq 0$ is a parameter; $f > 0$ is a constant; and $|\cdot|$ denotes the absolute distance between any pair of vertices of graph G' projected onto $\mathbf{w}$. Note that $\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} = \sum_{i,j=1}^{n} \left\| \mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j \right\|^2 \mathbf{V}_{ij}$. The constraint in (2) demands the distance between any pair of points of G' to be greater than f. The variable ξ_{ij} is used to measure the amount by which the constraint $\left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right|\mathbf{V}_{ij}' \geq f$ is violated. We can arbitrarily select the nonnegative constant f, and changing it to any other positive constant ω results in $\mathbf{w}$ being replaced by $\omega\mathbf{w}$.

The problem (5) has an equivalent *loss+regularization* formulation

$$\min \ \frac{1}{2}\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} + \delta\sum_{i,j=1}^{n}(f - \left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right|\mathbf{V}_{ij}')_+ \quad (6)$$

where the subscript "+" means the positive part ($z_+ = \max(z, 0)$). The loss function $(f-t)_+$ is called symmetric Hingle loss, where $t = \left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right|\mathbf{V}_{ij}'$.

B. Optimization

For DR in multi-class settings, the absolute function in (5) will be retained, such that the constraint is nonlinear. Fortunately, it can be viewed as a difference of two convex functions $\left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right|\mathbf{V}_{ij}'$ and $1-\xi_{ij}$. Therefore, we can solve the problem with the constrained "concave-convex" procedure (CCP) [16]. Considering that the CCP is rarely used in dimension reduction, we simply introduce this algorithm.

The CCP is designed for solving the optimization problems with a concave-convex function and concave-convex constraints and aims at solving the following optimization problem [16]

$$\begin{aligned} &\min_{\mathbf{z}} \ f_0(\mathbf{z}) - g_0(\mathbf{z}), \\ &\text{s.t.} \ f_i(\mathbf{z}) - g_i(\mathbf{z}) \leq \pi_i, i = 1, 2, ..., l, \end{aligned}$$

where f_i and g_i are two real-values convex functions on a vector space $\mathcal{Z}$ for all $i = 1, 2, ..., n$ and $\pi_i \in \mathbb{R}$. Denote by $T_1\{f, \mathbf{z}\}(\mathbf{z}')$ the first order Taylor expansion of f at location $\mathbf{z}$, that is $T_1\{f,\mathbf{z}\}(\mathbf{z}') = f(\mathbf{z}) + \partial_{\mathbf{z}} f(\mathbf{z})(\mathbf{z}' - \mathbf{z})$, where $\partial_{\mathbf{z}} f(\mathbf{z})$ is the gradient of the function f at $\mathbf{z}$. For non-smooth functions, $\partial_{\mathbf{z}} f(\mathbf{z})$ can be replaced by the subgradient. Initialize $\mathbf{z}_0$ with a random value or a best guess. The CCP calculates $\mathbf{z}_{t+1}$ from $\mathbf{z}_t$ by replacing $g_i(\mathbf{z})$ with $T_1\{g_i, \mathbf{z}_t\}(\mathbf{z})$, and then sets $\mathbf{z}_{t+1}$ to the solution of the following convex optimization problem

$$\begin{aligned} &\min_{\mathbf{z}} \ f_0(\mathbf{z}) - T_1\{g_0, \mathbf{z}_t\}(\mathbf{z}), \\ &\text{s.t.} \ f_i(\mathbf{z}) - T_1\{g_i, \mathbf{z}_t\}(\mathbf{z}) \leq c_i, i = 1, 2, ..., l. \end{aligned}$$

The above recursive procedure continues until $\mathbf{z}_t$ converges. Smola *et al.* [16] has proved the fast convergence of CCP. Clearly, the minimization (5) satisfies the condition of the CCP: simply define

$$f_0(\mathbf{w}, \xi_{ij}) = \frac{1}{2}\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} + \delta\sum_{i,j=1}^{n}\xi_{ij}, \ f_i(\mathbf{w}, \xi_{ij}) = f - \xi_{ij},$$

$$g_0(\mathbf{w}, \xi_{ij}) = 0, \text{ and } g_i(\mathbf{w}, \xi_{ij}) = \left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right|\mathbf{V}_{ij}'.$$

We also set π_i equal to zero for all i. We now solve the optimization (5) with the CCP. It is important to notice that while $\left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right|\mathbf{V}_{ij}'$ is a convex function with respect to $\mathbf{w}$, it is non-smooth. Thus, the gradient can be replaced by its subgradient [16]. Initiate the $\mathbf{w}_0$, and then the CCP calculates $\mathbf{w}_{t+1}$ from $\mathbf{w}_t$ by replacing $\left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right|\mathbf{V}_{ij}'$ with the first order Taylor expansion, i.e.

$$\begin{aligned} &\left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right|\mathbf{V}_{ij}' + \text{sign}\left[\mathbf{w}_t^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)\right](\mathbf{w}^{\mathrm{T}} - \mathbf{w}_t^{\mathrm{T}})(\mathbf{x}_i - \mathbf{x}_j)\mathbf{V}_{ij}' \\ &= |\mathbf{w}_t^{\mathrm{T}}(\boldsymbol{\mu}^{(l)} - \boldsymbol{\mu})| + \text{sign}\left[\mathbf{w}_t^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)\right][\mathbf{w}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)]\mathbf{V}_{ij}' \\ &\quad - \text{sign}\left[\mathbf{w}_t^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)\right][\mathbf{w}_t^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)]\mathbf{V}_{ij}' \\ &= \text{sign}\left[\mathbf{w}_t^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)\right]\left[\mathbf{w}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)\right]\mathbf{V}_{ij}' \end{aligned} \quad (7)$$

Substituting (7) back into the problem (5), we arrive at

$$\begin{aligned} &\min \ \frac{1}{2}\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} + \delta\sum_{i,j=1}^{n}\xi_{ij} \\ &\text{s.t.} \ \mathbf{S}_{ij}\mathbf{w} + \xi_{ij} \geq f, \xi_{ij} \geq 0. \end{aligned} \quad (8)$$

in which $\mathbf{S}_{ij} = \text{sign}\left[\mathbf{w}_t^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)\right](\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{V}_{ij}'$. We rewrite the optimization problem (8) in matrix form as

$$\begin{aligned} &\min \ \frac{1}{2}\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} + \delta\mathbf{e}^T\boldsymbol{\xi} \\ &\text{s.t.} \ \mathbf{F}\mathbf{w} + \boldsymbol{\xi} \geq f\mathbf{e}, \boldsymbol{\xi} \geq \mathbf{0}. \end{aligned} \quad (9)$$

where

$$\begin{aligned} \mathbf{F} = \Big\{&\text{sign}\left[\mathbf{w}_t^{\mathrm{T}}(\mathbf{x}_1 - \mathbf{x}_1)\right](\mathbf{x}_1 - \mathbf{x}_1)\mathbf{V}_{11}', ..., \text{sign}\left[\mathbf{w}_t^{\mathrm{T}}(\mathbf{x}_1 - \mathbf{x}_n)\right](\mathbf{x}_1 - \mathbf{x}_n)\mathbf{V}_{1n}', ..., \\ &\text{sign}\left[\mathbf{w}_t^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}_1)\right](\mathbf{x}_n - \mathbf{x}_1)\mathbf{V}_{n1}', ..., \text{sign}\left[\mathbf{w}_t^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}_n)\right](\mathbf{x}_n - \mathbf{x}_n)\mathbf{V}_{nn}'\Big\} \end{aligned}$$

and $\mathbf{e}$ is a column vector of ones of appropriate dimensions. The label of the newly-constructed $\mathbf{k}_{ij}$ mentioned above is computed as $\text{sign}\left[\mathbf{w}_t^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)\right]$ at the tth iteration. The model in (9) is a constrained convex optimization problem, which can be solved using its dual problem

$$\begin{aligned} &\min \ \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{F}(\mathbf{X}\mathbf{L}\mathbf{X}^T)^{-1}\mathbf{F}^T\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \\ &\text{s.t.} \ 0 \leq \boldsymbol{\alpha} \leq \delta\mathbf{e}, \end{aligned} \quad (10)$$

in which $\boldsymbol{\alpha}$ is the Lagrange multiplier vector and $\mathbf{X}\mathbf{L}\mathbf{X}^T$ is assumed to be nonsingular. After $\boldsymbol{\alpha}$ is computed, the solution $\mathbf{w}$ is calculated as $\mathbf{w} = (\mathbf{X}\mathbf{L}\mathbf{X}^T)^{-1}\mathbf{F}^T\boldsymbol{\alpha}$, which is

obtained by setting the Lagrange multiplier function of (9) with respect to the variable $\mathbf{w}$ equal to zero. It is easy to check that the optimization problem (10) is also convex if $\mathbf{XLX}^T$ is nonsingular.

According to the CCP, the solution $\mathbf{w}$ obtained from the minimization (9) is then replaced with $\mathbf{w}_{t+1}$. RSM obtains the final solution by solving the problems as defined in (9), iteratively. The aforementioned computation aids the generation of only one projection vector. In the following, we show how to generate multiple projection vectors by a recursive procedure.

C. Produce the Multiple Projection Vectors

We develop a recursive procedure to extract more discriminant projection vectors. Instead of calculating all the projection vectors once, the projection vectors will be obtained step by step. At each step, we calculate only a new projection vector based on the previously-obtained projection vectors. Before the next projection vector is computed, the training samples are updated, such that all the information represented by the "old" projection vectors will be eliminated. The first RSM projection vector $\mathbf{w}_1$ is calculated by (5). Let $\mathbf{W}_{p-1}=\left[\mathbf{w}_1,\mathbf{w}_2,...,\mathbf{w}_{p-1}\right]\in \mathrm{R}^{d\times(p-1)}$ be the matrix representing the previously-computed normalized $p-1$ ($1<p\le r$) projection vectors. Before the pth projection vector $\mathbf{w}_p$ is solved, we discard the information represented by $\mathbf{W}_{p-1}$ from all the training samples as follows

$$\mathbf{X}_p=\mathbf{X}-\mathbf{W}_{p-1}(\mathbf{W}_{p-1}^T\mathbf{X}) \tag{11}$$

Based on the new training set, we find the pth feature vector $\mathbf{w}_p$ by optimizing the following problem

$$\begin{aligned} &\min \ \frac{1}{2}\mathbf{w}_p^T\mathbf{X}_p\mathbf{L}\mathbf{X}_p^T\mathbf{w}_P+\delta\sum_{i,j=1}^{n}\xi_{ij} \\ &\text{s.t.} \quad \left|\mathbf{w}_p^T\mathbf{x}_{p,i}-\mathbf{w}_2^T\mathbf{x}_{p,j}\right|\mathbf{V}_{ij}^{'}+\xi_{ij}\ge f, \end{aligned} \tag{12}$$

which has the similar problem as defined in (5), just replacing $\mathbf{X}$ with $\mathbf{X}_p$. The solution $\mathbf{w}_p$ can be found with the CCP, similar to finding the solution $\mathbf{w}_1$. RSM bears the similar idea as suggested by Recursive Fisher Linear Discriminant (RFDA) [18], which is to generate new sample sets by projecting the samples into a subspace that is orthogonal to previously-computed projection vectors. RFDA can be viewed a variant of FDA, which solves a similar eigenvalue problem defined in FDA at each iteration. However, Out RSM algorithm casts RFDA as a SVM-type problem at each iteration. Clearly, RFDA is a special example of our approach.

In next section we will describe the algorithm of RSM in detail and give some theoretical proofs.

IV. Algorithm and Theorem

In summary, the algorithmic procedure of RSM formally stated below:

Step 1. Construct the graphs G and G' using the original sample set $\mathbf{X}$.

Step 2. Use RSM to extract the first projection vector $\mathbf{w}_1$ based on the set $\mathbf{X}$.

Step 3. Generate the new set $\mathbf{X}^{(p)}$ in which the information represented by all the previously extracted projection vectors is discarded by $\mathbf{X}_p=\mathbf{X}-\mathbf{W}_{p-1}(\mathbf{W}_{p-1}^T\mathbf{X})$, where $p\ge 2$.

Step 4. Use RSM to find another projection vector $\mathbf{w}_p$. This solution is similar to step 2.

Step 5. Go to step 3 if needed to extract more projection vectors.

After accomplishing the training of RSM, we obtain a projection matrix $\mathbf{W}\in\mathbf{R}^{d\times r}$. Depending on applications, some postprocessing, such as the nearest rule for classification, is applied to the projected samples to complete classification tasks.

It is observed that the optimization problem (5) is solved under the assumption that the matrix $\mathbf{XLX}^T$ is nonsingular. However, this assumption does not hold in many real applications, such as face recognition problems where sometimes the number of samples in the training set tends to be much smaller than that number of projection vectors in each sample, such that the matrix $\mathbf{XLX}^T$ is singular. A popular method to overcome this problem is to first project the sample set to the PCA space. Hereafter, we suppose that the matrix $\mathbf{XLX}^T$ is singular.

Given three sets

$$S_1=\left\{(\mathbf{x}_i,\mathbf{x}_j)\,|\,\text{there is larger similarity between } \mathbf{x}_i \text{ and } \mathbf{x}_j, 1\le i,j\le n\right\}$$

$$S_2=\left\{(\mathbf{x}_i,\mathbf{x}_j)\,|\,\text{there is smaller similarity between } \mathbf{x}_i \text{ and } \mathbf{x}_j, 1\le i,j\le n\right\}$$

and $S=\left\{(\mathbf{x}_i,\mathbf{x}_j)\,|\,i,j=1,...,n\right\}$, it is easy to conclude that $S_1\cap S_2=\varnothing$ and $S_1\cup S_2\subseteq S$, according to [10].

Definition 1: Given a nonzero edge weight $\mathbf{K}_{ij}$ put between $\mathbf{x}_i$ and $\mathbf{x}_j$, the weights $\mathbf{V}$ and $\mathbf{V}'$ can be defined as [4, 8, 14]

$$\mathbf{V}_{ij}=\begin{cases}\mathbf{K}_{ij}, & \text{if } (\mathbf{x}_i,\mathbf{x}_j)\in S_1\\ 0, & \text{otherwise}\end{cases},$$

$$\mathbf{V}_{ij}^{'}=\begin{cases}\mathbf{K}_{ij}, & \text{if } (\mathbf{x}_i,\mathbf{x}_j)\in S_2\\ 0, & \text{otherwise}\end{cases}$$

Typically, $\mathbf{K}_{ij}$ is set as the Gaussian kernel, namely, $\mathbf{K}_{ij}=\exp(-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2/\sigma)$, where σ is a variance.

Proposition 1: Denote by

$$\mathbf{S}_T=\sum_{i,j=1}^{n}\mathbf{K}_{ij}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^T$$

the generation matrix of the global scatter matrix in PCA. Based on above definitions on $\mathbf{V}$ and $\mathbf{V}'$, it is easy to check that

$$\begin{aligned}\mathbf{S}_T &= \sum_{i,j=1}^{n} \mathbf{K}_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \\ &= \mathbf{S}_{\mathrm{L}} + \mathbf{S}_{\mathrm{N}} + \mathbf{S}_{\mathrm{R}}\end{aligned} \tag{14}$$

where $\mathbf{S}_{\mathrm{N}} = \sum_{i,j=1}^{n} \mathbf{V}'_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$, $\mathbf{S}_{\mathrm{R}} = \sum_{i,j=1}^{n} \mathbf{A}_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$, $\mathbf{S}_{\mathrm{L}} = \sum_{i,j=1}^{n} \mathbf{V}_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T = \mathbf{X}\mathbf{L}\mathbf{X}^T$, and $\mathbf{A}_{ij} = \mathbf{K}_{ij} - \mathbf{V}_{ij} - \mathbf{V}'_{ij}$.

Let $\mathbb{B} = \mathrm{span}\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_q\}$ be the subspace and denote by $\mathbb{B}^{\perp} = \mathrm{span}\{\boldsymbol{\beta}_{q+1}, ..., \boldsymbol{\beta}_d\}$ its orthogonal complement, where $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_q$ are the first q eigenvectors of $\mathbf{S}_T$ corresponding to positive eigenvalues. Obviously, $\mathbb{B}^{\perp}$ is the null space of $\mathbf{S}_T$. Based on the equation (5), we have the following theorem.

Proposition 2: Let $\mathbf{w} = \mathbf{u} + \boldsymbol{\theta}$ be a decomposition of $\mathbf{w}$ ($\mathbf{w} \in \mathbb{R}^n$) into a part $\mathbf{u} \in \mathbb{B}$ and a part $\boldsymbol{\theta} \in \mathbb{B}^{\perp}$, then the constrained optimization problem (5) is equivalent to

$$\begin{aligned}&\min_{\mathbf{u}} \ \frac{1}{2}\mathbf{u}^T \mathbf{X}\mathbf{L}\mathbf{X}^T \mathbf{u} + \delta \sum_{i,j=1}^{n} \xi_{ij} \\ &\text{s.t.} \quad \left|\mathbf{u}^T \mathbf{x}_i - \mathbf{u}^T \mathbf{x}_j\right| \mathbf{V}'_{ij} + \xi_{ij} \geq f, \xi_{ij} \geq 0\end{aligned} \tag{15}$$

The proof is provided in **Appendix A**.

Proposition 2 discloses the fact that the solution of (5) can be produced in the subspace $\mathbb{B}$ without any loss of the information. Let $\mathbf{P}$ denote a transformation matrix of q dimensions, each column vector of which is corresponding to a non-zero eigenvalue of $\mathbf{S}_T$. By linear algebra theory, $\mathbb{B}$ is isomorphic to the q-dimensional Euclidean space $\mathbb{R}^q$ [14]. The isomorphic mapping is exactly the transformation matrix $\mathbf{P}$, one has $\mathbf{u} = \mathbf{P}\boldsymbol{\eta}$, $\mathbf{u} \in \mathbb{B}$, $\boldsymbol{\eta} \in \mathbb{R}^q$, where $\mathbf{P} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_q)$. By the above mapping, the formulation (15) can be re-formulated as the following problem

$$\begin{aligned}&\min_{\boldsymbol{\eta}} \ \frac{1}{2}\boldsymbol{\eta}^T \tilde{\mathbf{X}}\mathbf{L}\tilde{\mathbf{X}}^T \boldsymbol{\eta} + \delta \sum_{i,j=1}^{n} \xi_{ij} \\ &\text{s.t.} \ \left|\boldsymbol{\eta}^T \tilde{\mathbf{x}}_i - \boldsymbol{\eta}^T \tilde{\mathbf{x}}_j\right| \mathbf{V}'_{ij} + \xi_{ij} \geq f, \xi_{ij} \geq 0\end{aligned} \tag{16}$$

where $\tilde{\mathbf{X}} = \mathbf{P}^T \mathbf{X}$ is the PCA transformation of data matrix $\mathbf{X}$. Therefore, $\boldsymbol{\eta}$ can be generated in the PCA subspace. If $\boldsymbol{\eta}^*$ is the solution to (16), then, $\mathbf{u}^* = \mathbf{P}\boldsymbol{\eta}^*$ is the first RSM optimal feature. With the recursive procedure in our algorithm, the solution $\boldsymbol{\eta}_p^*$ to (16) on the set $\mathbf{X}^{(2)}$ can be calculated. Then, the r optimal projection vectors of RSM are $\mathbf{u}_p^* = \mathbf{P}_p \boldsymbol{\eta}_p^*$, $p = 1, 2, ..., r$.

V. Algorithm and Theorem

We generalize FDA and UDP by RSM to evaluate the performance of the proposed subspace framework. The resulting schemes are termed as RSM/FDA and RSVM/UDP, respectively. In this way, PCA, FDA, RFDA, LPP, UDP, and the proposed RSM/FDA and RSM/UDP are used to extract features. Note that PCA, UDP and RSS/UDP are unsupervised and the others are supervised. In all the experiments, two face databases YALE [19] and ORL [20] are used.

A. Face recognition on YALE database

The YALE database [18] was constructed by the Yale Center for Computational Vision and Control. It contains 165 grayscale images of 15 individuals. All the images demonstrate variation in lighting condition (center light, left light, and right light), facial expression (normal, happy, sad, sleepy, surprised, and winking). Original images were manually normalized, aligned, cropped and scaled to 32×32 pixels.

In our experiment, the first 2, 3 and 4 images per individual are used for training, respectively, and the remaining images for testing. Before implementing all the algorithms, PCA is first applied to throw away the components corresponding to zeroes eigenvalues. After dimension reduction, the Nearest Neighbor classifier is used for classification. The maximum recognition rate of each method and the corresponding dimension is listed in TableI. Fig. 1 plots the recognition rate versus the variation of dimensions.

TABLE I. The maximum recognition rates (%) on YALE database. The numbers in parentheses are the optimal dimensions after dimension reduction.

Training Size	PCA	FDA	RFDA	RSM/FDA	LPP	UDP	RSM/UDP
2 Train	52.6(25)	60.7(13)	64.4(15)	66.7(25)	48.2(28)	48.2(29)	57.1(26)
3 Train	59.2(24)	65.8(14)	70.8(28)	72.5 (27)	49.2(43)	55.0(37)	60.0(38)
4 Train	63.8(47)	71.4(14)	74.3(35)	79.1 (36)	52.4(59)	61.9(55)	66.7(43)

As can be seen, our supervised algorithm RSM/FDA outperforms all other six algorithms. The supervised algorithms FDA, RFDA and RSM/FDA outperform the unsupervised PCA, LPP, UDP and RSM/UDP. UDP performs better than LPP, which is consistent with [17]. Among all the unsupervised algorithms PCA, LPP, UDP and RSM/UDP, RSM/UDP performs the best.

B. Face recognition on YALE database

The ORL face database [19] contains 400 images of 40 individuals and has become a standard database for testing. The images were captured at different times with the different variations, including expression like open or closed eyes and smiling or non-smiling, and facial details like

glasses or no glasses. The images were aligned, cropped and scaled to 32×32 pixels. Each image is represented by a 32×32 (i.e., 1024) dimensional vector in image space. 8 images per individual are randomly selected for training, and the remaining 3 images are used for testing.

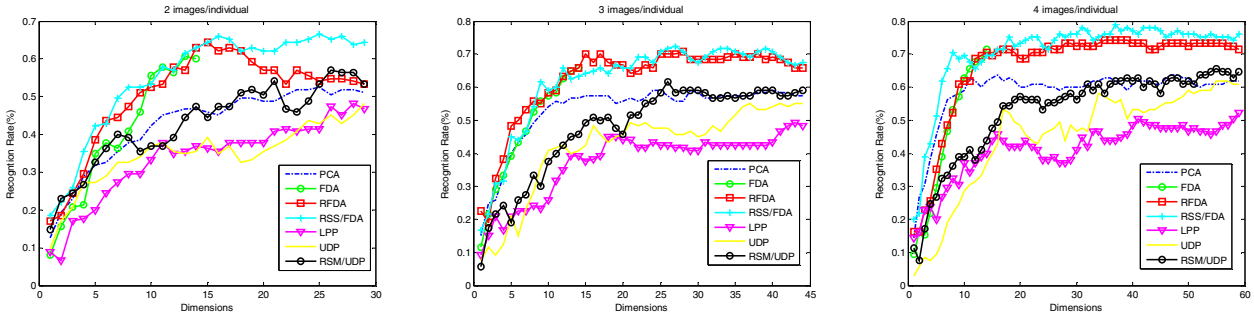

Fig1. Recognition rate versus the variation of dimensions.

TABLE II. THE MAXIMUM RECOGNITION RATES (%) ON ORL DATABASE. THE NUMBERS IN PARENTHESES ARE THE OPTIMAL DIMENSIONS ATER DIMENSION REDUCTION.

Training Size	PCA	FDA	RFDA	RSM/FDA	LPP	UDP	RSM/UDP
2 Train	71.6(44)	82.5 (39)	82.2(47)	83.5 (41)	69.7(77)	76.9(73)	77.2(65)
3 Train	76.8(63)	87.9 (36)	88.9(39)	90.0(40)	71.1(116)	79.3(74)	82.5(93)
4 Train	85.8(103)	91.3(39)	95.0(41)	95.8(39)	80.8 (156)	87.1(157)	91.3(125)

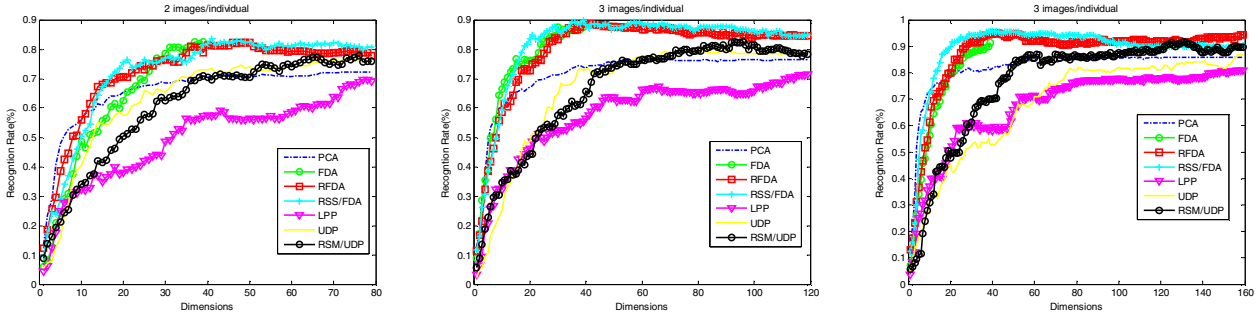

Fig.2. Recognition rate versus the variation of dimensions.

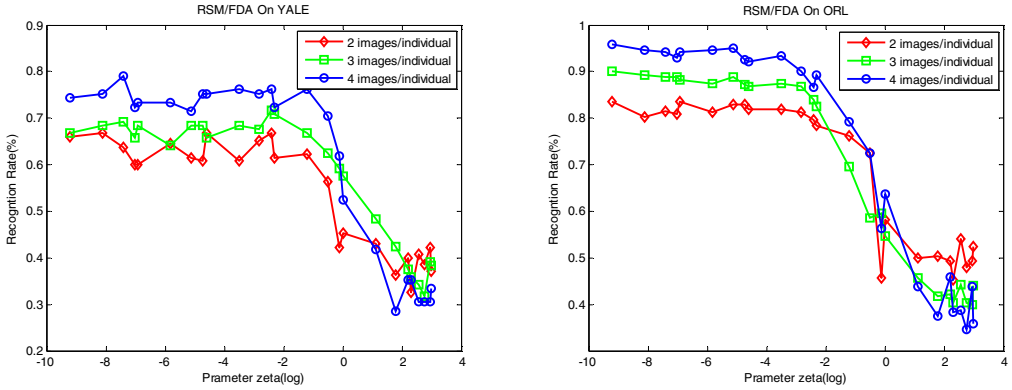

Fig.3. Recognition rate of RSM/FDA with Respect to different values of the parameter δ .

The experimental design is the same as the previous experiment. The best result obtained in the optimal subspace and the corresponding dimension of each method is listed TableII. As can be seen, our supervised RSM/FDA performs the best for all the cases. UDP obtains better result than other two unsupervised methods PCA and LPP, but is worse than its generalized version RSM/UDP. These results indicate the effectiveness of the proposed RSM.

C. *Parameter Selection*

In this section, we evaluate the performance of our algorithm with the different values of the parameter δ . In above experiments, the parameter is determined using the global-to-local strategy [14] to make the recognition result optimal. We select RSM/FDA as an instance.

Fig.3 shows the recognition rate of our algorithm RSM/FDA on YALE and ORL with respect to different values of this parameter. As can be seen, our algorithm is not sensitive to δ when it is selected from 0.0001 to 0.1.

VI. Conclusion

We have introduced a novel subspace learning framework, called Recursive Soft Margin Subspace Learning (RSM). Different from traditional large margin subspace learning algorithms involving the eigen-decomposition, RSM casts them as related SVM-type problems, respectively. Doing so makes these traditional algorithms possess the soft-margin concept, and thus further help improve the discriminant powers of the traditional large margin subspace learning algorithms by allowing some interclass pairwise distances to violate the maximization constraint. To achieve more projection vectors, a recursive procedure is designed. Theoretically, we reveal some nature of RSM. The experiments on two face databases YALE and ORL indicates the effectiveness of RSM. From the RSM model, it can be seen that it is easy to directly develop the sparse RSVM model, which is our future work.

Acknowledgment

The authors are extremely thankful to Scientific Research Foundation for Advanced Talents and Returned Overseas Scholars of Nanjing Forestry University, China National Funds for Distinguished Young Scientists (31125008), Jiangsu Qing Lan Project, Jiangsu talent peaks of six fields Project, Jiangsu Science Foundation (BK2012399), and National Science Foundations of China (61101197 and 61272220) for support.

References

[1] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. Pattern Analysis and Machine Intelligence, 19(7): 711-720, 1997.

[2] I. Joliffe. Principal Component Analysis. Springer-Verlag, 1986.

[3] Y. Chang, C. Hu, and M. Turk. Manifold of Facial Expression. Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures, Oct. 2003.

[4] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face Recognition Using Laplacianfaces. IEEE Trans. Pattern Analysis and Machine Intelligence, 27(3): 328-340, 2005.

[5] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. Proc. Neural Inf. Process. Syst., 321–328, 2004.

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

[7] S.T. Roweis, and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, 290:2323-2326, 2000.

[8] M. Belkin, and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation, 15(6):1373-1396, 2003.

[9] Z. Zhang, W. S. C. Tommy, and M. B. Zhao. M-Isomap: Orthogonal Constrained Marginal Isomap for Nonlinear Dimensionality Reduction. IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics (TSMC-B), 2012, to appear.

[10] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph Embedding and Extensions: A General Framework for Dimensionality eduction. IEEE Trans. Pattern Analysis and Machine Intelligence, 29(1): 40-51, 2007.

[11] M. Sugiyama, Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis. Journal of Machine Learning Research 8:1027-1061, 2007.

[12] D. Cai, X. He, K. Zhou, J. W. Han and H. J. Bao, Locality Sensitive Discriminant Analysis, Proc. 2007 Int. Joint Conf. on Artificial Intelligence (IJCAI'07), Hyderabad, India, Jan. 2007.

[13] X. He, D. Cai, and J. W. Han. Learning a Maximum Margin Subspace for Image Retrieval. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 20(2):189-201,2008.

[14] J. Yang, D. Zhang, J.Y. Yang, and B. Niu, "Globally Maximizing, Locally Minimizing: Unsupervised Discriminant Projection with Applications to Face and Palm Biometrics", IEEE Trans. Pattern Analysis and Machine Intelligence, 29(4): 650-664, 2007.

[15] C. Cortes and V. Vapnik. Support Vector Networks. Machine Learning, 20: 273-297, 1995.

[16] A. J. Smola, S. Vishwanathan, and T. Hofmann. Kernel methods for missing variables, In AISTATS, 2005.

[17] Ronan Collobert, Fabian Sinz, Jason Weston, Large Scale Transductive SVMs, Journal of Machine Learning Research 7 (2006) 1687–1712.

[18] C. Xiang, X. Fan, T. Lee, Face recognition using recursive Fisher linear discriminant, IEEE Trans Image Process. 15(8): 2097-105, 2006.

[19] Yale. Face Database. (Online), available from: /http://cvc.yale.edu/projects/yalefaces/yalefaces.html..

[20] ORL. Face Database. (online), available from: /http://www.uk.research.att.com/facedatabase.html.

[21] Framework for Semi-Supervised and Unsupervised Dimension Reduction," *IEEE Trans. Image Process.*, no. 11, 2010, pp.1921-1932.

[22] F. P. Nie, D. Xu, X.L. Li, and S.M Xiang, "Semisupervised Dimensionality Reduction and Classification Through Virtual Label Regression", *IEEE Trans. on Systems, Man, and Cybernetics—Part B: Cybernetics*, vol. 41, no. 3, 2011, pp. 675-685.

[23] D. Cai, "Spectral regression: A regression framework for efficient regularized subspace learning", Ph.D. dissertation, Dept. Comput. Sci., Univ. Illinois Urbana-Champaign, Urbana, May, 2009.

[24] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from examples", *J. Mach. Learn. Res.*, vol. 7, Nov. 2006, pp. 2399-2434.

[25] X. Zhu,"Semi-supervised learning literature survey," University Wisconsin Madison, 2007.

[26] D. Y. Zhou, J.Weston, and A. Gretton et al., "Ranking on Data Manifolds", in *Advances in Neural Information Processing Systems*, 2004.

[27] C. Liu, H. Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *Int. J. Comput. Vision.,* vol. 75, no. 1, 2000, pp. 115–134.

[28] E. Kokiopoulou and Y. Saad, "Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 29, no. 12, 2007, pp. 2143–2156,.

[29] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.,* vol. 84, no. 405, 1989, pp. 165–175.

[30] T. P. Zhang, B. Fang, Y.Y. Tang, Z, W. Shang, and B. Xu, "Generalized Discriminant Analysis: A Matrix Exponential Approach", *IEEE Trans. on Systems, Man, and Cybernetics—Part B: Cybernetics*, vol. 40, no. 1, 2010, 186-197.

Appendix A: The proof of proposition 2

Every $\boldsymbol{\theta}$ can be decomposed as a linear combination of the orthogonal eigenvectors of $\mathbf{S}_T$ that correspond to zero eigenvalues. Since $\boldsymbol{\theta} \in \mathbb{B}^{\perp}$, we have $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_T\boldsymbol{\theta} = 0$ and $\mathbf{S}_T\boldsymbol{\theta} = 0$. Sin $\mathbf{S}_T = \mathbf{S}_{\mathrm{L}} + \mathbf{S}_{\mathrm{N}} + \mathbf{S}_{\mathrm{R}}$, one can get $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_T\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_{\mathrm{L}}\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_{\mathrm{N}}\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_{\mathrm{R}}\boldsymbol{\theta} = 0$. We further get $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_L\boldsymbol{\theta} = 0$, $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\theta} = 0$ and $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_{\mathrm{R}}\boldsymbol{\theta} = 0$, since $\mathbf{S}_{\mathrm{L}}$, $\mathbf{S}_{\mathrm{N}}$ and $\mathbf{S}_{\mathrm{R}}$ are positive semi-definite, which implies $\mathbf{S}_L\boldsymbol{\theta} = 0$, $\mathbf{S}_N\boldsymbol{\theta} = 0$ and $\mathbf{S}_{\mathrm{R}}\boldsymbol{\theta} = 0$. Therefore,

$$\mathbf{w}^T\mathbf{S}_L\mathbf{w} = (\mathbf{u}+\boldsymbol{\theta})^{\mathrm{T}}\mathbf{S}_L(\mathbf{u}+\boldsymbol{\theta}) = \mathbf{u}^{\mathrm{T}}\mathbf{S}_L\mathbf{u}$$

$$\mathbf{w}^T\mathbf{S}_N\mathbf{w} = (\mathbf{u}+\boldsymbol{\theta})^T\mathbf{S}_N(\mathbf{u}+\boldsymbol{\theta}) = \mathbf{u}^T\mathbf{S}_N\mathbf{u}$$

Since $\left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right|\mathbf{V}'_{ij} = \text{sqrt}(\mathbf{w}^T\mathbf{S}_N\mathbf{w})$, we further conclude $\left|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right|\mathbf{V}'_{ij} = \left|\mathbf{u}^T\mathbf{x}_i - \mathbf{u}^T\mathbf{x}_j\right|\mathbf{V}'_{ij}$. Therefore, the projection axis $\mathbf{w}$ needing to be estimated can be replaced with $\mathbf{u}$. Thus, optimization problem (5) is equivalent to that in (15).■□