Rank-one Projections with Adaptive Margins for Face Recognition

Dong Xu[1,4*], Stephen Lin[2], Shuicheng Yan[3,5], Xiaoou Tang[2,3]

[1] MOE-Microsoft Key Laboratory of Multimedia Computing and Communication & Department of EEIS
University of Science and Technology of China, Hefei, Anhui, P. R. China

[2] Microsoft Research Asia, Beijing, P. R. China

[3] Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong

[4] Department of Electrical Engineering, Columbia University, US

[5] Beckman Institute, University of Illinois at Urbana-Champaign, US

Contact: dongxu@mail.ustc.edu.cn

Abstract

In supervised dimensionality reduction, tensor representations of images have recently been employed to enhance classification of high-dimensional data with small training sets. To handle tensor data, this approach has been formulated with tight restrictions on projection directions that, along with convergence issues and the assumption of Gaussian distributed class data, limits its face recognition performance. To overcome these problems, we propose a method of rank-one projections with adaptive margins (RPAM) that gives a provably convergent solution for tensor data over a more general class of projections, while accounting for margins between samples of different classes. In contrast to previous margin based works which determine margin sample pairs within the original high dimensional space, RPAM instead aims to maximize the margins defined in the expected lower dimensional feature subspace by progressive margin refinement after each rank-one projection. In addition to handling tensor data, vector-based variants of RPAM are presented for linear mappings and for nonlinear mappings using kernel tricks. Comprehensive experimental results demonstrate that RPAM brings significant improvement in face recognition over previous subspace learning techniques.

1. Introduction

Computer vision and pattern recognition has witnessed growing interest in dimensionality reduction techniques for classification. Among them, supervised methods such as Linear Discriminant Analysis (LDA) [1] and its variants [5] [11] have been particularly popular owing to their simplicity in computation and effectiveness in classification. In LDA, projections of high dimensional image data to a lower dimensional feature space are computed in a manner that seeks to maximize inter-class scatter while minimizing the scatter within each class. Despite the success of LDA in many applications, it often suffers from the *small sample size* problem when dealing with high dimensional face data [11]. This problem is exacerbated in LDA by rasterization of 2D image data into 1D vectors prior to processing, which may conceal higher order structure in images, e.g., concatenation of rows can effectively obscure correlations along columns.

Previous supervised techniques [14] [17], as well as some unsupervised methods [8] [16] [12] [13], address this problem by processing data as higher order tensors. Representation of data as tensors not only preserves image structure, but can significantly reduce the number of projection parameters to be learned [14] [17]. With fewer parameters to determine from small training sets, we say that tensor-based techniques offer greater *learnability* in dimensionality reduction. In gaining this learnability, DATER [14] and 2DLDA [17] compute projection matrices that must be in the form of Kronecker products of matrices, but this restriction of the solution space consequently limits the *potential discriminability* of the learned projection matrix, particularly in cases where greater training data is available.

In this paper, we present a supervised method called rank-one projections with adaptive margins (RPAM) that provides greater potential discriminability with the high learnability of tensor-based techniques. For higher discriminability, the projection matrix is computed as a series of rank-one projection vectors that are in the form of Kronecker products of vectors, rather than Kronecker products of matrices, such that a broader range of solutions becomes possible with the use of tensor data. By using rank-one pro-

*This work was performed when Dong Xu was a visiting student at Microsoft Research Asia

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

jections, a further benefit is that the solution is provably convergent, in contrast to the previous supervised tensor-based techniques [14] [17].

An additional advantage of employing a rank-one strategy in supervised learning is that it provides a platform for better handling of class margins. It is commonly believed that data samples that lie along class margins play an important role in pattern classification. For example, Support Vector Machines (SVMs) [10] utilize margin samples, referred to as support vectors, for constructing hyperplanes that partition a space into different classes. In dimensionality reduction, Non-parametric Linear Discriminant Analysis (NPLDA) [2] and a variant called Marginal Fisher Analysis (MFA) [15] were proposed to break the assumption of Gaussian distributed data in traditional LDA by placing larger weights on pairs of margin points between different classes in the computation of inter-class scatter. In NPLDA and MFA, the margins are defined by the closest pairs of points between different classes in the original feature space, as exemplified by the red lines in Fig. 1(a). But for dimensionality reduction as in NPLDA or MFA, although the defined pairs of margin samples may be well separated in the eventual dimensionality reduced space, the classes themselves may not be adequately partitioned because pairs of non-margin points in the original feature space may not be well separated after projection. This will lead to significant degradation in the classification ability of supervised subspace learning algorithms.

Ideally, the margin pairs of the optimal dimensionality reduced space, as shown in Fig. 1(b), should be used to guide the computation of projections. Since this information is generally indeterminable from examination of the original feature space, we take advantage of the iterative rank-one procedure in RPAM for adaptive refinement of margins. Specifically, our method initially utilizes the margins computed in the original feature space, and then iteratively adapts the margins to those that exist after each rank-one projection, which should provide a better estimate of the margins in the dimensionality-reduced feature space. With the incorporation of this rank-one adaptive margin technique, significant improvements can be gained in face recognition performance.

2. Motivations

Before formally describing RPAM, we discuss in greater detail the two motivations of this work in the context of supervised dimensionality reduction with tensor data. To facilitate this discussion, we first review some fundamental definitions on tensors [3] [4] and describe a property of tensor vectorization.

2.1. Tensor definitions

Definition-1:(*tensor inner product, norm and distance*). The inner product of two tensors $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2 \times \ldots \times m_n}$ and $\mathbf{Y} \in \mathbb{R}^{m_1 \times m_2 \times \ldots \times m_n}$ is defined as $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i_1=1,\ldots,i_n=1}^{m_1,\ldots,m_n} \mathbf{X}_{i_1,\ldots,i_n} \mathbf{Y}_{i_1,\ldots,i_n}$. The norm of tensor $\mathbf{X}$ is therefore defined to be $\|\mathbf{X}\| = \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}$, and the tensor distance between tensors $\mathbf{X}$ and $\mathbf{Y}$ is computed as $D(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|$.

Definition-2:(*k-mode product).* The k-mode product of tensor $\mathbf{X}$ with matrix $U \in \mathbb{R}^{m_k \times m'_k}$, *i.e.*, $\mathbf{Y} = \mathbf{X} \times_k U$, is defined as $\mathbf{Y}_{i_1,\ldots,i_{k-1},i,i_{k+1},\ldots,i_n} = \sum_{j=1}^{m_k} \mathbf{X}_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_n} \times U_{j,i}, i = 1, ..., m'_k$.

Definition-3:(*k-mode unfolding).* The k-mode unfolding of an n-th order tensor $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2 \times \ldots \times m_n}$ into a matrix $X^k \in \mathbb{R}^{m_k \times \prod_{i \neq k} m_i}$, *i.e.*, $X^k \Longleftarrow_k \mathbf{X}$, is defined as $X^k_{i_k,j} = \mathbf{X}_{i_1,\ldots,i_n}$, $j = 1 + \sum_{l=1,l\neq k}^{n} (i_l - 1) \prod_{o=l+1,o\neq k}^{n} m_o$.

Lemma-1: Take arbitrary tensors $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2 \times \ldots \times m_n}$ $\mathbf{Y} \in \mathbb{R}^{m'_1 \times m'_2 \times \ldots \times m'_n}$ and projection matrices $U^k \in \mathbb{R}^{m_k \times m'_k}$, $k = 1, \ldots, n$. Suppose $\mathbf{X}$ and $\mathbf{Y}$ are unfolded into matrices and then vectorized, where x and y are the unfolded vectors of $\mathbf{X}$ and $\mathbf{Y}$ respectively. We then have

$$\mathbf{Y} = \mathbf{X} \times_1 U^1 \times \ldots \times_n U^n \Longleftrightarrow y = P^T x \quad \text{with}$$
$$P = U^n \otimes U^{n-1} \ldots \otimes U^1 \in \mathbb{R}^{\prod_{k=1}^{n} m_k \times \prod_{k=1}^{n} m'_k}, \quad (1)$$

where $\otimes$ is the Kronecker product for which $A \otimes B = [A_{ij}B]$.

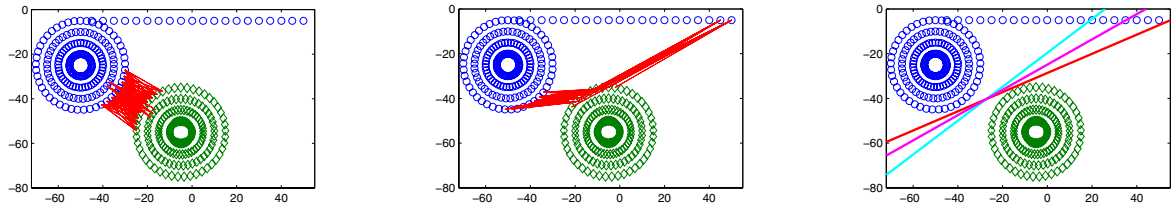
2.2. Learnability and potential discriminability

The basic objective of supervised dimensionality reduction is to learn a projection matrix that transforms the original high dimensional data to a lower dimension in which accurate classification can be achieved. For a projection matrix $P \in \mathbb{R}^{m \times m'}$ where $m = \prod_{k=1}^{n} m_k$ and $m' = \prod_{k=1}^{n} m'_k$, the distance between two data samples x, $y \in \mathbb{R}^m$ after dimensionality reduction becomes $\|P^T x - P^T y\|^2 = (x - y)^T P P^T (x - y)$. We define the similarity measure matrix S as

$$S = PP^T, \quad (2)$$

such that $\|P^T x - P^T y\|^2 = (x - y)^T S (x - y)$.

In solving for S, two factors influence its classification performance. One is its learnability, which is a problem in real applications such as face recognition that typically have small training sets and a large feature set. In such cases, finding optimal parameter values is more difficult for projection matrices with a larger number of parameters. The other factor is potential discriminability, which describes the ability to obtain the optimal measure matrix. When



(a) (b) (c)

Figure 1. Utility of adaptive margins. (a) margin pairs, indicated by red lines, of the original feature space; (b) margin pairs of the optimal dimensionality reduced space; (c) projection directions of LDA (magenta), MFA (cyan) and RPAM/L (red).

the examinable solution space of projection matrices is restricted, the potential discriminability becomes limited.

In vector-based algorithms [1], the discriminability is at its full potential since the entire possible solution space is examined. However, as mentioned previously, the small sample size problem degrades the learnability of these algorithms significantly.

While tensor-based algorithms [17] [14] have greater learnability because of a smaller number of parameters to estimate, their projection matrices are constrained to be a Kronecker product of smaller sized matrices as in Eq.(1), i.e., $P = U^n \otimes U^{n-1} \ldots \otimes U^1, \quad U^k \in \mathbb{R}^{m_k \times m'_k}, \; k = 1, \ldots, n$. Consequently, the similarity measure matrix is constrained to be of the form

$$S = S^n \otimes S^{n-1} \ldots \otimes S^1, \;\; S^k = U^k U^{k^T} \; \forall \, k, \quad (3)$$

which greatly limits the potential discriminability of the derived projection matrix. This restriction is particularly limiting for large training sets, since the additional data may not be fully exploited.

To increase potential discriminability while maintaining the high learnability associated with tensor representations, we utilize a series of rank-one projections, which has been used previously in unsupervised dimensionality reduction [8]. With this approach, each column of the projection matrix $P = [p_1, p_2, ..., p_d]$ has the form of a Kronecker product of unitary vectors:

$$p_j = u_j^n \otimes u_j^{n-1} \ldots \otimes u_j^1, \; u_j^k \in \mathbb{R}^{m_k}, \|u_j^k\| = 1 \; \forall \, j, k. \quad (4)$$

The resulting similarity measure matrix can then be expressed as

$$S = \sum_{j=1}^{d} u_j^n u_j^{n^T} \otimes u_j^{n-1} u_j^{n-1^T} \ldots \otimes u_j^1 u_j^{1^T}. \quad (5)$$

This more general form of projection matrix allows for greater potential discriminability with tensor data.

2.3. Effective margin analysis

For non-parametric separability of classes, a maximal distance between margin samples of different classes is targetted. Consider a sample set $X = [x_1, x_2, \ldots, x_N]$, $x_i \in \mathbb{R}^m$, where N is the total number of samples, and the class label of x_i is $l(x_i)$. Marginal Fisher Analysis (MFA) [15] solves for the projection matrix with the following optimization problem:

$$P^* = arg \max_P \frac{\sum_c \sum_{(i,j) \in N_{k_2}^-(c)} \|P^T x_i - P^T x_j\|^2}{\sum_i \sum_{j \in N_{k_1(i)}^+} \|P^T x_i - P^T x_j\|^2} \quad (6)$$

where $N_{k_1}^+(i)$ indicates the k_1 nearest neighbors of sample x_i within the same class, and $N_{k_2}^-(c)$ denotes a set of margin pairs that is computed as follows: for each class c, distances between its samples and samples in other classes are computed in the original feature space, then for the k_2 smallest distances, the corresponding pairs of points $\{(i,j), l(x_i) = c, l(x_j) \neq c\}$ are chosen.

These margin pairs are used in measuring separability among different classes, but as previously noted, margin pairs defined in the original feature space may not adequately represent the margins in the dimensionality reduced space. The margin pairs should ideally be computed in the optimal dimensionality reduced space according to distance $||P^T x_i - P^T x_j||$, where P is the optimal projection matrix. Our method attempts to approximate the set of optimal pairs $N_{k_2}^-(c, P)$ throughout the course of the algorithm.

3. Rank-one Projections with Adaptive Margins

With the goals of high learnability, large potential discriminability, and effective margin analysis, we present the following criterion for training samples represented as n-th order tensors $\mathbf{X_i} \in \mathbb{R}^{m_1 \times m_2 \times \ldots \times m_n}$ with corresponding unfolded vectors $x_i \in \mathbb{R}^{\Pi_{k=1}^n m_k}$, $i = 1, ..., N$:

$$P^* = \arg \max_P \frac{\sum_c \sum_{(i,j) \in N_{k_2}^-(c,P)} \|P^T x_i - P^T x_j\|^2}{\sum_i \sum_{j \in N_{k_1}^+(i)} \|P^T x_i - P^T x_j\|^2}$$

$$\text{with} \;\; P = [p_1, p_2, \ldots, p_d] \;\; \text{and} \;\; p_j = u_j^n \otimes u_j^{n-1} \otimes \ldots u_j^1. \quad (7)$$

$N_{k_1}^+(i)$ is defined as in Eq. (6); $N_{k_2}^-(c, P)$ represents the k_2 inter-class pairs of shortest distances in the expected lower

dimensional feature space projected by P; and each projection direction p_j is constrained to be the Kronecker product of n unitary vectors $u_j^n, u_j^{n-1}, \ldots, u_j^1$.

It can be seen that this criterion poses a chicken-and-egg problem, since P needs to be known to determine $N_{k_2}^-(c, P)$, and vice versa. No closed form solution exists, and in previous methods the margin set $N_{k_2}^-(c, P)$ is assumed to be that of the original high-dimensional feature space. To address this problem, we present a method to compute the solution in an iterative manner.

In solving for both the projection matrix P and the margin set $N_{k_2}^-(c, P)$, we adopt a greedy approach. Given the first $(l\text{-}1)$ rank-one projections, the l-th projection vector p_l is computed using the matrix $P^{l-1} = [p_1, p_2, ..., p_{l-1}]$ as an approximation of P in calculating the margin pairs:

$$p_l^* = \arg\max_{p_l} \frac{\sum_c \sum_{(i,j)\in N_{k_2}^-(c,P^{l-1})} \|{P^l}^T(x_i - x_j)\|^2}{\sum_i \sum_{j\in N_{k_1}^+(i)} \|{P^l}^T(x_i - x_j)\|^2}$$
$$s.t. \quad p_l = u_l^n \otimes u_l^{n-1} \otimes \ldots u_l^1. \tag{8}$$

To avoid redundancy among the rank-one projections, a projection vector p_l is computed in the complement space of the previous projections, yielding an objective function

$$\frac{\sum_c \sum_{(i,j)\in N_{k_2}^-(c,P^{l-1})} \|{p_l}^T x_i^l - {p_l}^T x_j^l\|^2 + a}{\sum_i \sum_{j\in N_{k_1}^+(i)} \|{p_l}^T x_i^l - {p_l}^T x_j^l\|^2 + b}, \tag{9}$$

where

$$x_i^l = x_i^{l-1} - p_{l-1} p_{l-1}^T x_i^{l-1}, \quad \text{with } x_i^1 = x_i, \tag{10}$$

$$a = \sum_c \sum_{(i,j)\in N_{k_2}^-(c,P^{l-1})} \sum_{o=1}^{l-1} \|p_o^T x_i - p_o^T x_j\|^2, \tag{11}$$

$$b = \sum_i \sum_{j\in N_{k_1}^+(i)} \sum_{o=1}^{l-1} \|p_o^T x_i - p_o^T x_j\|^2. \tag{12}$$

From Lemma-1, Eq. (9) can be rewritten in tensor form:

$$\frac{\sum_c \sum_{(i,j)\in N_{k_2}^-(c,P^{l-1})} \|(\mathbf{X}_i^l - \mathbf{X}_j^l) \times_k u_l^k|_{k=1}^n\|^2 + a}{\sum_i \sum_{j\in N_{k_1}^+(i)} \|(\mathbf{X}_i^l - \mathbf{X}_j^l) \times_k u_l^k|_{k=1}^n\|^2 + b}, \tag{13}$$

where $\mathbf{X}_i^l$ is the corresponding tensor representation of x_i^l and $\times_k u_l^k|_{k=1}^n$ is equivalent to $\times_1 u_l^1 \times_2 u_l^2 ... \times_n u_l^n$.

To our knowledge, there exists no closed form solution for Eq. (13), so we propose an iterative algorithm for determining a local minimum. If $(u_l^1, \ldots, u_l^{k-1}, u_l^{k+1}, \ldots, u_l^n)$ are given and we define $y_i^k = \mathbf{X}_i^l \times_1 u_l^1 \ldots \times_{k-1} u_l^{k-1} \times_{k+1} u_l^{k+1} \ldots \times_n u_l^n$, then the above objective function can be simplified to

$$u_l^{k*} = \arg\max_{u_l^k} \frac{\sum_c \sum_{(i,j)\in N_{k_2}^-(c,P^{l-1})} \|{u_l^k}^T(y_i^k - y_j^k)\|^2 + a}{\sum_i \sum_{j\in N_{k_1}^+(i)} \|{u_l^k}^T(y_i^k - y_j^k)\|^2 + b}$$
$$= \arg\max_{u_l^k} \frac{{u_l^k}^T S_b^k u_l^k}{{u_l^k}^T S_w^k u_l^k}, \tag{14}$$

where $S_b^k = \sum_c \sum_{(i,j)\in N_{k_2}^-(c,P^{l-1})} (y_i^k - y_j^k)(y_i^k - y_j^k)^T + aI$ and $S_w^k = \sum_i \sum_{j\in N_{k_1}^+(i)} (y_i^k - y_j^k)(y_i^k - y_j^k)^T + bI$.

This objective function can be solved by the generalized eigenvalue decomposition [15] method. Therefore, we can obtain a local optimum of Eq. (9) by iteratively optimizing one projection vector while fixing the other projection vectors.

The detailed procedure of rank-one projections with adaptive margins (RPAM) is given in Fig. 2.

Convergence Analysis: Unlike the iterative algorithms of 2DLDA and DATER, the iterative algorithm of RPAM can be proven to converge to a local optimum as follows:

Given the sample set $\mathbf{X_i} \in \mathbb{R}^{m_1 \times m_2 \times \ldots \times m_n}$, $i = 1, \ldots, N$, their class labels $c_i \in \{1, 2, \ldots, N_c\}$, the expected subspace dimension d, and the maximum iteration number T_{max}:

1. Initialize $P^0 = I$, $x_i^1 = x_i \; \forall \; i$;
2. For $l = 1, \ldots, d$ Do
   a) Compute $N_{k_2}^-(c, P^{l-1})$ from P^{l-1};
   b) For $t = 1, 2, \ldots, T_{max}$ Do
   - If t=1, initialize $u_l^k(0) \; \forall \; k$ as arbitrary column orthogonal matrices;
   - For $k = 1, 2, \ldots, n$, Do
     $y_i^k = \mathbf{X}_i^l \times_o u_l^o(t)|_{o=1}^{k-1} \times_o u_l^o(t\text{-}1)|_{o=k+1}^n$;
     $S_b^k u_l^k(t) = \lambda_0^k S_w^k u_l^k(t)$;
     End

   End
   c) $p_l = u_l^n(t) \otimes u_l^{n-1}(t) \otimes \ldots u_l^1(t)$;
   d) $x_i^{l+1} = x_i^l - p_l p_l^T x_i^l \; \forall \; i$;
   End
3. Output projection matrix $P = [p_1, ..., p_d]$.

Figure 2. Procedure for rank-one projections with adaptive margins (RPAM).

Proof. Each step of 2DLDA [17] and DATER [14] involves an optimization problem $\arg\max_U \frac{Tr(U^{k^T} S_b^k U^k)}{Tr(U^{k^T} S_w^k U^k)}$, which is similar to Eq. (14) but where U^k is a projection *matrix*. Since this function is difficult to optimize, these two algorithms alter the objective function to a more tractable form $arg\max_U Tr((U^{k^T} S_w^k U^k)^{-1}(U^{k^T} S_b^k U^k))$ for which generalized eigenvalue decomposition can be applied. This alteration of the objective function, however, results in a convergence problem that is demonstrated in Section 5.

In contrast, optimization of Eq. (14) involves the solution of only a projection *vector*. Since $u^{k^T} S_w^k u^k$ and $u^{k^T} S_b^k u^k$ are scalar values and do not involve the *Trace* operation as needed for matrices, the objective function $arg\max_u (u^{k^T} S_w^k u^k)^{-1}(u^{k^T} S_b^k u^k)$ can be optimized directly with generalized eigenvalue decomposition, without changing the objective function.

We define $f(u_l^1(t), u_l^2(t), \ldots, u_l^n(t))$ as

$$\frac{\sum_i \sum_{j\in N_{k_1}^+(i)} \|(\mathbf{X}_i^l - \mathbf{X}_j^l) \times_k u_l^k(t)|_{k=1}^n\|^2 + b}{\sum_c \sum_{(i,j)\in N_{k_2}^-(c,P^{l-1})} \|(\mathbf{X}_i^l - \mathbf{X}_j^l) \times_k u_l^k(t)|_{k=1}^n\|^2 + a}.$$

In each step of RPAM, f does not increase:

$$f(u_l^1(t), u_l^2(t), \ldots, u_l^n(t)) \geq f(u_l^1(t+1), u_l^2(t), \ldots, u_l^n(t)),$$

$$\ldots, \geq f(u_l^1(t+1), u_l^2(t+1), \ldots, u_l^n(t+1)). \quad (15)$$

Furthermore, $f(u_l^1(t), u_l^2(t), \ldots, u_l^n(t))$ has a lower bound of zero. So the iterative algorithm for RPAM converges to a local optimum in computing each projection vector. □

4. Vector-based variants

Most previous algorithms for dimensionality reduction address linear mappings of vector data. Since a vector is simply a first-order tensor, RPAM can also process vectors. For clarity, we denote the special case of RPAM for linear mappings of vector data as RPAM/L, and the tensor-based version as RPAM/T.

The kernel trick [7] has been widely applied to extend linear dimensionality reduction algorithms into nonlinear ones. The intuition of the kernel trick is to map data from the original feature space to a higher dimensional Hilbert space $\phi : x \rightarrow F$ in which the data may be linearly separable. In this new feature space, linear dimensionality reduction algorithms can then be applied.

We denote the data set after transformation as $\phi(X) = [\phi(x_1), \phi(x_2), \ldots, \phi(x_N)]$ and the Gram matrix as $K = \phi(X)^T \phi(X)$ with elements computed as inner products $K_{ij} = s(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$. In kernel analysis, the non-linear projection directions P are of the form $P = \phi(X)A$ where $A = [\alpha_1, \ldots, \alpha_d]$ represents a matrix

---

Given the sample set $x_i \in \mathbb{R}^m, i = 1, \ldots, N$, their class labels $c_i \in \{1, 2, \ldots, N_c\}$, the final lower dimension d and the iteration number T_{max}:

1. Initialize $P^0 = I$, $K_i^1 = K_i \; \forall \; i$;
2. For $l = 1, \ldots, d$ Do
   a) For $t = 1, 2, \ldots, T_{max}$ Do
   - If $t > 1$, compute $N_{k_2}^-(c, P^{l-1})$ from $P^l(t\text{-}1)$, else compute $N_{k_2}^-(c, P^{l-1})$ from P^{l-1};
   - Compute the optimal $\alpha_l(t)$ from Eq.(18);

   End
   b) Output $\alpha_l = \alpha_l(t)$ and set $P^l = P^l(t)$;
   c) $K_i^{l+1} = (I - K\alpha_l\alpha_l^T)K_i^l \; \forall \; i$;
   End
3. Output $A = [\alpha_1, \alpha_2, \ldots, \alpha_d]$

Figure 3. Procedure for vector-based Kernel RPAM.

of linear combinations. From this, the objective function of Eq.(7) can be reformulated for vectors as

$$A^* = \arg\max_A \frac{\sum_c \sum_{(i,j)\in N_{k_2}^-(c,P)} \|A^T K_i - A^T K_j\|^2}{\sum_i \sum_{j\in N_{k_1}^+(i)} \|A^T K_i - A^T K_j\|^2}$$
$$s.t. \quad \alpha_j^T K \alpha_j = 1 \; \forall \; j, \quad (16)$$

where K_i is the i-th column vector of K.

From the first (l-1) projection directions $P^{l-1} = [\phi(X)\alpha_1, \ldots, \phi(X)\alpha_{l-1}]$, the l-th projection p_l is constrained to lie in the complement space of the space spanned by the previous projection vectors. The term $p_l^T x_i^l$ in Eq. (9) is changed to $p_l^T \phi^l(x_i)$, which can be computed as follows:

$$\begin{aligned} & p_l^T \phi^l(x_i) \\ = \; & p_l^T (I - p_{l-1} p_{l-1}^T)\phi^{l-1}(x_i) \\ = \; & (\phi(X)\alpha_l)^T (I - \phi(X)\alpha_{l-1}\alpha_{l-1}^T \phi(X)^T)\phi^{l-1}(x_i) \\ = \; & \alpha_l^T (K_i^{l-1} - K\alpha_{l-1}\alpha_{l-1}^T K_i^{l-1}) \\ = \; & \alpha_l^T (I - K\alpha_{l-1}\alpha_{l-1}^T) K_i^{l-1}. \end{aligned} \quad (17)$$

This is equivalent to resetting $K_i^l = (I - K\alpha_{l-1}\alpha_{l-1}^T)K_i^{l-1}$ with $K_i^1 = K_i$. Similar to Eq. (9), α_l can be learned by maximizing the objective function

$$\max_{\alpha_l} \frac{\sum_c \sum_{(i,j)\in N_{k_2}^-(c,P^{l-1})} \|\alpha_l^T K_i^l - \alpha_l^T K_j^l\|^2 + a}{\sum_i \sum_{j\in N_{k_1}^+(i)} \|\alpha_l^T K_i^l - \alpha_l^T K_j^l\|^2 + b}, \quad (18)$$

where

$$a = \sum_{c} \sum_{(i,j)\in N_{k_2}^{-}(c,P^{l-1})} \sum_{o=1}^{l-1} \|\alpha_o^T K_i - \alpha_o^T K_j\|^2,$$

$$b = \sum_{i} \sum_{j\in N_{k_1}^{+}(i)} \sum_{o=1}^{l-1} \|\alpha_o^T K_i - \alpha_o^T K_j\|^2.$$

As with Eq. (14), this objective function can be solved by the generalized eigenvalue decomposition method. The entire procedure, referred to as RPAM/K, is listed in Fig. 3.

In Fig. 3, we note a minor difference in RPAM/L and RPAM/K from RPAM/T in the computation of margin samples. In RPAM/L and RPAM/K, when $t > 1$, we use the most recently computed projection matrix $P^l(t-1) = [p^1, ..., p^{l-1}, p^l(t\text{-}1)]$ to replace P^{l-1} in approximating the optimal margin samples. Note for RPAM/K, $p^l(t-1) = \phi(X)\alpha_l(t-1)$. For RPAM/T, only P^{l-1} is used to compute $N_{k_2}^{-}(c, P^{l-1})$, since changing the margin in iterations of t will effectively change the objective function, and the iterative algorithm would not be provably convergent. Another difference of RPAM/K is that for the first projection direction α_1, $N_{k_1}^{+}(i)$ and $N_{k_2}^{-}(c, P^0)$ are determined from data points in the high-dimensional Hilbert space, where the distance between samples x_i and x_j is computed as $D(x_i, x_j) = s(x_i, x_i) + s(x_j, x_j) - 2s(x_i, x_j)$.

5. Results

In our results, we first demonstrate the effects of margin adaption. Then, two benchmark face databases, XM2VTS [6] and CMU PIE [9], are used to evaluate the effectiveness of RPAM in comparison to LDA, MFA and their variants.

For the face recognition experiments, preprocessing of images includes alignment by fixing the locations of the two eyes, size normalization to 64x64 resolution, and histogram equalization. In all the experiments, the gallery and probe data are transformed into 1D vectors, 2D matrices of size 64x64, and 3D tensors of size 16x16x24. In the 3D tensor representation, we utilize downsampled images and include a dimension that consists of 24 Gabor features at six orientations and four scales. The dimensionality reduced vectors, matrices and tensors are acquired via the learned subspaces, and the nearest neighbor criterion is used for final classification.

5.1. Artificial Data

To clearly illustrate the effects of adaptive margins, we examine the artificial two-class problem of Fig. 1. In Fig. 1(c), the solid lines represent the projection directions computed by LDA, MFA, and RPAM/L. Since the samples in Class 1 do not form a Gaussian distribution, LDA fails to find the optimal projection direction. Although MFA considers margin samples, the results are also not correct because the margin pairs in the original feature space, indicated by red lines in Fig. 1(a), do not adequately represent the margins of the optimal dimensionality reduced space, as illustrated in Fig. 1(b). By iteratively adapting the margin samples ($T_{max} = 8$ in this experiment), RPAM/L identifies the optimal solution.

Algorithm	Recognition Rate (%)
LDA/L [1]	90.9
MFA/L [15]	91.5
RPAM/L	**97.3**
LDA/2D [17]	88.8
RPAM/2D	**97.6**
LDA/3D [14]	89.2
RPAM/3D	**98.3**

Table 1. Recognition rates on the XM2VTS database.

5.2. Face recognition on XM2VTS database

The XM2VTS database [6] contains 295 people, and for each person there are four frontal face images taken during four separate sessions. In our experiments, we select the 295x3 images of the first three sessions as training data, the 295 images of the first session as the gallery set, and the 295 images from the fourth session as the probe set. Recognition rates are listed in Table 1. The results demonstrate that RPAM and its variants outperform LDA, MFA and their corresponding variants.

An important difference of RPAM/T from 2DLDA [17] and DATER [14] is the convergence characteristics, which are shown for a second-order matrix representation in Fig. 4. The horizontal axis indicates the number of iterations, and the vertical axis is the similarity of two successively estimated projection matrices or vectors, i.e., $Tr[Abs(U^{k^T}(t)U^k(t-1))]/m'_k$ for 2DLDA and $Abs(u_{10}^{k^T}(t)u_{10}^k(t-1))$, $k = 1, 2$, for RPAM/2D. The rank-one approach of RPAM/2D does not exhibit the convergence problems that are evident with 2DLDA.

5.3. Face recognition on CMU PIE database

The CMU PIE (Pose, Illumination, and Expression) database [9] contains more than 40,000 facial images of 68 people. The images were acquired over different poses, under variable illumination conditions and with different facial expressions. In this experiment, five near-frontal poses and four illumination conditions are used, such that each person has twenty images. We randomly choose four images per person for training and use the remaining sixteen images for testing. The results are given in Table 2. Again the

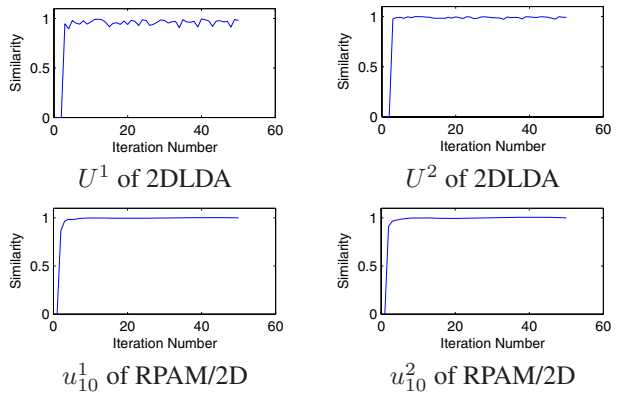


Figure 4. Convergence characteristics of 2DLDA and RPAM/2D. The vertical axis indicates the similarity of two successive projection matrices or vectors, and the horizontal axis gives the number of iterations.

Algorithm	Recognition Rate (%)
LDA/L [1]	76.2
MFA/L [15]	77.0
RPAM/L	**83.0**
LDA/K [7]	76.5
MFA/K [15]	78.2
RPAM/K	**82.8**
LDA/2D [17]	81.7
RPAM/2D	**88.0**
LDA/3D [14]	78.2
RPAM/3D	**95.5**

Table 2. Recognition rates on the CMU PIE database.

results demonstrate that RPAM and its variants outperform LDA, MFA and their corresponding variants.

6. Conclusion

In this paper, we have proposed a supervised subspace learning algorithm called Rank-one Projections with Adaptive Margins that overcomes previous shortcomings in tensor-based classification. The effectiveness of this approach is evidenced in experimental comparisons on benchmark databases with other dimensionality reduction methods. The presented greedy approach to margin adaptation provides only an approximate solution to the margin problem, and further investigation of this issue presents an interesting direction for future work.

References

[1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 711–720, 1997.

[2] K. Fukunaga and J. Mantock. Nonparametric discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 671–678, 1983.

[3] T. Kolda. Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, pages 243–255, 2001.

[4] L. Lathauwer, B. Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, pages 1253–1278, 2000.

[5] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, pages 467–476, 2002.

[6] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. *Proceedings of International Conference on Audio- and Video- Based Person Authentication*, 1999.

[7] K. Mtiller, S. Mika, G. Riitsch, K. Tsuda, and B. Scholkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, pages 181–201, 2001.

[8] A. Shashua and A. Levin. Linear image coding for regression and classification using the tensor-rank principle. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.

[9] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1615–1618, 2003.

[10] V. Vapnik. The nature of statistical learning theory. *Springer, New York*, 1995.

[11] X. Wang and X. Tang. Random sampling lda for face recognition. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.

[12] D. Xu, S. Yan, L. Zhang, Z. Liu, and H. Zhang. Coupled subspace analysis. *Microsoft Technique Report, MSR-TR-2004-106*, 2004.

[13] D. Xu, S. Yan, L. Zhang, H. Zhang, Z. Liu, and H. Shum. Concurrent subspace analysis. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[14] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang. Discriminant analysis with tensor representation. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[15] S. Yan, D. Xu, B. Zhang, and H. Zhang. Graph embedding: A general framework for dimensionality reduction. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[16] J. Ye. Generalized low rank approximations of matrices. *Proceedings of International Conference on Machine Learning*, 2004.

[17] J. Ye, R. Janardan, and Q. Li. Two-dimensional linear discriminant analysis. *Advances in Neural Information Processing Systems*, 2004.