Uncorrelated Multilinear Principal Component Analysis through Successive Variance Maximization

Haiping Lu HPLU@IEEE.ORG
Konstantinos N. Plataniotis KOSTAS@COMM.TORONTO.EDU
Department of Electrical and Computer Engineering, University of Toronto, ON, M5S 3G4, Canada

Anastasios N. Venetsanopoulos TASVENET@RYERSON.CA
Ryerson University, Toronto, ON, M5B 2K3, Canada

Abstract

Tensorial data are frequently encountered in various machine learning tasks today and dimensionality reduction is one of their most important applications. This paper extends the classical principal component analysis (PCA) to its multilinear version by proposing a novel unsupervised dimensionality reduction algorithm for tensorial data, named as uncorrelated multilinear PCA (UMPCA). UMPCA seeks a tensor-to-vector projection that captures most of the variation in the original tensorial input while producing uncorrelated features through successive variance maximization. We evaluate the UMPCA on a second-order tensorial problem, face recognition, and the experimental results show its superiority, especially in low-dimensional spaces, through the comparison with three other PCA-based algorithms.

1. Introduction

Various machine learning problems take multi-dimensional data as input, which are formally called tensors. The elements of a tensor are to be addressed by several indices and the number of indices used in the description defines the order of the tensor object, with each index defining one "mode" (Lathauwer et al., 2000). Many real-world data are naturally tensor objects. For example, matrix data such as gray-level images are second-order tensors, gray-scale video sequences and 3-D objects are third-order tensors. In addition, streaming data and mining data are frequently organized as third-order tensors. For instance, data in environmental sensor monitoring are often organized in three modes of time, location and type, and data in web graph mining are commonly organized in three modes of source, destination and text. Other applications involving tensorial data include data center monitoring, social network analysis, network forensics and face recognition (Faloutsos et al., 2007). In these practical applications, tensor objects are often specified in a high-dimensional tensor space, leading to the so-called curse of dimensionality. Nonetheless, the class of tensor objects in most applications are highly constrained to a subspace, a manifold of intrinsically low dimension (Shakhnarovich & Moghaddam, 2004), and feature extraction or dimensionality reduction is frequently employed to transform a high-dimensional data set into a low-dimensional space of equivalent representation while retaining most of the underlying structure (Law & Jain, 2006).

The PCA is a classical linear method for unsupervised dimensionality reduction that transforms a data set consisting of a large number of interrelated variables to a new set of uncorrelated variables, while retaining as much as possible the variations present in the original data set (Jolliffe, 2002). PCA on tensor objects requires their reshaping (vectorization) into vectors in a very high-dimensional space, which not only results in high computational and memory demands but also breaks the natural structure and correlation in the original data (Ye, 2005; Ye et al., 2004; Lu et al., 2008a). It is believed by many researchers that potentially more compact or useful representations can be obtained from the original form and PCA extensions operating directly on the tensor objects rather than their vectorized versions are emerging recently (Ye et al., 2004; Lu et al., 2008a; Xu et al., 2005).

In (Shashua & Levin, 2001), the tensor rank-one de-

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

composition (TROD) is used to represent a class of images based on variance maximization and (greedy) successive residue calculation. A two-dimensional PCA (2DPCA) is proposed in (Yang et al., 2004) that constructs an image covariance matrix using image matrices as inputs. However, linear transformation is applied only to the right side of image matrices so the image data is projected in one mode only, resulting in poor dimensionality reduction. A more general algorithm named generalized low rank approximation of matrices (GLRAM) was introduced in (Ye, 2005), which applies two linear transforms to both the left and right sides of input image matrices and results in a better dimensionality reduction than 2DPCA. GLRAM is developed from the perspective of approximation while the generalized PCA (GPCA) is proposed in (Ye et al., 2004) from the view of variation maximization, as an extension of PCA. Later, the concurrent subspaces analysis (CSA) is formulated in (Xu et al., 2005) for optimal reconstruction of general tensor objects, which can be considered as a generalization of GLRAM, and the multilinear PCA (MPCA) introduced in (Lu et al., 2008a) targets at variation maximization for general tensor objects in the extension of PCA to the multilinear case, which can be considered as a further generalization of GPCA.

However, none of the existing multilinear extensions of PCA mentioned above takes an important property of PCA into account, i.e., PCA derives uncorrelated features, which contain minimum redundancy and ensure independence among features. Instead, most of them produce orthogonal bases in each mode. Although uncorrelated features imply orthogonal projection bases in PCA, this is not necessarily the case for its multilinear extension. With this motivation, this paper investigates multilinear extension of PCA that can produce uncorrelated features. We propose a novel uncorrelated multilinear PCA (UMPCA) for unsupervised tensor object dimensionality reduction (feature extraction). UMPCA is based on the tensor-to-vector projection (TVP) (Lu et al., 2008b) and it follows the classical PCA derivation of successive variance maximization (Jolliffe, 2002). Thus, a number of elementary multilinear projections (EMPs) are solved to maximize the captured variance with the zero-correlation constraint. The solution is iterative in nature, as many other multilinear algorithms (Xu et al., 2005; Ye et al., 2004; Shashua & Levin, 2001).

The rest of this paper is organized as follows. Section 2 reviews basic multilinear notations and operations, as well as the concept of tensor-to-vector projection. In Sec. 3, the problem of UMPCA is formulated and the solution is derived as a sequential iterative process. Next, Sec. 4 evaluates the effectiveness of UMPCA in the popular face recognition task through comparison with PCA, MPCA and TROD. Finally, the conclusions are drawn in Sec. 5.

Table 1. Notations

Notations	Descriptions
$\mathcal{X}_m$, $m = 1, ..., M$	the m^{th} input tensor sample
$\mathbf{u}^{(n)}$, $n = 1, ..., N$	the n-mode projection vector
$\{{\mathbf{u}_p^{(n)}}^T, n = 1, ..., N\}$	the p^{th} EMP, where p is the index of the EMP
$\mathbf{y}_m$	the projection of $\mathcal{X}_m$ on the TVP $\{{\mathbf{u}_p^{(n)}}^T, n = 1, ..., N\}_{p=1}^P$
$\mathbf{y}_m(p) = y_{m_p} = \mathbf{g}_p(m)$	the projection of $\mathcal{X}_m$ on the p^{th} EMP $\{{\mathbf{u}_p^{(n)}}^T, n = 1, ..., N\}$
$\mathbf{g}_p$	the p^{th} coordinate vector

2. Multilinear Fundamentals

This section introduces the multilinear notations, operations and projections needed in the presentation of UMPCA, and for further pursuing of multilinear algebra, (Lathauwer et al., 2000) is a good reference. The important notations used in this paper are listed in Table 1 for handy reference.

2.1. Notations and basic multilinear operations

Due to the multilinear nature of tensor objects, new notations have been introduced in the literature for mathematical analysis. Following the notations in (Lathauwer et al., 2000), we denote vectors by lowercase boldface letters, e.g., $\mathbf{x}$; matrices by uppercase boldface letters, e.g., $\mathbf{U}$; and tensors by calligraphic letters, e.g., $\mathcal{A}$. Their elements are denoted with indices in parentheses. Indices are denoted by lowercase letters and span the range from 1 to the uppercase letter of the index, e.g., $n = 1, 2, ..., N$.

An N^{th}-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times ... \times I_N}$ is addressed by N indices i_n, $n = 1, ..., N$, and each i_n addresses the n-mode of $\mathcal{A}$. The n-mode product of a tensor $\mathcal{A}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{U}$, is a tensor with entries:

$$\begin{aligned}&(\mathcal{A} \times_n \mathbf{U})(i_1, ..., i_{n-1}, j_n, i_{n+1}, ..., i_N)\\ &\quad = \sum_{i_n} \mathcal{A}(i_1, i_2, ..., i_N) \cdot \mathbf{U}(j_n, i_n). \qquad (1)\end{aligned}$$

The scalar product of two tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times ... \times I_N}$ is defined as:

$$< \mathcal{A}, \mathcal{B} >= \sum_{i_1} ... \sum_{i_N} \mathcal{A}(i_1, ..., i_N) \cdot \mathcal{B}(i_1, ..., i_N). \quad (2)$$

A rank-one tensor $\mathcal{A}$ equals to the outer product of N

vectors: $\mathcal{A} = \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ ... \circ \mathbf{u}^{(N)}$, which means that $\mathcal{A}(i_1, i_2, ..., i_N) = \mathbf{u}^{(1)}(i_1) \cdot \mathbf{u}^{(2)}(i_2) \cdot ... \cdot \mathbf{u}^{(N)}(i_N)$ for all values of indices.

2.2. Tensor-to-vector projection

In order to extract uncorrelated features from tensorial data directly, we employ the TVP introduced in (Lu et al., 2008b), which is a more general form of the projection in (Shashua & Levin, 2001) and consists of multiple EMPs. An EMP is a multilinear projection $\{\mathbf{u}^{(1)^T}, \mathbf{u}^{(2)^T}, ..., \mathbf{u}^{(N)^T}\}$ consisting of one unit projection vector in each mode, i.e., $\| \mathbf{u}^{(n)} \| = 1$ for $n = 1, ..., N$, where $\| \cdot \|$ is the Euclidean norm for vectors. It projects a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times ... \times I_N}$ to a scalar y through the N unit projection vectors as

$$y = \mathcal{X} \times_1 \mathbf{u}^{(1)^T} \times_2 \mathbf{u}^{(2)^T} ... \times_N \mathbf{u}^{(N)^T} =< \mathcal{X}, \mathcal{U} >,$$

where $\mathcal{U} = \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ ... \circ \mathbf{u}^{(N)}$. An EMP can be viewed as a constrained linear projection since $< \mathcal{X}, \mathcal{U} >=< vec(\mathcal{X}), vec(\mathcal{U}) >= [vec(\mathcal{U})]^T vec(\mathcal{X})$, where $vec(\cdot)$ denotes the vectorized representation.

The TVP of a tensor object $\mathcal{X}$ to a vector $\mathbf{y} \in \mathbb{R}^P$ consists of P EMPs $\{\mathbf{u}_p^{(1)^T}, \mathbf{u}_p^{(2)^T}, ..., \mathbf{u}_p^{(N)^T}\}, p = 1, ..., P$, which can be written concisely as $\{\mathbf{u}_p^{(n)^T}, n = 1, ..., N\}_{p=1}^P$:

$$\mathbf{y} = \mathcal{X} \times_{n=1}^N \{\mathbf{u}_p^{(n)^T}, n = 1, ..., N\}_{p=1}^P, \tag{3}$$

where the p^{th} component of $\mathbf{y}$ is obtained from the p^{th} EMP as: $\mathbf{y}(p) = \mathcal{X} \times_1 \mathbf{u}_p^{(1)^T} \times_2 \mathbf{u}_p^{(2)^T} ... \times_N \mathbf{u}_p^{(N)^T}$. The TROD (Shashua & Levin, 2001) in fact seeks a TVP to maximize the captured variance, however, it takes a heuristic greedy approach. In the next section, we propose a systematic, more principled formulation by taking consideration of the correlation among features.

In addition, the TVP for dimensionality reduction here is related mathematically to the parallel factor analysis (PARAFAC) originated from psychometrics (Harshman, 1970), also known as the canonical decomposition (CANDECOMP) (Carroll & Chang, 1970), which is popular in factor analysis of multi-way data, i.e., tensors. However, they are developed from different perspectives. The PARAFAC in the factorization literature aims to decompose a higher-order tensor, often formed by arranging lower-order tensors, into a number of rank-one tensorial factors explaining the formation of the data. In contrast, the objective of the TVP for dimensionality reduction here is to learn a low-dimensional (subspace) representation of a class of tensor objects from a number of samples so that the underlying (class) structure is well captured.

3. Uncorrelated Multilinear PCA

This section proposes the UMPCA for unsupervised dimensionality reduction of tensor objects by first formulating the UMPCA objective function and then adopting the successive variance maximization approach and alternating projection method to solve the problem. In the presentation, for the convenience of discussion, the training samples are assumed to be zero-mean [1] so that the constraint of uncorrelated features is the same as orthogonal features (Koren & Carmel, 2004).

3.1. Problem formulation

Following the standard derivation of PCA given in (Jolliffe, 2002), we consider the variance of the principal components (PCs) one by one. In the TVP setting, the p^{th} PCs are $\{y_{m_p}, m = 1, ..., M\}$, where M is the number of training samples and y_{m_p} is the projection of the m^{th} sample $\mathcal{X}_m$ by the p^{th} EMP $\{\mathbf{u}_p^{(n)^T}, n = 1, ..., N\}$: $y_{m_p} = \mathcal{X}_m \times_{n=1}^N \{\mathbf{u}_p^{(n)^T}, n = 1, ..., N\}$. Accordingly, the variance is measure by their total scatter $S_{T_p}^{\mathbf{y}}$, which is defined as

$$S_{T_p}^{\mathbf{y}} = \sum_{m=1}^{M} (y_{m_p} - \bar{y}_p)^2, \tag{4}$$

where $\bar{y}_p = \frac{1}{M} \sum_m y_{m_p}$. In addition, let $\mathbf{g}_p$ denote the p^{th} coordinate vector, with its m^{th} component $\mathbf{g}_p(m) = y_{m_p}$. A formal definition of the unsupervised multilinear feature extraction problem to be solved in UMPCA is then given in the following:

A set of M tensor object samples $\{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_M\}$ are available for training. Each tensor object $\mathcal{X}_m \in \mathbb{R}^{I_1 \times I_2 \times ... \times I_N}$ assumes values in the tensor space $\mathbb{R}^{I_1} \bigotimes \mathbb{R}^{I_2} ... \bigotimes \mathbb{R}^{I_N}$, where I_n is the n-mode dimension of the tensor and $\bigotimes$ denotes the Kronecker product. The objective of the UMPCA is to find a TVP, which consists of P EMPs $\{\mathbf{u}_p^{(n)} \in \mathbb{R}^{I_n \times 1}, n = 1, ..., N\}_{p=1}^P$, mapping from the original tensor space $\mathbb{R}^{I_1} \bigotimes \mathbb{R}^{I_2} ... \bigotimes \mathbb{R}^{I_N}$ into a vector subspace $\mathbb{R}^P$ (with $P < \prod_{n=1}^N I_n$):

$$\mathbf{y}_m = \mathcal{X}_m \times_{n=1}^N \{\mathbf{u}_p^{(n)^T}, n = 1, ..., N\}_{p=1}^P, m = 1, ..., M, \tag{5}$$

such that the variance of the projected samples, measured by $S_{T_p}^{\mathbf{y}}$, is maximized in each EMP direction, subject to the constraint that the P coordinate vectors $\{\mathbf{g}_p \in \mathbb{R}^M, p = 1, ..., P\}$ are uncorrelated.

[1] When the training sample mean is not zero, it can be subtracted to make the training samples to be zero-mean.

In other words, the UMPCA objective is to determine a set of P EMPs $\{\mathbf{u}_p^{(n)^T}, n = 1, ..., N\}_{p=1}^P$ that maximize the variance while producing features with zero-correlation. Thus, the objective function for the p^{th} EMP is

$$\begin{aligned} \{\mathbf{u}_p^{(n)^T}, n = 1, ..., N\} &= \arg\max \sum_{m=1}^{M} (y_{m_p} - \overline{y}_p)^2, \\ \texttt{subject to} \quad & \mathbf{u}_p^{(n)^T}\mathbf{u}_p^{(n)} = 1 \text{ and} \\ \frac{\mathbf{g}_p^T \mathbf{g}_q}{\parallel \mathbf{g}_p \parallel \parallel \mathbf{g}_q \parallel} &= \delta_{pq}, \ p, q = 1, ..., P, \end{aligned} \tag{6}$$

where δ_{pq} is the Kronecker delta (defined as 1 for $p = q$ and as 0 otherwise).

3.2. The UMPCA algorithm

To solve the UMPCA problem (6), we follow the successive variance maximization approach in the derivation of PCA in (Jolliffe, 2002). The P EMPs $\{\mathbf{u}_p^{(n)^T}, n = 1, ..., N\}_{p=1}^P$ are determined one by one in P steps, with the p^{th} step obtaining the p^{th} EMP:

Step 1**:** Determine the first EMP $\{\mathbf{u}_1^{(n)^T}, n = 1, ..., N\}$ by maximizing $S_{T_1}^{\mathbf{y}}$ without any constraint.

Step 2**:** Determine the second EMP $\{\mathbf{u}_2^{(n)^T}, n = 1, ..., N\}$ by maximizing $S_{T_2}^{\mathbf{y}}$ subject to the constraint that $\mathbf{g}_2^T \mathbf{g}_1 = 0$.

Step $p(p = 3, ..., P)$**:** Determine the p^{th} EMP $\{\mathbf{u}_p^{(n)^T}, n = 1, ..., N\}$ by maximizing $S_{T_p}^{\mathbf{y}}$ subject to the constraint that $\mathbf{g}_p^T \mathbf{g}_q = 0$ for $q = 1, ..., p-1$.

In order to solve for the p^{th} EMP $\{\mathbf{u}_p^{(n)^T}, n = 1, ..., N\}$, we need to determine N sets of parameters corresponding to N projection vectors, $\mathbf{u}_p^{(1)}, \mathbf{u}_p^{(2)}, ...\mathbf{u}_p^{(N)}$, one in each mode. Unfortunately, simultaneous determination of these N sets of parameters in all modes is a complicated non-linear problem without an existing optimal solution, except when $N = 1$, which is the classical PCA where only one projection vector is to be solved. Therefore, we follow the approach in the alternating least square (ALS) algorithm (Harshman, 1970) to solve this multilinear problem. For each EMP to be determined, the parameters of the projection vector $\mathbf{u}_p^{(n^*)}$ for each mode n^* are estimated one mode by one mode separately, conditioned on $\{\mathbf{u}_p^{(n)}, n \neq n^*\}$, the parameter values of the projection vectors in the other modes.

To solve for $\mathbf{u}_p^{(n^*)}$ in the n^*-mode, assuming that $\{\mathbf{u}_p^{(n)}, n \neq n^*\}$ is given, the tensor samples are projected in these $(N-1)$ modes $\{n \neq n^*\}$ first to obtain the vectors

$$\begin{aligned} \tilde{\mathbf{y}}_{m_p}^{(n^*)} = \quad & \mathcal{X}_m \times_1 \mathbf{u}_p^{(1)^T} ... \times_{n^*-1} \mathbf{u}_p^{(n^*-1)^T} \\ & \times_{n^*+1} \mathbf{u}_p^{(n^*+1)^T} ... \times_N \mathbf{u}_p^{(N)^T}, \end{aligned} \tag{7}$$

where $\tilde{\mathbf{y}}_{m_p}^{(n^*)} \in \mathbb{R}^{I_{n^*}}$. This conditional subproblem then becomes to determine $\mathbf{u}_p^{(n^*)}$ that projects the vector samples $\{\tilde{\mathbf{y}}_{m_p}^{(n^*)}, m = 1, ..., M\}$ onto a line so that the variance is maximized, subject to the zero-correlation constraint, which is a PCA problem with the input samples $\{\tilde{\mathbf{y}}_{m_p}^{(n^*)}, m = 1, ..., M\}$. The corresponding total scatter matrix $\tilde{\mathbf{S}}_{T_p}^{(n^*)}$ is then defined as

$$\tilde{\mathbf{S}}_{T_p}^{(n^*)} = \sum_{m=1}^{M} (\tilde{\mathbf{y}}_{m_p}^{(n^*)} - \bar{\tilde{\mathbf{y}}}_p^{(n^*)})(\tilde{\mathbf{y}}_{m_p}^{(n^*)} - \bar{\tilde{\mathbf{y}}}_p^{(n^*)})^T, \tag{8}$$

where $\bar{\tilde{\mathbf{y}}}_p^{(n^*)} = \frac{1}{M}\sum_m \tilde{\mathbf{y}}_{m_p}^{(n^*)}$. With (8), we are ready to solve for the P EMPs. For $p = 1$, the $\mathbf{u}_1^{(n^*)}$ that maximizes the total scatter $\mathbf{u}_1^{(n^*)^T} \tilde{\mathbf{S}}_{T_1}^{(n^*)} \mathbf{u}_1^{(n^*)}$ in the projected space is obtained as the unit eigenvector of $\tilde{\mathbf{S}}_{T_1}^{(n^*)}$ associated with the largest eigenvalue. Next, we show how to determine the p^{th} $(p > 1)$ EMP given the first $(p-1)$ EMPs. Given the first $(p-1)$ EMPs, the p^{th} EMP aims to maximize the total scatter $S_{T_p}^{\mathbf{y}}$, subject to the constraint that features projected by the p^{th} EMP are uncorrelated with those projected by the first $(p-1)$ EMPs. Let $\tilde{\mathbf{Y}}_p^{(n^*)} \in \mathbb{R}^{I_{n^*} \times M}$ be a matrix with $\tilde{\mathbf{y}}_{m_p}^{(n^*)}$ as its m^{th} column, i.e., $\tilde{\mathbf{Y}}_p^{(n^*)} = \left[\tilde{\mathbf{y}}_{1_p}^{(n^*)}, \tilde{\mathbf{y}}_{2_p}^{(n^*)}, ..., \tilde{\mathbf{y}}_{M_p}^{(n^*)}\right]$, then the p^{th} coordinate vector is $\mathbf{g}_p = \tilde{\mathbf{Y}}_p^{(n^*)^T} \mathbf{u}_p^{(n^*)}$. The constraint that $\mathbf{g}_p$ is uncorrelated with $\{\mathbf{g}_q, q = 1, ..., p-1\}$ can be written as

$$\mathbf{g}_p^T \mathbf{g}_q = \mathbf{u}_p^{(n^*)^T} \tilde{\mathbf{Y}}_p^{(n^*)} \mathbf{g}_q = 0, q = 1, ..., p-1. \tag{9}$$

Thus, $\mathbf{u}_p^{(n^*)}$ $(p > 1)$ can be determined by solving the following constrained optimization problem:

$$\begin{aligned} & \mathbf{u}_p^{(n^*)} = \arg\max \mathbf{u}_p^{(n^*)^T} \tilde{\mathbf{S}}_{T_p}^{(n^*)} \mathbf{u}_p^{(n^*)}, \\ & \texttt{subject to } \mathbf{u}_p^{(n^*)^T} \mathbf{u}_p^{(n^*)} = 1 \text{ and} \\ & \mathbf{u}_p^{(n^*)^T} \tilde{\mathbf{Y}}_p^{(n^*)} \mathbf{g}_q = 0, q = 1, ..., p-1, \end{aligned} \tag{10}$$

The solution is given by the following theorem:

Theorem 1. *The solution to the problem (10) is the (unit-length) eigenvector corresponding to the largest eigenvalue of the following eigenvalue problem:*

$$\mathbf{\Psi}_p^{(n^*)} \tilde{\mathbf{S}}_{T_p}^{(n^*)} \mathbf{u} = \lambda \mathbf{u}, \tag{11}$$

where

$$\boldsymbol{\Psi}_p^{(n^*)} = \mathbf{I}_{I_{n^*}} - \tilde{\mathbf{Y}}_p^{(n^*)}\mathbf{G}_{p-1}\boldsymbol{\Phi}_p^{-1}\mathbf{G}_{p-1}^T\tilde{\mathbf{Y}}_p^{(n^*)^T}, \qquad (12)$$

$$\boldsymbol{\Phi}_p = \mathbf{G}_{p-1}^T\tilde{\mathbf{Y}}_p^{(n^*)^T}\tilde{\mathbf{Y}}_p^{(n^*)}\mathbf{G}_{p-1}, \qquad (13)$$

$$\mathbf{G}_{p-1} = [\mathbf{g}_1 \quad \mathbf{g}_2 \quad ...\mathbf{g}_{p-1}] \in \mathbb{R}^{M\times(p-1)}, \qquad (14)$$

and $\mathbf{I}_{I_{n^*}}$ *is an identity matrix of size* $I_{n^*} \times I_{n^*}$.

Proof. First, Lagrange multipliers can be used to transform the problem (10) to the following to include all the constraints:

$$\begin{aligned} F(\mathbf{u}_p^{(n^*)}) \;=\;& \mathbf{u}_p^{(n^*)^T}\tilde{\mathbf{S}}_{T_p}^{(n^*)}\mathbf{u}_p^{(n^*)} - \nu\left(\mathbf{u}_p^{(n^*)^T}\mathbf{u}_p^{(n^*)} - 1\right) \\ &- \sum_{q=1}^{p-1}\mu_q\mathbf{u}_p^{(n^*)^T}\tilde{\mathbf{Y}}_p^{(n^*)}\mathbf{g}_q, \qquad (15) \end{aligned}$$

where ν and $\{\mu_q, q = 1, ..., p-1\}$ are Lagrange multipliers.

The optimization is performed by setting the partial derivative of $F(\mathbf{u}_p^{(n^*)})$ with respect to $\mathbf{u}_p^{(n^*)}$ to zero:

$$\begin{aligned} \frac{\partial F(\mathbf{u}_p^{(n^*)})}{\partial \mathbf{u}_p^{(n^*)}} \;=\;& 2\tilde{\mathbf{S}}_{T_p}^{(n^*)}\mathbf{u}_p^{(n^*)} - 2\nu\mathbf{u}_p^{(n^*)} \\ &- \sum_{q=1}^{p-1}\mu_q\tilde{\mathbf{Y}}_p^{(n^*)}\mathbf{g}_q = 0. \qquad (16) \end{aligned}$$

Multiplying (16) by $\mathbf{u}_p^{(n^*)^T}$ results in

$$\begin{aligned} &2\mathbf{u}_p^{(n^*)^T}\tilde{\mathbf{S}}_{T_p}^{(n^*)}\mathbf{u}_p^{(n^*)} - 2\nu\mathbf{u}_p^{(n^*)^T}\mathbf{u}_p^{(n^*)} = 0 \\ &\Rightarrow \nu = \frac{\mathbf{u}_p^{(n^*)^T}\tilde{\mathbf{S}}_{T_p}^{(n^*)}\mathbf{u}_p^{(n^*)}}{\mathbf{u}_p^{(n^*)^T}\mathbf{u}_p^{(n^*)}}, \qquad (17) \end{aligned}$$

which indicates that ν is exactly the criterion to be maximized, with the constraint on the norm of the projection vector incorporated.

Next, a set of $(p-1)$ equations are obtained by multiplying (16) by $\mathbf{g}_q^T\tilde{\mathbf{Y}}_p^{(n^*)^T}$, $q = 1, ..., p-1$, respectively:

$$2\mathbf{g}_q^T\tilde{\mathbf{Y}}_p^{(n^*)^T}\tilde{\mathbf{S}}_{T_p}^{(n^*)}\mathbf{u}_p^{(n^*)} - \sum_{q=1}^{p-1}\mu_q\mathbf{g}_q^T\tilde{\mathbf{Y}}_p^{(n^*)^T}\cdot\tilde{\mathbf{Y}}_p^{(n^*)}\mathbf{g}_q = 0. \qquad (18)$$

Let

$$\boldsymbol{\mu}_{p-1} = [\mu_1 \; \mu_2 \; ... \; \mu_{p-1}]^T \qquad (19)$$

and use (13) and (14), then the $(p-1)$ equations of (18) can be represented in a single matrix equation as following:

$$2\mathbf{G}_{p-1}^T\tilde{\mathbf{Y}}_p^{(n^*)^T}\tilde{\mathbf{S}}_{T_p}^{(n^*)}\mathbf{u}_p^{(n^*)} - \boldsymbol{\Phi}_p\boldsymbol{\mu}_{p-1} = 0. \qquad (20)$$

Thus,

$$\boldsymbol{\mu}_{p-1} = 2\boldsymbol{\Phi}_p^{-1}\cdot\mathbf{G}_{p-1}^T\tilde{\mathbf{Y}}_p^{(n^*)^T}\tilde{\mathbf{S}}_{T_p}^{(n^*)}\mathbf{u}_p^{(n^*)}. \qquad (21)$$

Since from (14) and (19),

$$\sum_{q=1}^{p-1}\mu_q\tilde{\mathbf{Y}}_p^{(n^*)}\mathbf{g}_q = \tilde{\mathbf{Y}}_p^{(n^*)}\mathbf{G}_{p-1}\boldsymbol{\mu}_{p-1}, \qquad (22)$$

the equation (16) can be written as

$$\begin{aligned} &2\tilde{\mathbf{S}}_{T_p}^{(n^*)}\mathbf{u}_p^{(n^*)} - 2\nu\mathbf{u}_p^{(n^*)} - \tilde{\mathbf{Y}}_p^{(n^*)}\mathbf{G}_{p-1}\boldsymbol{\mu}_{p-1} = 0 \\ \Rightarrow &\nu\mathbf{u}_p^{(n^*)} = \tilde{\mathbf{S}}_{T_p}^{(n^*)}\mathbf{u}_p^{(n^*)} - \tilde{\mathbf{Y}}_p^{(n^*)}\mathbf{G}_{p-1}\frac{\boldsymbol{\mu}_{p-1}}{2} \\ &= \left[\mathbf{I}_{I_{n^*}} - \tilde{\mathbf{Y}}_p^{(n^*)}\mathbf{G}_{p-1}\boldsymbol{\Phi}_p^{-1}\mathbf{G}_{p-1}^T\tilde{\mathbf{Y}}_p^{(n^*)^T}\right]\tilde{\mathbf{S}}_{T_p}^{(n^*)}\mathbf{u}_p^{(n^*)}. \end{aligned}$$

Using the definition in (12), an eigenvalue problem is obtained as $\boldsymbol{\Psi}_p^{(n^*)}\tilde{\mathbf{S}}_{T_p}^{(n^*)}\mathbf{u} = \nu\mathbf{u}$. Since ν is the criterion to be maximized, the maximization is achieved by setting $\mathbf{u}_p^{(n)^*}$ to be the (unit) eigenvector corresponding to the largest eigenvalue of (11). □

By setting $\boldsymbol{\Psi}_1^{(n^*)} = \mathbf{I}_{I_{n^*}}$ and from Theorem 1, we have a unified solution for UMPCA: for $p = 1, ..., P$, $\mathbf{u}_p^{(n^*)}$ is obtained as the unit eigenvector of $\boldsymbol{\Psi}_p^{(n^*)}\tilde{\mathbf{S}}_{T_p}^{(n^*)}$ associated with the largest eigenvalue. Algorithm 1 summarizes the UMPCA developed here.

Algorithm 1 Uncorrelated Multilinear Principal Component Analysis (UMPCA)

```
Input: A set of tensor samples $\{\mathcal{X}_m \in \mathbb{R}^{I_1\times...\times I_N}, m = 1, ..., M\}$, the subspace dimensionality $P$, and the maximum number of iterations $K$.
for $p = 1$ to $P$ do
  for $n = 1$ to $N$ do
    Initialize $\mathbf{u}_{p_{(0)}}^{(n)} = \mathbf{1}/ \parallel \mathbf{1} \parallel$.
  end for
  for $k = 1$ to $K$ do
    for $n = 1$ to $N$ do
      Calculate $\tilde{\mathbf{y}}_{m_p}^{(n)} = \mathcal{X}_m \times_1 \mathbf{u}_{p_{(k)}}^{(1)^T} ... \times_{n-1} \mathbf{u}_{p_{(k)}}^{(n-1)^T} \times_{n+1} \mathbf{u}_{p_{(k-1)}}^{(n+1)^T} ... \times_N \mathbf{u}_{p_{(k-1)}}^{(N)^T}$, for $m = 1, ..., M$.
      Calculate $\boldsymbol{\Psi}_p^{(n)}$ and $\tilde{\mathbf{S}}_{T_p}^{(n)}$. Set $\mathbf{u}_{p_{(k)}}^{(n)}$ to be the (unit) eigenvector of $\boldsymbol{\Psi}_p^{(n)}\tilde{\mathbf{S}}_{T_p}^{(n)}$ associated with the largest eigenvalue.
    end for
  end for
  Set $\mathbf{u}_p^{(n)} = \mathbf{u}_{p_k}^{(n)}$ for all $n$.
  Calculate the coordinate vector $\mathbf{g}_p$.
end for
```

3.3. Initialization, projection order and termination

As an iterative algorithm, the UMPCA may be affected by the initialization method, the projection order and the termination conditions. Due to the space constraint, these issues, as well as the convergence and computational issues, are not studied here. Instead, we adopt simple implementation strategies for them. First, we use the uniform initialization for UMPCA, where all n-mode projection vectors are initialized to have unit length and the same value along the I_n dimensions in n-mode, which is equivalent to the all ones vector $\mathbf{1}$ with proper normalization. Second, as shown in Algorithm 1, the projection order, which is the mode ordering in computing the projection vectors, is from 1-mode to N-mode, as in other multilinear algorithms (Ye, 2005; Xu et al., 2005; Lu et al., 2008a). Third, the iteration is terminated by setting K, the maximum number of iterations.

4. Experimental Evaluation

The proposed UMPCA can potentially benefit various applications involving tensorial data, as mentioned in Sec. 1. Since face recognition has practical importance in security-related applications such as biometric authentication and surveillance, it has been used widely for evaluation of unsupervised learning algorithms (Shashua & Levin, 2001; Yang et al., 2004; Xu et al., 2005; Ye, 2005). Therefore, in this section, we focus on evaluating the effectiveness of UMPCA on this popular classification task through performance comparison with existing unsupervised dimensionality reduction algorithms.

4.1. The FERET database

The Facial Recognition Technology (FERET) database (Phillips et al., 2000) is widely used for testing face recognition performance, with 14,126 images from 1,199 subjects covering a wide range of variations in viewpoint, illumination, facial expression, races and ages. A subset of this database is selected in our experimental evaluation and it consists of those subjects with each subject having at least eight images with at most 15 degrees of pose variation, resulting in 721 face images from 70 subjects. Since our focus here is on the recognition of faces rather than their detection, all face images are manually cropped, aligned (with manually annotated coordinate information of eyes) and normalized to 80×80 pixels, with 256 gray levels per pixel. Figure 1 shows some sample face images from two subjects in this FERET subset.

Figure 1. Examples of face images from two subjects in the FERET subset used in our experimental evaluation.

4.2. Face recognition performance comparison

In the evaluation, we compare the performance of the UMPCA against three PCA-based unsupervised learning algorithms: the PCA (eigenface) algorithm (Turk & Pentland, 1991), the MPCA algorithm (Lu et al., 2008a)[2] and the TROD algorithm (Shashua & Levin, 2001). The number of iterations in TROD and UMPCA is set to ten, with the same (uniform) initialization used. For MPCA, we obtain the full projection and select the most descriptive P features for recognition. The features obtained by these four algorithms are arranged in descending variation captured (measured by respective total scatter). For classification of extracted features, we use the nearest neighbor classifier (NNC) with Euclidean distance measure.

Gray-level face images are naturally second-order tensors (matrices), i.e., $N = 2$. Therefore, they are input directly as 80×80 tensors to the multilinear algorithms (MPCA, TROD, UMPCA), while for PCA, they are vectorized to 6400×1 vectors as input. For each subject in a face recognition experiment, $L(= 1, 2, 3, 4, 5, 6, 7)$ samples are randomly selected for unsupervised training and the rest are used for testing. We report the results averaged over ten such random splits (repetitions).

Figures 2 and 3 show the detailed results[3] for $L = 1$ and $L = 7$, respectively. $L = 1$ is an extreme small sample size scenario where only one sample per class is available for training, the so-called one training sample (OTS) case important in practice (Wang et al., 2006), and $L = 7$ is the maximum number of training samples we can use in our experiments. Figures 2(a) and 3(a) plot the correct recognition rates against P, the dimensionality of the subspace for $P = 1, ..., 10$, and Figs 2(b) and 3(b) plot those for $P = 15, ..., 80$. From the figures, UMPCA outperforms the other three methods in both cases and across all dimensionality, indicating that the uncorrelated features extracted directly from the tensorial face data are more effective in classifi-

[2] Note that MPCA with $N = 2$ is equivalent to GPCA.

[3] Note that for PCA and UMPCA, there are at most 69 features when $L = 1$ (only 70 faces for training).

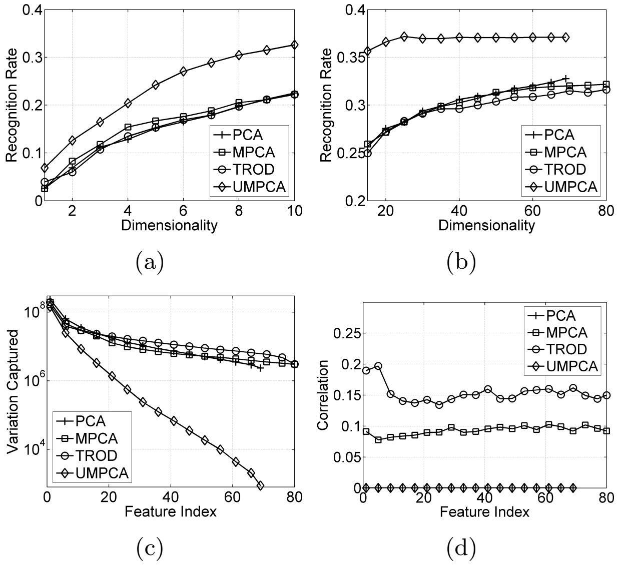

Figure 2. Detailed face recognition results on the FERET database for $L = 1$: (a) performance curves for the low-dimensional case, (b) performance curves for the high-dimensional case, (c) the variation captured by individual features and (d) the correlation among features.

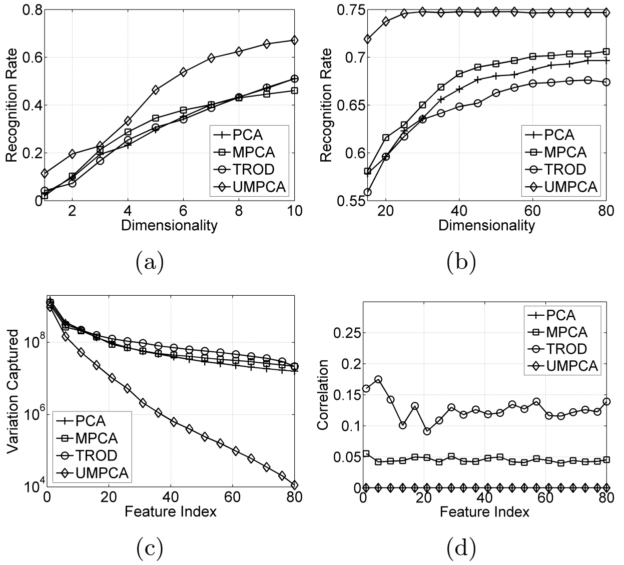

Figure 3. Detailed face recognition results on the FERET database for $L = 7$: (a) performance curves for the low-dimensional case, (b) performance curves for the high-dimensional case, (c) the variation captured by individual features and (d) the correlation among features.

cation. The figures also show that for UMPCA, the recognition rate saturates around $P = 30$, which can be explained by observing the variance captured by individual features as shown in Figs. 2(c) and 3(c) (in log scale). These figures show that the variance captured by UMPCA is considerably lower than those captured by the other methods, which is due to its constraints of zero-correlation and TVP. Despite capturing lower variance, UMPCA is superior in the recognition task performed. Nonetheless, when the variance captured is too low, those corresponding features are no longer descriptive enough to contribute in classification, leading to the saturation.

In addition, we also plot the average correlation of individual features with all the other features in Figs. 2(d) and 3(d). As supported by theoretical derivation, features extracted by PCA and UMPCA are uncorrelated. In contrast, features extracted by MPCA and TROD are correlated, with TROD features have higher correlation on average.

Table 2. Face recognition results on the FERET database: the recognition rates (in percentage) for various Ls and Ps.

L	P	1	5	10	20	50	80
2	PCA	2.8	20.2	32.0	39.1	43.6	45.1
	MPCA	2.6	21.4	28.1	38.9	44.6	**46.0**
	TROD	3.6	19.3	30.6	38.4	43.0	44.3
	UMPCA	**8.1**	**27.6**	**40.6**	**45.0**	**45.8**	45.7
3	PCA	2.7	23.9	37.1	45.9	51.3	52.6
	MPCA	2.3	25.9	34.8	45.5	52.0	53.3
	TROD	4.0	23.5	36.1	44.5	50.1	51.7
	UMPCA	**7.5**	**35.5**	**49.8**	**56.0**	**56.6**	**56.6**
4	PCA	2.7	25.5	41.7	49.4	56.8	57.9
	MPCA	2.3	28.7	39.4	50.2	57.5	58.9
	TROD	4.2	25.3	41.1	49.0	55.1	56.6
	UMPCA	**8.5**	**39.5**	**56.2**	**63.5**	**64.1**	**64.2**
5	PCA	3.0	28.9	47.1	55.6	63.9	64.6
	MPCA	2.6	33.0	43.2	56.8	64.3	65.8
	TROD	4.5	28.4	47.2	55.6	62.0	63.9
	UMPCA	**8.1**	**43.6**	**61.7**	**68.2**	**69.1**	**69.1**
6	PCA	2.8	30.3	49.0	58.5	66.7	68.1
	MPCA	2.2	33.5	45.7	59.7	67.9	69.7
	TROD	4.3	27.3	49.3	58.6	64.7	66.9
	UMPCA	**9.1**	**45.6**	**62.9**	**70.7**	**71.8**	**71.8**

The recognition results for $P = 1, 5, 10, 20, 50, 80$ are listed in Table 2 for $L = 2, 3, 4, 5, 6$, where the best recognition results among the four methods are shown in bold. More detailed results are omitted here to save space. From the table, UMPCA achieves superior recognition results in all cases except for $P = 80$ and $L = 2$, where the difference with the best results by MPCA is small (0.3%). In particular, for smaller P $(1, 5, 10, 20)$, UMPCA outperforms the other algorithms significantly, demonstrating its superior capability in classifying faces in low-dimensional spaces.

5. Conclusions

This paper proposes a novel uncorrelated multilinear PCA algorithm, where uncorrelated features are extracted directly from tensorial representation through a tensor-to-vector projection. The algorithm successively maximizes variance captured by each elementary projection while enforcing the zero-correlation constraint. The solution employs the alternating projection method and is iterative. Experiments on face recognition demonstrate that compared with other unsupervised learning algorithms including the PCA, MPCA and TROD, the UMPCA achieves the best results and it is particularly effective in low-dimensional spaces. Thus, face recognition through unsupervised learning benefits from the proposed UMPCA and in future research, it is worthwhile to investigate whether UMPCA can contribute in other unsupervised learning tasks, such as clustering.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments. This work is partially supported by the Ontario Centres of Excellence through the Communications and Information Technology Ontario Partnership Program and the Bell University Labs - at the University of Toronto.

References

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of “eckart-young” decomposition. *Psychometrika*, *35*, 283–319.

Faloutsos, C., Kolda, T. G., & Sun, J. (2007). Mining large time-evolving data using matrix and tensor tools. *Int. Conf. on Data Mining 2007 Tutorial.*

Harshman, R. A. (1970). Foundations of the parafac procedure: Models and conditions for an “explanatory” multi-modal factor analysis. *UCLA Working Papers in Phonetics*, *16*, 1–84.

Jolliffe, I. T. (2002). *Principal component analysis, second edition.* Springer Serires in Statistics.

Koren, Y., & Carmel, L. (2004). Robust linear dimensionality reduction. *IEEE Trans. Vis. Comput. Graphics*, *10*, 459–470.

Lathauwer, L. D., Moor, B. D., & Vandewalle, J. (2000). On the best rank-1 and rank-$(R_1, R_2, ..., R_N)$ approximation of higher-order tensors. *SIAM Journal of Matrix Analysis and Applications*, *21*, 1324–1342.

Law, M. H. C., & Jain, A. K. (2006). Incremental nonlinear dimensionality reduction by manifold learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, *28*, 377–391.

Lu, H., Plataniotis, K. N., & Venetsanopoulos, A. N. (2008a). MPCA: Multilinear principal component analysis of tensor objects. *IEEE Trans. Neural Netw.*, *19*, 18–39.

Lu, H., Plataniotis, K. N., & Venetsanopoulos, A. N. (2008b). Uncorrelated multilinear discriminant analysis with regularization and aggregation for tensor object recognition. *IEEE Trans. Neural Netw.* accepted pending minor revision.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. (2000). The FERET evaluation method for face recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, *22*, 1090–1104.

Shakhnarovich, G., & Moghaddam, B. (2004). Face recognition in subspaces. *Handbook of Face Recognition* (pp. 141–168). Springer-Verlag.

Shashua, A., & Levin, A. (2001). Linear image coding for regression and classification using the tensor-rank principle. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (pp. 42–49).

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neurosicence*, *3*, 71–86.

Wang, J., Plataniotis, K. N., Lu, J., & Venetsanopoulos, A. N. (2006). On solving the face recognition problem with one training sample per subject. *Pattern Recognition*, *39*, 1746–1762.

Xu, D., Yan, S., Zhang, L., Zhang, H.-J., Liu, Z., & Shum;, H.-Y. (2005). Concurrent subspaces analysis. *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition* (pp. 203–208).

Yang, J., Zhang, D., Frangi, A. F., & Yang, J. (2004). Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, *26*, 131–137.

Ye, J. (2005). Generalized low rank approximations of matrices. *Machine Learning*, *61*, 167–191.

Ye, J., Janardan, R., & Li, Q. (2004). GPCA: An efficient dimension reduction scheme for image compression and retrieval. *The* 10^{th} *ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (pp. 354–363).